Pontifícia Universidade Católica de Goiás
Programa de Mestrado em Engenharia de Produção e Sistemas

DESCOBERTA DE CONHECIMENTO NA PLATAFORMA LATTES: UM ESTUDO DE CASO NO INSTITUTO FEDERAL DE GOIÁS

RENATA DE SOUZA ALVES PAULA CAVALCANTE

2014

Pontifícia Universidade Católica de Goiás
Programa de Mestrado em Engenharia de Produção e Sistemas

DESCOBERTA DE CONHECIMENTO NA PLATAFORMA LATTES: UM ESTUDO DE CASO NO INSTITUTO FEDERAL DE GOIÁS

RENATA DE SOUZA ALVES PAULA CAVALCANTE

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Produção e Sistemas.

Orientadora: Maria José Pereira Dantas, *Doutora.*

Goiânia
Fevereiro 2014

Dados Internacionais de Catalogação da Publicação (CIP)
(Sistema de Bibliotecas PUC Goiás)

Cavalcante, Renata de Souza Alves Paula.
C376d Descoberta de conhecimento na Plataforma Lattes [manuscrito] : um estudo de caso no Instituto Federal de Goiás / Renata de Souza Alves Paula Cavalcante. – 2014.
217 f. : il.; grafs.; 30 cm.

Dissertação (mestrado) – Pontifícia Universidade Católica de Goiás, Programa de Mestrado em Engenharia de Produção e Sistemas, 2014.
"Orientadora: Profa. Dra. Maria José Pereira Dantas".

1. Mineração de dados (Computação). 2. Plataforma Lattes. I. Título.

CDU 004.45(043)

DESCOBERTA DE CONHECIMENTO NA PLATAFORMA LATTES: UM ESTUDO DE CASO NO INSTITUTO FEDERAL DE GOIÁS

Renata de Souza Alves Paula Cavalcante

Esta dissertação julgada adequada para obtenção do título de Mestre em Engenharia de Produção e Sistemas, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás em Fevereiro de 2014.

_____
Prof. Ricardo Luiz Machado, Dr.
Coordenador do Programa de Pós-Graduação em Engenharia de Produção e Sistemas

Banca Examinadora:

_____
Profª. Maria José Pereira Dantas, Dra.
Orientadora

_____
Prof. Sibelius Lellis Vieira, Dr.

_____
Prof. Francisco Ramos de Melo, Ph.D.

_____
Prof. José Elmo de Menezes, Dr.

Goiânia - Goiás
Fevereiro 2014

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida, pela Sua constante proteção em minha caminhada e pela renovação das forças quando ela me falta,

Aos meus pais Ronan e Maria de Lourdes, pelo amor incondicional e formação que recebi,

Carinhosamente, ao meu querido esposo Vinicius, pelo enorme apoio recebido, compreensão, paciência e revisões deste trabalho. Por estar sempre ao meu lado, sem medir esforços para me ajudar no que for preciso. Obrigada por me fazer acreditar que posso mais que imagino.

À minha orientadora Prof. Dra. Maria José, pelas horas de orientação, sugestões, dedicação e disponibilidade,

À minha amiga Elvia, pela disponibilidade em ajudar, com o empréstimo de livros,

À equipe da Pró-Reitoria de Pesquisa e Pós-Graduação do Instituto Federal de Goiás pela disponibilização dos dados e informações referentes ao estudo de caso deste trabalho, em especial ao Prof. Ruberley Rodrigues de Souza e a Viviane Margarida,

Aos colaboradores de TI do IFG, em especial a Roberval Lustosa,

Ao professor do IFG, Adelino Cândido Pimenta,

Ao colaborador Weliton, da Pró-Reitoria de Pós-Graduação e Pesquisa da PUC-GO.

“Estamos afogados em informação, mas morrendo de fome por conhecimento”.
(John Naisbett)

Resumo da Dissertação apresentada ao MEPROS/PUC Goiás como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia de Produção e Sistemas (M. Sc.).

DESCOBERTA DE CONHECIMENTO NA PLATAFORMA LATTES: UM ESTUDO DE CASO NO INSTITUTO FEDERAL DE GOIÁS

Renata de Souza Alves Paula Cavalcante

Fevereiro/2014

Orientadora: Profa. Maria José Pereira Dantas, Dra.

Ao longo do tempo, percebeu-se que a velocidade do acúmulo de informações era maior do que a velocidade de processamento e análise das mesmas. Não adianta uma organização ter muita informação, se não souber dela extrair conhecimento útil. É preciso que sejam feitas análises apuradas sobre os dados e descobrir quais são os padrões de comportamentos existentes nos mesmos. Assim, uma organização poderá realizar tomada de decisão de forma mais segura, baseada em fatos reais e não em meras suposições, inclusive no âmbito da gestão de Ciência e Tecnologia. Este trabalho teve como objetivo realizar um estudo de caso no Instituto Federal de Goiás (IFG), aplicando o processo de *Knowledge Discovery in Database* (KDD), na tentativa de identificar padrões que representem o perfil da produção científica dos docentes da instituição. A maior parte dos dados analisados foram extraídos da Plataforma Lattes (PL) e o período da pesquisa fixado no último triênio. Pretendeu-se obter conhecimento sobre a produtividade dos docentes e provê-los à Pró-Reitoria de Pesquisa e Pós-Graduação do IFG para auxiliar na sua gestão. A pesquisa aborda por meio de um levantamento bibliográfico os conceitos sobre Gestão do Conhecimento (GC), o processo de KDD, incluindo a Mineração de Dados (MD) com suas tarefas e técnicas, a produção científica, a PL e o contexto atual do IFG. Dessa forma, entre os resultados obtidos no trabalho, viu-se que a aplicação do KDD pode ser um poderoso instrumento para a gestão das informações nas instituições de ensino.

Palavras-chave: Descoberta de Conhecimento, Plataforma Lattes, Mineração de Dados.

Summary of Dissertation submitted to MEPROS/PUC Goiás as part of the requirements for the degree of Master of Engineering in Production Systems (M. Sc.)

KNOWLEDGE DISCOVERY IN LATTES PLATFORM: A CASE STUDY IN THE FEDERAL INSTITUTE OF GOIAS

Renata de Souza Alves Paula Cavalcante

February/2014

Advisor: Profa. Maria José Pereira Dantas*, Doctor.*

Over the last years, it was realized that the rate of accumulation of information is becoming larger than the speed of its process and analysis. An organization have no use in a lot of information if it doesn't know how to extract useful knowledge of it. It is necessary to do refined analysis on all data to find which patterns of behaviors exist in them. Thus, the process of decision making in an organization can be done more safely, based on real facts and not on mere assumptions, including decisions in the scope of Science and Technology Management. This study aimed to realize a case study at the Federal Institute of Goiás (IFG), applying the process of Knowledge Discovery in Database (KDD) in an attempt to identify patterns that represent the profile of the scientific production of the teachers of the institution. Most of the data analyzed was extracted from the Lattes Platform (PL) and the survey period was defined as the last triennium. The aim was to obtain knowledge about the productivity of teachers and provide this information to the Pró Reitoria de Pesquisa e Pós-Graduação of the IFG to assist their management. The research addresses through a bibliographic survey the concepts of Knowledge Management (KM), the process of KDD, including Data Mining (MD) with their tasks and techniques, the scientific production, the PL and the current context of the IFG. Thus, among the obtained results, it was seen that the application of KDD can be a powerful tool for information management in educational institutions.

Keywords*: Knowledge Discovery, Lattes Platform, Data Mining.*

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS E ABREVIATURAS

ACID	Atomicidade, Consistência, Isolamento e Durabilidade
AG	Algoritmos Genéticos
AGD	Algoritmo Genético para Descoberta de Regras Difusas
ANS	Atividade Não-Supervisionada
ARFF	*Attribute Relation File Format*
AS	Atividade Supervisionada
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CEFET	Centro Federal de Educação Tecnológica
CET	Ciências Exatas e da Terra
CH	Ciências Humanas
CL	Currículo Lattes
CNPJ	Cadastro Nacional de Pessoa Jurídica
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CPF	Cadastro de Pessoa Física
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
CT&I	Ciência, Tecnologia e Inovação
C&T	Ciência e Tecnologia
DCBD	Descoberta de Conhecimento em Banco de Dados
DE	Dedicação Exclusiva
DOI	*Digital Object Identifier*
DW	*Data Warehouse*
EAA	Experimento de Associação quanto aos Artigos em periódicos
EAD	Educação a distância
EAT	Experimento de Associação quanto aos Trabalhos completos

ECA	Experimento de Classificação quanto aos Artigos em periódicos
ECT	Experimento de Classificação quanto aos Trabalhos completos
EDM	*Educational Data Mining*
EUA	Estados Unidos da América
GC	Gestão do Conhecimento
GPL	*General Public License*
IA	Inteligência Artificial
IES	Instituição de Ensino Superior
IFs	Institutos Federais
IFAM	Instituto Federal de Educação, Ciência e Tecnologia do Amazonas
IFBA	Instituto Federal de Educação, Ciência e Tecnologia da Bahia
IFCE	Instituto Federal de Educação, Ciência e Tecnologia do Ceará
IFES	Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo
IFF	Instituto Federal de Educação, Ciência e Tecnologia Fluminense
IFG	Instituto Federal de Educação, Ciência e Tecnologia de Goiás
IFGOIANO	Instituto Federal Goiano
IFMT	Instituto Federal de Educação, Ciência e Tecnologia do Mato Grosso
IFRJ	Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro
IFRN	Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte
IFRR	Instituto Federal de Educação, Ciência e Tecnologia de Roraima
IFSC	Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina
IFSP	Instituto Federal de Educação, Ciência e Tecnologia de São Paulo
IFSUL	Instituto Federal de Educação, Ciência e Tecnologia Sul-Rio-Grandense
IFTO	Instituto Federal de Educação, Ciência e Tecnologia do Tocantins
INPI	Instituto Nacional de Propriedade Industrial
ISBN	*International Standard Book Number*
ISI	*Institute for Scientific Information*

ISSN	*International Standard Serial Number*
J48	Implementação do algoritmo C4.5 no WEKA
JDBC	*Java DataBase Connectivity*
KDD	*Knowledge Discovery in Databases*
MCT	Ministério da Ciência e Tecnologia
MD	Mineração de Dados
MGCTI	Mestrado em Gestão do Conhecimento e da Tecnologia da Informação
PHP	*Personal Home Page*
PIBIC	Programa Institucional de Bolsa de Iniciação Científica
PIBIC-Af	Programa Institucional de Bolsa de Iniciação Científica nas Ações Afirmativas
PIBITI	Programa Institucional de Bolsa de Iniciação em Desenvolvimento Tecnológico e Inovação
PIPECT	Programa Institucional de Incentivo à Participação em Eventos Científicos e Tecnológicos
PL	Plataforma Lattes
POLI	Escola Politécnica de Pernambuco
ProAPP	Programa de Apoio à Produtividade em Pesquisa
PROEJA	Programa de Educação para Jovens e Adultos
PROPPG	Pró-Reitoria de Pesquisa e Pós-Graduação
RBC	Raciocínio Baseado em Casos
RFEPCT	Rede Federal de Educação Profissional Científica e Tecnológica
RNA	Rede Neural Artificial
SciELO	*Scientific Electronic Library Online*
SGBD	Sistema Gerenciador de Banco de Dados
SIAPE	Sistema Integrado de Administração de Recursos Humanos
SQL	*Structured Query Language*
SUAP	Sistema Unificado de Administração Pública

TI	Tecnologia da Informação
UFLA	Universidade Federal de Lavras
URL	*Universal Resource Locator*
WEKA	*Waikato Environment for Knowledge Analysis*
XML	*Extensible Markup Language*

1. INTRODUÇÃO

O conhecimento para uma organização, seja ela do ramo educacional ou comercial, tem sido reconhecido como um dos mais importantes recursos. A Gestão do Conhecimento (GC) abrange toda a forma de gerar, armazenar, distribuir e utilizar o conhecimento, tornando necessária a utilização de tecnologias de informação para facilitar o processo, devido ao grande aumento no volume de dados nas últimas décadas.

As instituições de ensino superior (IES) são organizações voltadas para o conhecimento. Ao longo dos últimos anos, diversos autores vêm discutindo como avaliar a qualidade dos serviços prestados por essas instituições e nunca se questionou tanto a qualidade e os valores cobrados por esses serviços. Cabe às próprias IES gerarem soluções para gestão de políticas de ciência, tecnologia e inovação, que tenham um horizonte maior de planejamento a partir dessa enorme massa de dados ainda subutilizados (CARDOSO e MACHADO, 2008).

Devido à variada dinâmica produtiva existente no setor de Ciência, Tecnologia e Inovação (CT&I), gestores frequentemente necessitam de instrumentos informacionais capazes de monitorar e apontar as principais tendências do setor com a finalidade de formulação de políticas e estratégias de ação. A partir desta realidade, a criação e o uso de indicadores da atividade científica se inserem no contexto da CT&I como fatores indutores para o processo de tomada de decisão em instâncias governamentais e políticas, já que apontam o grau de maturidade, evolução e retrocessos de segmentos ligados ao setor (SANTANA *et al.*, 2011).

A informação está se tornando cada vez mais a principal matéria-prima de grandes organizações. Por isso faz-se necessário a aplicação de processos que acelerem a extração de informações de grandes bases de dados. Neste contexto, o processo KDD (*Knowledge Discovery in Databases* – Descoberta de Conhecimento

em Banco de Dados) pode ser utilizado para auxiliar a descoberta de conhecimento útil em grandes bases de dados (FAYYAD *et al.*, 1996).

Uma das principais etapas do processo de KDD, a Mineração de Dados (MD), também conhecida como *Data Mining*, é uma técnica de descoberta de conhecimento em grandes bases de dados. Ela pode prever tendências e comportamentos futuros, através da análise de dados já existentes, munindo os gestores de informações que auxiliem a tomada de decisões.

Portanto, pretende-se com este trabalho extrair conhecimento que identifique o(s) perfil(is) da produção científica dos docentes do Instituto Federal de Goiás (IFG) utilizando a técnica de Mineração de Dados. Os dados da produção científica analisados foram os que se encontravam disponíveis no Currículo Lattes (CL) dos docentes, pois a instituição ainda não dispõe de um sistema de informação que armazene e controle tal produtividade.

1.1 JUSTIFICATIVA

Na atual sociedade do conhecimento reconhece-se que o desenvolvimento de um país está estritamente ligado à Ciência, Tecnologia e Inovação (CT&I). Por isso, vem crescendo o interesse por estudos e análises que permitam visualizar algumas características de um determinado ramo da ciência, com vistas a aceitá-la como elemento propulsor de desenvolvimento da civilização na qual esta pode vir a afetar as estruturas sociais, econômicas e políticas de um país. O reconhecimento da ciência como estrutura chave para o desenvolvimento do país levou à busca incessante pela compreensão da sua natureza objetivando avaliar seus resultados e traçar novas políticas de investimentos (MORAIS e MACHADO, 2011).

Um dos grandes desafios da contemporaneidade é transformar o conhecimento em ações que beneficiem a comunidade e atendam às demandas da sociedade científica. A produção científica reveste-se no conjunto das atividades universitárias, porque é através dela que o conhecimento produzido no interior da universidade é disseminado, levando-o até as comunidades alternativas para um desenvolvimento sustentável. Nesta perspectiva, observa-se que os resultados obtidos na elaboração das pesquisas é, cada vez mais, foco de estudos e um dos principais instrumentos para a avaliação da ciência (MORAIS e MACHADO, 2011).

Segundo Cardoso e Machado (2008), a partir da observação de padrões nos dados estudados, decisões podem ser tomadas a curto e longo prazo, de modo a possibilitar a criação de indicadores para efeito comparativo entre as instituições de ensino superior (IES) e de apoio à gestão da política científica e tecnológica, aperfeiçoando o sistema de ensino superior do país.

Pasta (2011) afirma que para gerar conhecimento, não basta apenas ter a informação. As IES podem hoje serem consideradas como organizações. Uma das funções das Instituições de Ensino é a geração e disseminação de conhecimento, obtido por meio do processo de ensino e aprendizagem e, para que este processo aconteça numa forma dinâmica e eficaz, as IES estão cada vez mais buscando subsídios, ferramentas e técnicas para alcançar esse objetivo. Consequentemente, todo este conhecimento acumulado pode e deve ser utilizado para que, cada vez mais, as instituições busquem disponibilizar aos seus gestores, informações precisas e eficazes para tomada de decisões.

Nas instituições de ensino, onde o capital intelectual geralmente está inserido, pode-se utilizar todo o conhecimento descoberto para melhorar a qualidade dos serviços prestados e favorecer ainda mais os recursos para que mais conhecimentos aliados à sabedoria e à experiência sejam disseminados não só no meio acadêmico como para toda a comunidade. Esse investimento é importante, considerando também a forte competição dentro do mercado educacional, onde o número de instituições e

cursos criados a cada ano vem se tornando cada vez maior (OLIVEIRA e GARCIA, 2004).

Atualmente existe uma grande dificuldade das instituições de ensino para obter dados sobre a produção científica de seus colaboradores. A maior parte destes dados se encontra disponível publicamente na Plataforma Lattes (PL), sendo portanto, uma rica fonte de informações sobre a produção científica, tecnológica e bibliográfica dos pesquisadores do Brasil.

A utilização de técnicas de Mineração de Dados (MD) pode auxiliar na obtenção de informações importantes da PL e com isso orientar melhor gestão e investimento na área de Ciência e Tecnologia (C&T) da instituição. Porém, para Morais (2010), a determinação de padrões que são realmente úteis ainda requer uma grande interação com analistas humanos, o que torna o processo de extração do conhecimento uma tarefa não trivial.

Portanto, a realização deste trabalho justifica-se pela necessidade de se obter conhecimento organizacional em gestão de C&T para o IFG de modo a subsidiar as atividades relacionadas a esta atividade.

1.2 OBJETIVOS

O objetivo geral do trabalho é aplicar o processo de Descoberta de Conhecimento em Banco de Dados, sobretudo a etapa de Mineração de Dados, em um estudo de caso, para identificar padrões que representam o perfil da produção científica dos docentes do IFG. Pretende-se obter conhecimento real sobre a produtividade dos docentes, a partir dos dados dos currículos Lattes (CL), e provê-los à Pró-Reitoria de Pesquisa e Pós-Graduação (PROPPG) do IFG para auxiliar na sua gestão, como por exemplo, na definição de políticas de fomento à pesquisa e de

critérios para concessão de incentivos à mesma. Para atingir este objetivo, as tarefas de classificação e associação foram utilizadas.

1.3 ORGANIZAÇÃO DO TRABALHO

O presente trabalho é composto de seis capítulos. O Capítulo 2 abrange a fundamentação teórica, apresentando os conceitos da Gestão do Conhecimento (GC), do processo de Descoberta de Conhecimento em Banco de Dados (KDD), incluindo a etapa da Mineração de Dados (MD) com suas principais tarefas e técnicas, e a metodologia aplicada no estudo de caso, denominada *Cross Industry Standard Process for Data Mining* (CRISP-DM). O capítulo finaliza com a apresentação de alguns trabalhos correlatos à proposta deste.

O Capítulo 3 aborda a fundamentação teórica relativa ao contexto do domínio da informação utilizada neste trabalho: a produção científica, a Plataforma Lattes (PL), a estrutura do currículo Lattes (CL) e o Instituto Federal de Goiás (IFG).

O Capítulo 4 descreve os materiais e métodos empregados na pesquisa.

O Capítulo 5 apresenta o estudo de caso propriamente dito, estruturado de acordo com as seis etapas da metodologia CRISP-DM, onde em cada fase são descritos os processos realizados para a descoberta de conhecimento.

Por fim, o Capítulo 6 apresenta as conclusões obtidas na pesquisa, algumas dificuldades encontradas, sugestões para trabalhos futuros e contribuições do mesmo.

2. REFERENCIAL TEÓRICO

Este capítulo visa apresentar os conceitos referentes ao contexto da Gestão de Conhecimento, a Descoberta de Conhecimento em Banco de Dados (KDD), a Mineração de Dados (MD) e suas principais tarefas e técnicas. O capítulo aborda a importância e relevância da adesão de um processo para direcionar a extração de conhecimento mediante o grande volume de informações acumuladas nos bancos de dados atuais. O esclarecimento de tais conceitos é imprescindível para os fins metodológicos deste trabalho.

2.1 GESTÃO DE CONHECIMENTO

De acordo com Tarapanoff (2001, p. 36), as mudanças que vêm ocorrendo nas organizações atualmente convergem para a quebra de um paradigma histórico e, por meio dele, entramos na era da sociedade da informação e do conhecimento. A informação como principal matéria-prima das organizações é um insumo comparável à energia que alimenta um sistema; o conhecimento é utilizado na agregação de valor a produtos e serviços; a tecnologia constitui um elemento vital para as mudanças, em especial o emprego da tecnologia sobre acervos de informação. A rapidez, a efetividade e a qualidade constituem fatores decisivos de competitividade.

A informação é um recurso cada vez mais valorizado como viabilizador de decisões e de processos de conhecimento/inteligência nos mais diferentes campos e demandam novas teorias, novas habilidades de pensamento, novas capacidades para transformar dados caóticos em informação útil e novos níveis de inovação que sejam capazes de desenvolver aplicações práticas para informação obtida do ambiente interno ou externo à organização (MARCONDES, 2001).

As empresas estão preocupadas com o valor da informação durante o seu processo decisório. Por isso, usam a Gestão do Conhecimento (GC) com o objetivo de estruturá-lo por meio da utilização de diversas práticas que auxiliem na coleta das informações externas ao seu ambiente (AZARIAS *et al.*, 2009).

A Gestão do Conhecimento é a área que estuda o modo como as organizações entendem o que elas conhecem, o que elas necessitam conhecer e como elas podem tirar o máximo proveito do conhecimento (CARVALHO, 2000). Como o processo de GC é abrangente e complexo, torna-se necessária a utilização de tecnologias da informação, principalmente no que se refere à análise de grande quantidade de informação que é armazenada (CARDOSO e MACHADO, 2008).

Antes de chegar a uma definição do que seja gerenciar o conhecimento, é necessário fazer a distinção entre dado, informação e conhecimento. De acordo com Cardoso e Machado (2008), dados são fatos, imagens ou sons que podem ou não ser úteis ou pertinentes para uma atividade particular. São abstrações formais quantificadas, que podem ser armazenadas e processadas por computador. Informações são dados contextualizados, com forma e conteúdo apropriados para um uso particular. São abstrações informais (não podem ser formalizadas segundo uma teoria matemática ou lógica) que representam, por meio de palavras, sons ou imagens, algum significado para alguém. Conhecimento é uma combinação de instintos, ideias, informações, regras e procedimentos que guiam ações e decisões; tem embutido em si valores como sabedoria e *insights*. É a inteligência obtida pela experiência. Como exemplo, pode-se citar a experiência que um funcionário possui por ter trabalhado em determinadas atividades numa organização por muito tempo.

Para Goldschmidt e Passos (2005), os dados podem ser interpretados como itens elementares, captados e armazenados por recursos da Tecnologia da Informação (TI). As informações representam os dados processados, com significados e contextos bem definidos; por fim, está o conhecimento, que é considerado um

padrão ou conjunto de padrões, cuja formulação pode envolver e relacionar dados e informações.

Para se chegar à informação e ao conhecimento, podem ser consideradas algumas etapas: ao se atribuir algum significado a um dado, este se transforma em uma informação; se forem elaboradas interpretações de um conjunto de informações e de como estas podem ser utilizadas, constitui-se um conhecimento.

O conhecimento organizacional pode ser classificado em dois tipos: explícito e tácito. O conhecimento explícito pode ser articulado na linguagem formal, sobretudo em afirmações gramaticais, expressões matemáticas, especificações, manuais e assim por diante. Esse tipo de conhecimento pode ser então transmitido, formal e facilmente, entre os indivíduos. O conhecimento tácito é difícil de ser articulado na linguagem formal. É o conhecimento pessoal incorporado à experiência individual e envolve fatores intangíveis como crenças pessoais, perspectivas, sistemas de valor, *insights*, intuições, emoções, habilidades. Só pode ser avaliado por meio da ação (NONAKA e TAKEUCHI, 1997).

Considera-se os conhecimentos explícito e o tácito, unidades estruturais básicas que se complementam. Mais importante, a interação entre essas duas formas de conhecimento é a principal dinâmica da criação do conhecimento em uma organização. A criação do conhecimento organizacional é um processo em espiral em que a interação ocorre iterativamente (TARAPANOFF, 2001, p. 135).

Segundo Nonaka e Takeuchi (1997), para se tornar uma empresa que gera conhecimento a organização deve completar uma “espiral do conhecimento”, apresentada na Figura 1, espiral esta que vai de tácito para tácito, de explícito a explícito, de tácito a explícito, e finalmente, de explícito a tácito. Logo, o conhecimento deve ser articulado e então internalizado para tornar-se parte da base de conhecimento de cada pessoa. A espiral começa novamente depois de ter sido completada, porém em patamares cada vez mais elevados, ampliando assim a aplicação do conhecimento em outras áreas da organização.

As formas de interação entre o conhecimento tácito e o explícito, e entre o indivíduo e a organização, acontecem por meio de quatro processos principais da conversão do conhecimento que, juntos constituem a criação do conhecimento.

Figura 1 - Espiral do Conhecimento

Fonte: NONAKA e TAKEUCHI (1997, pág. 80)

.

Os quatro processos apresentados na Figura 1 são:

- Socialização é o compartilhamento do conhecimento tácito, por meio da observação, imitação ou prática (tácito para tácito);
- Externalização é a conversão do conhecimento tácito em explícito e sua comunicação ao grupo. O conhecimento se torna explícito expresso em forma de analogias, conceitos, hipóteses ou modelos (tácito para explícito);
- Combinação é o processo de padronização do conhecimento e sua documentação em um manual ou guia de trabalho a ser incorporado a um produto. O modo de conversão do conhecimento envolve a combinação de conjuntos diferentes de conhecimentos explícitos (explícito para explícito);
- Internalização é o processo de incorporação do conhecimento explícito em conhecimento tácito. Ocorre quando novos conhecimentos explícitos são compartilhados na organização e outras pessoas começam a utilizá-los para

aumentar, estender e reenquadrar seu próprio conhecimento tácito (explícito para tácito).

Na medida em que o conhecimento, tanto o tácito quanto o explícito, se torna um ativo central, produtivo e estratégico, o sucesso da organização depende cada vez mais da sua habilidade em coletar, produzir, manter e distribuir conhecimento (CARDOSO e MACHADO, 2008).

A Gestão do Conhecimento nas organizações é vista como a capacidade que as mesmas possuem, por meio de seus processos, de criar e utilizar o conhecimento. Para tal, é necessário o planejamento de pessoas, cultura, processos e tecnologias de modo que trabalhem em conjunto para atender as necessidades dos colaboradores, facilitando o aprendizado coletivo e o desenvolvimento de uma organização mais estruturada na geração do conhecimento (TERRA e ALMEIDA, 2003).

A Gestão do Conhecimento pode ser vista, então, como o conjunto de atividades que busca desenvolver e controlar todo tipo de conhecimento em uma organização, visando à utilização na consecução de seus objetivos. Esse conjunto de atividades deve ter, como principal meta, o apoio ao processo decisório em todos os níveis. Para isso, é preciso estabelecer políticas, procedimentos e tecnologias que sejam capazes de coletar, distribuir e utilizar efetivamente o conhecimento, bem como representar fator de mudança no comportamento organizacional (TARAPANOFF, 2001).

Para a criação de conhecimento explícito, diversas técnicas de descoberta de conhecimento podem ser utilizadas pelas organizações. Um dos maiores problemas enfrentados atualmente é o grande volume das bases de dados que as organizações possuem. O KDD pode ser utilizado como solução para este problema (TARAPANOFF, 2001).

2.2 DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS (KDD)

Para Cardoso e Machado (2008) a necessidade de informações disponíveis vem crescendo nos últimos anos e vários fatores contribuíram para esse aumento. O baixo custo de armazenagem pode ser visto como a principal causa do surgimento dessas enormes bases de dados. Outro fator é a disponibilidade de computadores de alto desempenho a um custo razoável. Como consequência, bancos de dados passam a conter verdadeiros tesouros de informação e, devido ao seu volume, ultrapassam a habilidade técnica e a capacidade humana na sua captação e interpretação.

As organizações têm se mostrado bastante eficientes em capturar, organizar e armazenar grandes quantidades de dados, obtidos de suas operações diárias. Porém, a maioria delas ainda não usa adequadamente essa grande massa de dados para transformá-la em conhecimentos que possam ser utilizados em suas próprias atividades. Com a geração de um volume cada vez maior de informação, é essencial tentar aproveitar o máximo possível desse investimento (AMORIM, 2006).

A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas. Portanto, torna-se imprescindível o desenvolvimento de ferramentas que auxiliem o homem, de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados para que se possa desenvolver e selecionar estratégias de ação em cada contexto de aplicação (GOLDSCHMIDT e PASSOS, 2005).

Num ambiente mutável, como o das organizações na atualidade, torna-se necessária a aplicação de técnicas e ferramentas automáticas que agilizem o processo de extração de informações relevantes de grandes volumes de dados. Uma metodologia emergente, que tenta solucionar o problema da análise de grandes

quantidades de dados e ultrapassa a habilidade e a capacidade humanas, é o *Knowledge Discovery in Databases* - KDD (CARDOSO e MACHADO, 2008).

Goldschmidt e Passos (2005, p. 2) também afirmam que para atender a este novo contexto, surge a área denominada *Knowledge Discovery in Databases* - KDD (em português, o acrônimo é o DCBD), que vem despertando grande interesse junto às comunidades científica e industrial.

A utilização do KDD acontece principalmente em organizações em que grandes bancos de dados são formados e há uma percepção do valor da sua análise de forma eficaz. Sem essa interpretação correta dos dados, importantes análises e decisões deixam de ser amparadas pelas informações contidas nos registros da empresa e passam a ser feita apenas na intuição de um tomador de decisão.

Os bancos de dados presentes nas organizações são utilizados para extrair informações úteis, mas o processo KDD provê mais do que isso, pois permite a determinação de padrões e modelos de forma que estes sejam os alicerces para a construção de conhecimento. Segundo Lemos (2003), isso ocorre porque o KDD é um processo interdisciplinar, que envolve diversas áreas do conhecimento que completam o processo de transformação dos dados, dentre elas estão: o aprendizado de máquina, bases de dados, a matemática e a estatística, sistemas especialistas e visualização de dados, como ilustra a Figura 2.

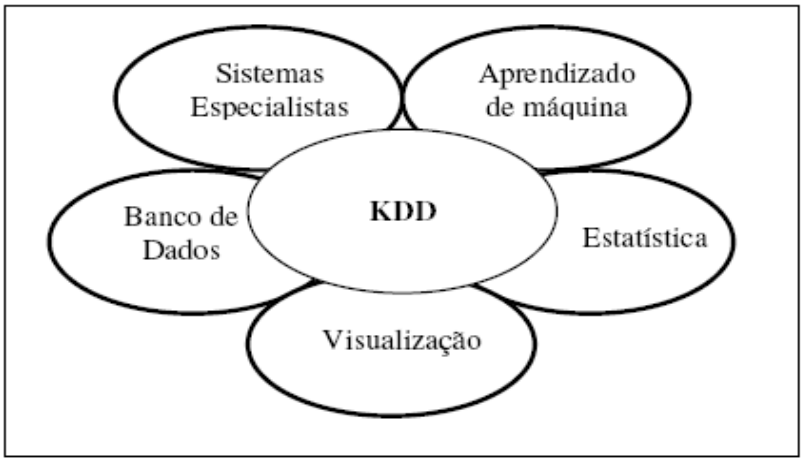


Figura 2 - Característica multidisciplinar do KDD

Fonte: Adaptado de LEMOS (2003)

O mecanismo de descoberta de conhecimento em Mineração de Dados (MD) consiste em uma série de etapas, iniciando com a definição dos objetivos para os quais é aplicado novo conhecimento até a exposição do mesmo a alta direção da organização como apoio a tomada de decisão. A MD propriamente dita é apenas uma destas etapas, conforme será mostrado na Figura 3.

No âmbito do KDD e da Mineração de Dados, é notável o trabalho pioneiro de Fayyad *et al*. (1996), resultado da compilação de vários artigos descritos por uma série de *Workshops* que foram realizados entre os anos de 1989 e 1994, que abordavam processos, modelos de classificação, clusterização e perspectivas estatísticas (ARAÚJO, 2007).

Segundo Fayyad *et al.* (1996, p. 6) o KDD é um processo que tem por objetivo a descoberta de conhecimento em banco de dados, sendo um processo não trivial de identificação de padrões, a fim de extrair informações implícitas e potencialmente úteis.

Para Fayyad *apud* Carvalho (2005a), os termos *Data Mining* e KDD (ou DCBD) muitas vezes são confundidos como sinônimos para identificar o processo de descoberta de conhecimento útil a partir de bancos de dados. O termo KDD foi estabelecido no primeiro *workshop* de KDD em 1989 para enfatizar que conhecimento é o produto final de uma descoberta baseada em dados.

Ainda no entendimento de Fayyad *et al.* (1996), o KDD se refere a todo o processo de descoberta de conhecimento mediante o processamento de dados, e envolve também a limpeza e preparação, incorporação de conhecimento e apresentação de resultados. O processo de KDD é interativo e iterativo, pois requer o envolvimento dos utilizadores nas tomadas de decisão e permite voltar às etapas anteriores. O processo é desenvolvido por uma sequência de etapas (Figura 3), não necessariamente por uma ordem linear, podendo ser realizadas por diversas vezes, bem como voltar aos passos anteriores para uma revisão do processo. A obtenção de

informação e de conhecimento válido e potencialmente útil só é possível com uma participação ativa dos utilizadores, ao nível das tomadas de decisão nas diferentes etapas que compõem o processo.

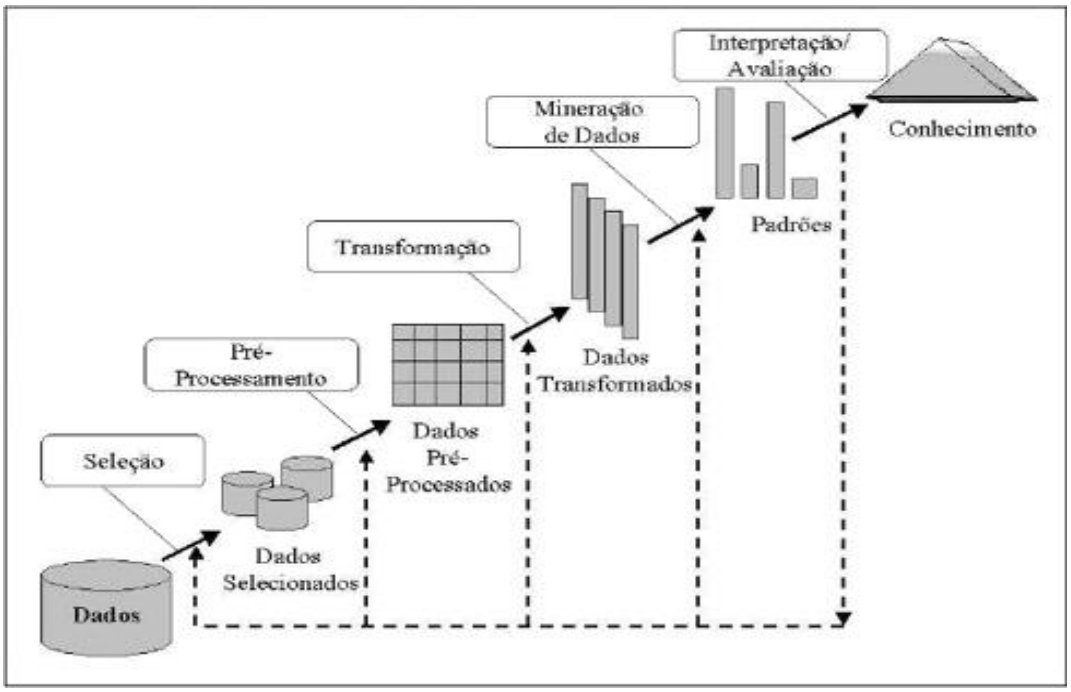


Figura 3 - Etapas do processo KDD

Fonte: FAYYAD *et al.* (1996, p. 10)

Damasceno (2010) explica as etapas da Figura 3, referentes ao processo KDD descrito por Fayyad *et al.* (1996), conforme descrito abaixo.

Na primeira etapa é necessária a definição dos objetivos do usuário, ou seja, definir o tipo de conhecimento que se deseja extrair do banco de dados. Após o levantamento de requisitos, deve-se criar o conjunto de dados no qual o processo irá trabalhar. Este conjunto de dados deve conter todas as informações necessárias para que os algoritmos de mineração possam alcançar seu objetivo. Essa etapa é conhecida como **Seleção**.

A segunda etapa é composta por tarefas de **Pré-Processamento**. Técnicas de pré-processamento são responsáveis pela remoção de ruídos (erros e exemplos fora do padrão), pela definição de estratégias para lidar com valores faltosos e pela formatação dos dados de acordo com os requisitos da ferramenta de mineração.

A terceira etapa, conhecida como **Transformação**, tem por objetivo localizar características úteis para representar os dados. É responsável também pela seleção dos melhores exemplos e atributos presentes no conjunto de dados. Após os dados terem sido limpos e pré-processados, aplica-se as técnicas de Mineração de Dados para alcançar os objetivos definidos na primeira etapa. Os objetivos identificados podem ser descritos como tarefas de classificação, regressão, agrupamento, predição, etc. É necessário escolher qual algoritmo de mineração deve ser utilizado após a determinação de qual tarefa de mineração será executada. As técnicas são escolhidas de acordo com as características dos dados e com os requisitos apresentados pelos usuários. Algumas técnicas de mineração contêm parâmetros que são utilizados em seu funcionamento, também faz parte desta etapa encontrar os melhores parâmetros, para que o método possa ser o mais preciso e ágil possível. Somente após todas essas tarefas terem sido realizadas, é hora da execução propriamente dita do algoritmo de mineração.

Na quarta etapa da **Mineração de Dados**, é que o algoritmo irá procurar por padrões utilizando as suas estratégias e todos os dados que foram informados.

A quinta etapa é a de **Interpretação/Avaliação** dos padrões identificados. Este passo inclui visualizar os padrões extraídos ou os modelos que resumem a estrutura e as informações presentes nos dados. Além da visualização, são utilizadas medidas tanto técnicas quanto subjetivas para avaliar os padrões extraídos. As medidas técnicas são informações referentes à precisão, erro médio, erro quadrático e taxas de falsos positivos e falsos negativos. Medidas subjetivas são referentes a informações como utilidade, entendimento ou complexidade dos padrões extraídos.

Ao final de todo o processo tem-se o “Conhecimento” em forma de padrões. Sendo assim, pode-se utilizar os padrões extraídos para os quais eles foram desejados. Os padrões podem ser utilizados sozinhos ou embutidos em outros sistemas.

A fase de preparação dos dados (que vai até a etapa de Transformação da Figura 3) absorve uma boa parte do tempo do processo *Knowledge Discovery in Databases* (KDD), consumindo aproximadamente 70%, além de ser uma fase de grande importância, pois nela são identificados os dados relevantes para a solução satisfatória do problema. Na verdade, apenas ter os dados não é suficiente; é necessário que eles estejam suficientemente corretos, adequados e tenham sido corretamente selecionados para que preencham todas as características desejadas. Para que um processo KDD obtenha sucesso, é necessário que os dados estejam disponíveis para o processamento e em condições de serem utilizados (AMARAL, 2001, p. 17).

Para que seja possível realizar a análise de grandes volumes de dados durante a fase de preparação dos dados, é importante mencionar outra área correlata à atividade de Mineração de Dados chamada *Data Warehouse* (DW). A criação de um DW é considerada como um dos primeiros passos para viabilizar a análise em grandes massas de dados (REZENDE, 2005, p. 308). Segundo Tarapanoff (2001, p. 271) o DW é um aliado no sentido de tornar a MD mais eficiente, além da análise de transações individualizadas, é possível tirar proveito da agregação e sumarização de coleções de dados-alvo.

Um *Data Warehouse* é uma forma de organizar os dados corporativos em um banco de dados paralelo, onde os dados se encontram bem estruturados e consolidados. Além de estar perfeitamente integrado ao ambiente operacional (que utilizam dados tipicamente armazenados, recuperados e atualizados pelos sistemas de informação da empresa), ele permite a segmentação dos dados e introduz mais atributos que podem aumentar o nível de informação, principalmente levando em conta o contexto histórico dos dados.

De acordo com Inmon e Hackathorn (1997), considerados pioneiros sobre o tema, o DW é uma coleção de dados integrados, orientados por assunto, variáveis com o tempo e não voláteis, usados para suporte ao processo gerencial de tomada de

decisões. O objetivo de um DW é fornecer uma imagem única da realidade do negócio. De uma forma geral, sistemas de *Data Warehouse* compreendem um conjunto de programas que extraem dados do ambiente de dados operacionais da empresa, um banco de dados que os mantém, e sistemas que fornecem estes dados aos seus usuários.

2.3 MINERAÇÃO DE DADOS (MD)

A MD é uma fase do processo KDD com objetivo de analisar informações de grandes conjuntos de dados no intuito de descobrir correlações e padrões que sejam úteis para os patrocinadores do projeto. Pode ser realizado de forma automática ou mais frequentemente, de forma semi-automática. Aborda a resolução de problemas através de análises de dados já presentes em banco de dados e os padrões descobertos devem ser significativos, na medida em que leva a alguma vantagem, normalmente econômica (WITTEN e FRANK, 2005).

Para Cabena *et al.* (1998), a definição de Mineração de Dados é dada de uma perspectiva de banco de dados: "Mineração de Dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados".

Para Tan *et al.* (2009, p. 4), a Mineração de Dados (MD) é uma parte integral do KDD, que é o processo geral de conversão de dados brutos em informações úteis, conforme mostrado na Figura 4. Este processo consiste de uma série de passos de transformação, do pré-processamento dos dados até o pós-processamento dos resultados da MD.

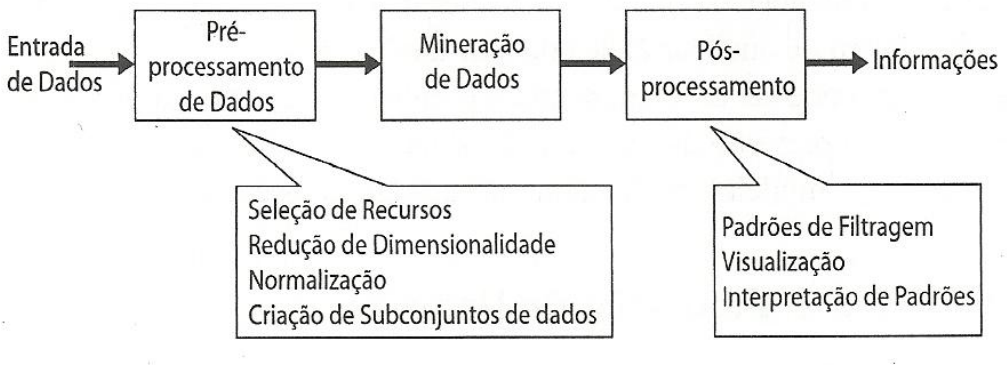


Figura 4 – Mineração de Dados no processo de KDD

Fonte: TAN *et al.* (2009, p.4)

Os dados de entrada podem ser armazenados em uma diversidade de formatos e podem ficar em um repositório central de dados ou serem distribuídos em múltiplos locais. O propósito do **pré-processamento** é transformar os dados de entrada brutos em um formato apropriado para análises subsequentes. Os passos envolvidos no pré-processamento de dados incluem a fusão de dados de múltiplas fontes, a limpeza dos dados para remoção de ruídos, observações duplicadas, a seleção de registros e características que sejam relevantes à tarefa de Mineração de Dados. Por causa das muitas formas através das quais os dados podem ser coletados e armazenados, o pré-processamento de dados talvez seja o passo mais trabalhoso e demorado no processo geral de descoberta de conhecimento.

"Fechar o laço" é a descoberta frequentemente usada para se referir ao processo de integrar os resultados da MD com os sistemas de apoio a decisões. Por exemplo, em aplicações de negócio, a compreensão permitida pelos resultados da MD pode ser integrada com ferramentas de administração de campanha de forma que promoções eficazes de vendas possam ser realizadas e testadas. Tal integração requer um passo de **pós-processamento** que assegure que apenas resultados válidos e úteis sejam incorporados ao sistema de apoio a decisões (TAN *et al.*, 2009).

Segundo Fayyad *et al.* (1996), embora a MD seja uma tecnologia poderosa na descoberta de informações ocultas nos bancos de dados, ela não elimina a

necessidade de conhecimento do negócio e o entendimento dos dados. Além disso, requer o entendimento da ferramenta escolhida, bem como do algoritmo utilizado na busca dos dados. Em consequência disto é necessário saber que tipo de informação e conhecimento se quer obter a partir da base de dados disponível.

Carvalho (2005b) afirma que ainda que as técnicas da MD sejam antigas, foi apenas nos últimos anos que passaram a ser usadas como exploração de dados, por vários motivos:

- **o volume de dados disponível atualmente é enorme: a** Mineração de Dados é uma técnica que só se aplica a grandes massas de dados, pois necessita disto para calibrar seus algoritmos e extrair dos dados conclusões confiáveis. Empresas de telefonia, cartões de crédito, bancos, entre outras, vem gerando a cada dia uma grande quantidade de dados sobre seus serviços e clientes. Estes dados são passíveis de análise por mineração;
- **os dados estão sendo organizados:** Com a tecnologia do *Data Warehouse*, os dados de várias fontes estão sendo organizados e padronizados de forma a possibilitar sua organização dirigida para o auxílio à decisão. As técnicas de MD necessitam de bancos de dados limpos, padronizados e organizados;
- **os recursos computacionais estão cada vez mais potentes:** A Mineração de Dados (MD) necessita de muitos recursos computacionais para operar seus algoritmos sobre grandes quantidades de dados. O aumento da potência computacional, devido ao avanço tecnológico e à queda dos preços dos computadores, facilita o uso da MD atualmente. O avanço da área de banco de dados, construindo bancos de dados distribuídos, também auxiliou em muito à MD;
- **a competição empresarial exige técnicas mais modernas de decisão:** As empresas da área de finanças, telecomunicações e seguro experimentam a cada dia mais competição. Como estas empresas sempre detiveram em seus bancos de dados uma enorme quantidade de informação, é natural que a MD

tenha se iniciado dentro de seus limites. Atualmente, outras empresas buscam adquirir dados para analisar melhor seus caminhos futuros através dos sistemas de apoio à decisão. Para empresas de serviços, a aquisição de dados é importante, pois precisam saber que serviço oferecer a quem. Para outras empresas, até a venda das informações pode ser um produto;

- **programas comerciais de MD já podem ser adquiridos:** As técnicas de MD são antigas conhecidas da Inteligência Artificial (IA), porém somente recentemente saíram dos laboratórios para as empresas. Alguns pacotes já podem ser encontrados no comércio, contendo algumas destas técnicas. As técnicas mais recentes, no entanto, ainda se encontram no campo acadêmico, sendo necessário que a empresa se dirija a uma universidade que realize pesquisa para obter ajuda.

Conforme Polito (1997) *apud* Preto e Silveira (2010), MD é a técnica que permite buscar informações que estejam, aparentemente, escondidas e ajudam a agilizar e/ou fortalecer as tomadas de decisões. Ele ainda afirma que as empresas que empregam MD estão muito a frente das outras, pois são capazes de: (1) Criar parâmetros para entender o comportamento dos consumidores; (2) Identificar afinidades entre as escolhas de produtos e serviços; (3) Prever hábitos de compras; (4) Analisar comportamentos habituais para se evitar fraudes.

As informações sobre a relação entre dados e, posteriormente a descoberta de novos conhecimentos, podem ser muito úteis para realizar atividades de tomada de decisão. Por exemplo, ao minerar os dados de um estoque de supermercado poderia-se descobrir que todas as sextas-feiras uma marca específica de cerveja se esgota nas prateleiras e, portanto, um gerente que obtém esta “nova informação” poderia planejar o estoque do supermercado para aumentar a quantidade de cervejas desta marca às sextas-feiras. Analogamente, é possível minerar dados de alunos para verificar a relação entre uma abordagem pedagógica e o aprendizado do aluno. Através desta informação o professor poderia compreender se sua abordagem

realmente está ajudando o aluno e desenvolver novos métodos de ensino mais eficazes. A MD tem sido aplicada em diversas áreas do conhecimento, como por exemplo, vendas, bioinformática, e ações contra-terrorismo (BAKER *et al.*, 2011).

2.3.1 Tarefas de Mineração de Dados

No processo de Mineração de Dados há dois conceitos importantes para o seu melhor entendimento e aplicação: a tarefa e a técnica de mineração a ser utilizada. De acordo com o tipo de informações e conhecimento almejados será feita a escolha do tipo de tarefa e técnica a utilizar.

Segundo Amo (2004), a tarefa consiste na especificação daquilo que se deseja encontrar nos dados, que tipo de regularidades ou categoria de padrões se tem interesse de encontrar, ou que tipo de padrões poderiam surpreender. Já a técnica de mineração consiste na especificação de métodos (algoritmos) que garantam como descobrir padrões interessantes. Redes Neurais e Algoritmos Genéticos são exemplos de técnicas que podem ser utilizadas. Por sua vez, uma técnica pode utilizar de algoritmos para implementar um determinado tipo de tarefa. A interação entre esses elementos está esquematizada na Figura 5.

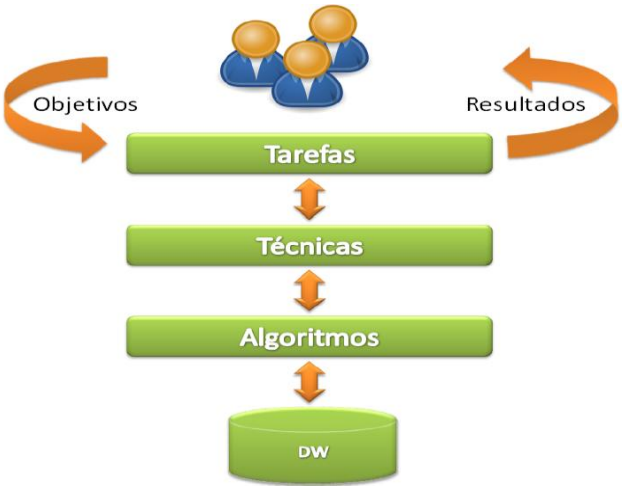


Figura 5 - Interação entre tarefas, técnicas e algoritmos de MD

Fonte: NEGREIROS e LIMA (2009, p. 7)

A definição da tarefa utilizada está condicionada aos objetivos do projeto de mineração e da disposição do domínio dos dados. A Mineração de Dados (MD) também envolve a utilização de diversas técnicas, materializada por algoritmos computacionais, necessárias para realizar as tarefas de mineração.

Tan *et al.* (2009, p.8) afirmam que as tarefas de MD são geralmente divididas em duas categorias principais:

- **Tarefas de Previsão**: o objetivo destas tarefas é prever o valor de um determinado atributo baseado nos valores de outros atributos. O atributo a ser previsto é comumente conhecido como a “variável dependente” ou “alvo”, enquanto que os atributos usados para fazer a previsão são conhecidos como as “variáveis independentes” ou “explicativas”;
- **Tarefas Descritivas**: o objetivo é derivar padrões (correlações, tendências, grupos, trajetórias e anomalias) que resumam os relacionamentos subjacentes nos dados. As tarefas descritivas da Mineração de Dados (MD) são muitas vezes exploratórias em sua natureza e frequentemente requerem técnicas de pós-processamento para validar e explicar os resultados.

A Figura 6 ilustra as duas principais categorias das tarefas de MD e a Figura 7 ilustra quatro das tarefas centrais da Mineração de Dados.

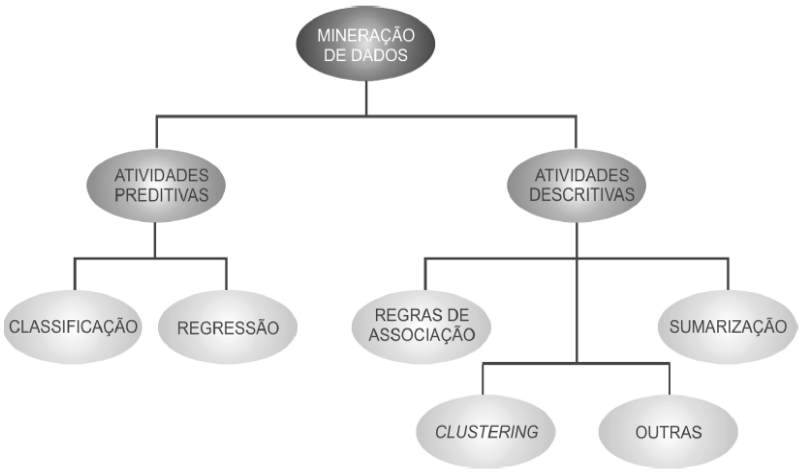
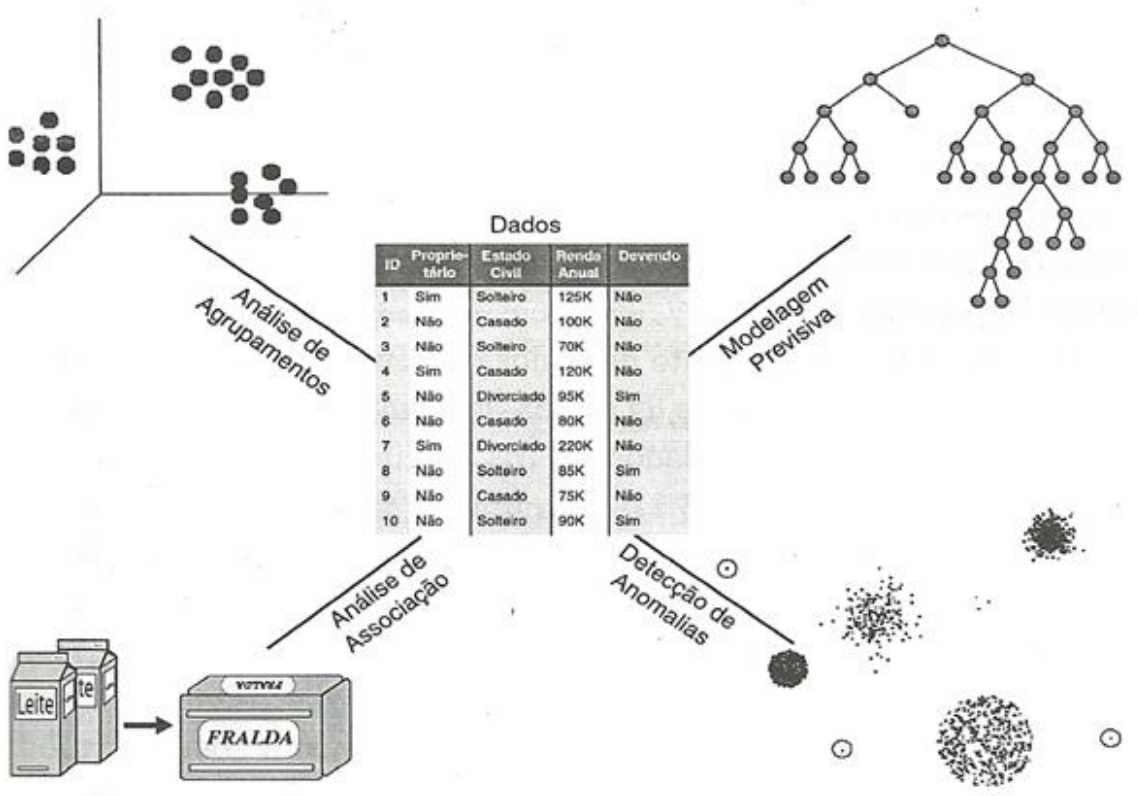


Figura 6 - Categorias das tarefas de Mineração de Dados
Fonte: DOMINGUES (2004, p.18)



ID	Proprietário	Estado Civil	Renda Anual	Devendo
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	80K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

Figura 7 - Quatro das tarefas centrais da Mineração de Dados

Fonte: TAN *et al.* (2009, p.9)

Segundo Pasta (2011), na MD existem dois tipos de aprendizados indutivos chamados de Aprendizagem Supervisionada (AS) e Aprendizagem Não-Supervisionada (ANS). A AS é direcionada a tomada de decisões e é por meio dela que se realizam inferências nos dados com objetivo de realizar predições, nas quais há o uso de atributos para previsão do valor futuro, enquanto na ANS as atividades são descritivas, permitindo a descoberta de padrões e a geração de novos conhecimentos.

Nas tarefas da ANS, o rótulo da classe e o número de classes que serão treinadas são desconhecidos. O objetivo destas tarefas é identificar padrões de comportamento semelhantes nos dados. Exemplos são as tarefas de Associação e de Agrupamento. Nas tarefas da AS trabalha-se com algoritmos preditivos e as classes já são especificadas, como ocorre nas tarefas de Classificação.

2.3.1.1 Associação

Uma regra de associação é um padrão da forma X → Y, em que X e Y são conjuntos de valores, ou seja, encontrar itens que determinem a presença de outros em uma mesma transação e estabelecer regras que correlacionam a presença de um conjunto de itens com outro intervalo de valores para outro conjunto de variáveis. Como exemplo, sempre que se orienta um aluno de doutorado, é publicado algum documento; descobrir regras de associação entre alunos de doutorado e número de publicações pode ser útil para melhorar a distribuição de orientados por professor (CARDOSO e MACHADO, 2008).

A tarefa de associação abrange a busca por itens que frequentemente ocorram de forma simultânea em transações do banco de dados. O exemplo mais clássico e didático da aplicação desta tarefa é o da cesta de compras. Do conjunto de dados da Tabela 1 pode ser extraída a seguinte regra:

{Fraldas → Cerveja}

Tabela 1 - Exemplo de transações de cestas de compras

TID	Itens
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cerveja, Cola}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Cola}

Fonte: TAN *et al.* (2009, p.390)

A confiança da regra sugere que há um relacionamento forte entre a venda de fraldas e cerveja. Durante um processo de descoberta de associações em sua vasta base de dados, uma grande rede de mercados norte-americana descobriu que um número razoável de compradores de fralda também comprava cerveja na véspera de finais de semana com jogos transmitidos pela televisão. Com uma análise mais detalhada sobre os dados, pode-se perceber que tais compradores eram, na realidade, homens que, ao comprarem fraldas para seus filhos, compravam também

cerveja para consumo enquanto cuidavam das crianças e assistiam aos jogos na televisão. Esta empresa utilizou o novo conhecimento para aproximar as gôndolas de fraldas e cervejas na rede de mercados, incrementando assim a venda conjunta dos dois produtos (GOLDSCHMIDT e PASSOS, 2005, p.13).

2.3.1.2 Classificação

Segundo Romão (2002), a classificação é uma das tarefas mais referenciadas na literatura de MD. Seu objetivo é descobrir um relacionamento entre um atributo meta (cujo valor, ou classe, será previsto) e um conjunto de atributos previsores. O sistema deve descobrir este relacionamento a partir de exemplos com classe conhecida. O relacionamento descoberto será usado para prever o valor do atributo meta (ou a classe) para exemplos cujas classes são desconhecidas.

A classificação é um processo de aprendizado em que um objeto é mapeado em uma das classes pré-definidas. A partir de um conjunto de atributos previamente escolhidos, o algoritmo de classificação procura estabelecer relações entre os dados, classificando os registros de acordo com as características de cada um, confrontando-os com as características das classes previamente determinadas. A classificação pode, então, identificar a qual classe este objeto pertence, a partir de seu conteúdo. Para tal, é necessário que as classes tenham sido previamente descritas, expressando suas características por meio de definições, fórmulas e/ou atributos (PAULA, 2004).

Como exemplo da tarefa de classificação, considere uma financeira que possua um histórico com os dados de seus clientes e o comportamento desses clientes em relação ao pagamento de empréstimos contraídos previamente. Considere dois tipos de clientes: os que pagam em dia e os inadimplentes. São as classes do problema. A aplicação da tarefa de classificação consiste em descobrir uma função que mapeie corretamente os clientes, a partir de seus dados, em uma destas classes.

Tal função, uma vez descoberta, pode ser utilizada para prever o comportamento de novos clientes que desejem contrair empréstimos junto à financeira. Essa função pode ser incorporada a um sistema de apoio à decisão que auxilie na filtragem e concessão de empréstimos somente a clientes bons pagadores (GOLDSCHMIDT e PASSOS, 2005, p.13).

Para Larose (2006), diversos outros exemplos de classificação podem ser encontrados: (1) Sistemas bancários: determinar se uma transação de cartão de crédito é fraudulenta ou não; (2) Educação: identificação de necessidades especiais para novos alunos; (3) Medicina: diagnóstico de doenças específicas; (4) Leis: determinar se um testamento foi realmente escrito pela pessoa falecida ou por outra; (5) Segurança: identificar se determinado comportamento financeiro ou pessoal de um indivíduo indica alguma ameaça de terrorista.

2.3.1.3 Agrupamento (*Clustering*)

O agrupamento, clusterização ou ainda análise de grupos é a tarefa que agrupa registros que apresentam similaridades. Os objetos semelhantes são aproximados formando um *cluster*, enquanto os objetos com pouca ou nenhuma semelhança são distanciados e inseridos em outros *clusters.*

O agrupamento é utilizado para separar os registros de uma base de dados em subconjuntos, de tal forma que os elementos de um *cluster* compartilhem de propriedades comuns que os distingam de elementos em outros *clusters*. O objetivo dessa tarefa é maximizar a similaridade *intracluster* e minimizar a similaridade *intercluster*. Diferente da tarefa de classificação, que tem rótulos predefinidos, a clusterização precisa automaticamente identificar os grupos de dados aos quais o usuário deverá atribuir rótulos, como por exemplo, uma empresa de telecomunicações pode realizar um processo de clusterização para obter grupos de clientes que

compartilhem o mesmo perfil de compra de serviços. (GOLDSCHMIDT e PASSOS, 2005, p.14).

Outro exemplo de aplicação de agrupamento é a descoberta de subpopulações homogêneas para consumidores do mercado e identificação de subcategorias do espectro de medidas. A Figura 8 mostra um possível agrupamento desse conjunto de dados. Nesse caso, todos os pontos passam a ser representados por x para indicar que os membros das classes não são mais conhecidos. Outro ponto importante é que os *clusters* se sobrepõem, permitindo assim que pontos pertençam a mais de um *cluster* (AMARAL, 2001, p. 26).

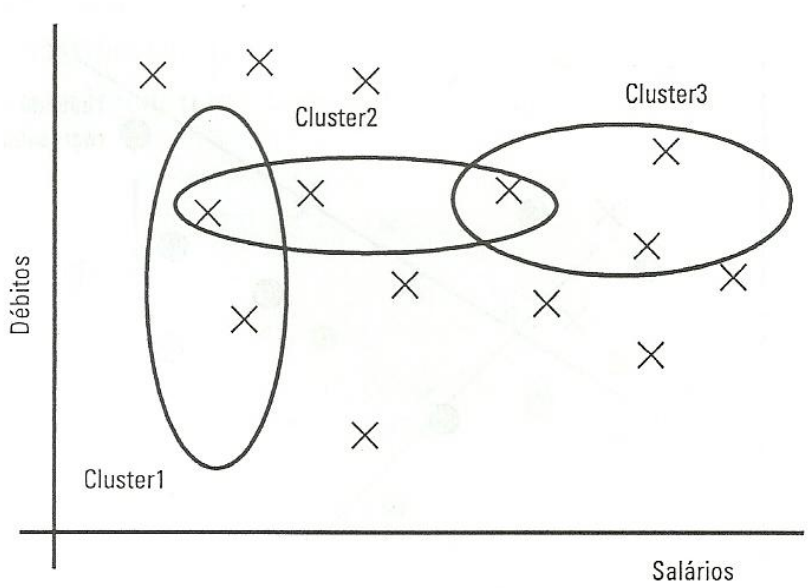


Figura 8 - Divisão do conjunto de dados de empréstimos em três grupos
Fonte: AMARAL, (2001, p.26)

2.3.1.4 Regressão (Estimação)

Para Fayyad *et al.* (1996, p.13) a regressão compreende a busca por uma função que mapeie um item de dado para uma variável de predição real estimada. Como exemplos de tarefas de estimativa, tem-se: a predição da soma da biomassa presente em uma floresta; estimativa da probabilidade de um paciente sobreviver, dado o resultado de um conjunto de diagnósticos de exames; predição do risco de

determinados investimentos; definição do limite do cartão de crédito para cliente em um banco; estimar o número de filhos em uma família; estimar a renda total de uma família; prever a demanda de um consumidor para um novo produto, dentre outros.

A tarefa de regressão é uma técnica de modelagem preditiva onde a variável alvo a ser avaliada é contínua. Seu objetivo é encontrar uma função alvo que possa ajustar os dados de entrada com um erro mínimo (TAN *et al*., 2009, p.873). É similar à tarefa de classificação, porém, é restrita apenas a atributos numéricos. No campo da análise estatística, os métodos de estimativa são largamente utilizados para determinar pontos estatísticos, intervalos de confiança, regressão linear simples, correlações e regressão múltipla (LAROSE, 2006).

Os dados da Tabela 2 correspondem às medidas de fluxo de calor e temperatura da pele de uma pessoa durante o sono. Suponha que se esteja interessado em prever a temperatura da pele de uma pessoa baseado nas medidas de fluxo de calor geradas por um sensor de calor. A Figura 9 ilustra a aplicação da técnica de regressão linear para encontrar a função que relaciona estas duas variáveis.

Tabela 2 - Medidas de fluxo de calor e temperatura da pele de uma pessoa

Fluxo de Calor	Temperatura da Pele	Fluxo de Calor	Temperatura da Pele	Fluxo de Calor	Temperatura da Pele
10.858	31.002	6.3221	31.581	4.3917	32.221
10.617	31.021	6.0325	31.618	4.2951	32.259
10.183	31.058	5.7429	31.674	4.2469	32.296
9.7003	31.095	5.5016	31.712	4.0056	32.334
9.652	31.133	5.2603	31.768	3.716	32.391
10.086	31.188	5.1638	31.825	3.523	32.448
9.459	31.226	5.0673	31.862	3.4265	32.505
8.3972	31.263	4.9708	31.919	3.3782	32.543
7.6251	31.319	4.8743	31.975	3.4265	32.6
7.1907	31.356	4.7777	32.013	3.3782	32.657
7.046	31.412	4.7295	32.07	3.3299	32.696
6.9494	31.468	4.633	32.126	3.3299	32.753
6.7081	31.524	4.4882	32.164	3.4265	32.791

Fonte: TAN *et al.* (2009, p.874)

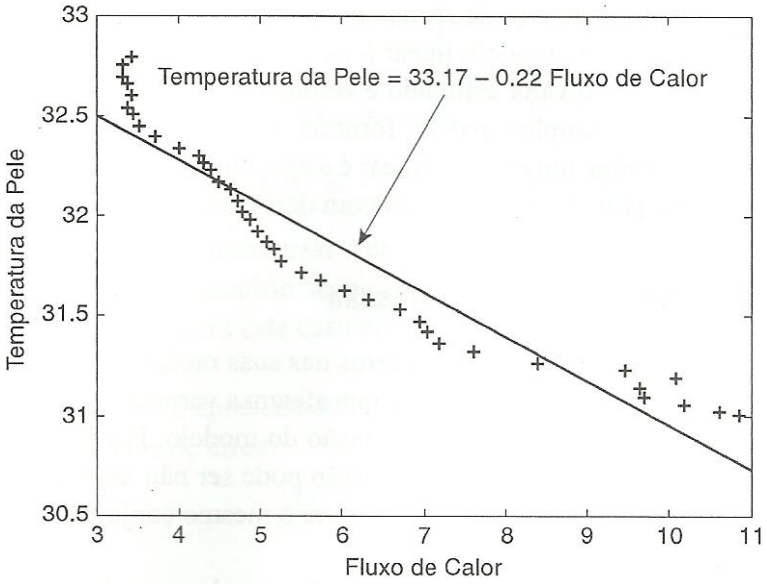


Figura 9 - Modelo linear que ajusta os dados da Tabela 2
Fonte: TAN *et al.* (2009, p.875)

2.3.1.5 Sumarização

Essa tarefa consiste em procurar identificar e indicar características comuns entre conjuntos de dados. Como exemplo, considere um banco de dados com informações sobre clientes que assinam um determinado tipo de revista semanal. A tarefa de sumarização deve buscar por características que sejam comuns a boa parte dos clientes, tais como: são assinantes da revista X, homens na faixa etária de 25 a 45 anos, com nível superior e que trabalham na área de finanças. Tal informação poderia ser utilizada pela equipe de marketing da revista para direcionar a oferta para novos assinantes. É muito comum aplicar a tarefa de sumarização a cada um dos agrupamentos obtidos pela tarefa de clusterização (GOLDSCHMIDT e PASSOS, 2005, p.14).

2.3.1.6 Detecção de Desvios ou *Outliers*

Para Côrtes *et al.* (2002), esta tarefa consiste em procurar conjuntos de dados que não obedecem ao comportamento ou modelo dos dados. Uma vez encontrados, podem ser tratados ou descartados para utilização no processo de MD. Trata-se de uma importante avaliação nos dados no sentido de descobrir probabilidades crescentes de desvios ou riscos associados aos vários objetivos traçados inicialmente na MD. Detectar esses desvios é muito análogo às técnicas utilizadas em análises estatísticas, onde são aplicados testes de significância que assumem uma distribuição, utilizando medidas estatísticas como a média aritmética e o desvio padrão para aferir essas diferenças.

Como exemplo, pode-se avaliar as vendas de uma determinada empresa para verificar o comportamento de suas vendas como um todo, bem como pode-se avaliar suas vendas por produtos, regiões e estados, podendo encontrar outro tipo de comportamento. A Figura 10 identifica visualmente a presença de *outliers*, onde os pontos externos aos polígonos são valores fora dos padrões da população (vendas) observada.

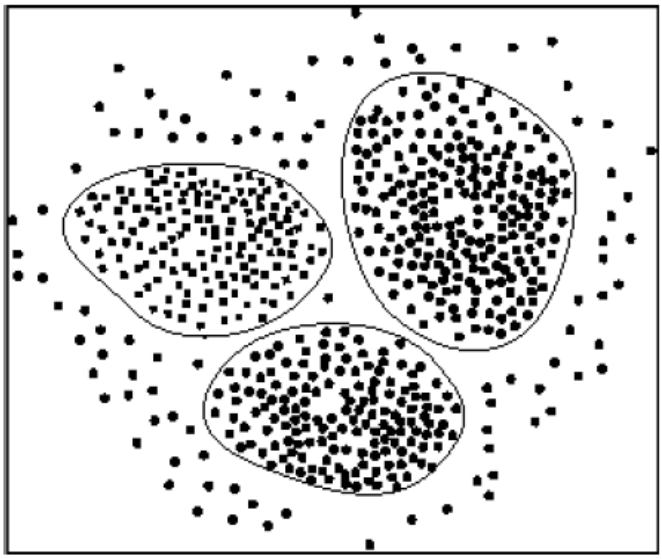


Figura 10 - Detecção de *Outliers* utilizando uma abordagem visual

Fonte: CÔRTES *et al.*(2002)

Dias (2001, p.11) sintetiza na Tabela 3 as principais tarefas de Mineração de Dados (MD), suas descrições e exemplos.

Tabela 3 - Síntese das principais tarefas de Mineração de Dados

TAREFA	DESCRIÇÃO	EXEMPLOS
Classificação	Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes.	• Classificar pedidos de crédito; • Esclarecer pedidos de seguros fraudulentos; • Identificar a melhor forma de tratamento de um paciente.
Estimativa ou Regressão	Usada para definir um valor para alguma variável contínua desconhecida.	• Estimar o número de filhos ou a renda total de uma família; • Estimar o valor em tempo de vida de um cliente; • Prever a demanda de um consumidor para um novo produto.
Associação	Usada para determinar quais itens tendem a co-ocorrerem (serem adquiridos juntos) em uma mesma transação.	• Determinar quais os produtos costumam ser colocados juntos em um carrinho de supermercado.
Segmentação ou Clusterização	Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos.	• Agrupar clientes por região do país; • Agrupar clientes com comportamento de compra similar; • Agrupar seções de usuários Web para prever comportamento futuro de usuário.
Sumarização	Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados.	• Tabular o significado e desvios padrão para todos os itens de dados; • Derivar regras de síntese.

Fonte: DIAS (2001, p.11)

2.3.2 Técnicas de Mineração de Dados

De acordo com o tipo de conhecimento e informação almejados pode ser escolhido uma técnica de Mineração de Dados (MD). Esta escolha tem como influência os objetivos e o tipo de problema que estão determinados.

Harrison (1998) afirma que não há uma técnica que resolva todos os problemas de MD. Existem diferentes métodos para diferentes propósitos, cada um com suas vantagens e desvantagens. A familiaridade com as técnicas é necessária para facilitar a escolha de uma tarefa de mineração conforme o problema a ser tratado. Assim para ser executada, cada tarefa possui várias técnicas de mineração envolvidas na busca

por padrões ocultos de dados, naturalmente umas mais indicadas que outras para o problema em questão. As técnicas de MD mais utilizadas são descritas nas subseções a seguir.

2.3.2.1 Árvores de Decisão

Uma árvore de decisão é uma árvore com o objetivo de separar as classes. Tuplas de classes diferentes tendem a ser alocadas em subconjuntos diferentes, cada um descrito por uma regra simples em um ou mais itens de dados. Essas regras podem ser expressas como declarações lógicas, em uma linguagem como SQL (*Structured Query Language*), de modo que possam ser aplicadas diretamente a novas tuplas. Uma das vantagens principais das árvores de decisão é o fato de que o modelo é bem explicável, uma vez que tem a forma de regras explícitas (HARRISON, 1998).

Segundo Rabelo (2007), árvore de decisão é uma técnica que utiliza a recursividade para o particionamento da base de dados. Cada nó não terminal desta árvore representa um teste ou decisão sobre o item de dado. Os algoritmos que implementam esta técnica são: CART, CHAID, C 4.5, C5.0, Quest, ID-3, SLIQ, SPRINT. A técnica de árvore de decisão, em geral, é apropriada para as tarefas de Classificação e Regressão.

Na árvore, o atributo mais importante é apresentado como o primeiro nó, e os atributos com relevância decrescente são mostrados nos nós subsequentes. As vantagens principais das Árvores de Decisão são que elas tomam decisões levando em consideração os atributos que são mais relevantes, além de serem compreensíveis para a maioria das pessoas. Ao escolher e apresentar os atributos em ordem de importância, as Árvores de Decisão permitem aos usuários conhecer quais fatores mais influenciam os seus trabalhos.

Os nós da árvore representam os atributos, os ramos recebem os valores possíveis para cada atributo e os nós folha representam as diferentes classes de um conjunto de treinamento. Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

Basgalupp (2010) apresenta um exemplo do funcionamento básico de uma árvore de decisão: o diagnóstico de pacientes com base nos dados da Tabela 4.

Tabela 4 - Conjunto de dados para diagnóstico da saúde de pacientes

Exemplo	Febre	Enjôo	Manchas	Dor	Diagnóstico
T1	sim	sim	pequenas	sim	doente
T2	não	não	grandes	não	saudável
T3	sim	sim	pequenas	não	saudável
T4	sim	não	grandes	sim	doente
T5	sim	não	pequenas	sim	saudável
T6	não	não	grandes	sim	doente

Fonte: BASGALUPP (2010)

Suponha que um paciente chegue ao consultório médico, para diagnosticá-lo, a primeira pergunta que pode ser feita ao paciente é se ele tem sentido dores. A cada pergunta respondida, outra pode ser realizada até que se chegue a uma conclusão sobre a classe do exemplo {doente ou saudável}. Essa série de perguntas e suas possíveis respostas podem ser organizadas na forma de uma árvore de decisão, a qual é uma estrutura hierárquica composta por nodos e arestas. A Figura 11 ilustra uma possível árvore de decisão para o problema descrito.

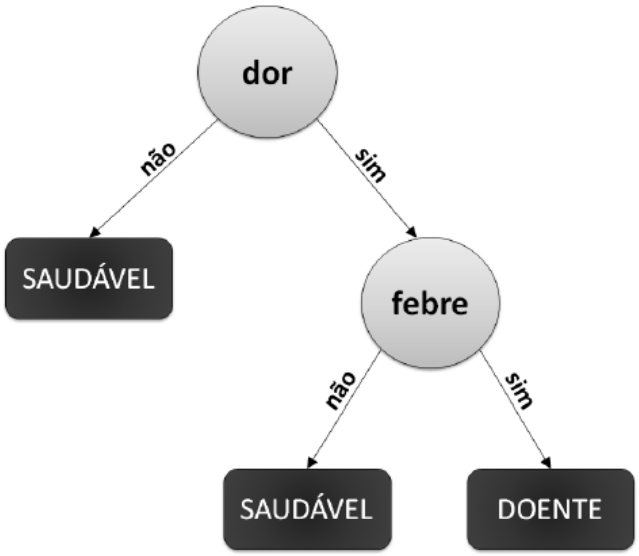


Figura 11 - Uma possível árvore de decisão para diagnosticar pacientes

Fonte: BASGALUPP (2010)

Dessa forma, é possível utilizar a árvore de decisão para classificar um novo paciente como saudável ou doente. Basta partir do nó raiz e ir percorrendo de acordo com as respostas aos teste dos nós internos até chegar ao nó folha, que indica a classe do paciente. Além da obtenção da classe, a grande vantagem é que a trajetória percorrida até a classe representa uma regra, facilitando a interpretabilidade do modelo pelo usuário, no caso um médico. Uma das regras para o exemplo dado seria: Se dor=sim e febre=sim, então o paciente está doente.

Para realizar uma classificação através de árvore de decisão é necessário dividir o conjunto de dados em dois conjuntos: **conjunto de dados de teste** e **conjunto de dados de treinamento**. O conjunto de dados de treinamento é utilizado para construir o modelo, nele os rótulos conhecidos são fornecidos aos registros. O conjunto de dados de teste é usado para testar o modelo criado e consiste de registros sem os rótulos de classes, pois são desconhecidos.

Segundo TAN *et al.* (2009), a avaliação do desempenho de um modelo de classificação é baseada nas contagens de registros de testes previstos correta e incorretamente pelo modelo. Estas contagens são tabuladas em uma tabela conhecida como matriz de confusão. A Tabela 5 mostra a matriz de confusão para um problema

de classificação binária. Cada entrada F_{ij} nesta tabela denota o número de registros da classe 0 previstos incorretamente como da classe 1. Baseado nas entradas da matriz de confusão, o número total de previsões corretas feitas pelo modelo é (F_{11} + F_{00}) e o número total de previsões incorretas é (F_{10} + F_{01}).

Tabela 5 - Matriz de confusão para um problema de 2 classes

		Classe Prevista	
		Classe = 1	*Classe = 0*
Classe Real	Classe = 1	F_{11}	F_{10}
	Classe = 0	F_{01}	F_{00}

Fonte: TAN *et al.* (2009)

Embora uma matriz de confusão forneça as informações necessárias para determinar o quão bem um modelo de classificação é executado, resumir estas informações com um único número tornaria mais conveniente comparar o desempenho de diferentes modelos. Isto pode ser feito usando uma **métrica de desempenho** como a **precisão**, que é definida da seguinte maneira:

$$\text{Precisão} = \frac{\text{Número de previsões corretas}}{\text{Número total de previsões}} = \frac{F_{11} + F_{00}}{F_{11}+F_{10}+F_{01}+F_{00}} \qquad \text{Eq. (1)}$$

A maioria dos algoritmos de classificação procura por modelos que atinjam a maior taxa de precisão quando aplicados ao conjunto de testes.

2.3.2.1.1 Algoritmo C4.5

O algoritmo C 4.5 constrói uma árvore de decisão, usando a abordagem *top-down*, em que o atributo mais significativo, ou seja, o mais generalizado, quando comparado a outros atributos do conjunto, é considerado o nó raiz da árvore. Na sequência da construção, o próximo nó da árvore será o segundo atributo mais significativo, e assim sucessivamente, até gerar o nó folha, que representa o atributo

alvo da instância. Após a construção, o algoritmo inicia um processo de poda, a fim de reduzir o excesso de ajustes (*overfitting*) aos dados de treinamento (BASGALUPP, 2010).

O algoritmo C4.5 utiliza uma propriedade estatística como critério de poda para a árvore de decisão. Essa propriedade é denominada **ganho de informação**. Ela mede como um determinado atributo separa os exemplos de treinamento de acordo com a classificação deles (GUARDA, 2009). O C4.5 usa o ganho de informação para selecionar, entre os candidatos, os atributos que serão utilizados a cada passo, enquanto constrói a árvore de decisão.

Ainda segundo Guarda (2009), para definir ganho de informação, é necessário definir também uma outra medida comumente usada em teoria de informação, chamada **entropia**, que caracteriza a impureza de uma coleção arbitrária de exemplos. Dada uma coleção S, que contém exemplos positivos e negativos de algum conceito objetivo, a entropia de S, relativa a esta classificação lógica é:

$$\text{Entropia} = -p_+ \log_2 p_+ - p_- \log_2 p_- \qquad \text{Eq. (2)}$$

Onde p_+ é a proporção de exemplos positivos em S e p_- é a proporção de exemplos negativos em S. Para ilustrar, suponha S uma coleção de 14 exemplos de algum conceito lógico que inclui 9 exemplos positivos e 5 exemplos negativos. Então a entropia de S relativa a esta classificação lógica é:

$$\text{Entropia}([9+, 5-]) = -(9/14)\log2(9/14) - (5/14)\log2(5/14) = 0.940$$

A entropia é 0 se todos os membros de S pertencem à mesma classe. A entropia é 1 quando a coleção contém um número igual de exemplos positivos e negativos. Se a coleção contém números desiguais de exemplos positivos e negativos, a entropia está entre 0 e 1.

Portanto, a entropia é uma medida de não homogeneidade de um conjunto de dados e o ganho de informação é justamente a medida usada pelo algoritmo para selecionar o melhor atributo a cada passo da construção da árvore. O melhor atributo é aquele com o maior ganho de informação.

De acordo com TAN *et al.* (2009), para determinar o quão boa é uma condição de teste realizada na árvore de decisão, é necessário comparar o grau de entropia do nó pai (antes da divisão) com o grau de entropia dos nós filhos (após a divisão). Quanto maior a diferença, melhor a condição do teste. O atributo que gerar uma maior diferença é escolhido como condição de teste.

Segundo Mitchel (1997) *apud* Martinhago (2005), a ideia básica do algoritmo C4.5 segue os seguintes passos:

a) escolher um atributo;

b) estender a árvore adicionando um ramo para cada valor do atributo;

c) passar os exemplos para as folhas (tendo em conta o valor do atributo escolhido);

d) para cada nó folha (se todos os exemplos são da mesma classe), associar esta classe ao nó folha, caso contrário, repetir os passos (a), (b) e (c).

Porém, a árvore assim construída pode estar ajustada demais aos dados de treinamento, fenômeno chamado ***overfitting***. Uma árvore de decisão **A** está ajustada demais aos dados, se existir uma árvore **A'** tal que **A** tem menor erro que **A'** no conjunto de treinamento, porém **A'** tem menor erro no conjunto de teste. Para corrigir o fato de uma árvore estar ajustada demais, deve-se executar um procedimento de poda da árvore.

De acordo com Witten *et al.* (2011), o J48 é uma versão implementada do tradicional algoritmo C4.5 desenvolvido por Ross Quinlan na década de 1970. Conforme Martinhago (2005), muitas pessoas na indústria de Mineração de Dados consideram o professor Ross Quinlan, da Universidade de Sydney, na Austrália, como o "pai das árvores de decisão". A contribuição de Quinlan foi um novo algoritmo chamado ID3, em 1983, e posteriormente surgiram as suas evoluções: ID4, ID6, C4.5, C5.0/See 5 e o J48.

Martinhago (2005) afirma que para a utilização do algoritmo J48 é necessário conhecer alguns parâmetros que podem ser modificados para proporcionar melhores resultados, tais como:

- parâmetro U: cria a árvore sem realizar poda;
- parâmetro C: indica o fator de confiança inicial (*confidence*) para a poda. O valor *default* do fator de confiança para o J48 é de 0.25 (25%);
- parâmetro M: indica o número mínimo de instâncias por folha. Valor *default é* 2;
- parâmetro R: usa a poda com redução de erro;
- parâmetro N: indica o número de partições para a poda com redução de erro, onde uma partição é utilizada como conjunto de poda. Valor default é 3;
- parâmetro B: usa árvore binária;
- parâmetro S: não utiliza subárvore de poda;
- parâmetro L: não apaga a árvore depois de construída.

O parâmetro C, chamado **fator de confiança**, indica ao algoritmo o nível da poda, quando menor o valor informado, maior será a poda realizada na árvore de decisão.

Sempre que a taxa de erro nos dados de teste seja significativamente superior à obtida nos dados de treino poderemos ajustar o fator de confiança. Nestes casos, deve-se diminuir o valor do fator de confiança para forçar uma poda maior e obter um modelo mais genérico. Nos casos em que os dados de teste não variam significativamente relativamente aos dados de treino, o fator de confiança pode ser aumentado de forma a obter-se um modelo com uma árvore mais detalhada (MONTEIRO, 2005).

A ferramenta de Mineração de Dados WEKA (apresentada na seção 2.6), após o processamento do algoritmo J48, apresenta entre os resultados uma medida estatística para avaliação dos modelos chamada índice *Kappa*. Segundo Landis e Koch (1977) *apud* Guimarães (2008), o teste de *Kappa* é uma medida de concordância

interobservador e mede o grau de concordância, além do que seria esperado tão somente pelo acaso. Para descrever se há ou não concordância entre dois métodos de classificação, utiliza-se a medida *Kappa* que é baseada no número de respostas concordantes, ou seja, no número de casos cujo resultado é o mesmo entre os avaliadores. Esta medida de concordância assume valor máximo igual a 1, que representa total concordância ou, ainda, pode assumir valores próximos e até abaixo de 0, os quais indicam nenhuma concordância.

A estatística *Kappa* é calculada em três etapas. Primeiro calcula-se um índice que represente a concordância esperada pelo acaso. Em segundo, calcula-se a concordância observada e por último, a estatística é calculada pela divisão da diferença entre a concordância observada e a esperada, pela diferença entre a concordância absoluta e a esperada pelo acaso. O resultado busca a maior diferença possível entre a concordância observada e a esperada (THOMPSON, 2001).

A Tabela 6 apresenta as faixas de valores obtidos pelo índice *Kappa* e suas respectivas interpretações de concordância com intervalo de confiança de 95%.

Tabela 6 - Índices *Kappa*

Valor *Kappa*	Concordância
<0	Inexistente
0	Pobre
0 a 0,20	Ligeira
0,21 a 0,40	Considerável
0,41 a 0,60	Moderada
0,61 a 0,80	Substancial
0,81 a 1	Excelente

Fonte: Adaptado de Landis e Koch (1977) *apud* Guimarães (2008)

2.3.2.2 Regras de Associação

Os algoritmos para a descoberta de Regras de Associação têm como objetivo procurar relações entre os dados de um conjunto de dados, que ocorrem com determinada frequência. Esta técnica é muito utilizada na área do comércio, na busca

de padrões de compra com intuito de orientar as ações dos gestores de vendas (PASTA, 2011).

Uma regra de associação é uma expressão representada na forma X → Y (X implica em Y) em que X e Y são conjuntos de itens da base de dados e X∩Y = Ø; X é o antecedente da regra (lado esquerdo) e Y é o consequente da regra (lado direito) e pode envolver qualquer número de itens em cada lado da regra (MARTINHAGO, 2005). A força de uma regra de associação pode ser medida em termos do seu **suporte** e **confiança**. O suporte determina a frequência na qual uma regra é aplicável a um determinado conjunto de dados, enquanto que a confiança determina a frequência na qual os itens em Y aparecem em transações que contenham X (TAN *et al*., 2009). As definições formais destas métricas são:

X => Y

Suporte (*sup*) = Número de registros com X e Y / Número total de registros

Confiança (*conf*) = Número de registros com X e Y/ Número de registros com X

Kampff (2009) destaca que a ordem de apresentação das regras estabelece uma lista de decisão, a ser aplicada em sequência. A regra que aparece primeiro na lista tem maior prioridade para predizer a classe. Quando um registro é classificado, nenhuma outra regra posterior de classificação será aplicada sobre ele.

Ainda segundo Dias (2001) a técnica de Regras de Associação é indicada para tarefas de Associação e alguns algoritmos que a implementam são: *Apriori*, *AprioriTid*, *AprioriHybrid*, AIS, SETM, DHP, DIC, *Eclat*, *Maxclique* e *Cumalte*.

Gonçalves (2005) afirma que uma regra de associação é interessante apenas quando os itens da regra apresentam **dependência positiva**, ou seja, quando o valor do suporte real da regra é maior do que o valor do suporte esperado. O **suporte esperado** é computado multiplicando-se o suporte dos itens que compõem a regra: *SupEsp* (*A* ∪ *B*) = *Sup* (*A*) *x Sup* (*B*). Portanto, por definição a **dependência positiva** entre itens é: Seja D uma base de dados de transações definida sobre um conjunto de

itens I. Sejam A ⊂ I e B ⊂ I dois conjuntos não vazios de itens, onde A ∩ B = Ø. Os conjuntos de itens A e B possuem dependência positiva se: Sup (A ∪ B) > SupEsp (A ∪ B).

Ainda de acordo com Gonçalves (2005), a medida de interesse *lift*, também conhecida como *interest*, é uma das mais utilizadas para avaliar dependências. Dada uma regra de associação A → B, esta medida indica o quanto mais frequente torna-se B quando A ocorre. Por definição: Seja D uma base de dados de transações. Seja A → B uma regra de associação obtida a partir de D. O valor do *lift* para A → B é computado por:

$$Lift\ (\mathrm{A}\ \rightarrow \mathrm{B}) = \frac{Conf\ (\mathrm{A}\rightarrow\mathrm{B})}{Sup(\mathrm{B})} \qquad \text{Eq. (3)}$$

Se *Lift* (A → B) = 1, então A e B são independentes. Se *Lift* (A → B) > 1, então A e B são positivamente dependentes. Se Se *Lift* (A → B) < 1, A e B são negativamente dependentes. Esta medida varia entre 0 e ∞ e possui interpretação bastante simples: quanto maior o valor do *lift*, mais interessante a regra.

2.3.2.2.1 Algoritmo *Apriori*

O algoritmo *Apriori* é um dos mais utilizados para regras de associação. Segundo Silva (2010), ele foi proposto em 1994 pela equipe de pesquisa do Projeto QUEST da IBM que originou o software *Intelligent Miner*. Este algoritmo faz buscas recursivas no banco de dados à procura dos conjuntos frequentes (conjuntos que satisfazem um suporte mínimo estabelecido).

Para Tan *et al.* (2009) o seu princípio se resume em: *se um conjunto de itens é frequente, então todos os seus subconjuntos também devem ser frequentes.* Para entender a ideia por trás do algoritmo, basta analisar a rede de conjuntos de itens mostrada na Figura 12. Suponha que {c,d,e} seja um conjunto de itens frequentes.

Claramente, qualquer transação que contenha {c,d,e} também deve conter seus subconjuntos {c,d}, {c,e}, {d,e}, {c}, {d} e {e}. Como resultado, se {c,d,e} for frequente, então todos os subconjuntos de {c,d,e} também devem ser frequentes (os conjuntos de itens sombreados na Figura 12).

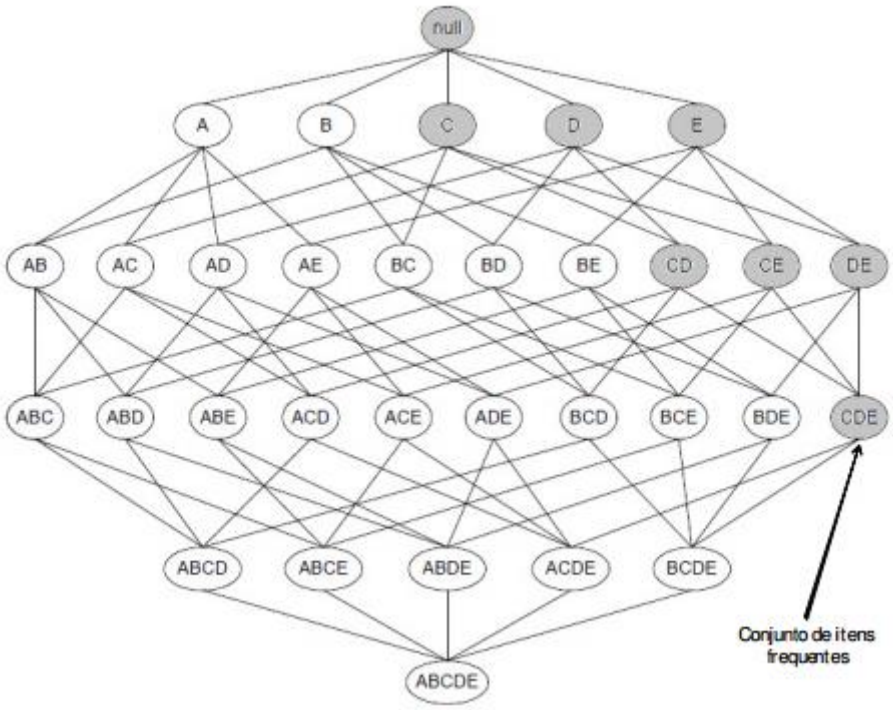


Figura 12 - Ilustração do algoritmo *Apriori*

Fonte: TAN *et al.* (2009)

De forma inversa, se um conjunto de itens como {a,b} for infrequente, então todos os seus superconjuntos devem ser infrequentes também. Conforme ilustrado na Figura 13, o subgrafo inteiro contendo os superconjuntos de {a,b} podem ser podados imediatamente, assim que {a,b} for descoberto como sendo infrequente também (conjuntos de itens sombreados na Figura 13). Esta estratégia de se diminuir o espaço de pesquisa exponencial baseado na medida de suporte é conhecida como **poda baseada em suporte**. Tal estratégia de poda é tornada possível por uma propriedade chave da medida de suporte saber que o suporte para um conjunto de itens nunca excede o suporte de seus subconjuntos. Esta propriedade também é conhecida como a propriedade **anti-monotônica** da medida do suporte.

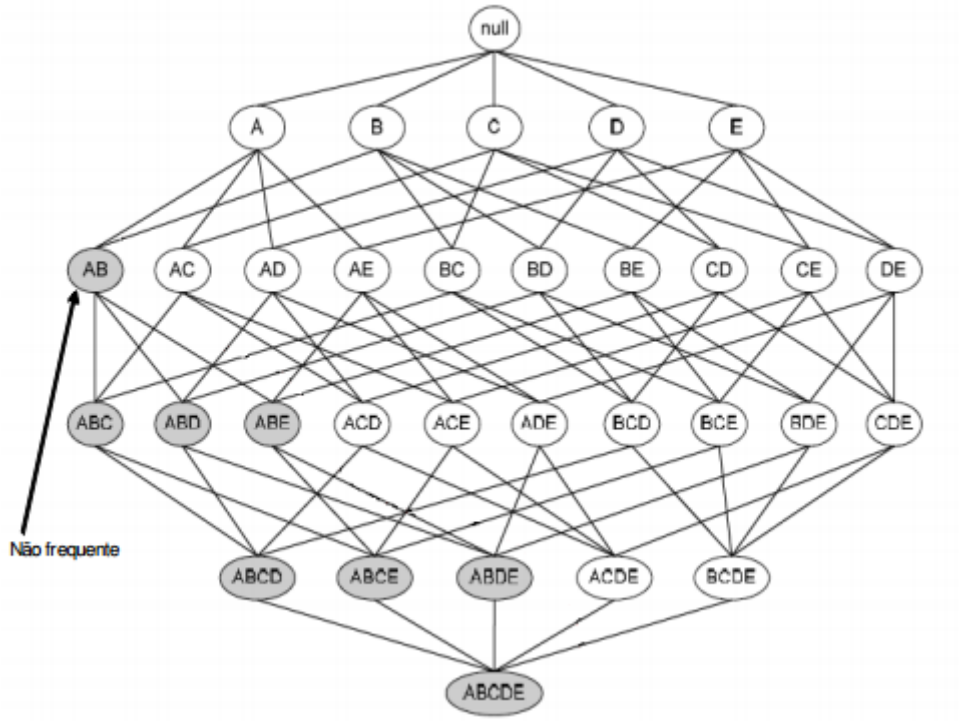


Figura 13 - Ilustração de poda baseada em suporte do algoritmo *Apriori*

Fonte: TAN *et al.* (2009)

Morais (2010) apresenta um exemplo da utilização do algoritmo *Apriori* para obtenção de regras de associação na base de dados da Tabela 7, considerando o suporte mínimo de 50% e a confiança mínima de 90%.

Tabela 7 - Base de dados com transações de clientes

Transações	Produtos Comprados
1	biscoito, manteiga, café, arroz
2	manteiga, feijão, ovos, pão, queijo
3	café, leite, ovos, pão
4	leite, café, manteiga, feijão, pão, arroz
5	leite, café, pão, queijo

Fonte: Morais (2010)

1 - Determinando todos os k-*itemsets* frequentes;

- Gera-se os candidatos a 1-*itemsets* frequentes, C_1:

sup({leite}) = 3/5 x 100 = 60%

sup({café}) = 4/5 x 100 = 80%

sup({pão}) = 4/5 x 100 = 80%

sup({biscoito}) = 1/5 x 100 = 20%

sup({manteiga}) = 2/5 x 100 = 40%

sup({queijo}) = 2/5 x 100 = 40%

sup({ovos}) = 2/5 x 100 = 40%

sup({arroz}) = 2/5 x 100 = 40%

sup({feijão}) = 2/5 x 100 = 40%

Logo, L1 = {{café}; {leite}; {pão}}

- Gera-se os candidatos a 2-itemsets frequentes, C2:

sup({café, leite}) = 3/5 x 100 = 60%

sup({café, pão}) = 3/5 x 100 = 60%

sup({leite, pão}) = 3/5 x 100 = 60%

Logo, L2 = {{café, leite}, {café, pão}, {leite, pão}}.

- Gera-se os candidatos a 3-itemsets frequentes, C3:

sup({café, leite, pão) = 3/5 x 100 = 60%

Logo, L_3 = {{café; leite; pão}}.

Como L_3 só possui um *itemset* o algoritmo para de iterar, pois não é possível gerar candidatos a 4-*itemsets*.

2 - Gerando regras de associação;

Para gerar as regras deve-se permutar os k-*itemsets* frequentes (k>=2) e selecionar as regras que possuem confiança maior ou igual a confiança mínima especificada pelo usuário. As regras de associação que obedecem as especificações são as que estão na Tabela 8.

Tabela 8 - Regras de Associação extraídas da Tabela 7

Regra de Associação	Suporte	Confiança
{leite} => {café}	60%	100%
{café} => {leite}	60%	75%
{leite} => {pão}	60%	100%
{pão} = {leite}	60%	75%
{café} => {pão}	60%	75%
{pão} => {café}	60%	75%
{café, leite} => {pão}	60%	100%
{leite, pão} => {café}	60%	100%
{café, pão} => {leite}	60%	100%

Fonte: Morais (2010)

Amo (2004) afirma que o *Apriori* possui 3 fases principais: (1) a fase da geração dos candidatos, (2) a fase da poda dos candidatos e (3) a fase do cálculo do suporte. As duas primeiras fases são realizadas na memória principal e não necessitam que o banco de dados seja varrido. A memória secundária só é utilizada caso o conjunto de *itemsets* candidatos seja muito grande e não caiba na memória principal. Mas, mesmo neste caso é bom salientar que o banco de dados, que normalmente nas aplicações é extremamente grande, não é utilizado. Somente na terceira fase, a fase do cálculo do suporte dos *itemsets* candidatos, é que o banco de dados é utilizado. Tanto na fase de geração de candidatos (Fase 1) quanto na fase da poda dos candidatos (Fase 2) a propriedade **anti-monotônica** é utilizada.

As regras geradas pelo algoritmo *Apriori* seguem o seguinte formato:

atributo=valor atributo=valor (nº instâncias) => atributo=valor (nº instâncias) conf:

(A) (B) (C) (D) (E) (F)

Uma regra nesse formato significa que das (C) ocorrências em que os atributos (A) e (B) estavam presentes, o atributo (D) também estava presente em (E) daquelas ocorrências, portanto gerando uma confiança (F).

2.3.2.3 Redes Neurais

As redes neurais representam uma metáfora do cérebro humano para o processamento da informação. Estes modelos são biologicamente inspirados, uma vez que funcionam como réplicas do nosso cérebro. A técnica das redes neurais tem se mostrado muito promissora em sistemas de previsão e de classificação de aplicações de negócios, devido à sua capacidade de aprender com os dados, da sua natureza não parametrizada e da sua capacidade de generalizar. A computação neural refere-se a uma metodologia de reconhecimento de padrões para a aprendizagem de máquina. O modelo resultante da computação neural é chamado de Rede Neural Artificial (RNA) ou somente de Rede Neural (TURBAN *et al*., 2010).

Estruturalmente, uma rede neural consiste de um número de elementos interconectados, denominados neurônios, organizados em camadas que aprendem pela modificação da comunicação entre as camadas. As redes neurais tentam construir representações internas de modelos ou padrões achados nos dados. A rede inicialmente recebe um conjunto de dados para ser treinada. Após isso, ela está pronta para fazer predições sobre novos dados inseridos, o que a torna adequada para as tarefas de classificação (NEGREIROS e LIMA, 2009).

Iyoda (2000) explica através da Figura 14 o modelo de um neurônio artificial.

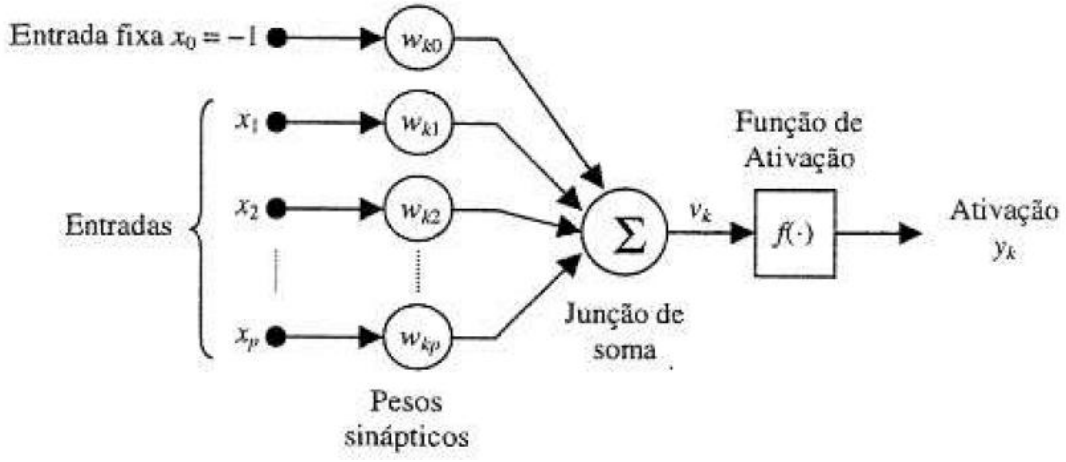


Figura 14 - Exemplo de Rede Neural

Fonte: Iyoda (2000)

Pode-se identificar três elementos básicos no modelo:

- Um conjunto de sinapses ou conexões de entrada, sendo cada entrada ponderada por um peso sináptico. Sendo assim, um sinal x_j na entrada da sinapse *j* conectada ao neurônio *k* é multiplicado pelo peso sináptico w_{kj}. Observe a ordem adotada para os índices (subscritos) na notação aqui empregada: o primeiro índice se refere ao neurônio em questão e o segundo índice ao terminal de entrada da sinapse ao qual o peso se refere. Quando uma entrada fixa está presente (entrada x_0 na Figura 14), então o peso sináptico correspondente é denominado peso de polarização;
- Uma função de soma, responsável pela combinação aditiva dos sinais de entrada, ponderados pelos respectivos pesos das sinapses do neurônio;
- Uma função de ativação geralmente não-linear e de formato sigmoidal, representando um efeito de saturação na ativação de saída y_k do neurônio.

O processo que busca a melhor calibração dos pesos é conhecido como processo de aprendizado ou treinamento da rede. Uma rede neural é treinada até que a saída do modelo resultante se torne próximo das informações desejadas. Esse aprendizado é realizado através de interações e ajustes nos pesos após as entradas informadas. O algoritmo de retropropagação (*Backpropagation*) é o mais utilizado para melhorar a predição dos dados.

2.3.2.4 Algoritmo *K-Means*

Um dos algoritmos de Agrupamento de dados mais conhecidos e utilizados, além de ser o que possui o maior número de variações é o *K-means* ou K-Médias, criado por J. McQueen em 1967 (PRASS, 2004). Ele é considerado um algoritmo não-supervisionado, pois realiza uma classificação automática dos dados sem a

necessidade de nenhuma pré-classificação existente, ao contrário do que ocorre na classificação onde as classes já são conhecidas *a priori*.

Conforme Tan *et al.* (2009), a técnica do *K-means* é simples. Primeiramente, o usuário informa ao algoritmo o número *K* de grupos (também chamados de centroides) que se deseja que ele agrupe os dados. Cada ponto (cada ponto é um dado) é atribuído ao centroide mais próximo e cada coleção de pontos atribuídos a um centroide é um grupo. O centroide de cada grupo é então atualizado iterativamente baseado nos pontos atribuídos aquele grupo até que os centroides não se alterem mais. O *K-means é* formalmente descrito pelo algoritmo apresentado no Quadro 1.

Algoritmo *K-means* básico
1: Selecione K pontos como centroides iniciais.
2: **repita**
3: Forme K grupos atribuindo cada ponto ao seu centroide mais próximo.
4: Recalcule o centroides de cada grupo
5: **até que** os centroides não mudem.

Quadro 1 - Algoritmo *K-means* básico

Fonte: TAN *et al.* (2009)

No primeiro passo, mostrado na Figura 15(a) os pontos são atribuídos aos centroides iniciais, que estão todos no grupo maior de pontos. Para este exemplo, usamos a média como o centroide. Após os pontos serem atribuídos a um centroide, ele é atualizado. Novamente, a figura para cada passo mostra o centroide no início do passo e a atribuição de pontos àqueles centroides. No segundo passo, os pontos são atribuídos aos centroides atualizados e os centroides são atualizados novamente. Nos passos 2, 3 e 4, que são mostrados na Figura 15 (b), (c) e (d), respectivamente, dois dos centroides se movem para os dois grupos pequenos de pontos na parte inferior das Figuras. Quando o algoritmo *K-means* termina na Figura 15(d), porque não há mais mudanças, os centroides identificaram os agrupamentos naturais dos pontos.

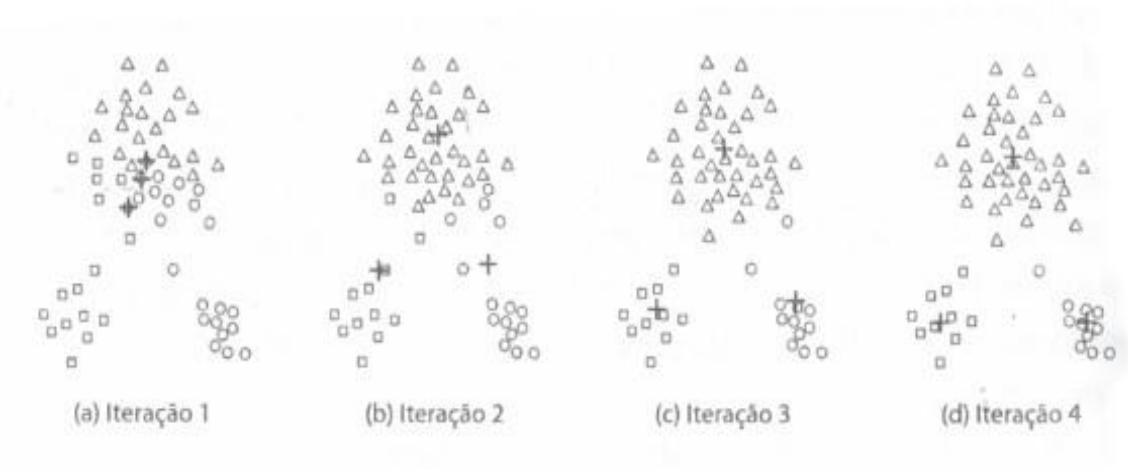


Figura 15 – Usando *K-means* para encontrar 3 grupos nos dados

Fonte: TAN *et al.* (2009)

Para algumas combinações de funções de proximidade e tipos de centroides, o *K-means* sempre converge para uma solução. O algoritmo atinge o estado no qual nenhum ponto muda de um grupo para outro e assim os centroides também não mudam. Devido ao fato da maioria da convergência ocorrer nos primeiros passos, a condição da linha 5 do algoritmo acima é muitas vezes substituída por uma condição mais fraca: repetir até que apenas 1% dos pontos mudem de grupo.

Embora existam técnicas de C*lustering* avançadas que podem sugerir o número ideal de *clusters* a serem usados. Geralmente, esse é um processo exploratório e o melhor número de *clusters* a ser usado normalmente é localizado por tentativa e erro.

Para calcular a distância entre os pontos de dados, o *K-means* necessita de uma função matemática. Pode haver diversos tipos de medidas de proximidade que sejam apropriadas para um determinado tipo de dados. Por exemplo, a distância de Manhattan pode ser usada para dados Euclidianos, enquanto que a medida de Jaccard é empregada frequentemente para documentos. A Distância Euclidiana é usada frequentemente para pontos de dados no espaço Euclidiano (TAN *et al.*, 2009).

Geralmente utiliza-se a “Distância Euclidiana” para calcular o quão “longe” uma ocorrência (ponto que representa um dado) está da outra. Segundo Seidel *et al.* (2008), as características de cada objeto são combinadas em uma medida de

semelhança, que pode ser de similaridade ou dissimilaridade, calculada para todos os pares de objetos, possibilitando a comparação de qualquer objeto com outro pela medida de similaridade e a associação dos objetos semelhantes por meio da análise de agrupamento. As medidas de distância representam a similaridade, que é representada pela proximidade entre as observações ao longo das variáveis.

A Distância Euclidiana é utilizada para calcular medidas específicas, assim como a Distância Euclidiana Simples e a Distância Euclidiana Quadrática ou Absoluta, que consiste na soma dos quadrados das diferenças, sem calcular a raiz quadrada. A Distância Euclidiana Quadrática é definida por:

$$DE = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \qquad \text{(Eq. 4)}$$

Onde:

X_{ij} é a *j*-ésima característica do *i*-ésimo indivíduo (cada característica do objeto tem sua medida de similaridade calculada para todos os pares de objetos);

Xi'j ' é a *j*-ésima característica do *i'*-ésimo indivíduo.

Quanto mais próximo de zero for a Distância Euclidiana, mais similares são os objetos comparados.

Para Linden (2009), o *K-means* é um algoritmo extremamente veloz, geralmente convergindo em poucas iterações para uma configuração estável, na qual nenhum elemento está designado para um *cluster* cujo centro não lhe seja o mais próximo. A grande vantagem desta heurística é a sua simplicidade: um programador experiente pode implementar uma versão própria em cerca de uma hora de trabalho.

Um eventual problema do *K-means* é que ele enfatiza a questão da homogeneidade e ignora a importante questão da boa separação dos *clusters*. Isto pode causar uma má separação dos grupos no caso de uma má inicialização dos centroides, inicialização esta que é feita de forma arbitrária (aleatória) no início da execução. Outro ponto que pode afetar a qualidade dos resultados é a escolha do

número de grupos feita pelo usuário. Um número pequeno demais de grupos pode causar a junção de dois *clusters* naturais, enquanto que um número grande demais pode fazer com que um *cluster* natural seja quebrado artificialmente em dois (LINDEN, 2009).

De acordo com Jain *et al*. (2010), algumas implementações do *K-means* só permitem atributos numéricos. Nesse caso, pode ser necessário converter o conjunto de dados em formato padrão ou converter dados categóricos em dados binários. Também pode ser necessário normalizar os valores dos atributos, que geralmente são medidos em escalas diferentes (por exemplo, idade e renda). A implementação do *K-means* no *software* WEKA, o algoritmo *SimpleKMeans,* trata automaticamente atributos categóricos e numéricos.

2.3.2.5 Algoritmos Genéticos

Os Algoritmos Genéticos (AG) são métodos evolutivos que começaram a ser desenvolvidos na década de 60 por John Holland na Universidade de Michigan. Este algoritmo é baseado nas teorias da Seleção Natural de Charles Darwin e nos mecanismos da genética (PIZZIRANI, 2003). Holland tinha como objetivo o estudo formal dos fenômenos da evolução, como ocorrem na natureza e o desenvolvimento de formas de importar tais fenômenos aos sistemas de computação (PASTA, 2011).

Sumathi e Sivanandam (2006) afirmam que os AG podem ser definidos como uma técnica de otimização baseada nos conceitos de combinação genética, mutação e seleção natural. A computação natural é uma aposta para a criação de modelos visando resolver os problemas atuais de otimização através de processos evolutivos e uma ferramenta de otimização, inspirada em fenômenos naturais.

O elemento principal dos AG é o indivíduo ou organismo, que representa uma solução possível no espaço de resposta. Ele é constituído por um ou mais

cromossomos, que por sua vez são formados por vários genes, como ilustra a Figura 16. A representação desses cromossomos se dá através de *strings* (conjunto de dados concatenados), nas quais cada dado representa um gene. Como na natureza, os genes contêm informações que determinam as características do indivíduo, estando nesse caso codificadas por meio de números binários ou outros alfabetos mais complexos. Cada gene do cromossomo representa uma variável de otimização, cujos valores possíveis são determinados pela codificação empregada no processo.

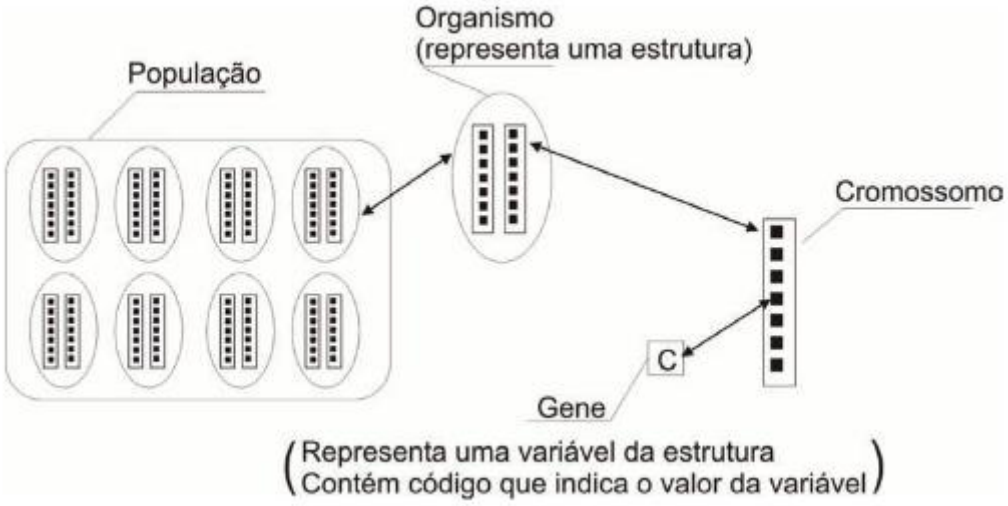


Figura 16 - Estrutura básica do algoritmo genético
Fonte: ALMEIDA (2006)

Para Amaral (2001), as características gerais dos AG são coincidentes com as características gerais da evolução das espécies:

- a evolução é um processo que ocorre basicamente nos cromossomos;
- o processo de seleção natural codifica as estruturas mais aptas às reprodução com mais frequência do que aquelas que não são aptas;
- o processo de reprodução se estabelece de três modos: Mutação, Reprodução, Cruzamento;
- a evolução genética não tem memória.

2.3.2.6 Raciocínio baseado em Casos (RBC)

Quando se tenta de resolver algum problema, uma das primeiras soluções está apoiada em experiências passadas. O Raciocínio Baseado em Casos (RBC) faz uso de soluções já utilizadas para a solução de determinado problema, procurando um caso mais similar ao proposto (PASTA, 2011).

Na resolução de problemas, aplicando o RBC, uma solução para um novo caso é obtida recuperando casos similares anteriormente analisados e derivando suas respectivas soluções de modo a se adequar ao novo problema. O processo se realiza quando um novo caso é apresentado ao sistema. Em face do novo problema, utiliza-se um conjunto de métricas de similaridade para determinar quais casos anteriores mais se assemelham ao caso proposto, bem como se determinam as características-chave utilizadas nessa comparação (FONSECA, 2008, p.16). A Figura 17 ilustra o ciclo básico do RBC.

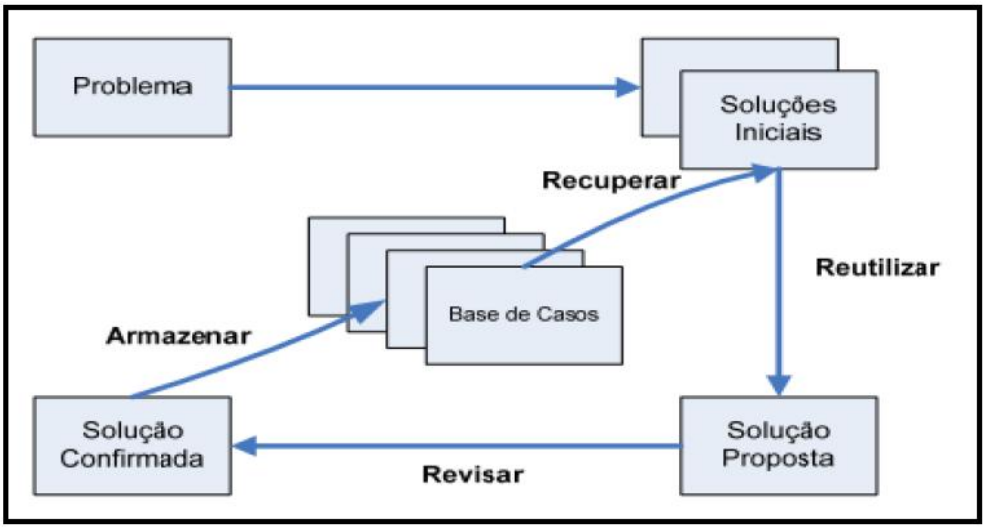


Figura 17 - Ciclo básico do Raciocínio baseado em Casos
Fonte: PASTA (2011)

Segundo Dias (2001), a técnica de RBC é indicada para as tarefas de classificação e segmentação e os algoritmos mais conhecidos que a implementam são: BIRCH, CLARANS, CLIQUE. Este mesmo autor resume na Tabela 9 as principais técnicas de Mineração de Dados (MD) e suas respectivas descrições, tarefas e algoritmos que as implementam.

Tabela 9 - Síntese das principais técnicas de Mineração de Dados

Técnica	Descrição	Tarefas	Algoritmos
Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, onde a cada nova geração, soluções melhores têm mais chance de ter "descendentes"	Classificação; Segmentação	Algoritmo Genético Simples; Genitor, CHC; Algoritmo de *Hillis*; GA-*Nuggets*; GA-PVMINER
Árvores de Decisão	Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos	Classificação; Regressão	CART, CHAID, C4.5, J48, C5.0, Quest, ID3, SLIQ e SPRINT
Descoberta de Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados	Associação	*Apriori*, *AprioriTid*, *AprioriHybrid*, AIS, SETM e DHP
Raciocínio Baseado em Casos	Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança	Classificação; Segmentação	BIRCH, CLARANS e CLIQUE
Redes Neurais Artificiais	Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa de conexões neuronais e dos pesos dessas conexões	Classificação; Segmentação	*Perceptron*, Rede MLP, Redes de *Kohonen*, Rede *Hopfield*, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede *Counterpropagation*, Rede RBF, Rede PNN, Rede *Time Delay*, *Neocognitron*, Rede BSB
K-Means	Algoritmo não supervisionado que classifica automaticamente os dados sem a necessidade de definição de classes	Agrupamento	*K-means, SimpleKMeans*

Fonte: Adaptado de DIAS (2001)

2.4 METODOLOGIA *CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING* (CRISP-DM)

A Mineração de Dados (MD) pode ser desenvolvida de modo não-sistemático, sem que haja nenhum cuidado em seu desenvolvimento, o que não é recomendado, pois acarreta em resultados não esperados ou imprecisos. Com intuito de evitar este tipo de situação, o uso de uma metodologia vem garantir que o processo de MD seja desenvolvido de modo sistemático e padronizado, o que acrescentará em resultados precisos e confiáveis (PASTA, 2011, p.68). Portanto, a eficiência do processo de MD está associada ao uso de uma metodologia denominada *Cross-Industry Standard Process for Data Mining* (CRISP-DM), que contém regras e padrões formalizados para orientar sua aplicação.

Esta metodologia foi proposta em meados da década de 1990 por um consórcio europeu de empresas, para servir como metodologia padrão não proprietária para MD. A Figura 18, ilustra esse processo proposto, que é uma sequência de seis etapas, que inicia com um bom entendimento do negócio e da necessidade do projeto de MD e finaliza com a implementação da solução que satisfaz a necessidade especificada (CHAPMAN *et al*., 2000).

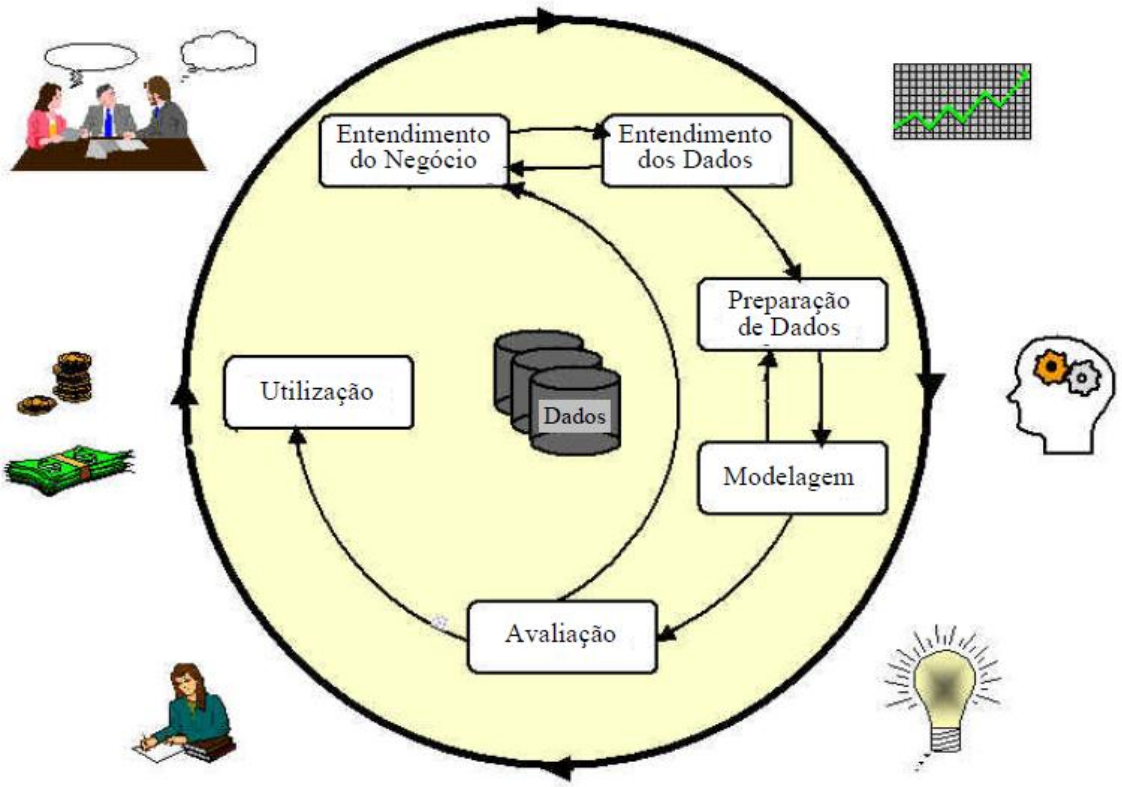


Figura 18 - Etapas do processo CRISP-DM

Fonte: Adaptada de CHAPMAN *et al.* (2000, p. 10)

A sequência dessas fases não é rigorosa, dependendo do resultado de cada fase ou de qual tarefa particular de uma fase precisa ser executada na próxima fase. As flechas indicam as dependências mais importantes e frequentes entre as fases. As fases da metodologia CRISP-DM são semelhantes as fases do processo de KDD.

O círculo externo na Figura 18 simboliza a natureza cíclica da MD. Um processo de MD continua após uma solução ter sido descoberta. Os processos de MD subsequentes se beneficiarão das experiências anteriores.

Conforme Chapman *et al*. (2000, p.10) traduzido por Dias (2001, p.24), cada etapa do processo CRISP-DM é definida da seguinte forma:

- **Entendimento do Negócio (ou *Business Understanding*)**: fase inicial do processo que visa o entendimento dos objetivos do projeto e dos requisitos sob o ponto de vista do negócio. Baseado no conhecimento adquirido, o problema de MD é definido e um plano preliminar é projetado para ativar os objetivos;

- **Entendimento dos Dados (ou *Data Understanding*)**: inicia com uma coleção de dados e procede com atividades que visam buscar familiaridade com os dados, identificar problemas de qualidade de dados, descobrir os primeiros discernimentos nos dados ou detectar subconjuntos interessantes para formar hipóteses da informação escondida;
- **Preparação de Dados (*Data Preparation*):** cobre todas as atividades de construção do *dataset* final. As tarefas de preparação de dados são, provavelmente, desempenhadas várias vezes e não em qualquer ordem prescrita. Estas tarefas incluem seleção de tabelas, registros e atributos, bem como transformação e limpeza dos dados para as ferramentas de modelagem;
- **Modelagem (*Modelling*)**: várias técnicas de modelagem são selecionadas e aplicadas e seus parâmetros são ajustados para valores ótimos. Geralmente, existem várias técnicas para o mesmo tipo de problema de MD. Algumas técnicas têm requisitos específicos na formação de dados. Portanto, retornar à fase de preparação de dados é frequentemente necessário.;
- **Avaliação (*Evaluation*):** o(s) modelo(s) construído(s) na fase anterior é avaliado e são revistos os passos executados na sua construção para se ter certeza de que o modelo representa os objetivos do negócio. O principal objetivo é determinar se existe alguma questão de negócio importante que não foi suficientemente considerada. Nesta fase, uma decisão sobre o uso dos resultados de MD deverá ser alcançada;
- **Utilização, ou Aplicação, (*Deployment*):** após o modelo ser construído e avaliado, ele pode ser usado de duas formas. Na primeira forma, o analista pode recomendar ações a serem tomadas baseando-se simplesmente na visão do modelo e de seus resultados. Na segunda forma, o modelo pode ser aplicado a diferentes conjuntos de dados.

O CRISP-DM foi projetado para fornecer orientação para os iniciantes em MD e para fornecer um modelo de processo genérico que pode ser especializado de acordo

com as necessidades de qualquer ramo de atividade ou da empresa. A metodologia CRISP-DM tem seu sucesso devido ao fato de ter sido desenvolvida à prática, não estar atrelada a nenhuma ferramenta específica de MD, mas sim a junção das melhores práticas que são utilizadas em um projeto de MD, aliada ao fato de atuar sobre todo o processo de MD (PASTA, 2011).

Para melhor entendimento de cada etapa do modelo CRISP-DM, e suas respectivas tarefas e saídas, a Tabela 10 apresenta um resumo dos conceitos anteriormente apresentados.

Tabela 10 - Etapas, Tarefas e Saídas da metodologia CRISP-DM

ETAPA	TAREFAS	SAÍDAS
Entendimento do Negócio	• Determinar os objetivos do negócio;	• Background; • Os objetivos do negócio; • Critérios de sucesso do negócio.
	• Avaliar a situação;	• Inventário dos recursos; • Requisitos,premissas e restrições; • Riscos e contingências; • Terminologia; • Custos e benefícios.
	• Determinar as metas da MD	• Metas da MD; • Critérios de sucesso da MD.
	• Produzir o plano do projeto	• Plano do projeto; • A avaliação inicial de ferramentas e técnicas.
Entendimento dos Dados	• Coletar os dados iniciais;	• Relatório da coleta inicial dos dados.
	• Descrever os dados;	• Relatório da descrição dos dados.
	• Explorar os dados;	• Relatório da exploração dos dados.
	• Verificar a qualidade dos dados;	• Relatório da qualidade dos dados.
Preparação dos Dados	• Selecionar os dados;	• Justificativa para inclusão/exclusão.
	• Limpar os dados;	• Relatório de limpeza dos dados.
	• Construção dos dados;	• Atributos derivados; • Registros gerados;
	• Integrar os dados;	• Dados mesclados.
	• Formatar os dados;	• Dados reformatados.
Modelagem	• Selecionar a técnica de modelagem;	• Técnica de modelagem; • Modelagem dos pressupostos.
	• Gerar o design do teste;	• Design do teste.
	• Construir o modelo;	• As definições de parâmetros; • Modelos; • Descrição do modelo resultante.
	• Avaliar o modelo.	• Modelo de avaliação; • Parâmetros revisados.
Avaliação	• Avaliar os resultados;	• Avaliação dos resultados de MD no que diz respeito aos critérios de sucesso empresarial; • Modelos aprovados.
	• Processo de revisão	• Revisão do processo.
	• Determinar os próximos passos.	• Lista de ações possíveis; • Decisão.
Utilização, Aplicação ou Desenvolvimento	• Implantação do plano;	• Plano de implantação.
	• Plano de manutenção e monitoramento;	• Plano de manutenção e monitoramento;
	• Produzir o relatório final;	• Relatório final; • Apresentação final.
	• Projeto de revisão.	• Documentação da experiência.

Fonte: Adaptado de CHAPMAN *et al.* (2000)

2.5 POSTGRESQL

O Sistema Gerenciador de Banco de Dados (SGBD) escolhido para auxiliar no desenvolvimento deste trabalho foi o PostgreSQL. Além de ser utilizado pelo sistema SUAP (uma das fontes de dados utilizada), é um *software* livre e robusto. Os dados gravados por um *script* desenvolvido para ler os arquivos XML dos currículos também foram inseridos diretamente no PostgreSQL. Assim foi possível manter a compatibilidade do banco e consolidar dados de origens diferentes em uma mesma base de dados.

O PostgreSQL é um SGBD objeto-relacional de código aberto. Tem mais de 15 anos de desenvolvimento ativo e uma arquitetura que ganhou reputação de confiabilidade, integridade de dados e conformidade a padrões. Roda em todos os grandes sistemas operacionais. É compatível com ACID (Atomicidade, Consistência, Isolamento e Durabilidade), tem suporte completo a chaves estrangeiras, junções (*joins*), visões, gatilhos e procedimentos armazenados. Inclui a maior parte dos tipos de dados. Suporta também o armazenamento de objetos binários, incluindo figuras, sons ou vídeos e possui uma excepcional documentação (POSTGRESQLBRASIL, 2013).

O PostgreSQL suporta ainda conjuntos de caracteres internacionais, codificação de caracteres *multibyte*, *Unicode* e sua ordenação por localização, sensibilidade a caixa (maiúsculas e minúsculas) e formatação. É altamente escalável, tanto na quantidade enorme de dados que pode gerenciar, quanto no número de usuários concorrentes que pode acomodar. Existem sistemas ativos com o PostgreSQL em ambiente de produção que gerenciam mais de 4TB de dados (POSTGRESQLBRASIL, 2013).

Além de toda a documentação oficial do SGBD disponível em inglês, a comunidade brasileira tem se esforçado para traduzir a documentação oficial para o nosso idioma. A versão do PostgreSQL utilizada neste trabalho foi a 8.4.

2.6 FERRAMENTA WEKA

Para realização das tarefas de Mineração de Dados neste trabalho foi utilizada uma ferramenta chamada WEKA (versão 3.7.5), acrônimo que significa *Waikato Enviroment for Knowledge Analysis*.

O WEKA é um *software* livre do tipo *open source* (disponível em WEKA, 2013), desenvolvido em linguagem Java, dentro das especificações da GPL (*General Public License*). O sistema foi desenvolvido por um grupo de pesquisadores da Universidade de Waikato na Nova Zelândia no ano de 1999 e ao longo dos anos se consolidou como uma ferramenta de MD muito utilizada em ambientes acadêmicos. Muitos *papers* científicos relatam experiências onde a ferramenta foi aplicada de forma bem sucedida sobre bases de diferentes domínios (GONÇALVES, 2011).

Para Jain *et al*. (2010), as principais características responsáveis pelo sucesso do WEKA são: a disponibilização de vários algoritmos diferentes para MD e aprendizagem de máquina, ser *open source*, independente de plataforma, facilmente utilizada por pessoas que não são especialistas em MD, fornece instalações flexíveis para experimentos de *script,* tem-se mantido atualizada com a adição de novos algoritmos, conforme eles aparecem na literatura de pesquisa. Em suma, a equipe WEKA realizou uma excelente contribuição no campo da MD.

A WEKA pode ser utilizada no modo *console* ou interface gráfica, chamada *WEKA Explorer* (ilustrada na Figura 19). O símbolo do WEKA é o pássaro *Kiwi* (típico da região da Nova Zelândia).

Figura 19 - Interface Gráfica do WEKA

Fonte: Autoria própria

Conforme se observa na Figura 19, o WEKA fornece quatro aplicações principais:

- ***Explorer***: utilizada para pré-processamento e aplicação de técnicas de aprendizagem de máquina (durante a explicação da ferramenta maiores detalhes serão inseridos);
- ***Experimenter***: usada para fins de comparação entre as mais variadas técnicas de inteligência computacional; é um ambiente para realização de experimentos que permite a condução de testes estatísticos entre diferentes esquemas de aprendizado;
- ***KnowledgeFlow***: provê praticamente as mesmas funcionalidades que se pode obter com o Explorer, com uma pequena diferença: a sua interface suporta o estilo “arraste-e-solte” (*drag-and-drop*) de interação. O módulo *KnowledgeFlow* permite a realização de técnicas de aprendizado incremental;
- ***Simple CLI***: permite a realização de experimentos por meio de comandos (*prompt*), ideal para máquinas com pouco poder de processamento e que enfrentam problemas de desempenho para trabalhar com telas gráficas. Além disso, a interface *Simple CLI* supre a necessidade de sistemas que não possuem seus próprios aplicativos de linha de comando (ou ainda, auxilia o

desenvolvedor que está trabalhando em sistemas cujas interfaces de linha de comando não são suficientemente poderosas).

Das opções supracitadas, no corrente trabalho somente é utilizado a aplicação ***Explorer***.

Segundo Souza Filho (2006), os dois grandes atrativos do WEKA são a sua facilidade de uso, por conta de um conjunto de interfaces gráficas intuitivas e amigáveis, e a possibilidade de o pesquisador aprimorar a ferramenta para servir a propósitos mais específicos (por se tratar de um software de código aberto, permite a realização de alterações por quem esteja disposto a levá-las adiante).

Para Pasta (2011), o WEKA é composto por dois pacotes: um pacote autônomo, para manipulação direta dos algoritmos, usando o formato de dados próprio, e um pacote de classes em Java que implementam estes algoritmos. Nessa segunda forma, é possível desenvolver uma aplicação em linguagem Java que faça uso destes algoritmos e aplicá-los em quaisquer bancos de dados através de uma conexão JDBC (*Java DataBase Connectivity*).

Na aba *Preprocess* (vide Figura 20) pode-se escolher entre diversas aplicações de pré-processamento como discretização, normalização, transformação e combinação. As bases de dados poderão ser carregadas a partir de arquivos ARFF (*Attribute Relation File Format*), a partir de uma URL (botão *Open URL*) ou a partir de um Banco de Dados (botão *Open DB*). Além disso, o WEKA oferece diversas opções de filtros para o pré-processamento. Ao abrir pela primeira vez o programa, todas as abas além da *Preprocess* estarão desabilitadas e só se tornarão ativas a partir do momento que um arquivo de dados válido for carregado. Ao forçar uma sequência lógica de passos a seguir, os desenvolvedores do WEKA aproveitaram para simplificar a interação com o *software*, traçando uma maneira correta para sua utilização, evitando, dessa forma, que a maioria dos potenciais erros relacionados ao tratamento dos dados ocorra (SOUZA FILHO, 2006).

Através da aba *PreProcess* também existe a possibilidade de importar dados via URL (*Universal Resource Locator*) ou de um banco de dados em formato de linguagem SQL (*Structured Query Language*), através de conexão via JDBC.

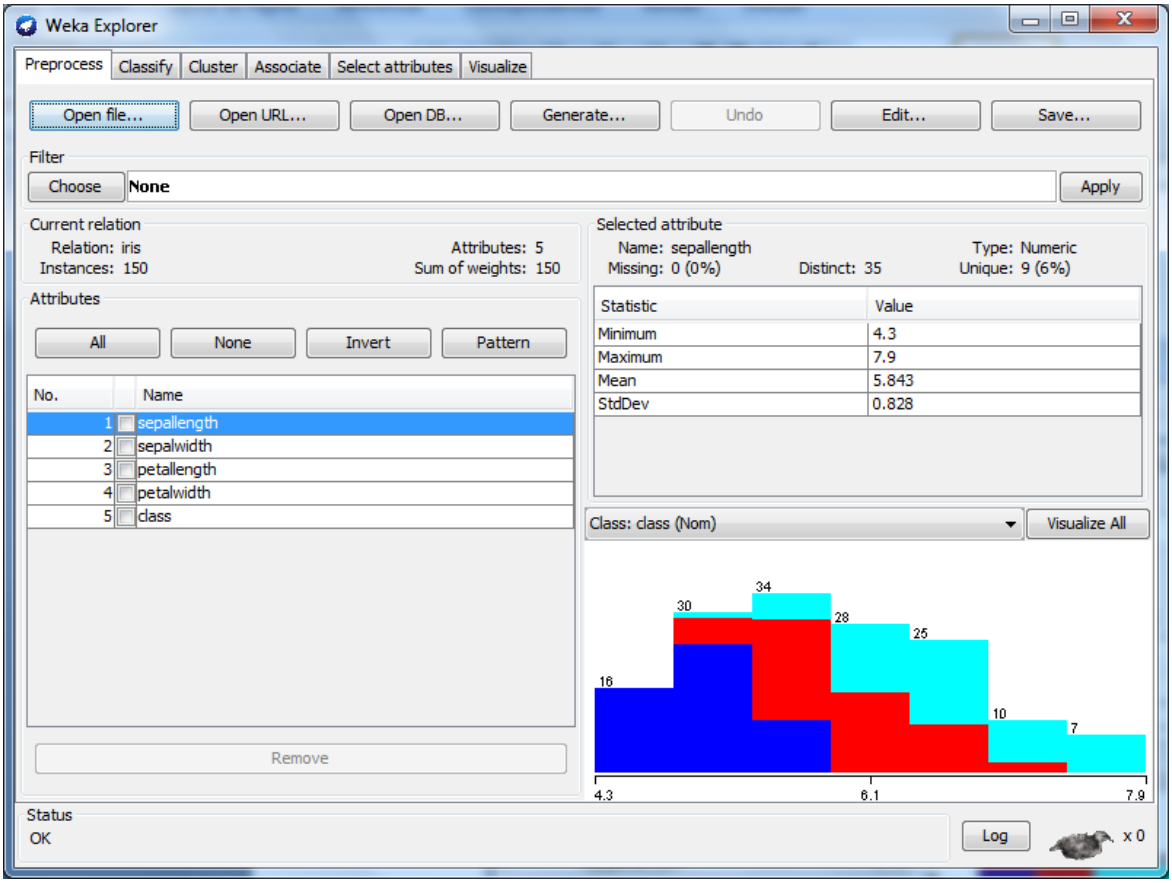


Figura 20 - Aba para Pré-Processamento dos dados no WEKA

Fonte: Autoria própria

Vários formatos de dados podem ser lidos no WEKA (csv, c4.5, binário), mas o formato padrão de dados utilizado na ferramenta é o ARFF (vide Figura 21). Ele é um arquivo texto que contém um conjunto de registros, precedido por um cabeçalho onde é declarado o nome da relação e todos os atributos com seus respectivos tipos de dados. A ordem da declaração indica a posição de cada atributo na seção DATA. Os atributos numéricos devem ser indicados com as palavras *numeric* ou *real*, e para os categóricos é necessário informar uma lista entre chaves indicando todos os valores possíveis para o atributo (como no atributo *class* da Figura 21). Também são suportados os tipos *date* e *string*.

Os dados devem ser colocados abaixo do parâmetro @DATA. Cada linha nesta seção representa um registro, e cada registro é representado por uma única linha, os dados são separados por vírgula e os valores nulos são indicados pelo ponto de interrogação (?). O WEKA trata o último atributo especificado no cabeçalho como o atributo “classe” e os demais como “preditivos” (GONÇALVES, 2011).

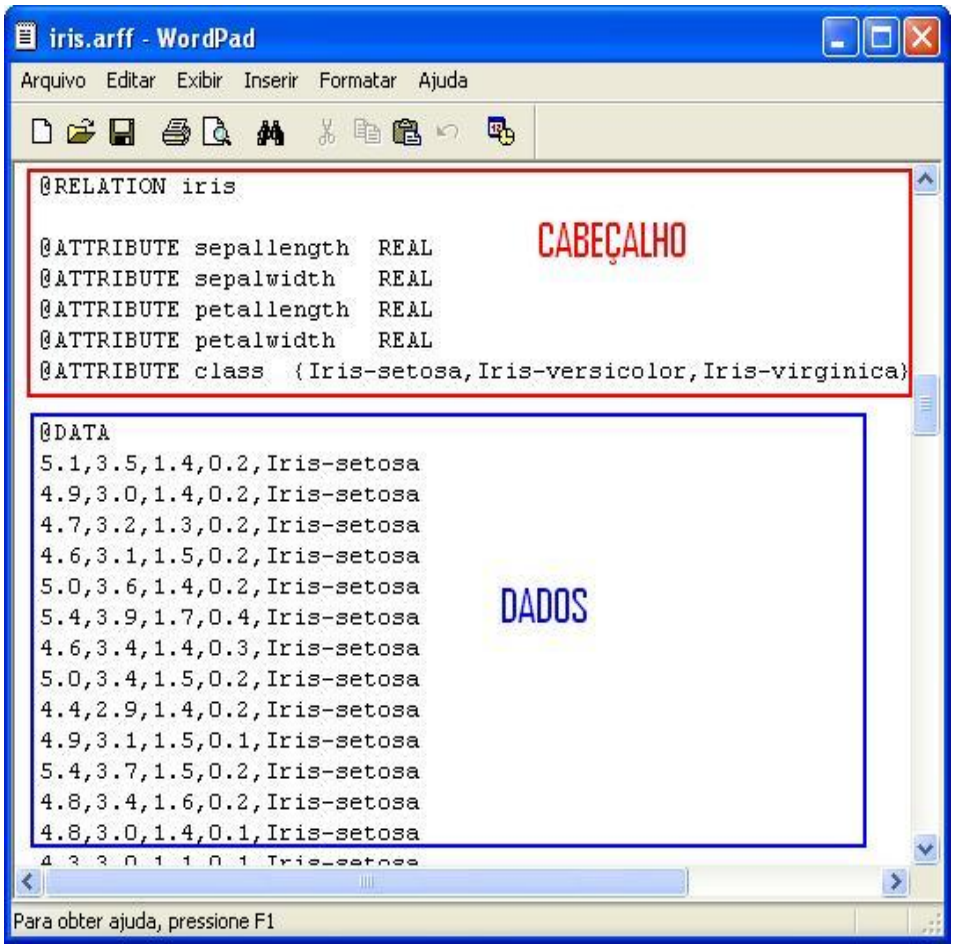


Figura 21 - Exemplo de arquivo ARFF

Fonte: Autoria própria

A ferramenta WEKA disponibiliza na aba *Classify,* quatro opções para seleção do conjunto de dados de teste, conforme mostra a Figura 22:

- *Use training set:* testa o mesmo conjunto de dados que foi utilizado para construir o modelo, sendo por isso a mais otimista delas;
- *Supplied test set*: testa o conjunto de dados informado em arquivo, de acordo com algum modelo que tenha sido previamente carregado;

- *Cross-validation*: significa validação cruzada. Segundo Santos *et al.* (2009), a técnica de Validação Cruzada consiste em dividir a base de dados em x partes (campo *folds*). Destas, x-1 partes são utilizadas para o treinamento e uma serve como base de testes. O processo é repetido x vezes, de forma que cada parte seja usada uma vez como conjunto de testes. Ao final, a correção total é calculada pela média dos resultados obtidos em cada etapa, obtendo-se assim uma estimativa da qualidade do modelo de conhecimento gerado e permitindo análises estatísticas;
- *Percentage Split*: o valor informado no campo "%" corresponde a porcentagem dos dados que será usada para construir o modelo, e o restante da diferença dos 100% corresponde a porcentagem dos dados utilizada para testar o modelo.

Figura 22 - Opções de escolha para o conjunto de teste na aba Classify do WEKA

Fonte: Autoria Própria

As características do WEKA e as técnicas implementadas são apresentadas em *Data Mining: Practical Machine Learning Tools and Techniques* (WITTEN *et al.*, 2011). Os autores deste livro são os próprios idealizadores da ferramenta.

Segundo Jain *et al.* (2010), o *SIGKDD Service Award* é o maior prêmio na área de MD e descoberta de conhecimento, destinado a um indivíduo ou grupo que tenha executado serviços significativos na área de MD. O prêmio SIGKDD do ano de 2005 foi concedido à equipe WEKA pelo desenvolvimento de seu *software.* O desenvolvimento do WEKA foi financiado por uma doação da Fundação do Governo da Nova Zelândia para investimento em Pesquisa, Ciência e Tecnologia.

Portanto, a escolha da ferramenta WEKA para a realização desse trabalho deve-se ao fato da mesma ser uma ferramenta gratuita, de interface simples e intuitiva, que implementa algoritmos confiáveis para tarefas de classificação, associação e agrupamento, possui ampla aplicabilidade (pois lida com atributos numéricos e categóricos) e tem se mostrado uma das ferramentas de MD mais utilizadas atualmente.

Os trabalhos correlatos citados de Cervi *et al.* (2009), Paula (2004), Romão (2002) e Morais (2010) também utilizaram a ferramenta WEKA para a Mineração de Dados.

3. PRODUÇÃO CIENTÍFICA

A pesquisa científica configura-se como atividade de caráter relevante em uma sociedade, visto que constitui uma das bases para a construção de um raciocínio crítico e reflexivo. O desenvolvimento da produção e publicação de trabalhos científicos possibilita o aprimoramento de estudos nas diversas áreas do conhecimento, servindo de referência para a consecução de novas pesquisas (VIEIRA *et al.*, 2011).

Para Leite Filho (2010), a produção científica inclui a produção de conhecimento através da pesquisa. Entende-se por pesquisa a busca sistemática, crítica e controlada de um maior conhecimento das relações existentes na realidade. Uma definição mais ampla de produção científica inclui trabalhos que possuem rigor científico no tratamento dos temas, incluindo-se neste universo, monografias, dissertações, teses e artigos.

Borba e Murcia (2006) mencionam que as pesquisas cujo enfoque é investigar as tendências e traçar o perfil de uma determinada área, em uma determinada época, têm ganhado força nos últimos anos, principalmente com a publicação de diversos trabalhos a respeito do assunto em vários meios de comunicação científica. Traçar o perfil de pesquisadores ou de uma determinada área, em uma determinada época, sob determinada perspectiva, reflete o comportamento de uma ordem, visando conhecer seus trajetos e fazer projeções de futuras tendências de estudos.

No Brasil, tem-se observado um incremento da pesquisa e da publicação científica, decorrente do aumento de professores e pesquisadores titulados, aumento na participação dos docentes em congressos nacionais e internacionais, expansão dos cursos de pós-graduação (*lato* e *stricto sensu*) e da pressão exercida pelos órgãos governamentais para que os docentes vinculados aos programas de pós-graduação tenham publicações científicas relevantes, pois um dos critérios de avaliação do

Ministério da Educação para as universidades inclui a produção intelectual dos docentes e pesquisadores (LEITE FILHO, 2010).

Cruz (2010) analisou, em seu artigo, algumas características do sistema de CT&I no Brasil. Para sintetizar a situação do sistema nacional, 3 indicadores de resultados foram tomados: o número de artigos científicos publicados em revistas de circulação internacional da base do *Institute For Scientific Information* (ISI), o número de doutores formados (dados mantidos pela CAPES/MEC) e o número de patentes obtidas por organizações no país no Escritório de Patentes dos EUA. Os dois primeiros indicadores, artigos e doutores formados, dão uma boa ideia da situação do sistema acadêmico de pesquisa, permitindo comparações internacionais elucidativas. Eles se relacionam com várias outras dimensões do sistema acadêmico, como abrangência, acesso, qualidade da educação básica, por serem afetados por estas. O indicador sobre o número de patentes permite que se forme uma ideia sobre a competitividade internacional das empresas em um mundo globalizado, no qual a criação de ideias é o principal criador de riqueza para a indústria e os serviços.

Ainda segundo Cruz (2010), houve em 2008, um aumento importante em relação a 2007, na quantidade de artigos científicos e outros tipos de publicações em revistas de circulação internacional cadastradas pelo ISI (vide Figura 23), pois este passou a cadastrar mais revistas editadas no Brasil, o que também é um bom indicador de interesse mundial pela ciência aqui produzida. No entanto, exatamente por esta razão (mudança da base), a série de número de publicações obtida desta forma não deve ser usada para a análise da evolução do sistema de produção de ciência no país. No mesmo ano de 2008, titularam-se no Brasil 10.711 doutores e organizações sediadas no Brasil obtiveram 101 patentes no Escritório de Patentes dos EUA. A Figura 23 mostra a evolução da quantidade de publicações de artigos científicos originados no Brasil e cadastrados no *Science Citation Index* do ISI, em comparação com as publicações na Espanha, Coréia e Índia.

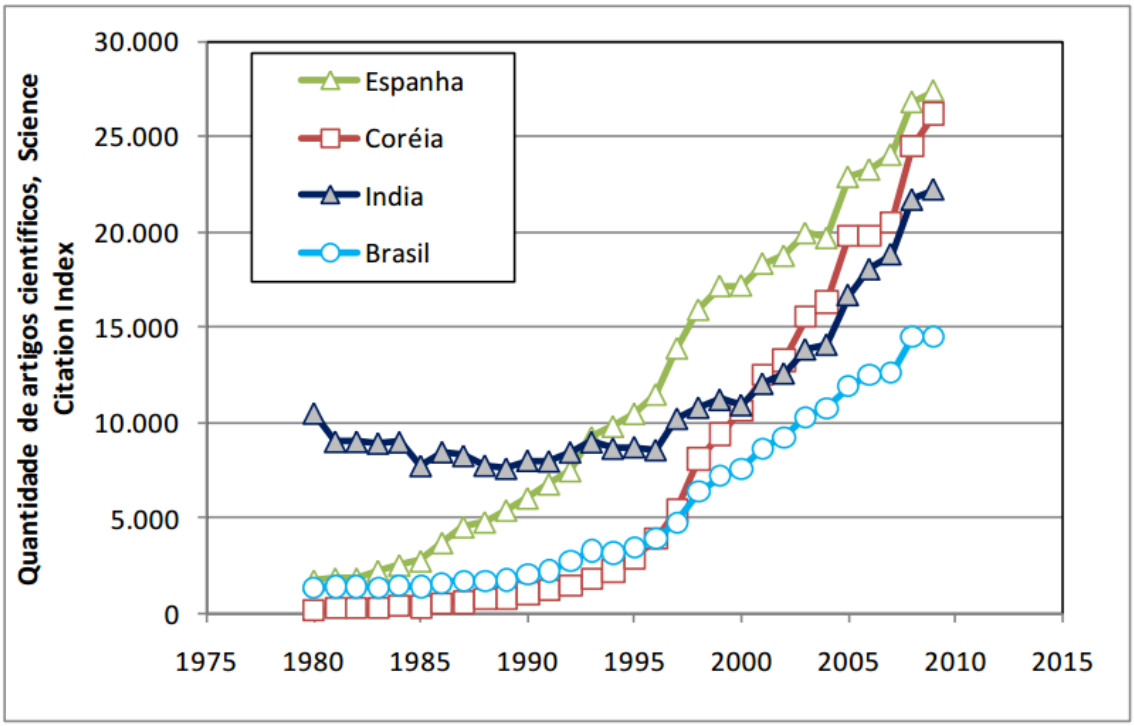


Figura 23 - Evolução da quantidade de publicações do tipo "*Article*" originadas no Brasil e cadastradas no *Science Citation Index* do ISI

Fonte: Cruz (2010)

Através da Figura 23 pode ser verificado que as taxas de crescimento anual foram bastante altas desde 1994, com a exceção das variações de 2006 para 2007 e 2008 para 2009. Apesar das taxas altas de crescimento, a produção científica brasileira colocou-se abaixo dos países comparados: Espanha, Coréia e Índia. Em particular, chamou a atenção à divergência em relação à trajetória da Coréia: até 1996 a produção científica brasileira superava a deste país, mas a partir de 1997 a Coréia superou o Brasil. Porém, deve ser mencionado que a capacidade de produção científica brasileira excedeu bastante a dos demais países da América Latina.

A Figura 24 apresenta a evolução do número de doutores formados anualmente no Brasil em comparação com as trajetórias de outros países. Um dos desafios relacionou-se com a mudança de tendência que pode ser observada a partir de 2003: de 1995 a 2002, a taxa de crescimento do número de doutores formados anualmente foi de 14% a.a., caindo para 5,4% a.a. a partir de 2003. Outro desafio ligado à formação de doutores foi a pequena intensidade de convivência internacional

dos titulados. A pós-graduação no Brasil avançou muito ao criar oportunidades para doutoramento no país, especialmente a partir da década de 80 do século passado. Mas uma consequência imprevista desta “nacionalização” foi a redução da intensidade de criação de redes e parcerias internacionais. O isolamento, agravado pela barreira linguística, prejudicou o progresso da ciência no Brasil e também a qualidade da formação dos doutores titulados, pois, como é bem sabido, a ciência avança mais e melhor quando há mais interação entre os cientistas, especialmente com os melhores entre eles.

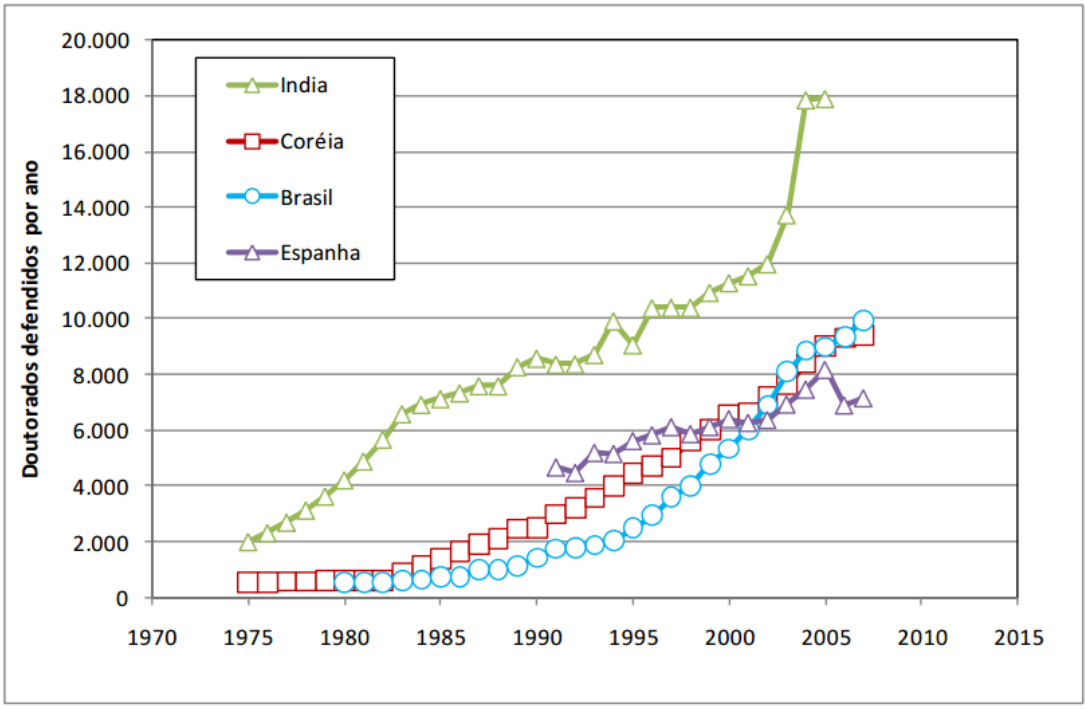


Figura 24 - Evolução da quantidade doutores formados anualmente

Fonte: Cruz (2010)

No que diz respeito à capacidade inovativa das empresas localizadas no país, a situação foi bem menos favorável do que a analisada quanto à produção científica ou à formação de doutores, conforme mostra a Figura 25.

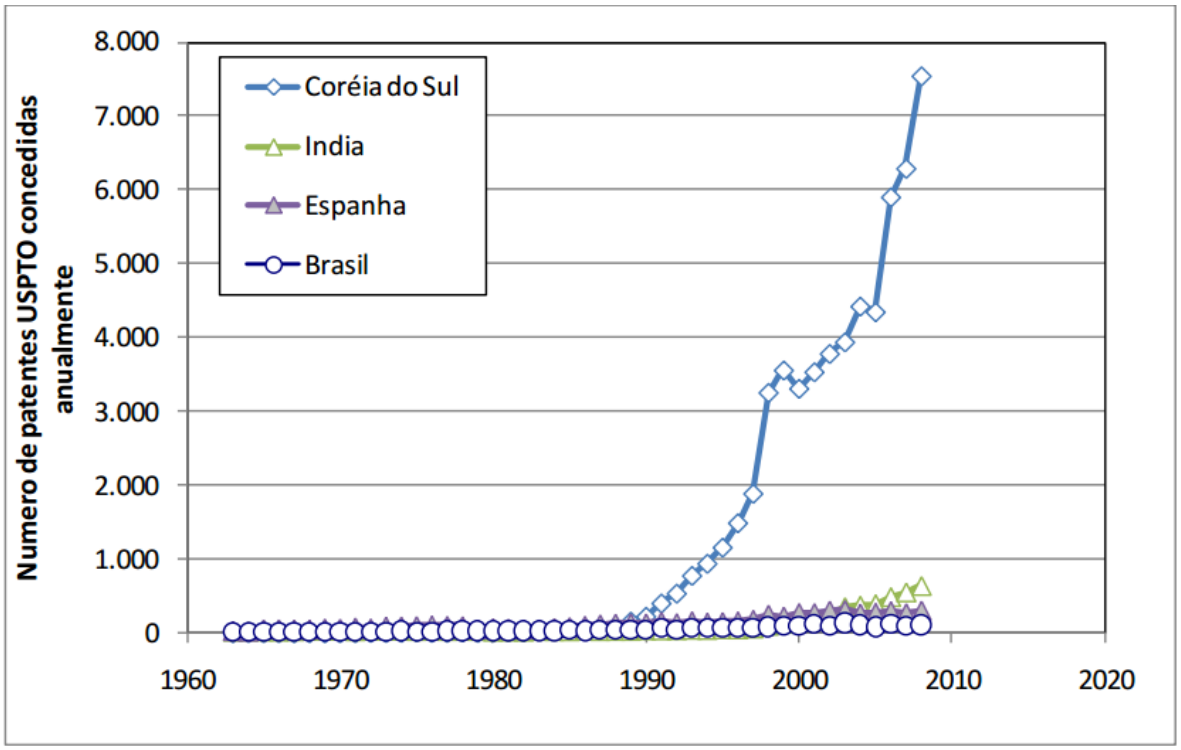


Figura 25 - Evolução na quantidade de patentes concedidas no Escritório de Patentes dos EUA à Coréia, Espanha, Índia e Brasil

Fonte: Cruz (2010)

Neste indicador, a Coréia superou os demais países: em 2008, as empresas coreanas obtiveram 7.549 patentes nos EUA enquanto as sediadas no Brasil apenas 101. É preciso esclarecer que a principal origem de patentes são empresas e não universidades. No caso dos EUA, das 87.901 patentes concedidas a organizações no país em 2003, apenas 4% foram para universidades. As demais foram quase na totalidade para empresas. Portanto, quando se fala da quantidade de patentes obtidas, está-se falando da capacidade da empresa daquele país de criar conhecimentos e incorporá-los efetivamente a seus produtos e processos.

Oito universidades responderam por aproximadamente 2/3 dos artigos científicos publicados em periódicos internacionais, conforme mostrado na Tabela 11. A Universidade de São Paulo, que possuía um corpo docente de 5.420 professores com doutorado, gerou, em 2008, 26% dos artigos científicos internacionais do país, seguida pela Universidade Estadual de Campinas que, com um corpo docente de

1.700 professores publicou, no mesmo ano, 9% da produção científica do Brasil (CRUZ, 2010).

Tabela 11 - Número de artigos científicos publicados pelas 8 principais universidades de pesquisa no Brasil, comparado com a produção científica total do país

	2000	2001	2002	2003	2004	2005	2006	2007	2008
USP	2.568	2.651	3.141	3.606	3.763	3.955	3.924	4.642	4.844
UNICAMP	1.111	1.110	1.350	1.418	1.517	1.594	1.601	1.645	1.636
UFRJ	1.041	1.036	1.086	1.185	1.200	1.287	1.214	1.332	1.416
UNIFESP	335	456	461	390	658	871	778	986	1.074
UFRGS	446	592	644	717	750	836	864	935	1.037
UFMG	484	546	559	677	632	762	799	865	959
UNESP	364	280	446	547	438	461	491	417	544
UFSC	243	255	308	197	351	372	393	409	530
Total	6.592	6.926	7.995	8.737	9.309	10.138	10.064	11.231	12.040
Brasil	9.786	10.330	11.662	13.512	13.904	14.880	14.955	16.638	18.783
Total/Brasil	67%	67%	69%	65%	67%	68%	67%	68%	64%

Contagem de itens constantes na base do ISI Science Citation Index em CD-ROM da Unicamp, incluindo Artigos, resenhas, revisões e outros itens

Fonte: Cruz (2010)

Ribeiro *et al.* (2010) realizou em seu artigo, uma análise comparativa de indicadores de produção científica da Rede Federal de Educação Profissional Científica e Tecnológica (RFEPCT). Segundo as autoras, o número de cursos de pós-graduações está diretamente relacionado com a quantidade e qualidade da produção científica no Brasil. Sendo assim, é de fundamental importância a existência de cursos de pós-graduação na RFEPCT, em todas as regiões do país, já que uma das propostas desta Rede constitui-se em incentivar a pesquisa e o desenvolvimento técnico-científico do Brasil. Percebeu-se que na RFEPCT havia uma discrepância entre o número de cursos e pós-graduação *stricto sensu* (mestrados e doutorados) entre as regiões geográficas (Figura 26) nos Institutos Federais (IFs) estudados.

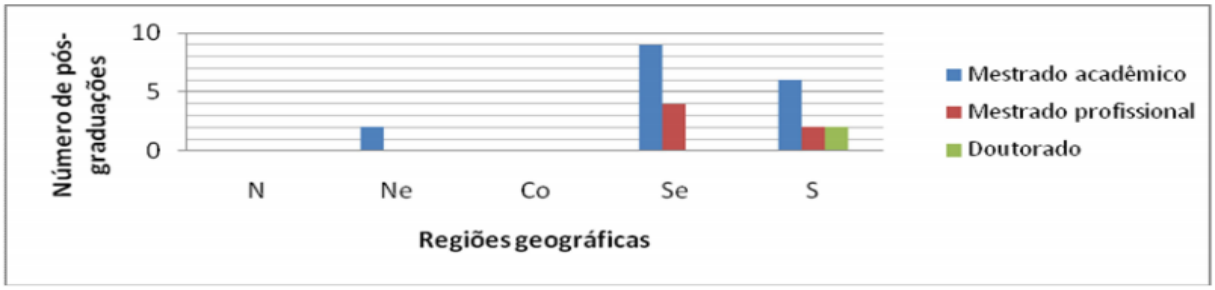


Figura 26 - Número de pós-graduações da RFEPCT, segundo a região geográfica (dados da CAPES em março de 2010)

Fonte: Ribeiro *et al*. (2010)

De acordo com a Figura 26, até início de 2010, a região Sudeste contou com o maior número de cursos de pós-graduação *stricto sensu* (total de 13 cursos), seguida da região Sul (10 cursos) e do Nordeste (2 cursos). As instituições pertencentes à RFEPCT das regiões Norte e Centro-Oeste ainda não ofereciam cursos nesta modalidade. O número de pós-graduações *stricto sensu* na modalidade mestrado ainda era muito baixo em toda a RFEPCT. Metade das instituições analisadas (IFAM, IFRR, IFTO, IFBA, IFCE, IFRN, IFGOIANO, IFMT, IFES, IFF, IFRJ, IFSP, IFSC, IFSUL) não apresentaram curso nessa modalidade, constituindo-se assim num fator limitante para o desenvolvimento de pesquisa e inovação.

Quanto maior o quantitativo de docentes com o título de mestrado e doutorado, maior a probabilidade de realização de publicações com alto índice de citações e ou de depositar patentes de invenção. Perceptivelmente, a distribuição de docentes doutores distribuiu-se de forma bastante heterogênea nas instituições em análise (Figura 27), sendo encontrados em menor número nos IFs da região Norte (RIBEIRO *et al*., 2010).

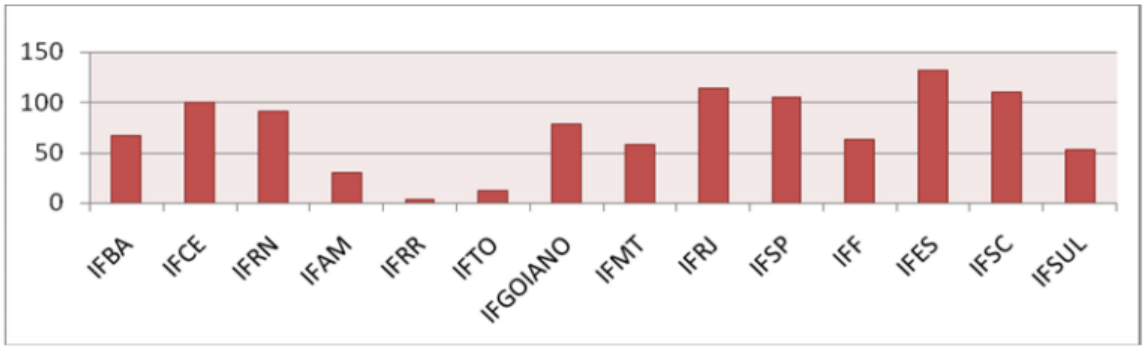


Figura 27 - Quantitativo de doutores dos IFs em estudo com título de doutorado

Fonte: Ribeiro *et al.* (2010)

Para Ribeiro *et al.* (2010), por meio da análise e discussão de fatores inerentes à produção científica, percebeu-se que ao longo dos cem anos de existência, a RFEPCT progrediu acompanhando os avanços do país. Em consonância com os interesses relativos ao desenvolvimento técnico-científico e econômico do país, uma série de esforços foi realizada para propagar a importância da pesquisa e inovação

para a nação. Isso pode ser constatado não pelas mudanças de nomenclatura institucional que ocorreram (desde Escola de Aprendizes e Artífices até Institutos Federais de Educação, Ciência e Tecnologia), mas pelos resultados apresentados, que revelam significativa melhora na área de produção científica. Em contrapartida, ainda podem ser dados largos passos.

Para Leite Filho (2010, p. 3), “a veiculação pela qual se processa as comunicações científicas pode ser descrita através de livros, periódicos, teses, dissertações, anais e congressos”. A publicação de trabalhos científicos é fundamental para o processo de difusão do conhecimento, permitindo o progresso da ciência e servindo de subsídio para a consecução de novas pesquisas. Nesse sentido, a existência de meios de divulgação e discussão de trabalhos científicos é essencial para o desenvolvimento das ciências. Neste contexto, reside a importância dos trabalhos que investigam a produção intelectual em uma determinada área do conhecimento.

3.1 PLATAFORMA LATTES (PL)

A Plataforma Lattes (PL) é um conjunto de sistemas que representa a experiência do CNPq na integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um único Sistema de Informações. A PL é uma base de dados pública. Sua dimensão atual se estende não só às ações de planejamento, gestão e operacionalização de fomento do CNPq, mas também de outras agências de fomento federais e estaduais, das fundações estaduais de apoio à ciência e tecnologia, das instituições de ensino superior e dos institutos de pesquisa. Além disso, se tornou estratégica não só para as atividades de planejamento e gestão, mas também para a formulação das políticas do Ministério da Ciência e Tecnologia (MCT) e de outros órgãos governamentais da área de CT&I (CNPq, 2013a).

Conforme descrito em CNPq (2013a), os subsistemas que integram a PL, são:

- **Currículo Lattes (CL):** se tornou um padrão nacional no registro da vida pregressa e atual dos estudantes e pesquisadores do país, e é hoje adotado pela maioria das instituições de fomento, universidades e institutos de pesquisa do país. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, se tornou elemento indispensável e compulsório à análise de mérito e competência dos pleitos de financiamentos na área de C&T;
- **Diretório dos Grupos de Pesquisa:** é um projeto desenvolvido no CNPq desde 1992, constituindo-se em bases de dados que contêm informações sobre os grupos de pesquisa em atividade no país. O Diretório mantém uma base corrente, cujas informações são atualizadas continuamente pelos líderes de grupos, pesquisadores, estudantes e dirigentes de pesquisa das instituições participantes, e o CNPq realiza censos bianuais, que são fotografias dessa base corrente. As informações contidas nessas bases dizem respeito aos recursos humanos constituintes dos grupos (pesquisadores, estudantes e técnicos), às linhas de pesquisa em andamento, às especialidades do conhecimento, aos setores de aplicação envolvidos, à produção científica e tecnológica e aos padrões de interação com o setor produtivo. Além disso, cada grupo é situado no espaço (região, UF e instituição) e no tempo.

  Os grupos de pesquisa inventariados estão localizados em universidades, instituições isoladas de ensino superior, institutos de pesquisa científica, institutos tecnológicos e laboratórios de pesquisa e desenvolvimento de empresas estatais ou ex-estatais. Os levantamentos não incluem os grupos localizados nas empresas do setor produtivo;
- **Diretório de Instituições:** foi concebido para promover as organizações do Sistema Nacional de CT&I à condição de usuárias da PL. Ele registra todas e quaisquer organizações ou entidades que estabelecem algum tipo de

> relacionamento com o CNPq (instituições nas quais os estudantes e pesquisadores apoiados pelo CNPq desenvolvem suas atividades; instituições onde os grupos de pesquisa estão abrigados, usuárias de serviços prestados pela Agência, como o credenciamento para importação pela Lei 8.010/90; instituições que pleiteiam participar desses programas e serviços, etc). A disponibilização pública dos dados da Plataforma na internet dão maior transparência e mais confiabilidade às atividades de fomento do CNPq e das agências que a utilizam, fortalecem o intercâmbio entre pesquisadores e instituições e é fonte inesgotável de informações para estudos e pesquisas. Na medida em que suas informações são recorrentes e cumulativas, têm também o importante papel de preservar a memória da atividade de pesquisa no país.

O Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) é uma agência do Ministério da Ciência e Tecnologia (MCT). Suas principais atribuições são o fomento a pesquisa científica e tecnológica e o incentivo a formação de pesquisadores brasileiros. O Conselho foi criado em 1951, e desde então desempenha papel primordial na formulação e condução das políticas de ciência, tecnologia e inovação. Sua atuação contribui para o desenvolvimento nacional e o reconhecimento das instituições de pesquisa e pesquisadores brasileiros pela comunidade científica internacional (CNPq, 2013b).

O nome da plataforma é uma homenagem a um dos maiores cientistas brasileiros, o físico Césare Mansueto Giulio Lattes, mais conhecido como César Lattes. Tal cientista tornou-se um ícone na produção científica mundial e um símbolo, para o Brasil, que serviu de inspiração e estímulo para as gerações seguintes (CNPq, 2013b).

A Plataforma Lattes (PL) oferece ao usuário, várias estatísticas sobre a base de currículos cadastrados. Através do Painel Lattes (acessível em CNPq, 2013a) pode-se obter informações qualificadas e atualizadas sobre a atuação de pesquisadores em CT&I cadastrada por intermédio do CL. Estas informações são

recursos valiosos na análise dos diversos aspectos da CT&I. As informações são organizadas em forma de painéis com dados qualificados por geografia, sexo, instituição, faixa etária, área de atuação, entre outros, e extraída da base de currículos atualizados nos últimos 48 meses.

Segundo CNPq (2013a), informações extraídas em 30 de Junho de 2013 contabilizaram um total de 2.601.696 currículos cadastrados na PL, onde 1.009.318 destes eram de estudantes.

3.1.1 Histórico da Plataforma Lattes

Conforme descrito em CNPq (2013a) desde os anos de 1980, já havia entre os dirigentes do CNPq a preocupação pela utilização de um formulário padrão para registro dos currículos dos pesquisadores brasileiros. Os objetivos deste formulário seriam, além de permitir a avaliação curricular do pesquisador, a criação de uma base de dados que possibilitasse a seleção de consultores e especialistas, e a geração de estatísticas sobre a distribuição da pesquisa científica no Brasil. Foi então criado um sistema denominado Banco de Currículos que, à época, contava com formulário de captação de dados em papel e etapas de enquadramento e digitação de dados em um sistema informatizado.

No final dos anos de 1980, o CNPq já disponibilizava às universidades e instituições de pesquisa do país, através da rede BITNET, precursora da Internet no Brasil, buscas sobre a base de currículos de pesquisadores brasileiros. À época, a base de dados contava com cerca de 30.000 currículos.

No início dos anos de 1990, o CNPq desenvolveu formulário eletrônico para a captação de dados curriculares para o Sistema Operacional DOS, denominado BCUR. Os pesquisadores preenchiam o formulário e o enviavam em disquete ao CNPq, que os carregava na base de dados.

Com a disseminação do Sistema Operacional *Windows* no meio acadêmico, o CNPq disponibilizou, juntamente com os formulários eletrônicos para automatização dos programas de bolsas à pós-graduação e habilitação de orientadores, o Currículo Vitae do Orientador para o ambiente *Windows*. Devido ao estágio ainda inicial do uso da Internet no Brasil, a rede foi utilizada apenas como meio para o envio de dados gerados de forma *off-line* pelos respectivos formulários eletrônicos. Pouco tempo depois, uma outra versão de formulário eletrônico para cadastramento de dados curriculares foi desenvolvida pelo MCT e denominado Cadastro Nacional de Competência em Ciência e Tecnologia - CNCT.

Ao final dos anos de 1990, o CNPq contratou os grupos universitários Stela, vinculado à Universidade Federal de Santa Catarina, e C.E.S.A.R, da Universidade Federal de Pernambuco, para que, juntamente com profissionais da empresa Multisoft, e técnicos das Superintendências de Informática e Planejamento, desenvolvessem uma única versão de currículo capaz de integrar as já existentes.

Assim, em agosto de 1999, o CNPq lançou e padronizou o Currículo Lattes (CL) como sendo o formulário de currículo a ser utilizados no âmbito do MCT e CNPq.

Desde então, o CL vem aumentando sua abrangência, sendo utilizado pelas principais universidades, institutos, centros de pesquisa e fundações de amparo à pesquisa dos estados como instrumento para a avaliação de pesquisadores, professores e alunos.

No final do ano de 2002, e após o desenvolvimento de uma versão em língua espanhola do CL, o CNPq, juntamente com a Bireme/OPAS cria a rede *ScienTI*. Essa rede, formada por Organizações Nacionais de Ciência e Tecnologia e outros Organismos Internacionais, teria o objetivo de promover a padronização e a troca de informação, conhecimento e experiências entre os participantes na atividade de apoio a gestão da área científica e tecnológica em seus respectivos países. Como forma de incentivar a criação das bases nacionais de currículos, o CNPq passou a licenciar gratuitamente o software e fornecer consultoria técnica para a implantação do CL nos

países da América Latina. Assim, o CL foi implantado em países como Colômbia, Equador, Chile, Perú, Argentina, além de Portugal, Moçambique e outros que se encontram em processo de implantação.

Em julho de 2005, a presidência do CNPq cria a Comissão para Avaliação do Lattes, composta por pesquisadores de diversas áreas do conhecimento, com o objetivo de avaliar, reformular e aprimorar a PL, corrigindo possíveis desvios e promovendo o aperfeiçoamento da ferramenta.

A atualização da PL visou sempre torná-la mais racional, prática e confiável. As críticas e sugestões consideradas necessárias, devem ser encaminhadas ao CNPq que adotará as iniciativas necessárias para que as mesmas sejam utilizadas como refinamento para outras mudanças.

Conforme CNPq (2013b), a mais nova versão da PL foi lançada oficialmente dia 23 de Julho de 2012, durante a Conferência "Ciência, Tecnologia e Inovação como protagonistas do Desenvolvimento Sustentável", proferida pelo ministro da Ciência, Tecnologia e Inovação, Marco Antônio Raupp, na 64ª Reunião da Sociedade Brasileira para o Progresso da Ciência (SBPC), em São Luís (MA). A nova versão traz funcionalidades que introduzem possibilidades de analisar de forma sistemática as atividades de ciência e tecnologia.

A PL passou a contar com abas em que a comunidade científica e tecnológica poderá registrar informações sobre inovação, educação e popularização da ciência e tecnologia e patentes e registros, que ganharam um módulo específico. As informações disponibilizadas neste sistema deixam de ser somente declaratórias e acrescentam o elemento de confiabilidade dos dados.

Para aqueles que informam suas inovações, é necessário indicar o CNPJ da empresa onde foi desenvolvido o projeto, entre outras informações. Após preencher os dados, um e-mail é encaminhado ao responsável pela empresa onde ocorreu a colaboração para certificação do projeto. Já na aba Patentes, o pesquisador pode incluir o número da patente e todos os dados serão recuperados na base de dados do

Instituto Nacional de Propriedade Industrial (INPI). Após isso, ao lado da patente indicada, aparece o símbolo de certificação deste instituto.

Em virtude das mudanças na nova versão da PL, os critérios de avaliação de projetos do CNPq passam a considerar o mérito científico do projeto; a relevância, originalidade e repercussão da produção científica do proponente; a formação de recursos humanos; a contribuição científica, tecnológica e de inovação incluindo patentes; a inserção internacional da pesquisa; a contribuição em educação e popularização da ciência entre outros quesitos.

3.1.2 Estrutura do Currículo Lattes (CL)

O CL 2.0 (sua última versão), um dos subsistemas da PL e objeto específico desse trabalho, está estruturado de forma hierárquica. Os níveis dessa hierarquia são:

- **Apresentação**: módulo inicial do sistema, é composto por um resumo do currículo do usuário, com a última data de atualização do mesmo, além de mais três *links* que direcionam para o currículo completo, a rede de colaboração e os indicadores de pesquisa desse usuário;
- **Dados Gerais**: este módulo (menu) agrupa os dados de identificação, endereços, idiomas, prêmios e títulos, além de outras informações relevantes;
- **Formação:** agrupa toda a formação acadêmica (titulação) e formação complementar do usuário;
- **Atuação:** agrupa todo o histórico de atuações profissionais do usuário até o vínculo atual, assim como suas linhas de pesquisa, suas áreas de atuação, entre outras informações como: membro de corpo editorial, membro de comitê de assessoramento, revisor de periódico, revisor de projeto de agência de fomento, etc. Vale lembrar que cada currículo poderá listar submenus

diferentes em cada módulo, pois o submenu só é listado se as informações tiverem sido preenchidas;

- **Projetos:** agrupa os projetos do usuário, sejam eles de pesquisa, de extensão, de desenvolvimento tecnológico ou outros;
- **Produções:** agrupa todas as produções científicas do usuário. Por isso, ela está dividida em três colunas:

  - **Produção Bibliográfica**: compreende os artigos completos publicados em periódicos, artigos aceitos para publicação, livros e capítulos, texto em jornal ou revista, trabalhos publicados em anais de eventos, apresentação de trabalho e palestra, partitura musical, tradução, prefácio, posfácio, além de outras produções bibliográficas;

  - **Produção Técnica**: incluem assessoria e consultoria, extensão tecnológica, programa de computador, produtos, processos ou técnicas, trabalhos técnicos, cartas, mapas, cursos de curta duração, desenvolvimento de material didático, editoração, manutenção de obra artística, maquete, entrevistas, programas e comentários na mídia, relatório de pesquisa, redes sociais, websites, blogs, entre outras produções técnicas;

  - **Produção Artística/Cultural**: incluem as artes cênicas e visuais, músicas, entre outras produções artísticas e culturais;
- **Patentes e Registros**: agrupa as patentes, programas de computador, cultivares protegidas e registradas, desenho industrial e marca registrada, topografia de circuito integrado registrado entre outros. O sistema deverá solicitar o número da patente para que os dados sejam recuperados de forma automática nos cadastros do INPI;
- **Inovação**: agrupa projetos com potencial de inovação;
- **Educação e Popularização de C&T**: menu criado para separar as produções que são consideradas como sendo de educação e de popularização da C&T;

- **Eventos**: agrupa as participações e organizações do usuário em eventos, congressos, exposições, feiras, olimpíadas, etc;
- **Orientações**: agrupa as orientações e supervisões concluídas e em andamento;
- **Bancas**: agrupa as participações em bancas de trabalhos de conclusão e em bancas de comissões julgadoras;
- **Citações**: agrupa as citações que o usuário teve em outros trabalhos de acordo com as bases bibliográficas como ISI, SciELO, SCOPUS, etc.

A Figura 28 e Figura 29 ilustram as opções de menu e submenus da versão do Currículo Lattes (CL) 2.0 supracitadas.

Figura 28 - Menus do CL 2.0
Fonte: CNPq, 2013c

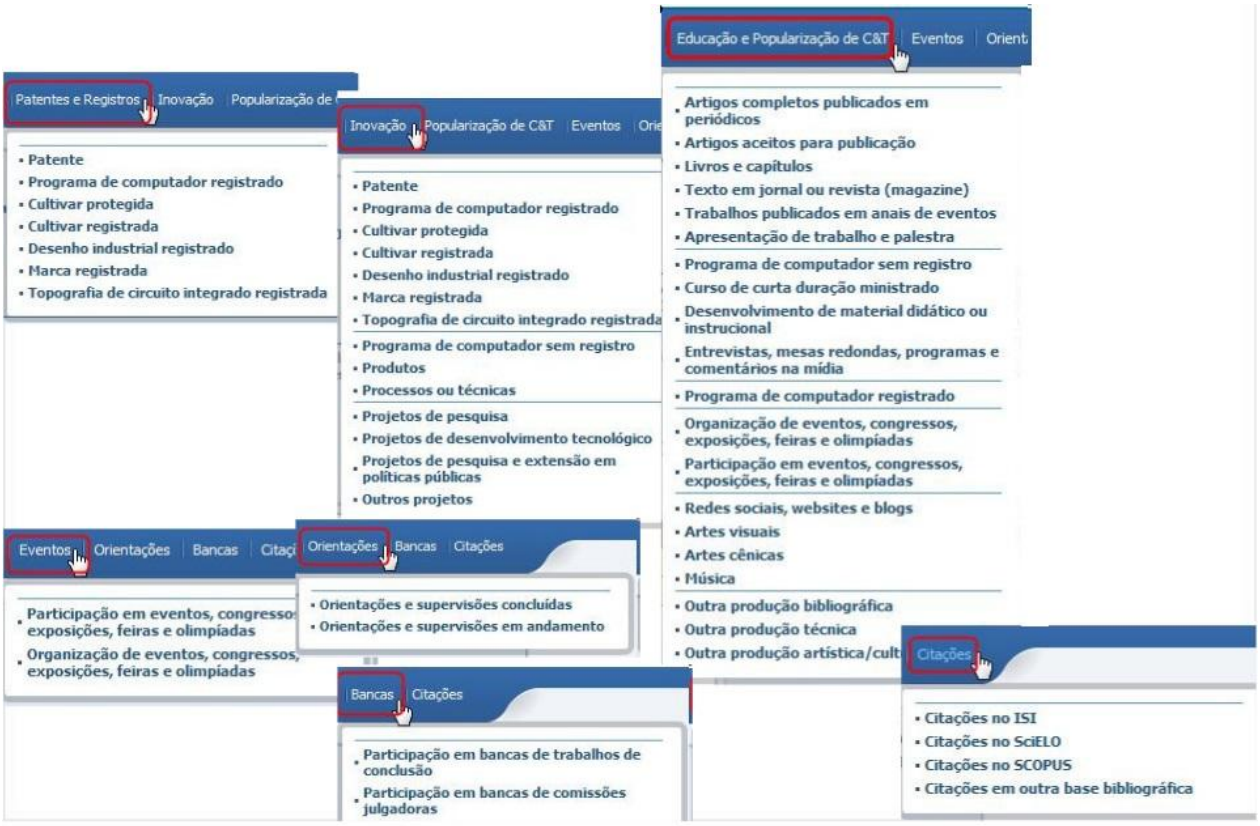


Figura 29 - Continuação dos Menus do CL 2.0
Fonte: CNPq, 2013c

Algumas das principais novidades da nova versão do CL 2.0 são:

- visualização da rede de colaboradores do usuário;
- identificação de coautores nas citações bibliográficas;
- inclusão de novos campos: passaporte, novas opções de contato (*facebook*, ICQ, *Google Talk*, *skype*, *Linkedin*, *Twitter*, etc), libras no menu idioma, etc;
- Subdivisão dos menus “Dados Gerais” e “Projetos”, para facilitar a usabilidade e preenchimento;
- certificação dos projetos pela empresa ou pelo coordenador, garantindo a originalidade do mesmo;
- preenchimento automático da opção “possui vínculo empregatício” em “Atuação Profissional”;
- inclusão de novos módulos: membro de comitê de assessoramento, revisor de projeto de agência de fomento, alteração no módulo de formação acadêmica

para contemplar o tipo de formação sanduíche, Patentes e Registros, Inovação (para projetos considerados inovação), Educação e Popularização de C&T, Citações, etc;

- Agrupamento dos menus de tipos de produção científica em um único menu “Produções” e adição do módulo de produção artística e cultural;
- nova busca textual;
- exibição de forma gráfica dos números referentes a citações, produção bibliográfica, produção técnica, orientações concluídas e todas as produções.

3.2 TRABALHOS CORRELATOS

Esta dissertação apresenta o uso da Mineração de Dados (MD) sobre os dados da Plataforma Lattes, com o intuito de descobrir conhecimento a respeito da produção científica dos docentes do IFG. Foram encontrados na literatura diversos trabalhos afins, o que realça a importância da MD no auxílio à tomada de decisão dentro do contexto da gestão de ambientes educacionais. Portanto, nesta seção serão apresentados alguns dos principais trabalhos encontrados, relacionados ao tema proposto.

Cardoso e Machado (2008) utilizaram a Plataforma Lattes (PL) como base para a aplicação e análise de uma ferramenta de MD com o objetivo de extrair informações a respeito da produção científica de colaboradores e principalmente professores da Universidade Federal de Lavras (UFLA). Foram selecionados 575 currículos para a pesquisa.

Sobre as regras de associação, as autoras realizaram 4 exemplos:

- Associação entre a quantidade de publicações contidas na PL, desenvolvidas por pessoas que trabalhavam na UFLA e as pessoas que não trabalhavam. Como resultado, obtiveram uma amostra com 1.977 publicações, das quais

55% eram publicações de pessoas que não estavam atuando na UFLA na época da publicação e o restante, 45% de pessoas que estavam atuando;

- As autoras analisaram os resultados obtidos no exemplo anterior, referente aos 55% de pessoas que tiveram alguma publicação, mas não estavam atuando na Universidade Federal de Lavras (UFLA). Como resultado obtiveram uma quantidade de 1.062 publicações. As autoras alertam que “uma pessoa, ao receber afastamento para fazer pós-graduação, por exemplo, não está atuando na UFLA durante o período do afastamento”;
- Associação entre as publicações cadastradas e o tempo de serviço de seus autores na UFLA, tendo como resultado a caracterização de que a maioria das publicações foi realizada após o ingresso do autor na UFLA;
- As autoras fizeram a junção de duas situações: o local de realização de uma pós-graduação, se no exterior ou no Brasil e o número de publicações feitas. O resultado revelou que: “A média de publicações no exterior de pessoas que cursaram a pós-graduação fora do Brasil é maior numa razão de 2,71 com relação às pessoas que cursaram pós-graduação no Brasil.”

As autoras elaboraram ainda mais quatro análises:

- Análises de regras de associação e de padrão sequencial: onde analisaram o tempo decorrido entre a conclusão do mestrado e o início do doutorado realizado pelas pessoas que trabalhavam na UFLA. O resultado mostrou que a maioria das pessoas leva de zero a três anos de intervalo entre esses dois tipos de pós-graduação.
- Análises de padrões sequenciais: duas consultas foram realizadas: a primeira analisou a relação entre o tempo de cadastro do currículo na Plataforma Lattes e o tempo de vínculo profissional com a instituição e a segunda, analisou a relação temporal entre o tempo de serviço e o ano de início das pesquisas realizadas pelo colaborador;

- Análises de *clusters:* através da identificação de um *cluster* considerado desconhecido, analisaram o tempo de duração das pesquisas realizadas pelos colaboradores da instituição.
- Análise de classificação e predição: esta teve por objetivo a análise entre as atividades exercidas e as publicações realizadas, onde buscaram saber em qual nível de atividade (ensino, pesquisa e direção) ocorriam mais publicações.

Cervi *et al.* (2009) analisaram o comportamento científico de 45 doutores brasileiros da área de ciência da computação e subárea banco de dados. O objetivo do trabalho era analisar a produção científica desses pesquisadores mediante uma abordagem temporal, utilizando como fonte de informação o histórico da produção dos mesmos cadastrados na PL. Os indicadores empregados foram o número de coautores, o número de publicações e a quantidade de orientações em nível de mestrado e doutorado de cada pesquisador. O intervalo analisado foi um período de 9 anos, compreendendo os anos de 2000 a 2008. Para a análise dos dados foram utilizadas técnicas de mineração de dados, como agrupamento e regressão, incorporados na ferramenta WEKA.

Os pesquisadores foram divididos em 3 grupos: Grupo 1, com 15 pesquisadores que concluíram o doutorado antes de 1994; Grupo 2, com também 15 pesquisadores que concluíram o doutorado entre 1995 e 2001; e o Grupo 3, com os 15 pesquisadores restantes que concluíram o doutorado entre 2002 e 2008. Foram realizados alguns experimentos envolvendo tarefas de Agrupamento e Regressão:

- Agrupamento por coautores: observou-se que os 3 grupos de pesquisadores não apresentavam semelhança com os 3 agrupamentos encontrados, revelando que o tempo de conclusão de doutorado não tinha influência no número de coautores destes pesquisadores.
- Agrupamento por produção: observou-se que os 3 grupos de pesquisadores não apresentavam semelhança com os 3 agrupamentos encontrados, revelando que o tempo de conclusão do doutorado não tinha influência no

número de publicações, uma vez que pesquisadores do Grupo 3 publicaram mais que os do Grupo 1 e 2.

- Agrupamento por orientações: observou-se que os 3 grupos de pesquisadores não apresentavam semelhança com os 3 agrupamentos encontrados, revelando que os pesquisadores do Grupo 1 possuíam um número maior de orientações do que a grande maioria dos pesquisadores dos Grupos 2 e 3.
- Regressão Linear Simples: o objetivo deste experimento era descobrir o número da produção dos pesquisadores a partir do número de orientações. Foi utilizado o número de artigos publicados em periódicos somado com o número de trabalhos publicados em eventos como “variável dependente” e o número de orientações de mestrado e de doutorado como “variável independente”. O resultado mostrou que para a grande maioria dos pesquisadores, o número real da produção não correspondia ao número da predição de sua produção.
- Regressão Linear Múltipla: O objetivo deste experimento era descobrir o número da produção dos pesquisadores a partir do número de coautores e o número de orientações de mestrado e de doutorado. Foi utilizado o número de artigos publicados em periódicos somado com o número de trabalhos publicados em eventos dos 45 pesquisadores como “variável dependente” e o número de coautores juntamente com o número de orientações de mestrado e de doutorado como “variável independente”. O resultado mostrou que o número real da produção dos pesquisadores era compatível com o número da predição de sua produção.

Mota *et al.* (2010) realizaram um estudo sobre as atividades de pesquisa da Rede Federal de Educação Profissional Científica e Tecnológica (RFEPCT) na última década. Os indicadores avaliados na pesquisa foram os referentes aos grupos de pesquisa, gestão da pesquisa e cursos de pós-graduação. A pesquisa foi realizada a partir da MD nas bases de informação do CNPq e da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Os sistemas escolhidos como fontes

privilegiadas foram a PL, o Diretório dos Grupos de Pesquisa, o Portal de Periódico e o Sistema de Avaliação de Pós-Graduação da CAPES, além de motores de busca da internet. Alguns resultados encontrados foram: o crescimento de 884,5% dos grupos de pesquisa na RFEPCT entre os anos de 2000 a 2008; 46% dos grupos de pesquisa da RFEPCT são da área de engenharia e a maior parte se encontram nas regiões sul e sudeste; a pesquisa revelou o caráter ainda embrionário das políticas de pesquisa, pós-graduação e inovação tecnológica na Rede Federal. Para as autoras, as diretrizes para as atividades de pesquisa nos Institutos Federais são geralmente locais e isso evidencia a ausência de um planejamento central para a área. Ocorre uma fragmentação dos recursos materiais e humanos que a Rede Federal, tanto localmente como globalmente, dispõe para a indução dessas atividades, o que não contribui com o crescimento qualitativo da pesquisa interna. Obviamente isso reflete o grau de maturidade da comunidade de pesquisa e o grau de sua respectiva institucionalização. O que foi possível observar com esse trabalho, além da relação intrínseca entre alguns elementos da política de pesquisa, foi a ausência de originalidade nas ações e no modelo adotado pelos Institutos Federais em relação à política científica e tecnológica incorporada nas universidades brasileiras.

Paula (2004) realizou dois estudos de caso. No primeiro, buscou-se caracterizar orientadores responsáveis por casos de sucesso no Programa de Iniciação Científica do CNPq. No segundo, o objetivo foi a identificação de linhas de pesquisa de um grupo de docentes, coerente com a sua produção científica e tecnológica, tomada da base de dados de currículos da Plataforma Lattes (PL). Foram aplicadas técnicas de MD no primeiro estudo de caso (ferramenta WEKA) e de Mineração de Texto no segundo (ferramenta Eurekha).

Para o primeiro estudo de caso, alguns resultados sobre as características do perfil do orientador do PIBIC foram: 75% eram do sexo masculino, 66% dos orientadores possuíam pós-doutorado; 58% tinham idade entre 40 a 50 anos enquanto

42% entre 50 a 60 anos; 73% dos orientadores eram de instituições federais enquanto 27% eram de estaduais; entre outras características.

No segundo estudo de caso, utilizou-se os dados curriculares dos docentes do programa de Mestrado em Gestão do Conhecimento e da Tecnologia da Informação (MGCTI) da Universidade Católica de Brasília para gerar agrupamentos que pudessem contribuir para a identificação das linhas de pesquisa que poderiam ser relacionadas ao programa. A produção científica e tecnológica dos docentes foi avaliada em quatro períodos: de 1998 a 1999, de 2000 a 2001, de 2002 a 2004 e no período total de 1998 a 2004 que corresponde ao período de existência do programa MGCTI. No primeiro período foram identificados dois agrupamentos, sendo que o primeiro agrupamento estava mais relacionado com temas do KDD e Inteligência Artificial, ou seja, da Tecnologia da Informação (TI); enquanto que o segundo estava mais relacionado com a GC. No segundo período foi identificado um agrupamento mais relacionado a temas da TI. No terceiro período foram identificados quatro agrupamentos onde se percebeu uma maior definição das áreas de interesse e uma maior caracterização interdisciplinar do MGCTI. No quarto período a identificação multidisciplinar do programa foi confirmada, a partir da identificação de cinco agrupamentos.

Romão (2002) propôs o desenvolvimento de um novo algoritmo híbrido de Mineração de Dados, denominado AGD (Algoritmo Genético para Descoberta de Regras Difusas), baseado na combinação de algoritmos genéticos e conjuntos difusos para extrair conhecimento correto, compreensível e relevante à tomada de decisão em gestão de Ciência e Tecnologia. O caráter “relevante” do conhecimento extraído foi baseado em impressões gerais do usuário. O protótipo implementado teve como estudo de caso a região sul do Brasil e banco de dados fornecido pelo CNPq. Na tese do autor, o objetivo era resolver uma generalização da tarefa de classificação, conhecida como modelagem de dependência, para extrair conhecimento na forma de regras de previsão.

A qualidade do algoritmo implementado foi avaliada diante dos dados devidamente preparados tendo como referencial para comparação, o algoritmo J48 (utilizou-se a ferramenta de domínio público WEKA). O protótipo apresentou eficiência aproximadamente equivalente ao J48 quanto à taxa de acerto, mas forneceu conhecimento mais compreensível, fruto do uso de regras com poucas condições e termos linguísticos difusos. Os resultados experimentais, em forma de regras de produção difusas, foram apresentados a usuários potenciais, através de entrevistas, que avaliaram o conhecimento novo obtido. Os entrevistados classificaram 45% das regras como muito relevantes, 31% de médio interesse e 24% de baixo interesse. A avaliação subjetiva, considerada satisfatória, foi próxima do grau de interesse fornecido pelo protótipo, calculado como contradição às impressões gerais fornecidas pelos usuários, confirmando a utilidade e relevância do novo algoritmo implementado.

Esta tese apresentou contribuições para os Pró-Reitores de Pesquisa das Universidades, para as agências de fomento a C&T da região Sul e para a comunidade científica da área de MD que estuda a combinação híbrida de técnicas de Inteligência Artificial (IA) para aproveitar as melhores características de cada uma delas.

Morais (2010) apresentou uma ferramenta capaz de extrair dados automaticamente de currículos da Plataforma Lattes e ainda utilizou técnicas de Mineração de Dados nos currículos extraídos dos professores da Escola Politécnica de Pernambuco (POLI) para descobrir informações que auxiliassem na tomada de decisão dos gestores da POLI com relação a investimentos nos cursos e nos docentes da instituição.

Os experimentos foram realizados com 129 professores do quadro efetivo da POLI divididos em cinco grupos: básico (34), mecânica (21), elétrica (31), civil (30) e computação (13). As tarefas utilizadas foram Agrupamento e Associação, e a ferramenta de MD foi a WEKA.

Realizou-se vários experimentos de agrupamentos de acordo com a quantidade de publicações dos docentes, o primeiro separou os docentes em 2 *clusters*, o segundo em 3 *clusters*, o terceiro em 4 *clusters* e o quarto experimento em 5 *clusters*. De acordo com os resultados, os departamentos de elétrica, mecânica e básico possuíam poucos professores com perfil de pesquisador; o departamento de mecânica praticamente não possuía professor com perfil de pesquisador enquanto o departamento de computação não possuía professor sem viés de pesquisa.

Algumas regras de associação interessantes também foram encontradas: os atributos de publicações e orientações mostraram forte relacionamento (publicações normalmente são produzidas em conjunto com alunos orientados); outras regras extraídas indicaram relacionamento entre os atributos departamento e orientações e entre departamento e publicações, assim como na tarefa agrupamento, estas regras também indicaram que os professores dos departamentos de elétrica, mecânica e básico possuíam poucas publicações e orientações; o departamento de civil tinham poucas publicações com DOI; o departamento de civil possuía poucos periódicos com fator de impacto e outra regra revelou que artigos publicados em periódicos, trabalhos completos e resumos publicados em anais de congressos influenciam muito o número de publicações gerais de um professor.

Baker e Yacef (2009) escreveram um artigo sobre o estado da *Educational Data Mining* (EDM) no ano de 2009, realizando uma revisão do assunto e destacando as visões futuras desta área. Os autores relataram a origem da comunidade de EDM no ano de 2009, as conferências internacionais sobre o tema, a criação do *Journal of Educational Data Mining* (JEDM), revista na qual este artigo foi publicado, a definição de EDM e os métodos mais utilizados. Além disso, eles analisaram a pesquisa de outros autores (ROMERO e VENTURA, 2007 *apud* BAKER *et al*., 2009) onde foi observada a proporção de cada tipo de método de EDM encontrado em artigos publicados sobre o assunto entre os anos de 1995 a 2005, e em seguida realizaram o

mesmo estudo em artigos publicados sobre EDM nos anos de 2008 e 2009, para identificar se houve alteração nas tendências anteriormente observadas.

Na primeira pesquisa que os autores citam, dos 60 artigos analisados, 43% envolveram métodos de Mineração de Relações (*Relationship Mining*), 28% envolveram a Predição, o restante envolveram métodos menos comuns. Na pesquisa realizada pelos próprios autores, um padrão bem diferente foi identificado, o método de Mineração de Relações caiu para o quinto lugar com apenas 9% dos artigos analisados, o método de Previsão passou a ocupar o primeiro lugar, representando 42% dos artigos, os métodos *Human Judgment* (decisão humana) e Agrupamento representaram respectivamente 12% e 15%, aproximadamente os mesmos valores na primeira pesquisa. Um novo método observado entre os anos de 2008 a 2009, é o método de descoberta com modelos que representou 19% dos artigos.

Segundo os autores, a comunidade de EDM estava focada na América do Norte, Europa Ocidental e Austrália/Nova Zelândia, com menor participação de outras regiões.

3.3 O INSTITUTO FEDERAL DE GOIÁS (IFG)

Conforme descrito em IFG (2013a), a história do Instituto Federal de Goiás possui uma longa trajetória, com origem no início do século passado, no dia 23 de setembro de 1909, quando, por meio do Decreto n.º 7.566, o então presidente Nilo Peçanha criou 19 Escolas de Aprendizes Artífices, uma em cada Estado do País. Em Goiás, a Escola foi criada na antiga capital do Estado, Vila Boa, atualmente cidade de Goiás. Na época, o objetivo era capacitar os alunos em cursos e oficinas de forjas e serralheria, sapataria, alfaiataria, marcenaria e empalhação, selaria e correaria.

Em 1942, com a construção de Goiânia, a escola foi transferida para a nova capital, se transformando em palco do primeiro batismo cultural da cidade. A Instituição recebeu então o nome de Escola Técnica de Goiânia, com a criação de cursos técnicos na área industrial, integrados ao ensino médio.

Com a Lei n.º 3.552, em 1959, a instituição alcançou a condição de autarquia federal, adquirindo autonomia administrativa, patrimonial, financeira, didático-pedagógica e disciplinar, recebendo a denominação de Escola Técnica Federal de Goiás, em agosto de 1965.

Por meio do decreto sem número, de 22 de março de 1999, a Escola Técnica Federal de Goiás foi transformada em Centro Federal de Educação Tecnológica de Goiás (CEFET-GO).

Em 29 de dezembro de 2008, foi então criado por meio da Lei nº 11.892, o atual Instituto Federal de Educação, Ciência e Tecnologia de Goiás (IFG), atendendo a uma proposta do governo federal, que desde 2003 editava novas medidas para a educação profissional e tecnológica. O IFG é autarquia federal de regime especial vinculada ao Ministério da Educação.

Os institutos federais, ao longo de suas histórias, foram e continuam sendo ambientes de formação e de realização de ações políticas, artísticas e culturais, reafirmando sua identidade como centro formador de ideias, conhecimentos, artistas, lideranças e, principalmente, profissionais qualificados e conscientes de suas responsabilidades com a vida e com a sociedade (IFG, 2013a).

Com a mudança para Instituto Federal, Goiás ficou com duas novas instituições: o IFG, formado pelo CEFET Goiás e o Instituto Federal Goiano (IFGOIANO), formado pela fusão dos CEFETs de Rio Verde e de Urutaí e da Escola Agrotécnica Federal de Ceres.

O IFG é uma autarquia federal detentora de autonomia administrativa, patrimonial, financeira, didático-pedagógica e disciplinar, equiparado às universidades federais. É uma instituição de educação superior, básica e profissional, pluricurricular e

multicâmpus, especializada na oferta de educação profissional, tecnológica e gratuita em diferentes modalidades de ensino (IFG, 2013a).

O IFG tem por finalidade formar e qualificar profissionais para os diversos setores da economia, bem como realizar pesquisas e promover o desenvolvimento tecnológico de novos processos, produtos e serviços, em estreita articulação com os setores produtivos e com a sociedade, oferecendo mecanismos para a educação continuada.

A instituição oferece desde educação integrada ao ensino médio à pós-graduação. Na educação superior, conta com os cursos de tecnologia, especialmente na área industrial, e os de bacharelado e licenciatura. Na educação profissional técnica de nível médio, o IFG atua, na forma integrada, atendendo também ao público de jovens e adultos, por meio do PROEJA. Atualmente são ofertados ainda cursos de mestrado profissional e especialização *lato sensu*, além dos cursos de extensão, de formação profissional de trabalhadores e da comunidade (Pronatec), de Formação Inicial e Continuada (FIC), que são cursos de menor duração, e os cursos de educação a distância. A Tabela 12 apresenta todos os cursos oferecidos pelo IFG em cada um de seus câmpus.

O IFG está iniciando sua carreira na abertura de cursos de mestrado. Atualmente, existem apenas dois, o Mestrado em Tecnologia de Processos Sustentáveis no câmpus Goiânia e o Mestrado em Educação para Ciências e Matemática no câmpus de Jataí. Nenhum deles possui turmas formadas ainda.

O IFG atende cerca de onze mil alunos nos seus dez campi, distribuídos nas cidades de Anápolis, Formosa, Goiânia, Inhumas, Itumbiara, Jataí, Luziânia, Uruaçu, Aparecida de Goiânia e Cidade de Goiás. Futuramente, o IFG terá o seu segundo câmpus em Goiânia, o Goiânia Oeste, e chegará em Águas Lindas de Goiás, Valparaíso, Novo Gama e Senador Canedo.

Existem atualmente 20 grupos de pesquisa do IFG cadastrados no Diretório dos Grupos de Pesquisa do CNPq. São 7 na área de pesquisa de Educação, 2 em

Engenharia Mecânica, 1 em Ciência e Tecnologia de Alimentos, 1 em Ecologia, 1 em Educação Física, 1 em Recursos Florestais e Engenharia Florestal, 1 em Ciência da Informação, 1 em Sociologia, 1 em Química, 1 em Engenharia Elétrica, 1 em Engenharia Civil, 1 em Artes e outro em Linguística.

O IFG vai continuar mantendo a tradição da Escola Técnica Federal de Goiás e do CEFET Goiás ao oferecer educação pública, gratuita e de qualidade para os jovens e os trabalhadores de Goiás. Inserido na RFEPCT, pretende ampliar sua inserção social contribuindo para o desenvolvimento social e econômico do Estado.

Tabela 12 - Cursos oferecidos em cada câmpus do IFG

	Técnico Integrado ao Ensino Médio	Técnico Subsequente (pós-médio)	PROEJA (cursos técnicos)	Curso Superior de Tecnologia	Curso Superior de Licenciatura	Curso Superior de Bacharelado	Especialização *Lato Sensu*	Mestrado
Goiânia	Controle Ambiental; Edificações; Eletrônica; Eletrotécnica; Instrumento Musical; Mineração; Informática para internet; Trânsito; Telecomunicações	Eletrotécnica; Mecânica; Mineração	Cozinha; Informática; Transporte Rodoviário	Agrimensura; Geoprocessamento; Construção de Edifícios; Estradas; Gestão de Turismo; Hotelaria; Processos Químicos; Redes de Telecomunicações; Saneamento Ambiental; Transporte Terrestre	Física; História; Matemática; Música	Engenharia Ambiental; Engenhria Civil; Engenharia de Controle e Automação; Engenharia Elétrica; Engenharia Mecânica; Química; Turismo; Sistemas de Informação; Engenharia de Transportes	Especialização em Matemática; Especialização em Políticas e Gestão da Educação Profissional e Tecnológica	Tecnologia de Processos Sustentáveis
Jataí	Edificações; Eletrotécnica; Agrimensura; Informática	Agrimensura Açúcar e Álcool (EAD)	Edificações	Análise e Desenvolvimento de Sistemas	Física	Engenharia Civil; Engenharia Elétrica	Ensino de Ciências e Matemática	Educação para Ciências e Matemática
Inhumas	Informática; Química; Alimentos		Panificação; Manutenção e Suporte em Informática;		Química	Sistemas de Informação; Informática; Ciência e Tecnologia de Alimentos		
Itumbiara	Eletrotécnica; Química; Automação Industrial	Automação Industrial; Eletrotécnica;			Química	Engenharia Elétrica		
Uruaçu	Edificações; Informática; Química;	Cerâmica (EAD)	Informática; Comércio		Química	Engenharia Civil		
Anápolis	Comércio Exterior; Edificações; Química	Edificações (EAD); Química (EAD)	Secretaria Escolar; Transporte de Carga	Tecnologia em Logística	Ciências Sociais; Química	Engenharia Civil da Mobilidade		
Formosa	Biotecnologia; Informática para Internet; Controle Ambiental; Edificações; Saneamento Ambiental	Edificações	Manutenção e Suporte em Informática; Edificações	Análise e Desenvolvimento de Sistemas	Biologia; Ciências Sociais	Engenharia Civil		
Aparecida de Goiânia	Agroindústria; Edificações; Química		Panificação; Modelagem do Vestuário		Dança	Engenharia Civil		
Cidade de Goiás	Edificações; Informática para Internet							
Luziânia	Edificações; Informática para Internet; Química; Mecânica	Edificações	Manutenção e Suporte em Informática;	Análise e Desenvolvimento de Sistemas	Química	Sistemas de Informação		

Fonte: Autoria própria

3.3.1 Programas de incentivo à Pesquisa

A PROPPG é o órgão administrativo do IFG, dirigido por um Pró-Reitor, responsável pela execução e gestão da política institucional referente à pesquisa e pós-graduação.

Assim, cabe à PROPPG propor, conduzir, planejar, coordenar, executar e avaliar as políticas institucionais de pesquisa, inovação e pós-graduação, no âmbito de todos os *campi;* implementar os planos de formação e aperfeiçoamento do corpo docente e técnico-administrativo em nível de pós-graduação; implementar e coordenar os programas e planos de concessão de bolsas de pesquisa e de pós-graduação; planejar, avaliar e supervisionar a elaboração de propostas de implementação, alteração ou extinção de cursos de pós-graduação.

Portanto, a PROPPG dispõe de alguns programas de incentivo à atividade de pesquisa dentro da instituição que serão descritos a seguir.

O **PIBIC, PIBIC-EM e PIBIC-Af** (Programa Institucional de Bolsas de Iniciação Científica (para alunos de cursos superiores), Programa Institucional de Bolsas de Iniciação Científica (para alunos do Ensino Médio) e Programa Institucional de Bolsas de Iniciação Científica nas Ações Afirmativas. Conforme IFG (2013b), estes programas tem por objetivo, dentre outros, despertar a vocação e desenvolver o pensamento científico do estudante de graduação, contribuir para a formação de recursos humanos para atividades de pesquisa; e fomentar a pesquisa científica no IFG, visando a ampliação da participação de servidores docentes e técnico-administrativos e estudantes para melhorar e consolidar a posição da Instituição junto à sociedade acadêmica e científica.

As bolsas do PIBIC-Af são destinadas, exclusivamente, aos estudantes que ingressaram no IFG por meio do Sistema de Cotas (o IFG reserva 50% das vagas de

cada curso para alunos oriundos da rede pública de ensino. Os demais alunos concorrem na modalidade livre concorrência).

O **PIBITI** (Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação) visa conceder bolsas de Iniciação Tecnológica destinadas a estudantes dos cursos superiores do IFG.

O PIBIC/PIBIC-Af e o PIBITI são programas voltados para o aluno e se destinam a complementar o ensino, oferecendo a eles a oportunidade de descobrir como o conhecimento científico e tecnológico é construído. Esse objetivo é conseguido pela participação do estudante nas atividades teóricas e práticas no ambiente de pesquisa. Essa vivência possibilitará ao aluno ver e entender o mundo sob o prisma da ciência. Para tanto, é necessário que professores/pesquisadores dediquem parte de seu tempo ao ensino conceitual e prático da pesquisa ao estudante.

O orientador poderá inscrever o estudante no PIBIC, PIBIC-Af e PIBITI, de duas maneiras: submetendo um projeto de pesquisa cuja execução seja de responsabilidade do estudante, sob sua orientação; ou solicitando a inclusão do estudante em um projeto de pesquisa cadastrado, ou em processo de cadastramento, na PROPPG, desde que a vigência deste projeto englobe o período previsto no Edital;

O **PIPECT** (Programa Institucional de Incentivo à Participação em Eventos Científicos e Tecnológicos) tem o objetivo de viabilizar condições para que servidores do IFG possam participar de eventos científicos e tecnológicos nacionais e internacionais, expondo os resultados de pesquisas realizadas no IFG e possibilitando a troca de experiências com pesquisadores de outras instituições. Este programa destina-se exclusivamente a servidores, docentes e técnico-administrativos, do quadro efetivo do IFG, que possuem trabalhos científicos aceitos para serem apresentados e publicados no evento. A concessão do auxílio é feita na forma de bolsa de incentivo, destinada a custear despesas com: taxa de inscrição; impressão de pôster (se for o caso); hospedagem; alimentação; e passagens.

o **ProAPP** (Programa de Apoio à Produtividade em Pesquisa) conforme IFG (2013c), é um programa de incentivo à produtividade em pesquisa através da concessão de bolsas aos servidores do IFG, portadores do título de mestre ou doutor, de acordo com os preceitos estabelecidos em regulamentação própria aprovado pela Resolução Nº 14, de 20 de dezembro de 2011, do Conselho Superior do IFG.

Os objetivos do ProAPP são: fomentar a pesquisa científica e tecnológica no IFG, ampliando sua produção acadêmico-científico-cultural; possibilitar o envolvimento de forma direta de estudantes no mundo da pesquisa, por meio de sua participação nas pesquisas desenvolvidas pelos servidores; estimular iniciativas inovadoras e propiciar a geração e a transformação do conhecimento, de forma a atender as necessidades e interesses da sociedade; promover a geração de produtos e/ou processos inovadores que resultem em propriedade intelectual; contribuir para a transformação e consolidação do IFG como centro de referência em pesquisa.

Para concorrer ao ProAPP o candidato deve apresentar um projeto de pesquisa, que será avaliado conforme seu mérito técnico-científico, cultural e social, e sua produção intelectual dos últimos 3 anos.

Dos quatro programas apresentados, o ProAPP e o PIPECT são destinados somente aos servidores da instituição. Os critérios para análise e julgamento de mérito para seleção dos mesmos são estabelecidos em edital de acordo com a produção científica de cada um (vide IFG, 2013c e IFG, 2013d). O formulário para análise técnica da produtividade do pesquisador para o programa ProAPP está ilustrado no Anexo A e o formulário para análise do currículo do servidor para o programa PIPECT, no Anexo B.

4. MATERIAL E MÉTODOS

Neste capítulo apresenta-se o material utilizado e os procedimentos realizados no desenvolvimento desta pesquisa. Foram utilizadas a pesquisa bibliográfica, a pesquisa documental e a metodologia de estudo de caso.

A pesquisa bibliográfica embasou a aquisição de conhecimento sobre os temas relacionados ao trabalho, envolvendo consultas a livros de referência, artigos científicos, dissertações e teses já publicados na área.

Quanto à pesquisa documental, o material empregado foi:

- os dados de 839 currículos Lattes de docentes do IFG, extraídos de arquivos XML (*Extensible Markup Language*). Os docentes selecionados para o estudo foram os de situação “ativo permamente” (descartados os docentes substitutos, de contrato temporário, aposentados e desativados no sistema SUAP - Sistema Unificado de Administração Pública), estando eles distribuídos nos 10 câmpus do Instituto (além da Reitoria);
- dados do banco de dados do sistema SUAP;
- dados dos grupos de pesquisa do IFG, cadastrados no CNPq;
- dados de planilhas do IFG sobre os docentes que atuam nos cursos de pós-graduação;
- editais dos programas ProAPP e PIPECT.

A metodologia de estudo de caso, juntamente com a metodologia CRISP-DM, empregada para orientar o processo de KDD, serão detalhadas no capítulo 5.

Algumas ferramentas computacionais foram utilizadas como auxílio na aplicação dos métodos e serão apresentadas a seguir.

4.1 PROBLEMA DA PESQUISA

O problema a ser explorado refere-se à identificação de padrões, que representam o perfil da produtividade científica dos docentes ativos do IFG. A identificação de tais padrões pode ser encontrada através da aplicabilidade do processo de KDD.

O principal fator que influenciou na escolha do material a ser estudado foi a carência de informações da Pró-Reitoria de Pesquisa e Pós-Graduação sobre a produtividade científica dos pesquisadores. O IFG ainda não dispõe de um sistema de informação que controle tal produção. A proposta do estudo é identificar padrões e médias dessa produtividade, minerando os dados das publicações do último triênio (2011 a 2013) disponíveis nos currículos Lattes.

4.2 HIPÓTESE

A aplicação do processo de KDD utilizando as tarefas de Classificação e Associação nos dados dos currículos Lattes poderá identificar o perfil da produtividade científica dos docentes do IFG, bem como de padrões dessa produção.

4.3 RECURSOS UTILIZADOS

Para realização do estudo de caso algumas ferramentas computacionais foram utilizadas: o **PostgresSQL** (apresentado na seção 2.5 do Referencial Teórico) para limpeza e preparação dos dados e o *software* **WEKA** (apresentado na seção 2.6) para realizar a etapa de Mineração de Dados.

5. ESTUDO DE CASO

Apresenta-se neste capítulo, o estudo de caso realizado com os dados dos currículos Lattes (CL) dos docentes ativos do Instituto Federal de Goiás (IFG). A organização das seções seguintes segue as fases definidas da metodologia de KDD, denominada CRISP-DM.

5.1 ENTENDIMENTO DO NEGÓCIO

Para o entendimento do negócio da instituição onde se realizou o estudo de caso do processo de Mineração de Dados, foram realizados vários contatos (via *e-mail* e telefone), visitas e reuniões presenciais com as pessoas responsáveis pela Pró-Reitoria de Pesquisa e Pós-Graduação (PROPPG) do Instituto Federal de Goiás (IFG), inclusive com o Pró-Reitor de Pesquisa e Pós-Graduação.

Foram realizadas também, leituras sobre informações da instituição: sua história, cursos oferecidos e corpo docente dos mesmos, programas de incentivo à atividade de pesquisa e seus editais, entre outros, para melhor compreender o contexto atual no qual o IFG se insere.

Nesta fase foi analisada a situação do IFG, no que diz respeito as informações de sua produção científica, e detectada a carência de conhecimento real sobre a mesma. Após a definição dos objetivos do projeto de MD junto à equipe do IFG, iniciou-se os planejamentos das metas para atingi-los.

As informações sobre a produção científica e tecnológica que existiam disponíveis no IFG para a realização do KDD eram *a priori* os dados dos currículos Lattes (CL) dos docentes existentes no banco de dados do sistema SUAP (Sistema Unificado de Administração Pública), entre outras informações não cadastradas neste

mesmo banco, como por exemplo, os núcleos de pesquisa do CNPq, nos quais os docentes participam.

O sistema SUAP é um sistema de informação *web* que vem sendo desenvolvido pela equipe de TI do Instituto Federal do Rio Grande do Norte (IFRN) desde 2007, com o objetivo de informatizar os processos administrativos do mesmo e facilitar a sua gestão. Como esse sistema vem apresentando bons resultados no IFRN, ele foi disponibilizado gratuitamente à Rede Federal de Educação Profissional Científica e Tecnológica (RFEPCT). Vários institutos estão adotando-o, entre eles o IFG.

Entre os módulos que já estão em operação no SUAP estão: Recursos Humanos, Protocolo, Almoxarifado, Patrimônio, Planejamento, Ponto Eletrônico, Assistência Estudantil, Ensino, Financeiro, Orçamento, Frotas, Chaves, Contratos e Convênios, Contracheques, Portaria, Materiais e Progressões. O módulo de Recursos Humanos está integrado com as informações disponibilizadas pelo Sistema Integrado de Administração de Recursos Humanos (SIAPE). Conforme descrito em SIAPEnet (2013), o SIAPE é um sistema *on-line*, de abrangência nacional, que constitui-se hoje na principal ferramenta para a gestão do pessoal civil do Governo Federal, realizando mensalmente o pagamento de cerca de 1 milhão e 300 mil servidores ativos, aposentados e pensionistas em 214 órgãos da administração pública federal direta, instituições federais de ensino, ex-territórios, federais, autarquias, fundações e empresas públicas.

Os IFs, além de utilizar, também podem aprimorar o sistema SUAP. Essas melhorias são enviadas ao IFRN e implantadas, desde que atendam às demandas de toda a RFEPCT.

O SUAP possui também o módulo CNPq, que faz conexão via *webservice* com a Plataforma Lattes (PL) e importa os dados dos Currículos Lattes (CL) dos servidores que possuírem vínculo naquela instituição. Quando se cadastra o CL, a pessoa informa no menu “Atuação”, submenu “Atuação Profissional”, a(s) sua (s)

instituição(ões) de vínculo. O *webservice* só consegue extrair os currículos cujo vínculo institucional seja o vínculo oficial reconhecido pelo CNPq da instituição conectada. O nome do vínculo institucional oficial é cadastrado no sistema de Diretórios de Instituições do CNPq pelos seus gestores.

Portanto, alguns currículos não foram capturados porque o vínculo institucional cadastrado pelo docente não era o vínculo reconhecido do IFG pelo CNPq.

Para Soares (2012), o *webservice* é uma solução utilizada na integração de sistemas e na comunicação entre aplicações diferentes. Com esta tecnologia é possível que novas aplicações possam interagir com aquelas que já existem e que sistemas desenvolvidos em plataformas diferentes sejam compatíveis. Os *webservices* são componentes que permitem às aplicações enviar e receber dados em formato XML. Cada aplicação pode ter a sua própria "linguagem", que é traduzida para uma linguagem universal, o formato XML.

Este recurso do *webservice* é disponibilizado gratuitamente pelo CNPq para as instituições de ensino cadastradas. Para acessá-lo é necessário informar um número IP a ser liberado para a conexão. Para isso, foram realizados alguns contatos (*e-mail* e telefone) com a equipe técnica da infraestrutura de TI do CNPq para a alteração do número IP pré-cadastrado.

No módulo CNPq é possível gerar no SUAP (Sistema Unificado de Administração Pública) alguns gráficos de indicadores da produção bibliográfica/técnica e das orientações concluídas e em andamento de seus servidores.

Alguns contatos via *e-mail* também foram feitos com a equipe de desenvolvimento do sistema SUAP para melhor compreender o módulo CNPq.

5.2 ENTENDIMENTO DOS DADOS

Na consulta realizada no sistema SUAP em 28-12-2013, constavam cadastrados 2.037 servidores: 1.204 docentes e 833 técnicos-administrativos. O gráfico da Figura 30 (extraído do módulo do CNPq do sistema SUAP na mesma data) mostra a quantidade de servidores que tiveram seus currículos importados pelo *webservice* e daqueles que não foram identificados nenhum currículo Lattes (CL). Os 607 servidores que não tiveram currículo Lattes importados, ou não o possuíam, ou não cadastraram o vínculo institucional oficial do IFG (conforme explicado na seção 5.1).

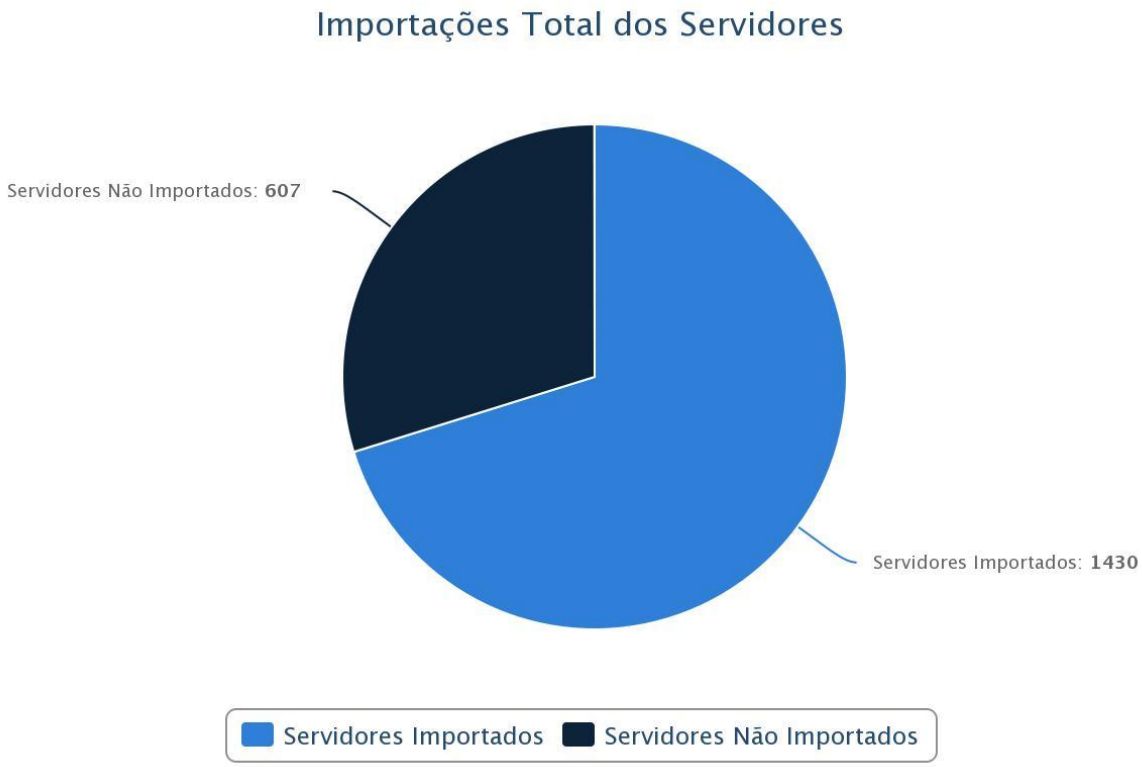


Figura 30 - Currículos de servidores do IFG importados do CNPq

Fonte: Sistema SUAP (dados em 28-12-2013)

Inicialmente pensou-se em minerar os dados dos CL de todos os servidores do IFG, o que somava um total de mais de 1.400 currículos, incluindo os técnicos-administrativos. Mas após análise preliminar dos dados observou-se que a produção

científica destes últimos era pouco representativa, optando-se então pela seleção dos currículos dos docentes, resultando em 1.033 currículos, como mostra a Figura 31. Observa-se também pela figura que aproximadamente 14% (171) dos docentes não tiveram seus CL incluídos na pesquisa, por não terem sido importados pelo *webservice* do CNPq.

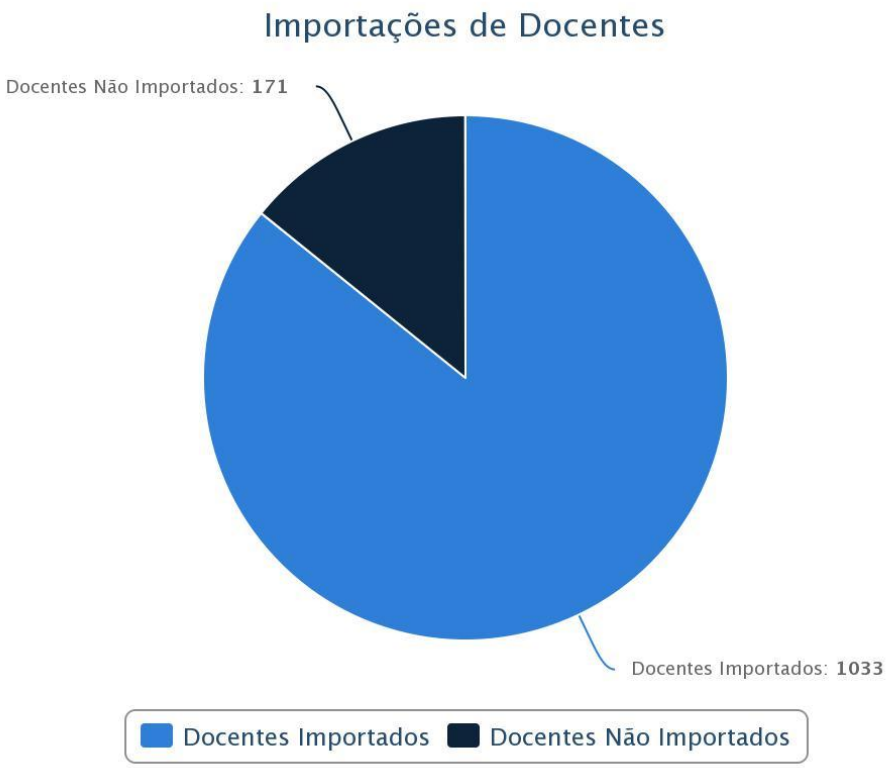


Figura 31 - Currículos de docentes do IFG importados do CNPq

Fonte: Sistema SUAP (dados em 28-12-2013)

Destes currículos foram novamente selecionados somente os docentes efetivos e que estavam em atividade no IFG. Foram descartados os professores substitutos, aposentados, e aqueles que por algum outro motivo estavam com status inativo no sistema SUAP. Portanto, a população envolvida no estudo foi de 839 docentes que dispunham de currículo Lattes cadastrado com vínculo no IFG.

Os dados coletados do banco de dados do SUAP eram úteis para a realização deste trabalho, mas *a posteriori* descobriu-se que muitas informações desejadas dos CL não existiam no banco de dados do SUAP, pois nem todas são importadas por ele.

Segundo os responsáveis pelo sistema, o desenvolvimento deste módulo precisou ser interrompido para atender demandas prioritárias de outros módulos.

Portanto, foi necessário criar outro mecanismo para obtenção dos dados integrais dos currículos Lattes. Para isso, foi desenvolvido um *script* em linguagem PHP (*Personal Home Page*) para ler os currículos em formato XML e salvar os dados necessários em um novo banco de dados. Esses arquivos XML foram adquiridos conectando-se via *webservice* no CNPq (da mesma forma que o sistema SUAP acessa-os). Esse processo é melhor detalhado na próxima subseção 5.3.

A escolha dos dados a serem submetidos ao processo de Mineração de Dados foi baseada nas informações mais relevantes que caracterizam o pesquisador, assim como nas produções científicas e tecnológicas avaliadas nos editais do Programa de Apoio à Produtividade em Pesquisa (ProAPP), pois este é o programa do IFG que está mais intimamente relacionado ao foco deste trabalho, a produtividade em pesquisa.

Os editais do programa ProAPP só consideram as publicações registradas no CL dos últimos 3 anos. Da mesma forma, o período avaliado das produções científicas e tecnológicas neste estudo está fixado no último triênio (2011 a 2013).

5.3 PREPARAÇÃO DOS DADOS

Esta fase do modelo CRISP-DM consome a maior parte do tempo de um projeto, pois envolve a seleção dos atributos para o processo de Mineração de Dados (MD), assim como a limpeza e transformações necessárias nos dados para o processamento na ferramenta de mineração.

O processo de extração dos arquivos XML dos currículos Lattes (CL) no CNPq foi realizado conforme as etapas descritas a seguir.

Primeiramente foi necessário recuperar o número identificador (NroIDCnpq) que cada currículo possui no CNPq. O NroIDCnpq é uma *string* de 16 posições. Para

recuperá-lo pela URL, era necessário passar como parâmetro o CPF do docente. Nesse momento, foi necessário consultar via SQL, o banco de dados do sistema SUAP para conseguir os CPFs dos docentes selecionados para o estudo de caso. A tabela do SUAP que contém o CPF do docente é a tabela **pessoa_fisica**. Segundo o "Guia de utilização de serviços de extração de currículos do CNPq", a URL para recuperar o NroIDCnpq do currículo, passando o CPF como parâmetro é:

http://servicosweb.cnpq.br/srvcurriculo/servlet/ServletID?cpf=00000000000

De posse dos CPFs (dos docentes selecionados para o estudo de caso), foram criados comandos utilizando *wget* (um aplicativo para transferência de arquivos sob o protocolo *HTTP - Hypertext Transfer Protocol*) seguidos da URL completa, (como mostrado acima). O resultado deste procedimento foi a obtenção de vários arquivos XML. O conteúdo de cada um desses arquivos é:

<?xml version="1.0" encoding="ISO-8859-1" ?>
<NRO_ID_CNPQ>XXXXXXXXXXXXXXXX</NRO_ID_CNPQ>

Caso tenha encontrado o NroIDCnpq com o CPF passado; e:

<?xml version="1.0" encoding="ISO-8859-1" ?>
<NRO_ID_CNPQ />

Caso não tenha encontrado o NroIDCnpq correspondente.

Depois de conseguir o NroIDCnpq de cada currículo, a próxima etapa foi realizar o *download* dos currículos em formato XML. A URL que recupera o currículo, passando o NroIDCnpq como parâmetro é:

http://servicosweb.cnpq.br/srvcurriculo/servlet/ServletZip?id=0000000000000000

Para isso, também foram criados vários comandos utilizando o *wget,* seguidos da URL citada. O formato da resposta, caso encontre o currículo, é um arquivo

compactado em formato “zip” de nome “**CUR_[nroIdCnpq].zip**”. Dentro do zip, existe um arquivo xml com o nome **[nroIdCnpq].xml**.

Por fim, esses arquivos “.zip” foram descompactados e processados pelo *script* de obtenção dos dados em formato XML. O código do *script* consta no Anexo C.

Foi possível extrair as informações do arquivo XML pelo fato de existirem funções PHP que reconhecem expressões regulares (padrões) como *strings*. O texto semi-estruturado do XML facilitou o estabelecimento desses padrões, uma vez que as *tags* puderam ser utilizadas como delimitadores para identificação dos dados de interesse.

Cada *tag* no arquivo XML dos currículos Lattes originou uma tabela no banco de dados criado, resultando em 280 tabelas.

O banco de dados do sistema de informação SUAP por sua vez, é formado por 460 tabelas. Dentre essas, 10 continham dados que interessavam a pesquisa. São elas: **pessoa, pessoa_fisica, servidor, setor, situacao, cargo_emprego, grupo_cargo_emprego, jornada_trabalho, unidadeorganizacional e cnpq_curriculovittaelattes**. A Figura 32 ilustra o diagrama de relacionamento entre as tabelas do banco de dados do sistema SUAP que foram utilizadas neste trabalho.

A tabela **cnpq_curriculovittaelattes** também foi necessária, pois continha um campo com o número identificador de cada currículo Lattes, sendo, portanto, o elo entre as tabelas originárias do SUAP e os dados extraídos dos arquivos XML.

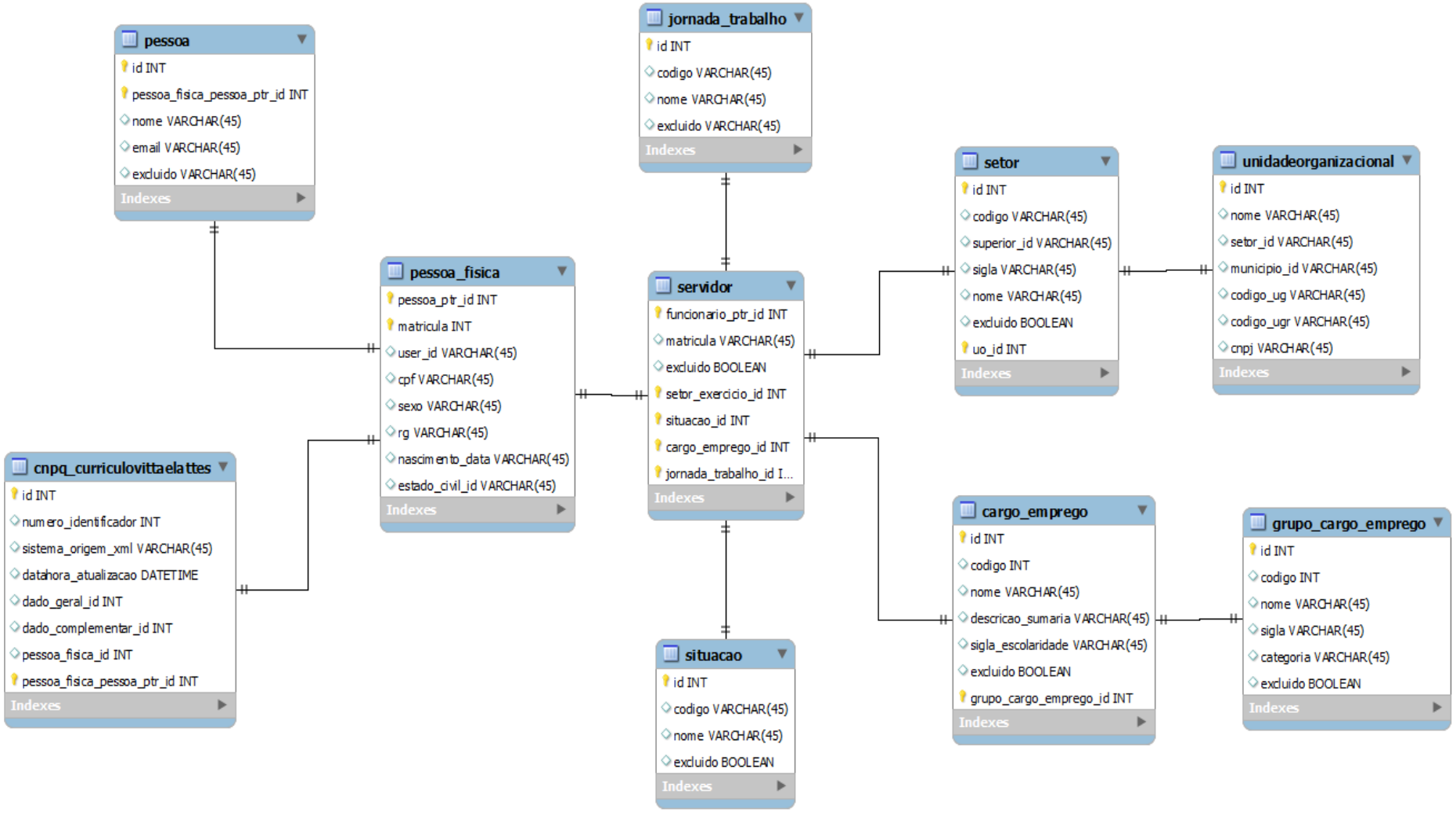


Figura 32 - Diagrama de relacionamento entre as tabelas utilizadas do SUAP

Fonte: Autoria Própria

A tabela **dados_adicionais** foi inserida no banco de dados para:

- inserir os dados dos docentes que participavam dos 20 núcleos de pesquisa do IFG cadastrados no CNPq. Estes dados foram adquiridos consultando o Diretório de Grupos de Pesquisa através do *site* do CNPq;
- inserir os dados dos docentes que atuam em algum dos 5 cursos de pós-graduação do IFG (3 especializações e 2 mestrados, vide Tabela 12). A lista do corpo docente de cada curso de pós-graduação foi disponibilizada pela PROPPG.

Os dados supracitados foram inseridos na tabela **dados_adicionais** através de comandos SQL. Estes são os únicos dados não originados do banco de dados do SUAP e dos arquivos XML dos currículos.

Para facilitar as consultas no banco de dados em meio a tantas tabelas, mascarando a sua complexidade, foram criadas 20 *view*s (visões). Uma *view* é uma maneira alternativa de observação dos dados das tabelas que compõem uma base de dados. Pode ser considerada como uma tabela virtual ou uma consulta armazenada, pois não armazena os dados físicos, e sim, a relação entre eles em linguagem SQL. O Anexo D apresenta a linguagem SQL de criação das *views*.

A Figura 33 apresenta o diagrama de relacionamento entre as *views* criadas. A demanda da PROPPG do IFG, era analisar as publicações de artigos em periódicos e trabalhos completos em anais de eventos, dentro do triênio. Porém, decidiu-se preparar de antemão, *views* de outros tipos de publicações para possíveis futuros experimentos. As *views* das publicações consideradas estão do lado direito da Figura 33. A *view* **v_titulacaomax** que consta no Anexo D não aparece na Figura 33 por ser uma derivação da *view* **v_titulacao**.

As 10 tabelas consultadas no banco de dados do sistema SUAP foram sintetizadas na *view* **suap_docente** (a que aparece no centro do diagrama). Com exceção da tabela **dados_adicionais** que foi criada para inserção de dados externos, o restante das *view*s retornam dados extraídos dos arquivos XML dos CL.

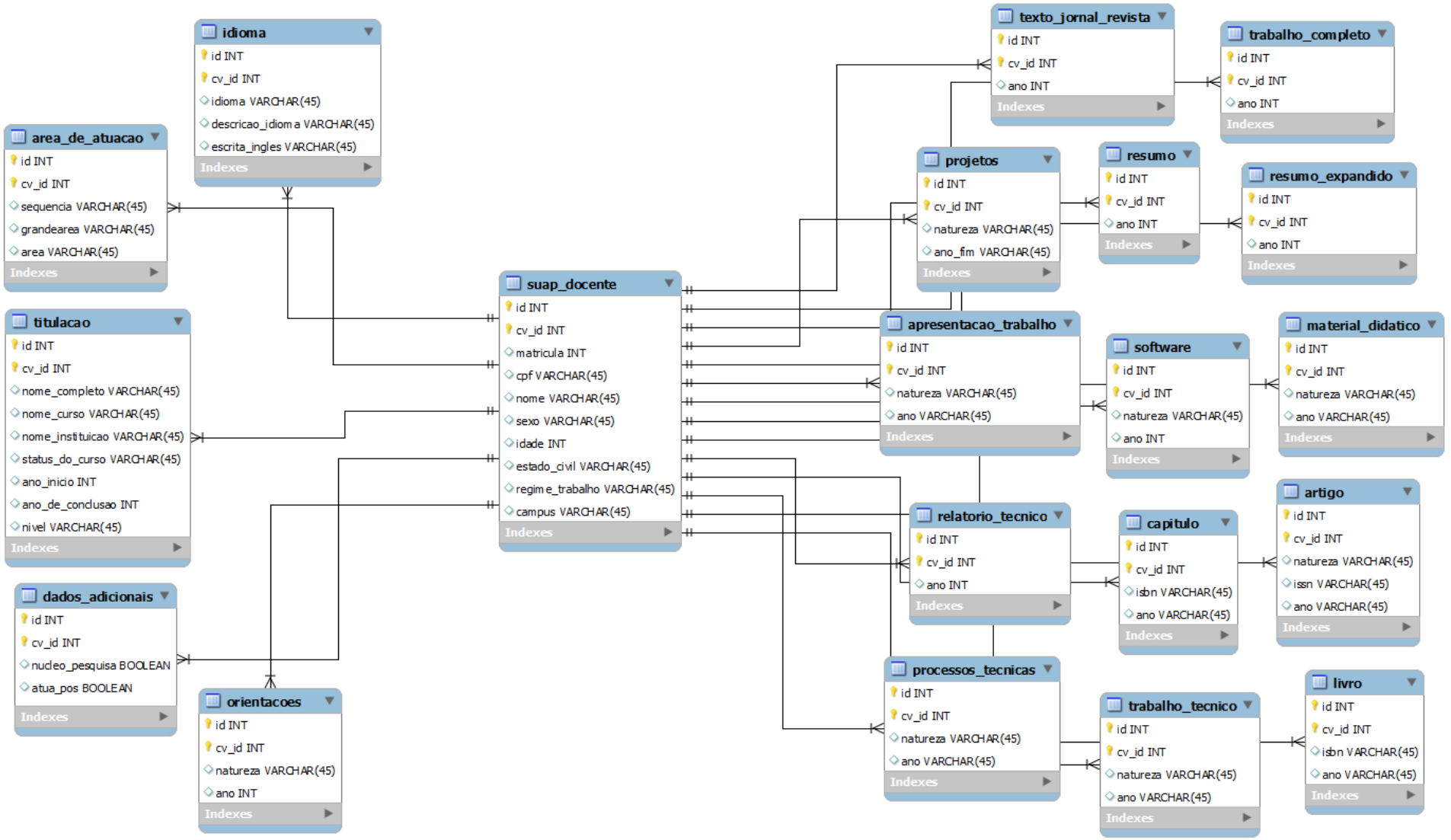


Figura 33 - Diagrama de relacionamento entre das *views* criadas
Fonte: Autoria própria

Várias transformações nos dados foram realizadas (via comandos SQL no momento de criação das *views)* para adequação dos mesmos aos formatos esperados pelos algoritmos que foram utilizados no WEKA:

- os dados quantitativos referentes à “idade do servidor”, “anos de conclusão da titulação máxima”, “projeto”, “orientação de graduação”, “orientação de iniciação científica” e “orientação de especialização” foram categorizados em faixas. Como não existia no IFG nenhum critério definido para segmentar tais atributos, eles foram estabelecidos baseando-se na distribuição de frequência realizada pelo *software* WEKA;
- os dados referentes à “regime de trabalho”, “sexo”, “titulação máxima”, “curso em andamento”, “escrita em inglês”, “câmpus” e “grande área” foram abreviados em siglas para facilitar a leitura dos resultados;
- os dados referentes a “orientação de mestrado”, “núcleo de pesquisa”, “atuação em pós-graduação” foram transformados em dados binários com valores “S” ou “N”.

Após o processo de organização e sintetização dos dados de interesse dentro das *views*, os atributos escolhidos foram:

- ***regimetrabalho***: atributo que representa o regime de trabalho que o docente possui no IFG: dedicação exclusiva (DE), 40 horas ou 20 horas;
- ***sexo***: atributo que representa o sexo do docente: Masculino ou Feminino;
- ***idade***: tempo de vida acumulado do docente. Este atributo foi discretizado em 4 faixas de valores: 20 a 29, 30 a 39, 40 a 49, 50 anos ou mais;
- ***titulacaomax***: atributo que representa a titulação máxima concluída do docente. Os valores que o atributo pode assumir são “G” para Graduação, “E” para Especialização, “M” para Mestrado, “D” para Doutorado ou “P” para Pós-Doutorado. Trinta docentes possuem mais de um curso em seu nível máximo de escolaridade, como por exemplo, 2 especializações, 2 mestrados, etc. Para

estes casos, foi considerado somente o último curso concluído de sua titulação máxima, ou seja, o mais recente;

- ***anosformacao***: atributo que representa o tempo de conclusão da última titulação máxima. Este atributo foi discretizado nos seguintes intervalos: 0anos (representa aqueles que possuem menos de um ano do término da titulação máxima), 1a3anos (entre 1 a 3 anos que concluiu a titulação máxima) e 4oumais (os que possuem mais de 4 anos que concluíram);
- ***cursoandamento***: atributo que indica se o docente possui algum curso de graduação ou pós-graduação em andamento. O domínio para o atributo é “G” para curso de Graduação, “E” para Especialização, “M” para Mestrado, “D” para Doutorado, “P” para Pós-Doutorado e “Nenhum” caso não esteja cursando nenhum destes;
- ***escritaingles***: atributo que indica o nível de escrita do docente na língua inglesa. O seu domínio é (P)ouco, (R)azoavelmente e (B)em. No currículo Lattes (CL) existem quatro atributos correspondentes a domínio no idioma em nível de: leitura, conversação, escrita e compreensão, bem como dezenas de idiomas diferentes. Se fossem considerados os quatro atributos relacionados a um idioma, obter-se-ia uma proporção relativamente alta de atributos sobre idioma. Assim, considerou-se apenas o nível de escrita. Este atributo é mais forte que nível de leitura uma vez que, para ter um bom nível de escrita, um bom nível de leitura geralmente é necessário. Pelo mesmo motivo, decidiu-se colocar informação sobre apenas uma língua estrangeira, inglês, já que o número de publicações científicas em inglês é maior do que as publicados em outra língua;
- ***campus***: representa o câmpus de lotação atual do servidor. Os valores possíveis para este atributo são: GYN(Goiânia), JAT (Jataí), INH (Inhumas), ITU (Itumbiara), APA (Aparecida de Goiânia), ANA (Anápolis), LUZ (Luziânia), FOR (Formosa), URU (Uruaçu) e REI (Reitoria). Apesar da Reitoria não ser um

câmpus, alguns docentes se encontram em cargos de direção e por isso estão lotados na mesma;

- ***grandearea***: representa a primeira grande área de conhecimento do CNPq informada pelo docente no menu *Atuação* e submenu *Áreas de Atuação* do CL. As grandes áreas de conhecimento consideradas pelo CNPq são oito: ENG (Engenharias), CET (Ciências Exatas e da Terra), CH (Ciências Humanas), CSA (Ciências Sociais Aplicadas), LLA (Linguística, Letras e Artes), CS (Ciências da Saúde), CB (Ciências Biológicas) e CA (Ciências Agrárias);
- ***area***: atributo que representa a primeira área de conhecimento do CNPq informada pelo docente no menu *Atuação* e submenu *Áreas de Atuação* do CL. A área de atuação é o segundo nível de especificação da *Área de Atuação* no currículo, aparecendo logo após a grande área do conhecimento, seguida pelos níveis da Subárea e Especialidade. Foram mantidas no domínio deste atributo somente as áreas de conhecimento com ocorrência de seis ou mais registros em cada. As demais foram agregadas em um único domínio denominado “Outras”. Esta transformação nos dados foi realizada na intenção de diminuir o domínio do atributo e facilitar a leitura dos algoritmos. As áreas de conhecimento do CNPq mais comuns entre os docentes e que foram mantidas no domínio do atributo constam na Tabela 13;
- ***atuaposgraduacao***: atributo que representa se o docente atua ou não em algum dos cinco cursos de pós-graduação do IFG. O atributo recebe o valor de “S” se o docente atua em pelo menos um dos cursos, e “N”, caso contrário;
- ***nucleopesquisa***: atributo que representa se o docente participa ou não de algum dos vinte núcleos de pesquisa do IFG no CNPq. O atributo recebe o valor de “S” se o docente participa de pelo menos um, e “N”, caso contrário;
- ***projeto*****:** atributo que representa a quantidade de projetos que o docente desenvolveu ou está desenvolvendo dentro do período da pesquisa. O domínio do atributo pode ser: Nenhum, Um, Dois, Três, Quatro, Cinco_oumais. Foram

contabilizados os projetos de pesquisa, desenvolvimento, extensão e de outras naturezas;

- ***orientacaograd:*** atributo que representa a quantidade de orientação(ões) de graduação concluída(s) no triênio. O domínio do atributo pode ser: Nenhuma, Uma, Duas, Três, Quatro_oumais;
- ***orientacaoic:*** atributo que representa a quantidade de orientação(ões) de iniciação científica concluída(s) no triênio. O domínio do atributo pode ser: Nenhuma, Uma, Duas, Três_oumais;
- ***orientacaoespec:*** atributo que representa a quantidade de orientação(ões) de especialização concluída(s) no triênio. O domínio do atributo pode ser: Nenhuma, Uma, Duas_oumais;
- ***orientacaomes:*** atributo que indica se o docente possui pelo menos uma orientação de mestrado concluída no triênio. O domínio do atributo pode ser: “S” (orientou pelo menos uma) ou “N” (não orientou nenhuma). Quanto às orientações de doutorado foi encontrado somente um registro no banco e por isso elas não foram consideradas***;***
- ***classe:*** atributo alvo (meta) estabelecido, que rotula os docentes conforme os critérios estabelecidos abaixo:

  - **classe A**: se o docente teve 3 ou mais artigos publicados em periódicos com ISSN, ou 3 ou mais trabalhos completos publicados em anais de eventos no triênio;

  - **classe B**: se o docente teve 2 artigos ou 2 trabalhos completos publicados no triênio;

  - **classe C**: se o docente teve 1 artigo ou 1 trabalho completo publicado no triênio;

  - **classe D**: se o docente não teve nenhum artigo ou trabalho completo publicado no período da pesquisa.

As análises quanto às publicações de artigos e trabalhos completos em eventos foram realizadas separadamente, ora considerando somente os artigos, ora somente trabalhos completos.

Os critérios de categorização dos docentes em classes foram definidos junto a equipe da PROPPG do IFG, segundo os seus interesses.

A Tabela 13 apresenta um resumo dos atributos escolhidos e suas características.

Tabela 13 - Atributos selecionados para a Mineração de Dados

	Atributos	Tipo	Domínio	Descrição
1	regimetrabalho	categórico	DE, 40H, 20H	DE (Dedicação Exclusiva), 40H (40 horas), 20H (20 horas)
2	sexo	categórico	M, F	M (Masculino), F (Feminino)
3	idade	categórico	20a29, 30a39, 40a49, 50oumais	Faixas etárias da idade do docente
4	titulacaomax	categórico	G, E, M, D, P	G (Graduação), E (Especialização), M (Mestrado), D (Doutorado), P (Pós-Doutorado)
5	anosformacao	categórico	0anos, 1a3anos, 4oumais	Faixas de tempo de conclusão da última titulação máxima
6	cursoandamento	categórico	G, E, M, D, P, Nenhum	G (Graduação), E (Especialização), M (Mestrado), D (Doutorado), P (Pós-Doutorado) e NENHUM (nenhum curso em andamento)
7	escritaingles	categórico	P, R, B	P (Pouco), R (Razoável) e B (Bem)
8	campus	categórico	GYN, APA, ITU, LUZ, INH, JAT, GOI, FOR, URU, ANA, REI	GYN (Goiânia), APA (Aparecida de Goiânia), ITU (Itumbiara), LUZ (Luziânia), INH (Inhumas), JAT (Jataí), GOI (Cidade de Goiás), FOR (Formosa), URU (Uruaçu), ANA (Anápolis), REI (Reitoria).
9	grandearea	categórico	ENG, CET, CH, CB, CS, LLA, CSA, CA	ENG (Engenharias), CET (Ciências Exatas e da Terra), CH (Ciências Humanas), CB (Ciências Biológicas), CS (Ciências da Saúde), LLA (Linguística, Letras e Artes), CSA (Ciências Sociais Aplicadas) e CA (Ciências Agrárias)
10	area	categórico	Química, Ciência da Computação, Engenharia Civil, Engenharia Elétrica, Agronomia, Ciência e Tecnologia de Alimentos, Educação Física, Matemática, Física, Geociências, Educação, História, Filosofia, Geografia, Sociologia, Administração, Arquitetura e Urbanismo, Turismo, Engenharia Mecânica, Engenharia de Transportes, Engenharia Sanitária, Letras, Linguística, Artes, Biologia Geral, Outras.	Áreas de atuação do CNPq com mais de 5 ocorrências entre os docentes
11	atuaposgraduacao	categórico	S, N	S (atua em algum dos 5 cursos de pós-graduação do IFG), N (não atua)
12	nucleopesquisa	categórico	S, N	S (participa de núcleo de pesquisa do CNPq) e N (não participa)
13	projeto	categórico	Nenhum, Um, Dois, Tres, Quatro, Cinco_oumais	Quantidade de projetos desenvolvidos pelo docente no triênio
14	orientacaograd	categórico	Nenhuma, Uma, Duas, Tres, Quatro_oumais	Quantidade de orientações de graduação realizadas pelo docente no triênio
15	orientacaoic	categórico	Nenhuma, Uma, Duas, Tres_oumais	Quantidade de orientações de iniciação científica realizadas pelo docente no triênio
16	orientacaoespec	categórico	Nenhuma, Uma, Duas_oumais	Quantidade de orientações de especialização realizadas pelo docente no triênio
17	orientacaomes	categórico	S, N	S (se o docente concluiu alguma orientação de mestrado no triênio), N (caso contrário)
18	classe	categórico	A, B, C, D	A (3 ou mais publicações no triênio), B (2 publicações), C (1 publicação), D (nenhuma publicação no triênio)

Fonte: Autoria Própria

Com a criação das *views* da Figura 33, facilitou-se o desenvolvimento de uma consulta SQL que retornasse todos os dados de interesse do estudo, a serem inseridos no arquivo ARFF. Essa SQL foi armazenada na *view* **v_arff** (vide Anexo E). Os dados retornados pela *view* **v_arff** foram então exportados diretamente do PostgreSQL para um arquivo de extensão ".**csv**". Arquivos com esta extensão contém dados separados por vírgula (***C**omma **S**eparated **V**alues*).

Em seguida, o arquivo ".csv" foi renomeado para extensão .ARFF, no qual se acrescentou o nome da relação (*relation* @lattes) e a definição de cada nome e tipo dos atributos. A Figura 34 mostra o texto de uns dos arquivos "lattes.arff" (com as classes definidas de acordo com a quantidade de artigos publicados no triênio) que foi processado pela ferramenta WEKA.

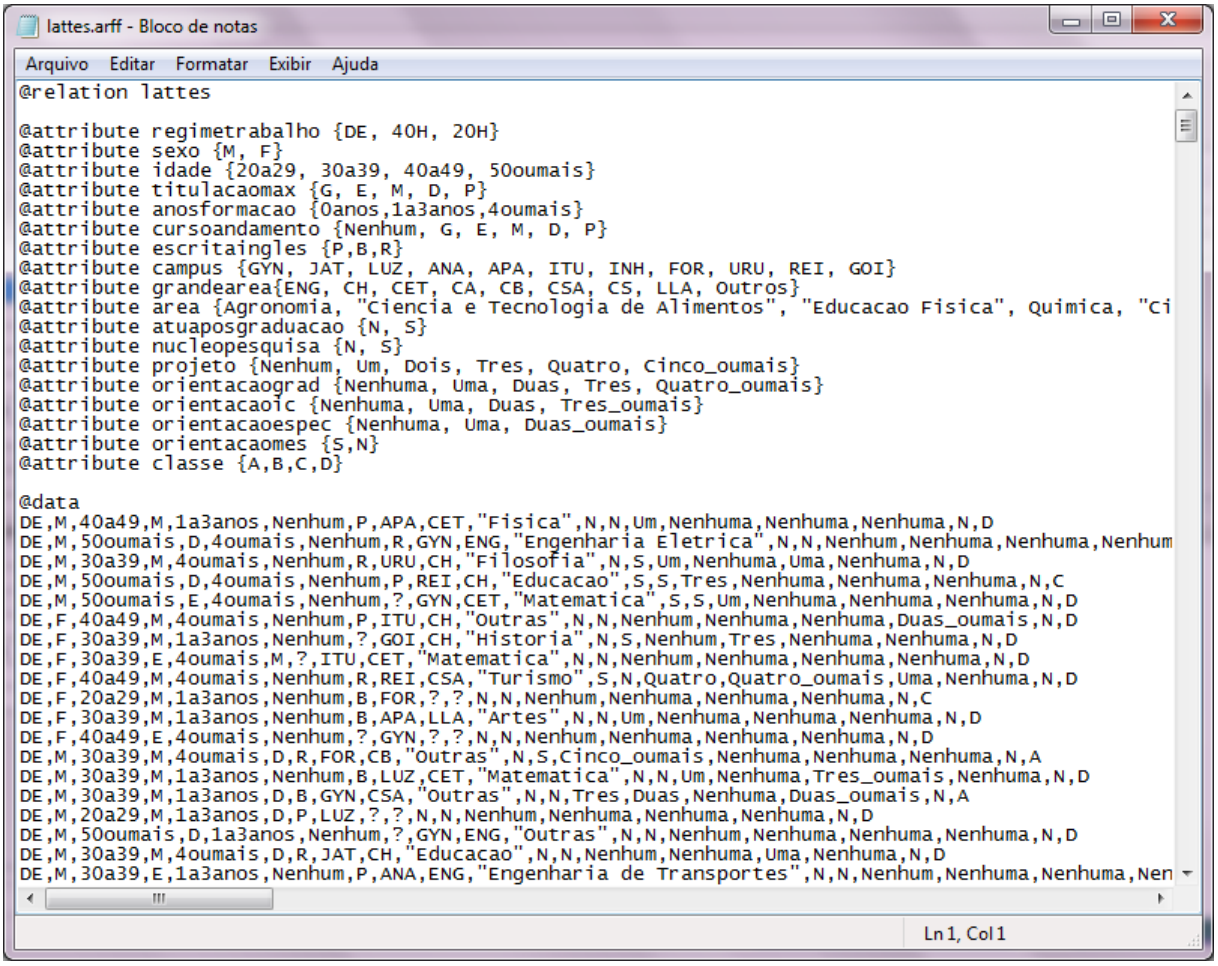


Figura 34 - Arquivo ARFF e seus atributos

Fonte: Autoria Própria

Em toda pesquisa é fundamental ter-se uma visão geral dos dados a serem analisados. O Anexo F apresenta uma análise descritiva dos dados da população envolvida, distribuídos por atributo.

5.4 MODELAGEM

Nesta seção são apresentados os experimentos de Mineração de Dados (MD) que foram realizados na ferramenta WEKA com diferentes entradas. Foram realizadas neste trabalho, as tarefas de Classificação e Associação.

Para as tarefas de Classificação e Associação foram realizados 3 tipos de experimentos: primeiramente, a classificação dos docentes segundo a publicação de artigos em periódicos; em seguida, a classificação dos docentes segundo a publicação de trabalhos completos, e por último, a classificação considerando vários tipos de publicações no triênio, estabelecendo pontuações para as publicações e alguns atributos.

Para melhor identificar os experimentos, eles serão designados pelas seguintes siglas:

- **ECA**: Experimento de Classificação de **artigos** em periódicos;
- **ECT**: Experimento de Classificação de **trabalhos completos** em anais de eventos;
- **EC**: Experimento de Classificação incluindo **vários tipos de publicações**, estabelecendo **pontuações** para cada tipo e alguns atributos;
- **EAA**: Experimento de Associação referente aos **artigos** em periódicos;
- **EAT**: Experimento de Associação referente aos **trabalhos completos** em anais de eventos;

A maioria das siglas mencionadas acima estarão acompanhadas por um número que indica a ordem de realização dos experimentos.

5.4.1 Classificação

Para a tarefa de Classificação foi utilizado o algoritmo J48 (apresentado na seção 2.3.2.1.1), por ser a implementação na ferramenta WEKA, do algoritmo de árvore de decisão baseado no C4.5. O J48 é capaz de manipular atributos binários, ordinais, nominais e contínuos, mas no intuito de melhorar a indução dos atributos na construção da árvore de decisão, atributos categóricos foram utilizados.

Na etapa de preparação dos dados algumas conversões foram realizadas para favorecer a legibilidade da árvore gerada pelo modelo, visto que na execução do algoritmo, ao criar um novo nível de ramificação da árvore, é realizado um produto cartesiano entre o novo atributo e o domínio de valores categóricos do mesmo.

No primeiro experimento de Classificação (ECA1) pretendeu-se classificar os docentes quanto ao número de **artigos publicados em periódicos**. Aqueles que produziram no último triênio, 3 ou mais artigos foram classificados como pertencentes à classe A, os que produziram 2 artigos à classe B, somente 1 artigo à classe C e os que não produziram nenhum artigo, à classe D.

Os artigos sem preenchimento de ISSN (*International Standard Serial Number* - código que constitui um identificador unívoco para cada título de publicação em série) foram desconsiderados. Os “artigos aceitos para publicação” que também constam nos CL não foram contabilizados.

A Figura 35 mostra a aba *Preprocess* do *software* WEKA, após ter sido carregado o arquivo **lattes.arff** da Figura 34. Para cada atributo selecionado, o WEKA mostra as dados estatísticos do mesmo.

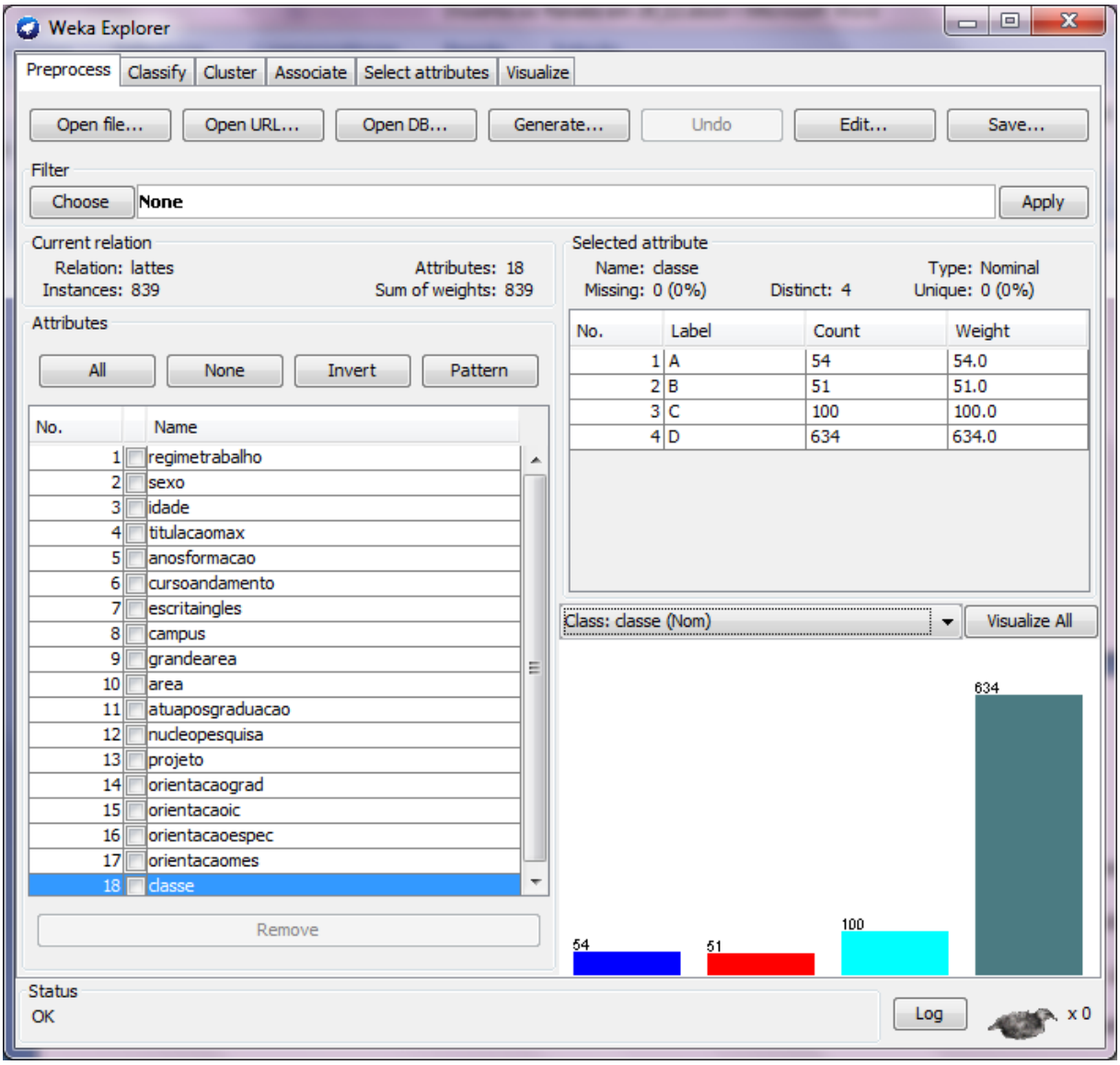


Figura 35 - Aba *Preprocess* do WEKA após a leitura do arquivo lattes.ARFF

Fonte: Autoria Própria

Na Figura 35, a primeira caixa em azul escuro representa a quantidade de docentes classe A, a caixa em vermelho representa os docentes classe B, a caixa em azul claro os docentes classe C e a caixa maior em cinza, representa os docentes classe D.

Pelo histograma apresentado na Figura 35, já é possível perceber que somente 54 docentes (6,43% da população) publicaram no último triênio, três ou mais artigos (classe A), enquanto 75,56% da população foi classificada como classe D. Ao clicar no botão *Visualize All*, à direita na Figura 35, a tela da Figura 36 é apresentada.

Figura 36 - Histogramas de distribuição das classes em relação aos atributos (Botão *Visualize All*)

Fonte: Autoria Própria

O WEKA apresenta interessantes histogramas sobre a distribuição de frequência de cada classe em relação a cada atributo de entrada. O último histograma da Figura 36 é o da **classe**.

A aba *Classify* (usada para tarefas de Classificação), a seleção do algoritmo J48, e os parâmetros utilizados para o mesmo, no ECA1 e ECT1, são apresentados na Figura 37.

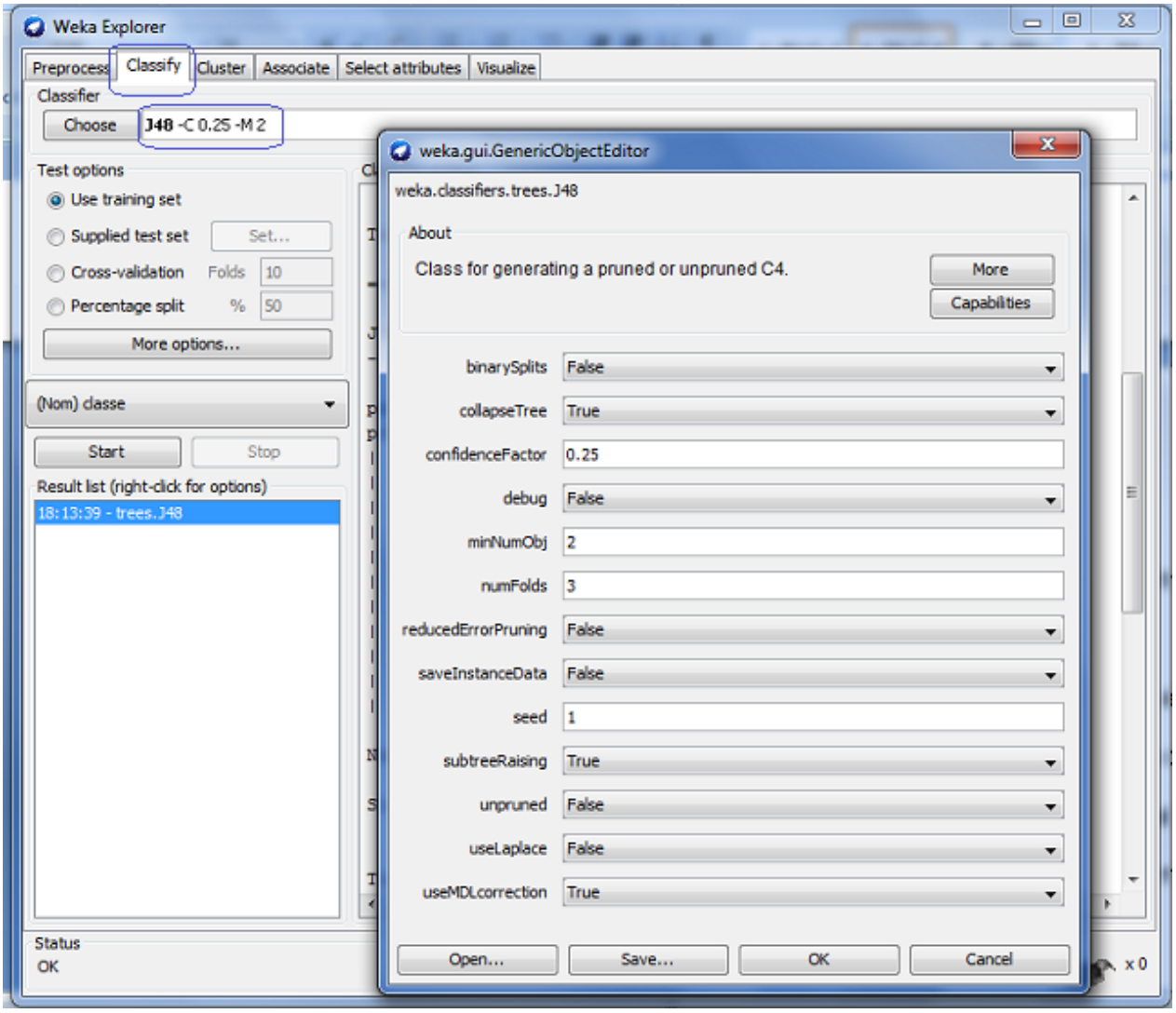


Figura 37 - Aba *Classify* e parâmetros do algoritmo J48

Fonte: Autoria Própria

Os parâmetros de configuração dos algoritmos definem o seu fluxo de execução e influenciam incisivamente nos resultados da mineração. Segundo Del-Fiaco (2012), os parâmetros do J48 no WEKA significam:

- *BinarySplits*: Usar divisor binário em atributos nominal ao construir árvores;
- *CollapseTree*: Remover atributos que não reduzem erro de treinamento;
- *ConfidenceFactor:* Fator de confiança utilizado na poda do algoritmo (quanto menor, mais poda é realizada);
- *Debug*: Mostrar informações adicionais no console;
- *MinNumObj*: Número mínimo de instâncias por folha;
- *NumFolds*: Determina o tamanho do conjunto de poda. Uma dobra é utilizada na poda, o resto para o crescimento da árvore;
- *ReducedErrorPruning*: Reduzir poda, se houver erro;
- *SaveInstanceData*: Salvar os dados de treinamento para visualização;
- *Seed*: Quantidade de descendência de dados para randomizar a redução de erros na poda;
- *SubtreeRaising*: Considerar a operação elevação de subárvore na poda;
- *Unpruned* : Se a poda é realizada;
- *UseLaplace*: Se a contagem de folhas for suavizada com base em um lugar;

No ECA1 foi utilizado o fator de confiança padrão do J48 no Weka de 25%, podendo realizar poda (*unpruned=False*) e a opção de teste do modelo selecionada foi "*Percentage Split*", com valor informado de 66%. Com fator de confiança de 25%, o algoritmo não criou árvore de decisão, classificou diretamente todos os docentes como classe D, informando que 75,78% das instâncias foram classificadas corretamente.

No ECT1, replicou-se a simulação do ECA1, exatamente com os mesmos atributos, parâmetros do J48 e opção de teste, porém agora, quanto ao número de **trabalhos completos** publicados em anais de eventos. Desta vez, 81 docentes foram classificados como classe A, representando 9,65% da população, conforme mostra a Figura 38. O algoritmo também não gerou árvore neste caso, classificou todas as 839 instâncias diretamente como classe D e informou uma taxa de acerto de 75,08%.

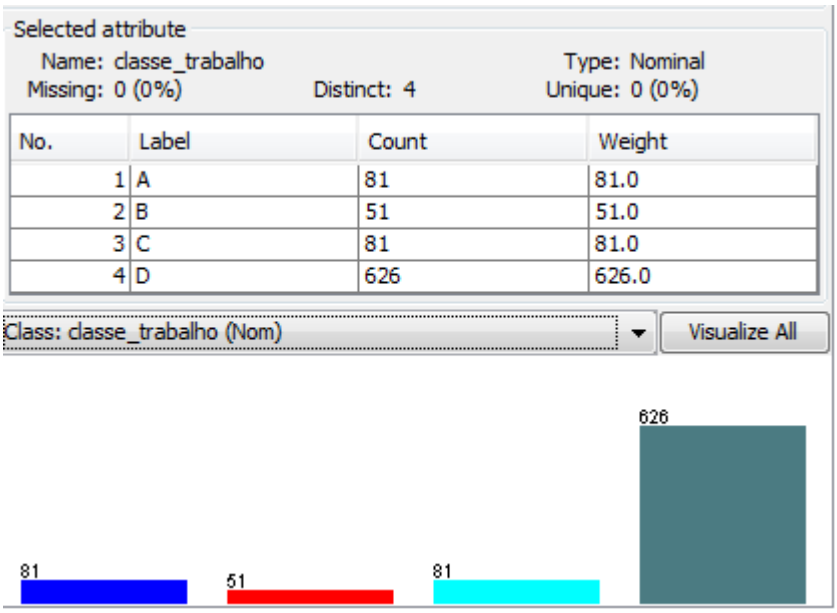


Figura 38 - Histograma das classes para trabalhos completos

Fonte: Autoria própria

No ECT1, houve uma pequena melhora na expressividade da classe A, porém o J48 ainda não conseguiu gerar árvore de decisão com fator de confiança de 25%.

Exaustivos experimentos foram realizados a procura do melhor modelo para a classificação dos docentes quanto à publicação de artigos e trabalhos completos. Os experimentos foram realizados alterando-se o fator de confiança, os parâmetros *unpruned*, *MinNumObjects, Numfolds* e as opções de teste do modelo.

A Tabela 14 apresenta os dados dos principais experimentos para a classificação de **artigos** e a Tabela 15 para os **trabalhos completos**. Todos os atributos apresentados na Tabela 13 foram incluídos nestes experimentos. A Figura 39 apresenta um comparativo entre o fator de confiança, o tamanho da árvore de decisão e o número de instâncias classificadas corretamente para os experimentos de ECA1 a ECA13. A Figura 40 apresenta o mesmo comparativo para os experimentos de ECT1 a ECT15.

Os experimentos que forneceram maior percentual de acerto foram aqueles com opção de teste *Use training set*, pois esta é a opção mais otimista delas. Desconsiderando estes últimos, os experimentos na linha cinza em cada tabela foram os que forneceram o maior número de instâncias classificadas corretamente.

Tabela 14 - Dados de experimentos para classificação de artigos

Experimento	Fator de Confiança	Tamanho da Árvore	Instâncias corretamente classificadas	Índice *Kappa*	*Unpruned*	*MinNumObject*	*NumFolds*	Opção de teste
ECA1	25%	1	75,78%	0	*false*	2	3	*Percentage Split 66%*
ECA2	32%	1	75,56%	0	*false*	2	3	*Use training set*
ECA3	33%	66	79,73%	0,33	*false*	2	3	*Use training set*
ECA4	33%	66	79,73%	0,33	*false*	2	5	*Use training set*
ECA5	40%	80	74,94%	0,24	*false*	2	3	*Supplied test set*
ECA6	33%	66	74,13%	0,12	*false*	2	3	*Cross-validation 10*
ECA7	33%	402	86,88%	0,64	*true*	2	3	*Use training set*
ECA8	60%	176	81,40%	0,43	*false*	5	3	*Use training set*
ECA9	60%	86	79,02%	0,33	*false*	10	3	*Use training set*
ECA10	65%	382	70,67%	0,18	*false*	2	3	*Cross-validation 10*
ECA11	80%	382	70,67%	0,18	*false*	2	3	*Cross-validation 10*
ECA12	100%	382	70,67%	0,18	*false*	2	3	*Cross-validation 10*
ECA13	33%	1	74,25%	0,09	*false*	3	3	*Cross-validation 10*

Fonte: Autoria própria

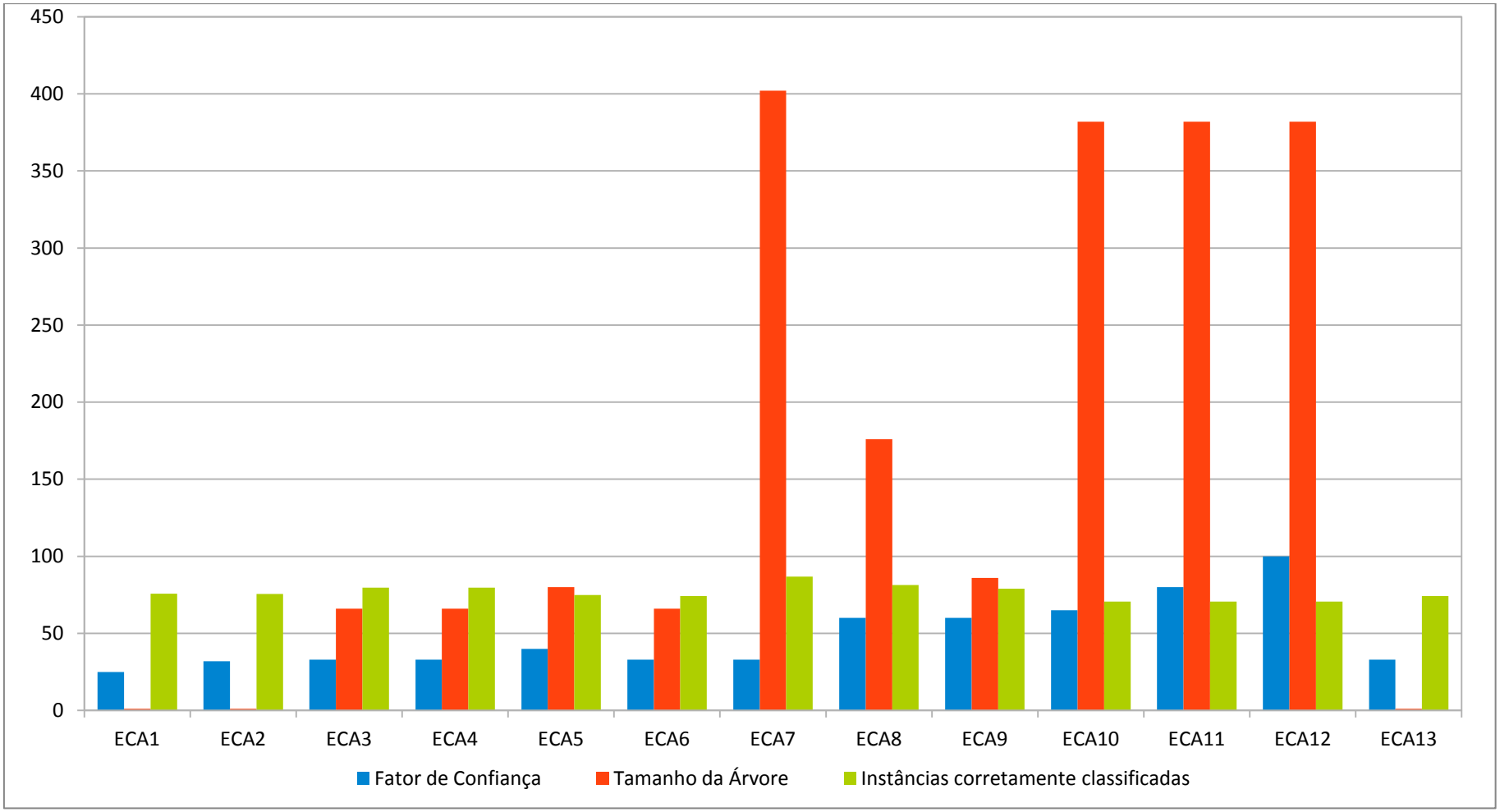


Figura 39 - Gráfico comparativo do ECA1 ao ECA13

Fonte: Autoria própria

Tabela 15 - Dados de experimentos para classificação de trabalhos completos

Experimento	Fator de Confiança	Tamanho da Árvore	Instâncias corretamente classificadas	Ìndice *Kappa*	*Unpruned*	*MinNumObject*	*NumFolds*	Opção de teste
ECT1	25%	1	75,43%	0	*false*	2	3	*Percentage Split 66%*
ECT2	36%	1	74,61%	0	*false*	2	3	*Use training set*
ECT3	37%	103	79,97%	0,34	*false*	2	3	*Use training set*
ECT4	37%	103	79,97%	0,34	*false*	2	5	*Use training set*
ECT5	37%	103	72,07%	0,10	*false*	2	3	*Percentage Split 50%*
ECT6	37%	103	73,53%	0,02	*false*	2	3	*Cross-validation 10*
ECT7	37%	465	87,84%	0,70	*true*	2	3	*Use training set*
ECT8	37%	465	68,05%	0,17	*true*	2	3	*Cross-validation 10*
ECT9	60%	465	68,05%	0,17	*true*	2	3	*Cross-validation 10*
ECT10	100%	427	64,43%	0,07	*false*	2	3	*Percentage Split 50%*
ECT11	37%	327	68,05%	0,16	*true*	3	3	*Cross-validation 10*
ECT12	37%	1	74,61%	0	*false*	3	3	*Cross-validation 10*
ECT13	37%	1	74,61%	0	*false*	4	3	*Cross-validation 10*
ECT14	37%	195	69,12%	0,13	*true*	5	3	*Cross-validation 10*
ECT15	37%	1	74,61%	0	*false*	6	3	*Cross-validation 10*

Fonte: Autoria própria

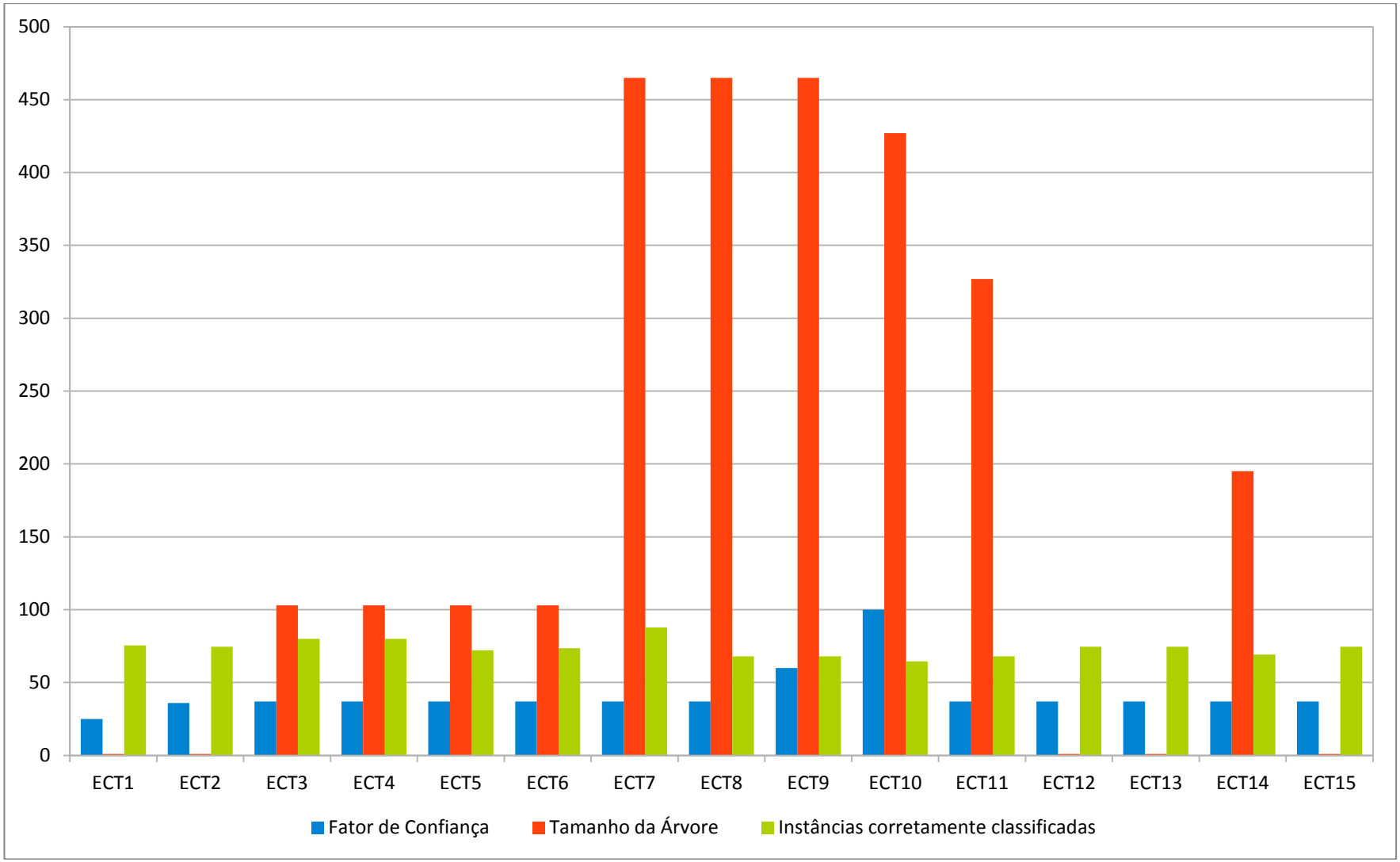


Figura 40 - Gráfico comparativo do ECT1 ao ECT15

Fonte: Autoria própria

No ECA5 foi utilizado a opção de teste do WEKA chamada *Supplied test set,* onde o conjunto original de dados foi dividido em 50% para treinamento (420 instâncias) e 50% para teste (419 instâncias). Esta divisão nos dados foi realizada pela aba *PreProcess* com o filtro *RemovePercentage.*

No ECA5 elevou-se o fator de confiança do J48 até 40% e no ECT5 até 37% para que fossem geradas árvores de decisão com nível de poda (ramificações) mais interessante.

Vale lembrar que o fator de confiança a ser estabelecido para o experimento, assim como outros parâmetros do algoritmo, irá depender do grau de detalhe que o analista de negócio deseja para a árvore de decisão. Quanto menor o fator de confiança, mais poda será realizada na árvore e mais genérica ela será, consequentemente menos atributos irão aparecer na mesma.

A Tabela 16 resume as informações das árvores de decisão geradas no ECA5 e ECT5 (experimentos que forneceram a melhor taxa de acerto). Devido à estrutura complexa das árvores de decisão idealizadas pelo algoritmo, não foi possível apresentá-las em uma única figura.

Tabela 16 - Dados das Árvores de Decisão em ECA5 e ECT5

Exp.	Fator de Confiança	Tamanho da Árvore	Nº de Folhas	Instâncias classificadas corretamente	Instâncias classificadas incorretamente	Nº de atributos que aparecem na árvore
ECA5	40%	80	64	74,94%	25,05%	10
ECT5	37%	103	85	72,07%	27,92%	11

Fonte: Autoria própria

Dentre os 18 atributos iniciais, o algoritmo elaborou um modelo utilizando 10 atributos no ECA5 e 11 atributos no ECT5.

Analisando as árvores de decisão geradas em ECA5 e ECT5 verifica-se a existência de folhas com resultados mais expressivos entre os demais. A Tabela 17 apresenta o percurso do nó raiz até a folha, para o(s) melhor(es) resultado(s) para as classes de A a D, gerados na árvore de decisão do ECA5. As instâncias classificadas

erroneamente foram subtraídas do total de instâncias classificadas em cada nó folha, e o restante das instâncias classificadas corretamente estão apresentadas na quarta coluna da Tabela 17.

Tabela 17 - Percurso das melhores folhas para as classes de A a D no ECA5

Folha	Classe	Percurso	Nº Instâncias classificadas corretamente	Precisão de acerto na folha
1	A	orientacaoespec=Duas_oumais, cursoandamento=D	3	75%
2	A	orientacaoespec=Duas_oumais, cursoandamento=Nenhum, orientacaomes=S	2	100%
3	A	orientacaoespec=Nenhuma, projeto = Quatro, titulacaomax=D, atuaposgraduacao=S	2	100%
4	B	orientacaoespec=Uma, anosformacao=4oumais	4	80%
5	C	orientacaoespec=Nenhuma, projeto=Nenhum, anosformacao=0anos, titulacaomax=D	6	86%
6	D	orientacaoespec=Nenhuma, projeto=Nenhum, anosformacao=4oumais	143	91%

Fonte: Autoria própria

A Tabela 17 apresenta para a classe A, os percursos das três melhores folhas, pois a quantidade de instâncias classificadas corretamente foram próximas.

Analisando a Tabela 17 percebe-se que o atributo **orientacaoespec** foi o que promoveu o maior ganho de informação nas 4 classes (A, B, C e D).

Nos experimentos de classificação deste trabalho, a melhor folha na árvore de decisão foi considerada aquela que classificou a maior quantidade de registros corretamente e com a melhor **taxa de precisão** de acerto na folha (conforme explicado na seção 2.3.2.1).

Portanto, quanto à publicação de artigos no último triênio no IFG, destacaram-se:

- na **classe A**: os docentes que estavam cursando doutorado e orientaram duas ou mais especializações (folha 1); os que orientaram duas ou mais especializações e pelo menos uma orientação de mestrado (folha 2), e os

doutores que desenvolveram 4 projetos e atuam nos cursos de pós-graduação (folha 3 da Tabela 17);

- na **classe B**: os docentes que orientaram uma especialização e possuem mais de 4 anos de formação da sua titulação máxima;
- na **classe C:** os doutores, com menos de um ano de formação, que não orientaram especialização e não desenvolveram projetos no triênio;
- na **classe D**: os docentes que não orientaram especialização, não desenvolveram projetos e que já possuem mais de 4 anos de formação da titulação máxima.

Como pode ser observado pelos experimentos, os atributos que influenciaram na publicação de artigos em periódicos, no último triênio no IFG, foram **orientacaoespec**, **cursoandamento**, **orientacaomes**, **titulacaomax**, **atuaposgraduacao**, **anosformacao** e **projeto**.

A Tabela 18 apresenta os percursos das melhores folhas para as classes de A a D, geradas na árvore de decisão do ECT5.

Tabela 18 - Percurso das melhores folhas para as classes de A a D no ECT5

Folha	Classe	Percurso	Nº Instâncias classificadas corretamente	Precisão de acerto na folha
1	A	projeto=Cinco_oumais, campus=ITU	4	100%
2	A	projeto=Um, orientacaoic=Nenhuma, anosformacao=0anos, titulacaomax=M	4	100%
3	B	projeto=Um, orientacaoic=Duas, orientacaograd=Uma	3	100%
4	B	projeto=Tres, area=História, cursoandamento=Nenhum	3	75%
5	C	projeto=Cinco_oumais, campus=GYN, atuaposgraduacao=S, idade=40a49anos	3	75%
6	D	projeto=Nenhum	406	86%

Fonte: Autoria própria

No ECT5, o atributo que forneceu o maior ganho de informação para as classes de A a D foi o de **projeto**. A Tabela 18 apresentou os percursos para as duas melhores folhas para as classes A e B e um percurso para as classes C e D.

Logo, quanto à publicação de trabalhos completos, destacaram-se:

- na **classe A**: os docentes que desenvolveram 5 ou mais projetos, lotados no câmpus de Itumbiara, e os mestres recém formados (com menos de 1 ano de conclusão do mestrado), que desenvolveram um projeto e não orientaram iniciação científica;
- na **classe B**: os docentes que desenvolveram pelo menos um projeto, tiveram duas orientações de iniciação científica e uma de graduação; na folha 4 da Tabela 18 também houve destaque para aqueles que desenvolveram 3 projetos no triênio, são da área de História e não estavam cursando nenhum curso no momento;
- na **classe C**: os docentes que desenvolveram 5 ou mais projetos, lotados no câmpus Goiânia, que atuam nos cursos de pós-graduação do IFG e possuem idade entre 40 a 49 anos;
- na **classe D**: os docentes que não desenvolveram nenhum projeto no triênio de 2011 a 2013, como também ocorreu em ECA5.

De acordo com os percursos das melhores folhas apresentados na Tabela 18, os atributos que influenciaram na publicação de trabalhos completos, no último triênio no IFG, foram **projeto**, **orientacaoic**, **anosformacao**, **titulacaomax**, **campus**, **orientacaograd**, **area**, **cursoandamento**, **atuaposgraducao** e **idade**.

Conforme apresentado, o número de docentes classificados nas classes A, B e C nos experimentos anteriores é muito baixo, o que indica que o número de publicações em artigos e trabalhos completos entre os docentes do IFG foi pequeno nos últimos 3 anos. Em consequência de tais resultados, resolveu-se realizar novos experimentos considerando outros tipos de publicações no mesmo período, afim de aumentar a proporção de docentes entre as classes e tentar descobrir novos padrões.

Portanto, além dos artigos e trabalhos completos, foram considerados nos mesmos experimentos, outros itens de publicação de maior importância e expressividade entre os dados dos docentes do IFG. Assim, o critério para

classificação entre as classes foi modificado. A quantidade de cada tipo de publicação considerada foi multiplicada por uma pontuação (peso), para distinguir a relevância entre elas. Além disso, alguns domínios de atributos também foram pontuados.

As pontuações estabelecidas foram baseadas nas pontuações dos programas ProAPP (Programa de Apoio a Produtividade em Pesquisa) e PIPECT (Programa Institucional de Incentivo à Participação de Eventos Científicos e Tecnológicos), conforme constam nos Anexos A e B. Para os itens de publicação considerados que não aparecem nos dois formulários dos anexos: apresentação de trabalho, processos ou técnicas, relatório de pesquisa e material didático ou instrucional foi atribuído 1 ponto para cada tipo.

Os itens de publicação que passaram a ser considerados e as respectivas pontuações estabelecidas foram:

- **livros publicados ou organizados com ISBN** - (10 pontos): o ISBN *(International Standard Book Number*) é um sistema que identifica numericamente os livros segundo o título, o autor, o país e a editora, individualizando-os inclusive por edição (ISBN, 2013). Os registros de livros publicados sem preenchimento do ISBN no período foram desconsiderados;
- **artigos completos publicados em periódicos com ISSN** - (10 pontos);
- **capítulos de livros com ISBN** - (5 pontos): os capítulos publicados em livros sem ISBN foram desconsiderados;
- **trabalhos completos publicado em anais de eventos** - (3 pontos)**;**
- **resumos expandidos (estendido) publicado em anais de eventos** - (2 pontos);
- **resumos publicados em anais de eventos** - (1 ponto);
- **texto em jornal ou revista** - (1 ponto);

- **trabalho técnico** - (1 ponto): foram inclusos neste item as produções técnicas das seguintes naturezas: elaboração de projeto, parecer, assessoria, relatório técnico, consultoria, entre outras;
- **relatório de pesquisa** - (1 ponto);
- **material didático ou instrucional** - (1 ponto): nesta categoria foram inclusos itens das seguintes naturezas: tutorial, livros, apostilas, vídeo aula, cartilha, manual, guia entre outros;
- **processos ou técnicas** - (1 ponto): podem ser de natureza pedagógica, analítica, processual, instrumental ou outras;
- ***software*** - (3 pontos);
- **apresentação de trabalho** - (1 ponto): podem ser das seguintes naturezas: congressos, simpósios, conferências, seminário, comunicação e outras;
- **orientação de mestrado** - (2 pontos);
- **orientações de graduação, especialização e iniciação científica** – (1 ponto) cada.

Os domínios dos atributos pontuados foram:

- atributo **regimetrabalho=DE** - (5 pontos);
- atributo **titulacaomax=P** ou **titulacaomax=D** (15 pontos) e **titulacaomax=M** (10 pontos);

A Tabela 19 resume os itens de publicação e os domínios de atributos que foram pontuados.

Tabela 19 - Itens pontuados no experimento EC1

Item pontuado	Pontuação
apresentação de trabalho	1 ponto
trabalho completo	3 pontos
artigo em periódico com ISSN	10 pontos
resumo	1 ponto
resumo expandido	2 pontos
trabalho técnico	1 ponto
capítulo de livro com ISBN	5 pontos
livro publicado ou organizado com ISBN	10 pontos
texto em jornal ou revista	1 ponto
material didático ou instrucional	1 ponto
relatório de pesquisa	1 ponto
software	3 pontos
processo ou técnica	1 ponto
orientação de mestrado	2 pontos
orientações de especialização, iniciação científica e graduação	1 ponto
regimetrabalho=DE	5 pontos
titulacaomax=P ou titulacaomax=D	15 pontos
titulacaomax=M	10 pontos

Fonte: Autoria própria

A Tabela 20 apresenta o quantitativo de docentes que tiveram pelo menos uma publicação de cada tipo que foi listada na tabela anterior, dentro do período da pesquisa no IFG. Os itens estão listados em ordem decrescente de quantidade.

Tabela 20 - Quantidade de docentes por tipo de publicação

Item de Publicação	Valor absoluto - %
Apresentação de trabalho	325 – 38,73%
Trabalho completo publicado em anais de eventos	213 – 25,38%
Artigo completo publicado em periódico com ISSN	205 – 24,43%
Resumo publicado em anais de eventos	136 – 16,20%
Resumo expandido publicado em anais de eventos	87 – 10,36%
Trabalho técnico	86 – 10,25%
Capítulo de livro com ISBN	66 – 7,86%
Livro publicado ou organizado com ISBN	38 – 4,52%
Material didático ou instrucional	27 – 3,21%
Texto em jornal ou revista	25 – 2,97%
Relatório técnico	18 – 2,14%
Software	15 – 1,78%
Processo ou técnica	4 – 0,47%

Fonte: Autoria própria

Os itens de publicação mais comuns no IFG no último triênio foram **apresentação de trabalho**, **trabalho completo** e **artigo**.

Para a definição do atributo **classe** a cada registro de docente, em todos os experimentos, primeiramente foi criada uma *view* chamada **v_publicacoes** (vide Anexo G), onde os itens de publicações são somados por docente. Esse somatório de cada tipo de publicação por docente é retornado para a *view* **v_arff**. Ao executar a v_arff, seu resultado é extraído para um arquivo **.csv**, dentro dessa planilha foi adicionada uma coluna onde o atributo **classe** é atribuído conforme o critério em questão.

Após o cálculo da pontuação final de cada docente no triênio, calculou-se também a média e desvio padrão das pontuações. A média obtida da população foi de 26,41 pontos e o desvio padrão de 25,64 pontos. Posteriormente, foram estabelecidas 5 classes conforme os critérios apresentados na Tabela 21.

Tabela 21 - Critérios para as classes dos experimentos com pontuações

Classe	Critério	Qtde de docentes em cada classe
A	**Pontuação** > Média + 2 Desvios Padrão	33
B	Média + 1 Desvio Padrão < **Pontuação** < Média + 2 Desvios Padrão	50
C	Média < **Pontuação** < Média + 1 Desvio Padrão	195
D	Média - 1 Desvio padrão < **Pontuação** < Média	553
E	**Pontuação** < Média - 1 Desvio padrão	8

Fonte: Autoria própria

O experimento EC1 detalhado a seguir, foi o que forneceu o melhor resultado, considerando todas as publicações da Tabela 19. Ele foi realizado com todos os atributos da Tabela 13, com fator de confiança de 35%, podendo realizar poda (*unpruned=false*), com número mínimo de 2 instâncias por folha (*minNumObj=2*) e opção de teste *Supplied test set* para testar o modelo. A divisão do conjunto de dados foi de 70% para treinamento (587 registros) e 30% para teste (252 registros). Os

demais parâmetros do J48 foram mantidos em seus valores *default*. A Figura 41 apresenta o resultado do EC1, obtido através do conjunto de teste.

```
User supplied test set
Relation:     lattes-weka.filters.unsupervised.attribute.Remove-R1,19-21-weka.filte
Instances:     unknown (yet). Reading incrementally
Attributes:   18

=== Summary ===

Correctly Classified Instances         204               80.9524 %
Incorrectly Classified Instances        48               19.0476 %
Kappa statistic                          0.5532
Mean absolute error                      0.1076
Root mean squared error                  0.2331
Coverage of cases (0.95 level)          98.4127 %
Total Number of Instances              252

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.875     0.008      0.778     0.875     0.824      0.995    A
                 0.364     0.012      0.571     0.364     0.444      0.89     B
                 0.419     0.016      0.897     0.419     0.571      0.874    C
                 0.976     0.47       0.809     0.976     0.885      0.896    D
                 1         0.004      0.667     1         0.8        0.999    E
Weighted Avg.    0.81      0.32       0.818     0.81      0.786      0.894

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
   7   0   1   0   0 |   a = A
   0   4   1   6   0 |   b = B
   1   2  26  33   0 |   c = C
   1   1   1 165   1 |   d = D
   0   0   0   0   2 |   e = E
```

Figura 41 - Resultado do EC1

Fonte: Autoria própria

Percebe-se que o modelo de classificação criado classificou quase 81% das instâncias de teste corretamente. As informações da árvore de decisão resultante estão apresentadas na Tabela 22.

Tabela 22 - Dados da árvore de decisão em EC1

Exp.	Fator de Confiança	Tamanho da Árvore	Nº de Folhas	Instâncias classificadas corretamente	Instâncias classificadas incorretamente	Índice *Kappa*	Nº de atributos que apareceram na árvore
EC1	35%	132	117	80,95%	19,04%	0,55	9

Fonte: Autoria própria

A Tabela 23 apresenta os percursos das melhores folhas para as classes de A a D da árvore de decisão do EC1.

Tabela 23 - Percurso das melhores folhas para as classes de A a D no EC1

Folha	Classe	Percurso	Nº Instâncias classificadas corretamente	Precisão de acerto na folha
1	A	titulacaomax=D, projeto=Tres, sexo=F	2	67%
2	A	titulacaomax=M, projeto=Tres, area=Educacao Fisica	2	64%
3	B	titulacaomax=D, projeto=Quatro,	4	67%
4	C	titulacaomax=D, projeto=Um, orientacaograd=Nenhuma	8	89%
5	D	titulacaomax=M, projeto=Nenhum, orientacaoespec=Nenhuma	165	84%
6	E	titulacaomax=E, regimetrabalho=40H, sexo=M	3	75%
7	E	titulacaomax=G, regimetrabalho=40H	2	100%

Fonte: Autoria própria

Portanto, considerando outros tipos de publicações, além de artigos e trabalhos completos, destacaram-se:

- na **classe A**: as doutoras e os mestres que desenvolveram 3 projetos no triênio, com destaque para a área de Educação Física;
- na **classe B**: os doutores que desenvolveram 4 projetos;
- na **classe C**: os doutores que desenvolveram um projeto e não orientaram graduação;
- na **classe D**: os mestres que não desenvolveram projetos e não orientaram especialização no período;
- na **classe E**: os especialistas e graduados de regime de trabalho 40 horas, havendo destaque para o sexo masculino entre os especialistas.

Pelos experimentos foi possível notar que as melhores folhas apresentadas para as classes menos representativas classificaram corretamente um pequeno número de instâncias. Tais classes, além de possuírem poucas instâncias, ainda são divididas entre os conjuntos de dados de treinamento e teste, e posteriormente as poucas instâncias presentes no conjunto de teste são distribuídas entre as folhas da árvore de decisão resultante. Como exemplo, observa-se pela primeira linha da matriz de confusão da Figura 41 que somente 8 das 33 instâncias da classe A em EC1 foram selecionadas para teste.

Logo, de acordo com a árvore de decisão gerada para o EC1, os atributos que mais influenciaram na classificação quando considerados todos os tipos de publicações da Tabela 19, promovendo maior ganho de informação, foram **titulacaomax** e **projeto**.

5.4.2 Associação

A tarefa de Associação também foi aplicada nos dados deste trabalho no intuito de descobrir regras de relação entre os atributos selecionados. O algoritmo escolhido para esta tarefa foi o *Apriori* (apresentado na seção 2.3.2.3.1), por ser o mais usado e considerado um dos mais eficientes entre os algoritmos associativos, além de estar disponível na ferramenta WEKA.

A Figura 42 exibe a tela do WEKA ao se selecionar a aba *Associate* e o algoritmo *Apriori,* clicando no botão *Choose. O*s valores *default* dos parâmetros do *Apriori* também aparecem na tela da direita na mesma figura.

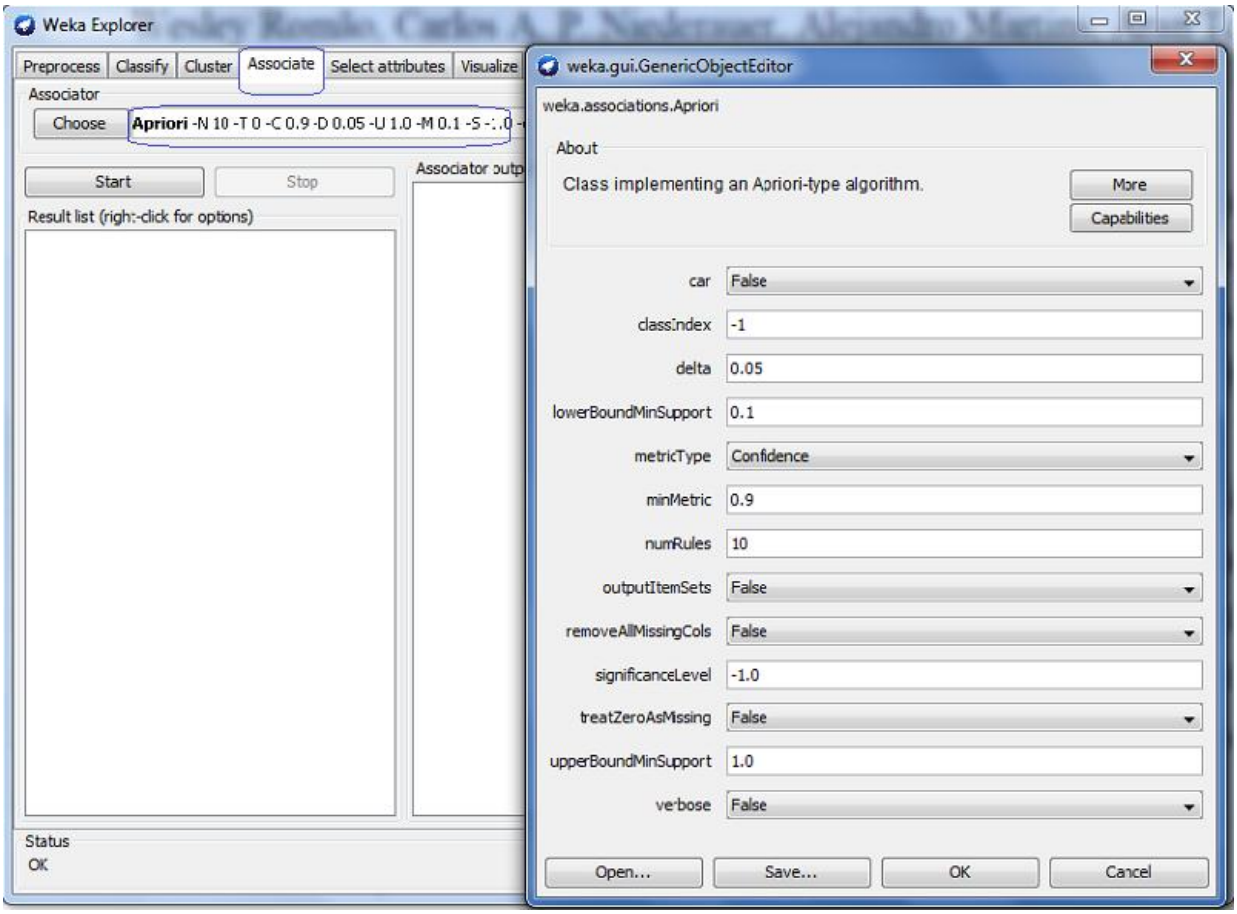


Figura 42 - Aba *Associate* e parâmetros do *Apriori* no WEKA

Fonte: Autoria própria

De acordo com Bueno e Viana (2012), os parâmetros do *Apriori* no WEKA significam:

- *car*: se verdadeiro considera que os dados já foram minerados;
- *classIndex*: índice do atributo classe. Se indicado como -1 , o último atributo é considerado classe;
- *delta*: o algoritmo diminui o suporte de confiança pelo valor especificado em delta. No caso de minerações mais detalhadas este valor deve ser pequeno, mas como consequência o tempo de interação aumenta;
- *lowerBoundMinSupport:* limite inferior para o suporte das regras, ou seja, suporte mínimo;
- *metricType*: o tipo de métrica a qual serão geradas as regras. Existem 4 tipos: *confidence*: é a confiança da regra, mede a probabilidade condicional de P(c)

dado A. Geralmente dá ênfase a regras que não estão relacionadas; *lift*: mede a distância para a independência entre A e C e pode variar entre 0 e infinito; *leverage*: mede o número de casos extras obtidos em relação ao esperado; *conviction*: tenta capturar o grau de implicação entre A e C, se nos resultados o valor for 1 indica independência.

- *minMetric*: a menor confiança aceita;
- *numRules*: determina o número de regras que será mostrada pelo *software*;
- *outputItemSets:* se verdadeiro, o *software* vai mostrar os conjuntos de *item sets* descobertos;
- *removeAllMissingCols:* se verdadeiro remove as colunas de valores dos atributos que estiverem nulos;
- *significanceLevel:* teste de significância (usado somente com a métrica *confidence*);
- *treatZeroasMissing:* se for verdadeiro, valores preenchidos com zero são tratados como valores vazios;
- *upperBoundMinSupport:* limite superior para o suporte;
- *verbose*: se verdadeiro mostra os detalhes da mineração, os passos do algoritmo.

Para a tarefa de Associação duas abordagens podem ser utilizadas. Uma pode realizar experimentos utilizando valores de **suporte** e **confiança mínimos**, em busca dos itens **mais frequentes**, conforme a proposta original do algoritmo *Apriori*. E a outra pode tratar de itens **menos frequentes**, acrescentando ao algoritmo um limite **máximo** para o **suporte**, pois conforme descrito em Romão *et al.* (1999), em se tratando de pesquisadores, é bastante provável que o decisor esteja interessado em descobrir o comportamento das minorias, posto que elas podem revelar pesquisadores com alta performance ou, por outro lado, com baixo rendimento.

Iniciou-se os experimentos de Associação relacionando os docentes quanto ao número de artigos em periódicos no triênio. A Tabela 24 apresenta o número de regras geradas pelo algoritmo com diferentes variações nos valores mínimos para o **suporte** e a **confiança**, na busca pelos itens mais frequentes. Os 18 atributos da Tabela 13 foram selecionados nos experimentos de EAA1 a EAA5. Percebe-se que o número de regras geradas é inversamente proporcional aos valores determinados para o suporte e confiança mínimos: a medida que estes últimos decrescem, aumenta-se o número de regras.

Tabela 24 - Nº de regras conforme valores mínimo para suporte e confiança

Experimento	Suporte Mínimo	Confiança Mínima	Número de Regras
EAA1	70%	80%	288
EAA2	60%	50%	1608
EAA3	50%	40%	5416
EAA4	50%	20%	5416
EAA5	40%	20%	21264

Fonte: Autoria Própria

No experimento EAA1, as 288 regras geradas apresentaram alguma redundância, como mostra a Figura 43, que exibe a tela apresentada no WEKA.

```
Best rules found:

 1. classe_artigo=D 634 ==> orientacaomes=N 634    <conf:(1)> lift:(1.01) lev:(0) [3] conv:(3.78)
 2. orientacaoespec=Nenhuma classe_artigo=D 614 ==> orientacaomes=N 614    <conf:(1)> lift:(1.01) lev:(0) [3] conv:(3.66)
 3. atuaposgraduacao=N classe_artigo=D 598 ==> orientacaomes=N 598    <conf:(1)> lift:(1.01) lev:(0) [3] conv:(3.56)
 4. regimetrabalho=DE atuaposgraduacao=N orientacaoespec=Nenhuma 690 ==> orientacaomes=N 689    <conf:(1)> lift:(1) lev:(0) [3] conv:(2.06)
 5. orientacaoespec=Nenhuma orientacaoic=Nenhuma 683 ==> orientacaomes=N 682    <conf:(1)> lift:(1) lev:(0) [3] conv:(2.04)
 6. atuaposgraduacao=N orientacaoespec=Nenhuma orientacaoic=Nenhuma 648 ==> orientacaomes=N 647    <conf:(1)> lift:(1) lev:(0) [2] conv:(1.93)
 7. nucleopesquisa=N orientacaoespec=Nenhuma 639 ==> orientacaomes=N 638    <conf:(1)> lift:(1) lev:(0) [2] conv:(1.9)
 8. regimetrabalho=DE orientacaoespec=Nenhuma orientacaoic=Nenhuma 631 ==> orientacaomes=N 630    <conf:(1)> lift:(1) lev:(0) [2] conv:(1.88)
 9. atuaposgraduacao=N nucleopesquisa=N orientacaoespec=Nenhuma 624 ==> orientacaomes=N 623    <conf:(1)> lift:(1) lev:(0) [2] conv:(1.86)
10. regimetrabalho=DE atuaposgraduacao=N orientacaoespec=Nenhuma orientacaoic=Nenhuma 600 ==> orientacaomes=N 599    <conf:(1)> lift:(1) lev:(0)
11. orientacaoic=Nenhuma orientacaograd=Nenhuma 593 ==> orientacaomes=N 592    <conf:(1)> lift:(1) lev:(0) [2] conv:(1.77)
12. atuaposgraduacao=N orientacaoespec=Nenhuma 744 ==> orientacaomes=N 742    <conf:(1)> lift:(1) lev:(0) [2] conv:(1.48)
13. regimetrabalho=DE orientacaoespec=Nenhuma 736 ==> orientacaomes=N 734    <conf:(1)> lift:(1) lev:(0) [2] conv:(1.46)
14. regimetrabalho=DE atuaposgraduacao=N 717 ==> orientacaomes=N 715    <conf:(1)> lift:(1) lev:(0) [2] conv:(1.42)
15. regimetrabalho=DE orientacaoic=Nenhuma 661 ==> orientacaomes=N 659    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.31)
16. regimetrabalho=DE orientacaograd=Nenhuma 630 ==> orientacaomes=N 628    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.25)
17. regimetrabalho=DE atuaposgraduacao=N orientacaoic=Nenhuma 623 ==> orientacaomes=N 621    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.24)
18. atuaposgraduacao=N orientacaograd=Nenhuma 622 ==> orientacaomes=N 620    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.24)
19. regimetrabalho=DE nucleopesquisa=N 609 ==> orientacaomes=N 607    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.21)
20. atuaposgraduacao=N orientacaoespec=Nenhuma orientacaograd=Nenhuma 609 ==> orientacaomes=N 607    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.21)
21. regimetrabalho=DE orientacaoespec=Nenhuma orientacaograd=Nenhuma 608 ==> orientacaomes=N 606    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.21)
22. regimetrabalho=DE atuaposgraduacao=N nucleopesquisa=N 595 ==> orientacaomes=N 593    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.18)
23. orientacaoespec=Nenhuma 796 ==> orientacaomes=N 793    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.19)
24. regimetrabalho=DE 776 ==> orientacaomes=N 773    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.16)
25. orientacaoic=Nenhuma 716 ==> orientacaomes=N 713    <conf:(1)> lift:(1) lev:(0) [1] conv:(1.07)
26. atuaposgraduacao=N orientacaoic=Nenhuma 674 ==> orientacaomes=N 671    <conf:(1)> lift:(1) lev:(0) [1] conv:(1)
27. orientacaograd=Nenhuma 671 ==> orientacaomes=N 668    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
28. nucleopesquisa=N 666 ==> orientacaomes=N 663    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.99)
29. atuaposgraduacao=N nucleopesquisa=N 648 ==> orientacaomes=N 645    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.97)
30. orientacaoespec=Nenhuma orientacaograd=Nenhuma 648 ==> orientacaomes=N 645    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.97)
```

Figura 43 - Regras geradas no experimento EAA1

Fonte: Autoria Própria

Com a redução dos valores para o suporte e a confiança em EAA2, o aumento de regras minerou outros itens. Além de regras idênticas as geradas em EAA1, outras regras menos frequentes apareceram.

As regras que aparecem na Figura 43, são regras relativas a maioria dos docentes da população e estão coerentes com os dados em estudo, pois a grande maioria dos docentes são de regime de trabalho DE, não orientaram, não atuam nos cursos de pós-graduação e não participam dos núcleos de pesquisa, conforme mostram as 30 primeiras regras do EAA1 mostradas na Figura 43.

Nos experimentos EAA3, EAA4 e EAA5, aumentaram as regras e a redundância entre elas. Porém, novas regras foram descobertas. Agora, as regras passaram a considerar outros atributos que antes não apareciam.

No EAT1 replicou-se o experimento EAA5 (com as mesmas configurações para o algoritmo, inclusive para os valores do suporte e confiança mínimos do EAA5 apresentados na Tabela 24), porém com a classificação quanto ao número de trabalhos completos publicados em anais de eventos. Observou-se no EAT1 que as regras geradas eram muito similares as regras do EAA5.

Através das primeiras regras mineradas já foi possível notar certa relação entre os atributos **orientacaomes**, **orientacaoespec**, **orientacaograd**, **orientacaoic**, **nucleopesquisa**, **atuaposgraduacao** e **projeto**. Na maioria das regras em que estes atributos aparecem juntos, quando o docente possui um deles com valor de “N” ou “Nenhum”, os demais também aparecem com valores similares, e geralmente pertencem a classe D.

O suporte e a confiança atuam como medidas de interesse no processo de mineração de regras de associação. O WEKA não informa o valor do suporte das regras, mas ele pode ser calculado dividindo o número do consequente da regra, pelo número total da população estudada (839 docentes). Tomando como exemplo a Regra 28 da Figura 43, o suporte é de 0,7902. A interpretação do **suporte** para a mesma regra, é a seguinte: 79,02% dos docentes da população não participam de núcleos de

pesquisa e não orientaram mestrado. A interpretação da **confiança** é: 100% dos docentes (o WEKA arredonda para cima) que não participam de núcleo de pesquisa, também não orientaram mestrado.

Visto que a mineração de bases de dados reais pode levar à geração de centenas de milhares de regras de Associação, grande parte delas óbvias e redundantes, buscou-se encontrar aquelas mais relevantes, que apresentassem real dependência entre os atributos. Inicialmente, tentou-se selecionar os 18 atributos da Tabela 13 simultaneamente. Porém, devido a infinidade de regras que o algoritmo gera ao fazer a combinação dos domínios de cada par de atributos (causando escassez de recursos computacionais e interrompendo o processamento do *software*), além da complexidade das mesmas, (exibindo vários atributos na mesma regra), por questões de simplicidade optou-se por selecionar os atributos de dois a dois. Por este motivo, as regras apresentadas na Tabela 25 possuem somente um item no antecedente e um item no consequente da mesma regra.

A Tabela 25 apresenta apenas alguns exemplos das regras encontradas com dependência positiva. Em todas elas, o valor do suporte da regra (**Sup(A U B)**) é maior que o valor dos suportes de cada item multiplicados (**SupEsp(A U B)**). Além disso, todos os valores de *lift* são maiores que 1. A ferramenta WEKA calcula o valor de *lift* para cada regra.

Tabela 25 - Regras de Associação com dependência positiva entre os itens

Nº da Regra	Regra	SupEsp(A U B)	Sup(A U B)	Conf.	*Lift*
1	atuaposgraduacao=S ==> nucleopesquisa=S	1,58%	5,60%	72%	3,51
2	titulacaomax=D ==> campus=GYN	5,97%	8,34%	47%	1,39
3	campus=ANA ==> titulacaomax=M	4,91%	6,79%	83%	1,38
4	campus=INH ==> projeto=Cinco_oumais	0,51%	1,07%	13%	2,09
5	campus=GYN ==> grandearea=ENG	6,06%	9,53%	28%	1,57
6	grandearea=CET ==> nucleopesquisa=S	5,91%	6,91%	24%	1,17
7	campus=JAT ==> atuaposgraduacao=S	0,68%	1,78%	20%	2,58
8	area=Matematica ==> atuaposgraduacao=S	0,56%	2,02%	28%	3,60
9	projeto=Cinco_oumais ==> nucleopesquisa=S	1,32%	3,93%	61%	2,96
10	campus=INH ==> nucleopesquisa=S	1,62%	3,45%	43%	2,10
11	campus=GOI ==> nucleopesquisa=S	0,53%	0,95%	36%	1,76
12	atuaposgraduacao=S ==> projeto=Cinco_oumais	0,49%	1,79%	23%	3,59
13	projeto=Cinco_oumais ==> sexo=F	2,46%	3,57%	56%	1,44
14	projeto=Quatro ==> sexo=F	1,38%	1,78%	48%	1,26
15	regimetrabalho=40H ==> grandearea=ENG	1,20%	2,14%	32%	1,75
16	orientacaograd=Quatro_oumais ==> area=Quimica	0,33%	0,95%	20%	2,77
17	sexo=F ==> grandearea=LLA	3,75%	7,38%	19%	1,96
18	sexo=M ==> grandearea=ENG	11,06%	14,66%	24%	1,32
19	campus=APA ==> escritaingles=B	1,43%	1,90%	26%	1,32
20	orientacaoespec=Uma ==> grandearea=CH	0,39%	0,95%	42%	2,36

Fonte: Autoria Própria

A interpretação do valor do *lift* para a Regra 1 da Tabela 25 é: que a participação em núcleos de pesquisa do CNPq é 3,51 vezes maior entre os docentes que atuam nos cursos de pós-graduação do IFG.

Para descobrir o perfil predominante de cada classe de docente quanto à publicação de artigos, também através de regras de associação, realizou-se mais 18 experimentos, combinando cada um dos 17 atributos da Tabela 13 (supostos atributos previsores) com o atributo **classe**. Tais experimentos foram executados com os

seguintes parâmetros para o *Apriori: lowerBoundMinSupport=0.0, minMetric=0.0 e numRules*=200. Os demais foram mantidos em seus valores padrões.

Em cada um desses experimentos, foram geradas quatro regras para cada combinação de domínio do atributo em questão com o atributo **classe**. A Figura 44 ilustra o experimento realizado com os atributos **sexo** e **classe**, no qual foram geradas 16 regras. As regras aparecem em ordem decrescente do valor de confiança. Nesse caso, é possível interpretar pelas regras 6 e 7 da figura, que dos 54 docentes da classe A, 29 são homens, gerando uma confiança de 54% e 25 são mulheres, com confiança de 46%. Portanto, apesar dos valores de confiança destas duas regras estarem próximos, nesse caso, os homens ainda são predominantes na classe A de artigos. Este mesmo procedimento foi repetido para cada atributo: a primeira regra listada onde o antecedente é a classe (A, B, C ou D), o valor do consequente da mesma regra, indica o domínio mais frequente do atributo em questão naquela classe. A Tabela 26 apresenta o perfil predominante encontrado para a classe A, a Tabela 27 da classe B, a Tabela 28 da classe C e a Tabela 29 para a classe D quanto à classificação de artigos em periódicos.

```
Best rules found:

 1. sexo=M 516 ==> classe_artigo=D 409    <conf:(0.79)> lift:(1.05) lev:(0.02) [19] conv:(1.17)
 2. sexo=F 323 ==> classe_artigo=D 225    <conf:(0.7)> lift:(0.92) lev:(-0.02) [-19] conv:(0.8)
 3. classe_artigo=D 634 ==> sexo=M 409    <conf:(0.65)> lift:(1.05) lev:(0.02) [19] conv:(1.08)
 4. classe_artigo=B 51 ==> sexo=F 30    <conf:(0.59)> lift:(1.53) lev:(0.01) [10] conv:(1.43)
 5. classe_artigo=C 100 ==> sexo=M 57    <conf:(0.57)> lift:(0.93) lev:(-0.01) [-4] conv:(0.87)
 6. classe_artigo=A 54 ==> sexo=M 29    <conf:(0.54)> lift:(0.87) lev:(-0.01) [-4] conv:(0.8)
 7. classe_artigo=A 54 ==> sexo=F 25    <conf:(0.46)> lift:(1.2) lev:(0.01) [4] conv:(1.11)
 8. classe_artigo=C 100 ==> sexo=F 43    <conf:(0.43)> lift:(1.12) lev:(0.01) [4] conv:(1.06)
 9. classe_artigo=B 51 ==> sexo=M 21    <conf:(0.41)> lift:(0.67) lev:(-0.01) [-10] conv:(0.63)
10. classe_artigo=D 634 ==> sexo=F 225    <conf:(0.35)> lift:(0.92) lev:(-0.02) [-19] conv:(0.95)
11. sexo=F 323 ==> classe_artigo=C 43    <conf:(0.13)> lift:(1.12) lev:(0.01) [4] conv:(1.01)
12. sexo=M 516 ==> classe_artigo=C 57    <conf:(0.11)> lift:(0.93) lev:(-0.01) [-4] conv:(0.99)
13. sexo=F 323 ==> classe_artigo=B 30    <conf:(0.09)> lift:(1.53) lev:(0.01) [10] conv:(1.03)
14. sexo=F 323 ==> classe_artigo=A 25    <conf:(0.08)> lift:(1.2) lev:(0.01) [4] conv:(1.01)
15. sexo=M 516 ==> classe_artigo=A 29    <conf:(0.06)> lift:(0.87) lev:(-0.01) [-4] conv:(0.99)
16. sexo=M 516 ==> classe_artigo=B 21    <conf:(0.04)> lift:(0.67) lev:(-0.01) [-10] conv:(0.98)
```

Figura 44 - Regras encontradas entre os atributos sexo e classe

Fonte: Autoria Própria

Tabela 26 - Perfil predominante da classe A para artigos

CLASSE A Artigos (54 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	6%	91%
sexo	**M**	3%	54%
idade	**30a39**	3%	52%
titulacaomax	**M**	4%	57%
anosformacao	**1a3anos**	3%	50%
cursoandamento	**Nenhum**	4%	67%
escritaingles	**R**	3%	43%
campus	**GYN**	1%	22%
grandearea	**CET**	1%	20%
area	**Outras**	1%	15%
area	**Química**	1%	11%
atuaposgraduacao	**N**	5%	81%
nucleopesquisa	**N**	4%	56%
projeto	**5_oumais**	2%	33%
orientacaomes	**N**	6%	93%
orientacaoespec	**Nenhuma**	5%	81%
orientacaoic	**Nenhuma**	4%	59%
orientacaograd	**Nenhuma**	4%	63%

Fonte: Autoria Própria

Na Tabela 26, aparecem duas linhas para o atributo **area**: foram encontradas 8 ocorrências para a área “Outras” e 6 ocorrências para a área de Química na classe A de artigos em periódicos.

Tabela 27 - Perfil predominante da classe B para artigos

CLASSE B Artigos (51 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	6%	92%
sexo	**F**	4%	59%
idade	**30a39**	3%	57%
titulacaomax	**M**	3%	55%
anosformacao	**4oumais**	3%	55%
cursoandamento	**Nenhum**	4%	71%
escritaingles	**R**	3%	57%
campus	**GYN**	3%	43%
grandearea	**CET**	2%	29%
area	**Química, Educação, Letras e Outras**	1%	10%
atuaposgraduacao	**N**	5%	86%
nucleopesquisa	**N**	5%	78%
projeto	**Nenhum**	2%	37%
orientacaomes	**N**	6%	100%
orientacaoespec	**Nenhuma**	5%	86%
orientacaoic	**Nenhuma**	5%	88%
orientacaograd	**Nenhuma**	4%	73%

Fonte: Autoria Própria

As áreas de Química, Educação, Letras e Outras apareceram empatadas na classe B de artigos com 5 registros cada uma.

Tabela 28 - Perfil predominante da classe C para artigos

CLASSE C Artigos (100 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	11%	95%
sexo	**M**	7%	57%
idade	**30a39**	6%	52%
titulacaomax	**M**	8%	65%
anosformacao	**4oumais**	7%	56%
cursoandamento	**Nenhum**	7%	58%
escritaingles	**R**	4%	36%
campus	**GYN**	3%	28%
grandearea	**CET**	4%	32%
area	**Química**	1%	12%
atuaposgraduacao	**N**	10%	88%
nucleopesquisa	**N**	8%	63%
projeto	**Nenhum**	4%	37%
orientacaomes	**N**	12%	99%
orientacaoespec	**Nenhuma**	11%	94%
orientacaoic	**Nenhuma**	9%	73%
orientacaograd	**Nenhuma**	10%	84%

Fonte: Autoria Própria

Tabela 29 - Perfil predominante da classe D para artigos

CLASSE D Artigos (634 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	70%	92%
sexo	**M**	49%	65%
idade	**30a39**	28%	38%
titulacaomax	**M**	45%	60%
anosformacao	**4oumais**	52%	68%
cursoandamento	**Nenhum**	55%	73%
escritaingles	**R**	28%	37%
campus	**GYN**	26%	35%
grandearea	**CET**	22%	29%
area	**Ciência da Computação**	7%	9%
atuaposgraduacao	**N**	71%	94%
nucleopesquisa	**N**	64%	84%
projeto	**Nenhum**	49%	65%
orientacaomes	**N**	76%	100%
orientacaoespec	**Nenhuma**	73%	97%
orientacaoic	**Nenhuma**	67%	89%
orientacaograd	**Nenhuma**	62%	81%

Fonte: Autoria Própria

O mesmo procedimento foi repetido para descobrir o perfil predominante dos docentes em cada classe quanto à classificação de trabalhos completos publicados

em anais de eventos. A Tabela 30 apresenta o perfil encontrado da classe A, a Tabela 31 o perfil da classe B, a Tabela 32 para a classe C e a Tabela 33 para a classe D.

Tabela 30 - Perfil predominante da classe A para trabalhos completos

CLASSE A Trabalhos completos (81 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	9%	91%
sexo	**M**	6%	59%
idade	**30a39**	5%	47%
titulacaomax	**M**	6%	67%
anosformacao	**4oumais**	4%	46%
cursoandamento	**Nenhum**	6%	60%
escritaingles	**R**	4%	43%
campus	**GYN**	3%	30%
grandearea	**ENG**	3%	27%
area	**Engenharia Elétrica**	2%	16%
atuaposgraduacao	**N**	9%	89%
nucleopesquisa	**N**	6%	63%
projeto	**Nenhum**	3%	26%
orientacaomes	**N**	9%	96%
orientacaoespec	**Nenhuma**	9%	89%
orientacaoic	**Nenhuma**	8%	79%
orientacaograd	**Nenhuma**	7%	73%

Fonte: Autoria Própria

Tabela 31 - Perfil predominante da classe B para trabalhos completos

CLASSE B Trabalhos completos (51 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	6%	96%
sexo	**F**	3%	53%
idade	**30a39**	3%	49%
titulacaomax	**M**	4%	63%
anosformacao	**4oumais**	3%	47%
cursoandamento	**Nenhum**	4%	71%
escritaingles	**R**	2%	39%
campus	**GYN**	2%	33%
grandearea	**CET**	2%	31%
area	**Ciência da Computação**	1%	20%
atuaposgraduacao	**N**	5%	86%
nucleopesquisa	**N**	4%	69%
projeto	**Nenhum**	2%	37%
orientacaomes	**N**	6%	100%
orientacaoespec	**Nenhuma**	5%	88%
orientacaoic	**Nenhuma**	4%	73%
orientacaograd	**Nenhuma**	4%	63%

Fonte: Autoria Própria

Tabela 32 - Perfil predominante da classe C para trabalhos completos

CLASSE C Trabalhos completos (81 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	9%	93%
sexo	**M**	6%	63%
idade	**30a39**	4%	46%
titulacaomax	**M**	5%	57%
anosformacao	**4oumais**	6%	58%
cursoandamento	**Nenhum**	6%	64%
escritaingles	**R**	4%	41%
campus	**GYN**	3%	33%
grandearea	**CET**	3%	31%
area	**Ciência da Computação e História**	1%	10%
atuaposgraduacao	**N**	8%	86%
nucleopesquisa	**N**	7%	68%
projeto	**Nenhum**	3%	35%
orientacaomes	**N**	10%	99%
orientacaoespec	**Nenhuma**	9%	98%
orientacaoic	**Nenhuma**	7%	74%
orientacaograd	**Nenhuma**	7%	68%

Fonte: Autoria Própria

Tabela 33 - Perfil predominante da classe D para trabalhos completos

CLASSE D Trabalhos completos (626 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	69%	92%
sexo	**M**	47%	63%
idade	**30a39**	30%	40%
titulacaomax	**M**	44%	59%
anosformacao	**4oumais**	52%	69%
cursoandamento	**Nenhum**	54%	72%
escritaingles	**R**	28%	37%
campus	**GYN**	26%	34%
grandearea	**CET**	22%	29%
area	**Outras**	7%	9%
area	**Matemática**	6%	8%
atuaposgraduacao	**N**	70%	94%
nucleopesquisa	**N**	63%	84%
projeto	**Nenhum**	48%	65%
orientacaomes	**N**	74%	100%
orientacaoespec	**Nenhuma**	72%	96%
orientacaoic	**Nenhuma**	66%	89%
orientacaograd	**Nenhuma**	63%	84%

Fonte: Autoria Própria

Comparando os valores do suporte das regras das classes D com os das classes A, percebe-se que eles são maiores nas classes D, pois estas agregam a grande maioria do total da população de docentes.

As próximas cinco tabelas apresentam respectivamente o perfil predominante das classes A, B, C, D e E, também quanto à publicação de vários tipos de publicações, para as quais foram estabelecidas pontuações diferentes.

Tabela 34 - Perfil predominante da classe A para várias publicações

CLASSE A Várias Publicações (33 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	4%	94%
sexo	**F**	2%	52%
idade	**30a39**	2%	52%
titulacaomax	**M**	2%	61%
anosformacao	**1a3anos**	2%	48%
cursoandamento	**Nenhum**	2%	61%
escritaingles	**R**	2%	39%
campus	**GYN e INH**	1%	18%
grandearea	**CET**	1%	24%
area	**Química, Educação e Linguística**	1%	15%
atuaposgraduacao	**N**	3%	79%
nucleopesquisa	**N**	2%	55%
projeto	**Cinco_oumais**	2%	39%
orientacaomes	**N**	4%	94%
orientacaoespec	**Nenhuma**	3%	73%
orientacaoic	**Nenhuma**	2%	61%
orientacaograd	**Nenhuma**	2%	58%

Fonte: Autoria Própria

A predominância do câmpus de lotação na classe A para várias publicações ficou empatada entre os câmpus Goiânia e Inhumas, com 6 registros de cada.

As áreas de Química, Educação e Linguística também apareceram empatadas entre as regras, com ocorrência de 5 docentes de cada área na classe A.

Tabela 35 - Perfil predominante da classe B para várias publicações

CLASSE B Várias Publicações (50 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	5%	90%
sexo	**F**	3%	52%
idade	**30a39**	3%	50%
titulacaomax	**M**	3%	54%
anosformacao	**4oumais**	3%	56%
cursoandamento	**Nenhum**	4%	70%
escritaingles	**R**	3%	44%
campus	**GYN**	2%	30%
grandearea	**CET**	1%	24%
area	**Outras**	1%	16%
area	**Química, Ciência da Comp. E Eng. Elétrica**	0,47%	8%
atuaposgraduacao	**N**	5%	82%
nucleopesquisa	**N**	4%	64%
projeto	**Cinco_oumais**	2%	26%
orientacaomes	**N**	6%	96%
orientacaoespec	**Nenhuma**	5%	88%
orientacaoic	**Nenhuma**	4%	70%
orientacaograd	**Nenhuma**	4%	64%

Fonte: Autoria Própria

Desta vez, apareceram empatadas na classe B, as áreas de Química, Ciência da Computação e Engenharia Elétrica, com 4 docentes de cada área.

Tabela 36 - Perfil predominante da classe C para várias publicações

CLASSE C Várias Publicações (195 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	22%	96%
sexo	**M**	13%	54%
idade	**30a39**	12%	53%
titulacaomax	**M**	15%	63%
anosformacao	**4oumais**	12%	52%
cursoandamento	**Nenhum**	16%	68%
escritaingles	**R**	9%	41%
campus	**GYN**	8%	34%
grandearea	**CET**	7%	30%
area	**Química**	3%	11%
atuaposgraduacao	**N**	20%	86%
nucleopesquisa	**N**	16%	67%
projeto	**Nenhum**	7%	29%
orientacaomes	**N**	23%	99%
orientacaoespec	**Nenhuma**	21%	90%
orientacaoic	**Nenhuma**	17%	75%
orientacaograd	**Nenhuma**	17%	71%

Fonte: Autoria Própria

Tabela 37 - Perfil predominante da classe D para várias publicações

CLASSE D Várias Publicações (553 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**DE**	61%	93%
sexo	**M**	43%	66%
idade	**30a39**	24%	36%
titulacaomax	**M**	40%	60%
anosformacao	**4oumais**	47%	71%
cursoandamento	**Nenhum**	47%	72%
escritaingles	**R**	24%	37%
campus	**GYN**	23%	35%
grandearea	**CET**	19%	29%
area	**Ciência da Computação**	6%	9%
atuaposgraduacao	**N**	63%	96%
nucleopesquisa	**N**	57%	86%
projeto	**Nenhum**	47%	71%
orientacaomes	**N**	66%	100%
orientacaoespec	**Nenhuma**	65%	99%
orientacaoic	**Nenhuma**	60%	92%
orientacaograd	**Nenhuma**	56%	86%

Fonte: Autoria Própria

Tabela 38 - Perfil predominante da classe E para várias publicações

CLASSE E Várias Publicações (8 Docentes)			
Atributo	**Perfil**	**Suporte**	**Confiança**
regimetrabalho	**40H**	1%	100%
sexo	**M**	1%	75%
idade	**50oumais**	0,47%	50%
titulacaomax	**E**	1%	75%
anosformacao	**4oumais**	1%	88%
cursoandamento	**Nenhum**	1%	75%
escritaingles	**R**	0,35%	38%
campus	**GYN**	0,47%	50%
grandearea	**CET, ENG e CB**	0,11%	13%
area	**Física, Eng. Civil, Biologia Geral**	0,11%	13%
atuaposgraduacao	**N**	1%	100%
nucleopesquisa	**N**	1%	100%
projeto	**Nenhum**	1%	88%
orientacaomes	**N**	1%	100%
orientacaoespec	**Nenhuma**	1%	100%
orientacaoic	**Nenhuma**	1%	100%
orientacaograd	**Nenhuma**	1%	100%

Fonte: Autoria Própria

Dos 8 docentes da classe E para várias publicações, somente 3 preencheram a grande área e a área de conhecimento no currículo Lattes (CL). Portanto, houve empate na predominância das grandes áreas de Ciências Exatas e da Terra, Engenharias e Ciências Biológicas e de suas respectivas áreas de Física, Engenharia Civil e Biologia Geral, com 1 docente em cada.

Apesar da Tabela 34 apresentar o domínio predominante para cada atributo na classe A para várias publicações, vale fazer algumas considerações:

- 15% dos docentes da classe A estavam lotados no câmpus de Jataí, 15% no câmpus Aparecida de Goiânia e outros 15% no câmpus de Itumbiara;
- 21% dos docentes da classe A são da grande área de Ciências Humanas;
- 25% dos docentes da área de Linguística e 19% da área de Educação Física estão na classe A;
- 45% dos docentes da classe A participam de núcleos de pesquisa;
- a classe A é predominante de docentes que não orientaram mestrado. Porém, 40% dos que orientaram estão classificados na classe A;
- 25% dos docentes que orientaram duas ou mais especializações estão na classe A;
- 21% dos docentes da classe A orientaram três ou mais iniciação científica.

5.5 AVALIAÇÃO

A avaliação de um modelo de Mineração de Dados é uma etapa complexa e desafiadora, devido as possibilidades de observações que podem ser concatenadas. Pode envolver os fatores tempo, recursos financeiros, recursos humanos e qualidade dos dados para determinar qual resultado pode ser considerado qualificado e eficaz. O modelo é avaliado como efetivo se responde aos objetivos tratados na fase de entendimento do problema (LAROSE, 2006).

Para avaliar modelos quanto à qualidade e eficácia é essencial a utilização de métodos estatísticos. Com relação à tarefa de Classificação, a avaliação pode ser realizada através dos conceitos de taxa de acertos, taxa de erros, falsos positivos e falsos negativos que aparecem na matriz de confusão.

A ferramenta WEKA utiliza ainda o método estatístico *Kappa* para realizar a avaliação do modelo criado pelo algoritmo J48 na Classificação. A estatística *Kappa* apareceu nas figuras dos resultados dos experimentos de classificação deste trabalho ou nas tabelas com dados dos experimentos.

Apesar de alguns experimentos não terem apresentado valores altos para a estatística *Kappa*, os valores obtidos para a estatística não invalidam os resultados do estudo. Eles são úteis para demonstrar até que ponto as técnicas ou a combinação de atributos utilizadas fornecem um nível de concordância satisfatório ao pesquisador.

Quanto à tarefa de Associação, além das medidas do suporte e confiança das regras, outra medida de interesse objetiva chamada *lift* foi usada para averiguar a dependência entre os atributos presentes nas regras.

No intuito de validar a dependência entre os atributos que apareceram nos percursos das melhores folhas para as classes A nos experimentos de Classificação ECA5, ECT5 e EC1, foram geradas regras de Associação no WEKA, alterando o parâmetro do J48 chamado *metricType.* Este parâmetro do algoritmo tem como métrica padrão para avaliação das regras a “confiança”, que foi alterada para a métrica chamada *lift*. A Tabela 39 mostra os valores encontrados para o suporte, a confiança e o *lift* em cada regra.

Tabela 39 - Medidas de interesse objetivas para regras de Associação em ECA5, ECT5 e EC1

Exp.	Folha	Regra	*Sup.*	*Conf.*	*Lift*
ECA5	1	cursoandamento=D orientacaoespec=Duas_oumais ==> classe_artigo=A	0,36%	60%	9,32
ECA5	2	cursoandamento=Nenhum orientacaomes=S orientacaoespec=Duas_oumais ==> classe_artigo=A	0,24%	100%	15,54
ECA5	3	titulacaomax=D atuaposgraduacao=S projeto=Quatro orientacaoespec=Nenhuma ==> classe_artigo=A	0,24%	100%	15,54
ECT5	1	campus=ITU projeto=Cinco_oumais ==> classe_trabalho=A	0,47%	100%	10,36
ECT5	2	titulacaomax=M anosformacao=0anos projeto=Um orientacaoic=Nenhuma ==> classe_trabalho=A	0,47%	100%	10,36
EC1	1	sexo=F titulacaomax=D projeto=Tres ==> classe_pontuacao=A	0,24%	67%	16,95
EC1	2	titulacaomax=M area=Educacao Fisica projeto=Tres ==> classe_pontuacao=A	0,24%	50%	12,71

Fonte: Autoria Própria

Todos os valores de *lift* apresentados na Tabela 39 são maiores que um. Isso comprova a **dependência positiva** entre os atributos que apareceram nos percursos das melhores folhas para as classes A nos experimentos ECA5, ECT5 e EC1.

5.6 UTILIZAÇÃO, IMPLANTAÇÃO OU DESENVOLVIMENTO

Através dos resultados das tarefas de Classificação e Associação foi possível notar algumas características importantes da produção científica do IFG:

- A maior parte da produtividade científica dos docentes do instituto concentra-se em produções bibliográficas. Apesar do IFG ser uma instituição tradicional de formação técnica, as produções técnicas são pouco representativas de modo geral. Segundo a equipe da Pró-Reitoria de Pesquisa e Pós-Graduação do IFG, muitas produções técnicas não são devidamente cadastradas no Lattes e/ou não são registradas nos órgãos competentes como deveriam;
- O percentual de docentes classificados nas classes mais produtivas foi muito baixo. Tal fato indica que o número de publicações consideradas neste trabalho entre os docentes do IFG foi pequeno no último triênio. A classe A de artigos representou 6,43% da população, a classe A de trabalhos completos representou 9,65%, e a classe A para várias publicações apenas 3,93% do total de docentes estudados;
- Alguns atributos que *a priori* eram considerados previsores para as classes de maior produtividade, nem sequer foram incluídos entre as principais regras. Os atributos regime de trabalho, proficiência de escrita em inglês, idade, grande área do CNPq e participação em núcleo de pesquisa não apareceram nos percursos das melhores folhas para a classe A nas árvores de decisão das

tarefas de Classificação, nem entre as melhores regras de Associação com maior grau de suporte e confiança;

- Os docentes que mais publicaram artigos no triênio foram aqueles que orientaram especialização e/ou mestrado ou que atuavam nos cursos de pós-graduação do instituto, a maioria doutores;
- O desenvolvimento de projetos, as orientações, o tempo de conclusão da titulação máxima, a atuação nos cursos de pós-graduação e a titulação máxima do docente se apresentaram como os principais atributos de influência na maior produtividade de artigos, trabalhos completos e de outros tipos de produções científicas;
- Os perfis predominantes encontrados através das regras de Associação para as classes A de artigos, trabalhos completos e várias publicações foram coincidentes para 12 dos 17 atributos previsores: regimetrabalho=DE, idade=30a39anos, titulacaomax=M, cursoandamento=Nenhum, escritaingles=R, campus=GYN (com destaque também para o câmpus de Inhumas nas regras para várias publicações), atuaposgraduacao=N, nucleopesquisa=N, orientacaomes=N, orientacaoespec=Nenhuma, orientacaoic=Nenhuma, orientacaograd=Nenhuma. A diferença entre os perfis surgiu em 5 dos atributos: sexo, anosformacao, grandearea, area e projeto. Os docentes classe A de artigos são predominantemente homens, com tempo de conclusão do mestrado de 1 a 3 anos, da grande área de Ciências Exatas e da Terra, da área de Outras e Química e tiveram 5 ou mais projetos desenvolvidos no triênio. Os docentes classe A de trabalhos completos são predominantemente homens, já possuem mais de 4 anos que concluíram o mestrado, da grande área de Engenharias, da área de Engenharia Elétrica e não desenvolveram nenhum projeto no período da pesquisa. Enquanto que os docentes classe A para várias publicações são em sua maioria mulheres, com tempo de conclusão do mestrado de 1 a 3 anos, da grande área de Ciências

Exatas e da Terra, distribuídos igualmente entre as áreas de Química, Educação e Linguística e desenvolveram 5 ou mais projetos no período;

- Os perfis identificados para a classe D nos experimentos para artigos e trabalhos completos foram praticamente o mesmo, com exceção da área do conhecimento. Este perfil, assim como o perfil da classe E para várias publicações, apresentam-se como os mais críticos, e devem ser observados pelos gestores do IFG, a fim de desenvolverem políticas de incentivo à produção científica focadas aos mesmos;
- Os perfis identificados para as classes menos produtivas de acordo com os critérios estabelecidos neste trabalho, destacaram os docentes que já se encontram na "zona de conforto", ou seja, aqueles com mais idade, maior tempo de conclusão da titulação máxima, que não desenvolveram projetos, não orientaram alunos no triênio, etc;
- A área de Química sobressaiu na classe A dos experimentos de Associação, tanto de artigos quanto para várias publicações. Para trabalhos completos, o destaque foi para a área de Engenharia Elétrica;
- Através dos resultados dos experimentos percebeu-se que o desenvolvimento de projetos influencia na produtividade científica do docente, pois geralmente todo projeto gera como artefato final algum tipo de publicação que pode ser cadastrada no currículo Lattes;
- A utilização conjunta de medidas de interesse objetivas de **suporte**, **confiança** e do ***lift*** nos experimentos de Associação diminui a chance da mineração de regras óbvias e irrelevantes e além disso, possibilita aos usuários a realização de análises alternativas sobre uma mesma regra, enriquecendo o poder de entendimento a respeito das associações. Porém, vale lembrar que também existem as medidas de interesse subjetivas na mineração de regras de associação. Uma regra costuma ser interessante subjetivamente quando é útil e surpreendente para um pesquisador/analista que a examina.

Enfim, várias constatações podem ser feitas através dos resultados apresentados neste trabalho e outras várias ainda podem ser descobertas realizando novos experimentos aproveitando-se do banco de dados criado.

6. CONCLUSÕES

Apesar de ter sido aplicado em uma área específica, a pesquisa científica do IFG, o trabalho demonstrou que o processo de *Knowledge Discovery in Database* pode ser um poderoso instrumento para a gestão das informações de Ciência, Tecnologia e Inovação nas instituições de ensino para apoio à tomada de decisão.

A geração de conhecimento em *Knowledge Discovery in Database* pode acontecer em todas as etapas e não somente na Mineração de Dados. No estudo preliminar realizado, as etapas iniciais de seleção, pré-processamento e transformação já forneceram informações relevantes.

A Plataforma Lattes armazena grande quantidade de informações passíveis de serem exploradas e capazes de revelar relações não explícitas. Uma vez devidamente atualizada, a Plataforma Lattes é uma enorme fonte de informação para a geração de conhecimento útil para a gestão das instituições de ensino.

Com relação às publicações, percebeu-se que elas se concentram nas produções bibliográficas, apesar do IFG ser tradicionalmente uma instituição de formação técnica. Ao atender à demanda da Pró-Reitoria de Pesquisa e Pós-Graduação (PROPPG) notou-se que tanto o número de artigos em periódicos, quanto o de trabalhos completos publicados no último triênio entre os docentes do IFG foi pequeno. Cabe à PROPPG tomar as iniciativas necessárias na tentativa de melhorar a produtividade científica e principalmente tecnológica da instituição.

As diversas regras de Associação que foram apresentadas mostraram que alguns dados que aparentemente não estão relacionados. Na realidade, possuem aspectos em comum, que podem ser explorados. Porém, há muitas outras descobertas que ainda podem ser feitas aproveitando-se o banco de dados criado.

Inúmeras regras de Associação foram geradas e grupos seletos de docentes foram descobertos. Por um lado, algumas das regras são previsíveis. Por outro lado, o

algoritmo exige o aprofundamento do conhecimento por parte do decisor, onde a sua sensibilidade e experiência são fundamentais. Este deve intervir no processo estipulando limites para o suporte e a confiança de acordo com o seu interesse.

Todo o processo de Descoberta de Conhecimento em Banco de Dados (KDD) deve contar com a presença de especialistas do negócio (decisor), com participação maior ou menor, dependendo da etapa. Sua presença é fundamental nas etapas de Mineração de Dados e Interpretação/Avaliação, onde os padrões obtidos devem ser avaliados buscando identificar conhecimento útil que possa ser incorporado à instituição.

A simples extração de padrões não acrescenta conhecimento à organização. Para tal, é necessário que os especialistas de negócio identifiquem, a partir dos resultados gerados, aqueles que são úteis e possuem valor agregado.

A partir dos diversos padrões de comportamento observados nas informações que foram apresentadas, decisões podem ser tomadas não somente a curto prazo, mas também a longo prazo, pois é possível prever de forma mais segura prováveis comportamentos futuros.

Diante dos resultados, pode-se perceber que é possível obter-se uma visão mais abrangente dos dados institucionais, pelo fato de ter sido disponibilizada uma grande quantidade de informações sobre a pesquisa científica da IFG. Portanto, é possível iniciar uma melhoria na gestão do conhecimento dessa instituição fazendo uso dessas informações, pois é exatamente essa a base da gestão do conhecimento: dados integrados, gerando informações analíticas e abrangentes.

6.1 DIFICULDADES ENCONTRADAS

Dentre as dificuldades encontradas durante o desenvolvimento deste trabalho algumas podem ser destacadas.

O desenvolvimento do *script* para importação dos dados dos currículos Lattes em formato XML, a construção do banco de dados e posterior organização das *views* concentrou as maiores dificuldades encontradas. Esta foi a etapa mais longa e trabalhosa, pois diante da enorme quantidade de tabelas e dados adquiridos era necessário selecionar e tratar somente aqueles que interessavam ao estudo.

Outro problema encontrado para realizar a análise dos dados foi a falta de padronização dos dados cadastrados. Muitos currículos são preenchidos de forma incorreta e/ou incompleta e nem todos os docentes atualizam seus currículos periodicamente.

6.2 SUGESTÕES DE TRABALHOS FUTUROS

Por fim, pode-se dizer que este trabalho foi apenas um passo para o desenvolvimento de um trabalho de mudança na gestão do conhecimento das atividades de CT&I do IFG. Algumas sugestões para trabalhos futuros são:

- estabelecimento de novos critérios de exploração dos dados pela Pró-Reitoria de Pesquisa e Pós-Graduação do IFG, gerando descoberta de novas informações e novo conhecimento;
- elaboração de normas para o preenchimento e atualização dos currículos Lattes das pessoas envolvidas com a pesquisa científica do instituto;
- criação de indicadores de Ciência Tecnologia e Inovação para o IFG, com o objetivo de auxiliar a elaboração de novas políticas de gestão;
- aplicação futura do *script* desenvolvido em currículos atualizados para comparação dos novos resultados com os resultados obtidos neste estudo;
- O *script* desenvolvido poderá ser incrementado, desenvolvendo-se um sistema de informação que mantenha e controle os dados da produção científica do IFG.

6.3 CONTRIBUIÇÕES

Este trabalho foi submetido em formato de artigo na Revista de Administração Pública (RAP) da Fundação Getúlio Vargas na data de 07 de Janeiro de 2014.

REFERÊNCIAS

ALMEIDA, F. S. **Otimização de Estruturas de Materiais Compósitos Laminados utilizando Algoritmos Genéticos**. Dissertação M. Sc, Programa de Pós-Graduação em Engenharia Civil, Universidade Federal do Rio Grande do Sul – UFRGS, Porto Alegre, RS, Brasil, 2006. Disponível em: <http://www.lume.ufrgs.br/bitstream/handle/10183/10615/000599796.pdf?sequence=1>. Acesso em: 08 maio. 2013.

AMO, S. **Técnicas de Mineração de Dados**. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, v. 24, 2004, Salvador, BA, Brasil. Disponível em: <http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf>. Acesso em: 11 fev. 2013.

AMORIM, T. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de base de dados**. Universidade Federal de Pernambuco, Recife, PE, Brasil. 2006. Disponível em: <http://www.cin.ufpe.br/~tg/2006-2/tmas.pdf>. Acesso em: 22 de dez. 2011.

ARAÚJO, R. F. **Suporte decisório inteligente no processo de compra e venda de Imóveis no contexto de entidades fechadas de previdência complementar: Estudo de caso da Fundação Ceres.** Dissertação de M.Sc, Programa de Pós-Graduação em Gestão do Conhecimento e da Tecnologia da Informação da Universidade Católica de Brasília, Brasília, DF, Brasil, 2007. Disponível em: <http://www.bdtd.ucb.br/tede/tde_arquivos/3/TDE-2008-05-27T133526Z-609/Publico/Texto%20Completol.pdf> Acesso em: 29 abr. 2013.

AZARIAS P.; MATOS, S. N.; SCANDELARI, L.. **Aplicação da mineração de dados para a Geração do Conhecimento: um experimento prático**. V Congresso Nacional de Excelência em Gestão. Niterói, RJ, Brasil, 2009. Disponível em: <www.excelenciaemgestao.org/Portals/2/documents/cneg5/anais/T8_0203_0548.pdf>. Acesso em: 09 jan. 2013.

BAKER, R. S. J.; CARVALHO, A. M. J. B.; ISOTANI, S. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. Revista Brasileira de Informática na Educação. Vol. 19, nº 2. 2011. Disponível em: <http://www.br-ie.org/pub/index.php/rbie/article/view/1301/1172> Acesso em: 26 jan. 2012.

BAKER, R. S. J.; YACEF, K. **The State of Educational Data Mining in 2009: A Review and Future Visions**. JEDM - Journal of Educational Data Mining, Vol. 1, Issue 1, P. 3 -17, 2009. Disponível em: <http://www.educationaldatamining.org/JEDM/images/articles/vol1/issue1/JEDMVol1Issue1_BakerYacef.pdf> Acesso em: 20 jul. 2013.

BASGALUPP, M. P. **LEGAL-Tree: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão**. Tese de D.Sc., ICMC/USP São Carlos, São Paulo, 2010. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-12052010-165344/pt-br.php> Acesso em: 14 out. 2013.

BORBA, J. A.; MURCIA, F. D. **Oportunidades para Pesquisa e Publicação em Contabilidade: Um Estudo Preliminar sobre as Revistas Acadêmicas de Língua Inglesa do Portal de Periódicos da CAPES**. Revista Brazilian Business Review(BBR), Vitória, vol. 3, n. 1, p. 88-103. 2006. Disponível em:

<http://redalyc.uaemex.mx/src/inicio/ArtPdfRed.jsp?iCve=123016269007> Acesso em: 20 mar. 2013.

BUENO, M. F.; VIANA, M. R. **Mineração de Dados: Aplicações, Eficiência e Usabilidade**. Anais do Congresso de Iniciação Científica do Inatel (Instituto Nacional de Telecomunicações) – INCITEL 2012. Santa Rita do Sapucaí, MG, Brasil, 2012. Disponível em: <http://www.inatel.br/ic/component/docman/doc_download/65-mineracao-de-dados-aplicacoes--eficiencia-e-usabilidade> Acesso em: 16 out. 2013.

CABENA, P.; HADJINIAN, P.; STADLER, R.; VERHEES, J.; ZANASI, A. **Discovering Data Mining**: **from Concept to Implementation**. Upper Saddle River, Prentice Hall PTR, New Jersey, 1998.

CARDOSO, O. N. P.; MACHADO, R. T. M. **Gestão do Conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras**. Revista de Administração Pública, vol. 42, nº 3, p. 495-528, 2008. Disponível em: <http://www.scielo.br/pdf/rap/v42n3/a04v42n3.pdf> Acesso em: 22 dez. 2011.

CARVALHO, R. B. **Aplicações de Softwares de Gestão do Conhecimento: Tipologia e Usos**. Dissertação de M. Sc, Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil, 2000. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/LHLS-69XQRP/mestrado___rodrigo_baroni_de_carvalho.pdf?sequence=1>. Acesso em: 18 mar. 2013.

CARVALHO, D. R. **Árvore de Decisão, Algoritmo Genético para tratar o problema de pequenos disjuntos em classificação de dados.** Tese de D. Sc., Ciências em Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005a. Disponível em: <http://www.ipardes.gov.br/webisis.docs/tese_deborah_carvalho.pdf> Acesso em: 29 abr. 2013.

CARVALHO, L. A. V. **Data Mining – A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração**. Editora Ciência Moderna Ltda, Rio de Janeiro, RJ, Brasil, 2005b.

CERVI, C. R.; GALANTE, R. OLIVEIRA, J. P. M. **Análise do Comportamento de Pesquisadores baseada em Dados de Produção Científica**. Universidade Federal do Rio Grande do Sul – UFRGS, 2009. Disponível em: <http://upf.br/~cervi/publications/seminario_de_andamento_2009.pdf> Acesso em: 13 maio 2013.

CHAPMAN, P.; CLINTON, J.; KERBER, R., *et al*. **CRISP-DM 1.0**, **Step-by-step data mining guide. Cross Industry Standard Process for Data Mining**, 2000. Disponível em:<ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/serManual/CRISP-DM.pdf>. Acesso em: 09 maio 2013.

CNPq. **Plataforma Lattes**. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 14 maio 2013a.

CNPq. **Conselho Nacional de Desenvolvimento Científico e Tecnológico**. Disponível em: <http:// http://cnpq.br/>. Acesso em: 14 maio 2013b.

CNPq. **Currículo Lattes 2.0**. Disponível em: <http://www.cnpq.br/documents/313759/4d62720f-12ef-4ef2-b94c-e996b472834b>. Acesso em: 17 maio 2013c.

CÔRTES, S. C.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de Dados - Funcionalidades, Técnicas e Abordagens**. PUC-RIO. 2002. Disponível em: <ftp://ftp.inf.puc-rio.br/pub/docs/techreports/02_10_cortes.pdf> Acesso em: 06 maio 2013.

CRUZ, C. H. B. **Ciência, Tecnologia e Inovação no Brasil: desafios para o período 2011 a 2015.** Revista Interesse Nacional, 2010. Disponível em: <http://www.ifi.unicamp.br/~brito/artigos/CTI-desafios-InteresseNacional-07082010-FINAL.pdf> Acesso em: 26 set. 2013.

DAMASCENO, M. **Introdução a Mineração de Dados utilizando o WEKA**. Instituto Federal do Rio Grande do Norte – IFRN, V CONNEPI (Congresso Norte-Nordeste de Pesquisa e Inovação), Macau, RN, Brasil, 2010. Disponível em: <http://connepi.ifal.edu.br/ocs/anais/conteudo/anais/files/conferences/1/schedConfs/1/papers/258/public/258-4653-1-PB.pdf> Acesso em: 01 maio 2013.

DEL-FIACO, R. C. **Aplicação da Mineração de Dados na Descoberta de Padrões do Perfil de alunos do curso de SI-UNUCET-UEG**. Dissertação de M. Sc, Mestrado em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás. Goiânia, GO, Brasil, 2012.

DIAS, M. M. **Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados.** Tese de D. Sc, Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil, 2001. Disponível em: <http://www.din.uem.br/~mmdias/documentos/tese_Madalena.pdf> Acesso em: 07 de Maio de 2013.

DOMINGUES, M. A. **Generalização de Regras de Associação**. Dissertação de M.Sc, Instituto de Ciências Matemáticas e de Computação, Universidade São Paulo-USP, São Carlos, SP, Brasil, 2004. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-10082004-154242/publico/dissertacao.pdf> Acesso em: 30 abr. 2013.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in Knowledge Discovery and Data Mining**. California: AAAI Press/The MIT Press, 1996.

GUARDA, Á. **Aprendizado de Máquina: Árvore de Decisão Indutiva**. Centro de Informática, Universidade Federal de Pernambuco – UFPE, Recife, PE, Brasil, 2009. Disponível em: <http://www.cin.ufpe.br/~pacm/SI/ArvoreDecisaoIndutiva.pdf> Acesso em: 25 nov. 2013.

GUIMARÃES, P. R. B. **Métodos Quantitativos Estatísticos**. Curitiba: IESDE Brasil S.A., 2008. 245 p. Disponível em: <http://people.ufpr.br/~prbg/public_html/ce003/LIVRO3.pdf> Acesso em: 01 nov. 2013.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: um guia prático**. Rio de Janeiro: Elsevier, 2005.

GONÇALVES, E. C. **Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas**. Instituto de Computação da Universidade Federal Fluminense. Niterói, 2005. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v4.1/art04.pdf> Acesso em: 16 jan. 2014.

GONÇALVES, E. C. **Data Mining com a Ferramenta WEKA**. III Fórum de Software Livre de Duque de Caxias. Escola Nacional de Ciências Estatísticas (ENCE). Rio de Janeiro. 2011. Disponível em: <http://pt.scribd.com/doc/82645623/31/WEKA-em-Acao-%E2%80%93-Mineracao-de-um-Classificador> Acesso em: 19 jun. 2013.

HARRISON, T.H. **Intranet Data Warehouse: Ferramentas e Técnicas para a utilização do Data Warehouse na Intranet.** 358 p., Editora Berkeley,1998.

IEDMS. **International Educational Data Mining Society**. Disponível em: <http://www.educationaldatamining.org>. Acesso em: 20 jul. 2013.

IFG. **Instituto Federal de Educação, Ciência e Tecnologia de Goiás**. Disponível em: <http://www.cefetgo.br/index.php/instituicao>. Acesso em: 17 maio 2013a.

IFG. **Instituto Federal de Educação, Ciência e Tecnologia de Goiás**. Edital Nº 005/2013-PROPPG, de 04 de Março de 2013. Programa Institucional de Bolsas de Iniciação Científica – PIBIC. Disponível em: <http://www.ifg.edu.br/dppg/images/Editais/edital_005_2013-proppg-ifg_pibic_pibic-af.pdf>. Acesso em: 21 maio 2013b.

IFG. **Instituto Federal de Educação, Ciência e Tecnologia de Goiás**. Edital Nº 008/2013-PROPPG. Programa de Apoio à Produtividade em Pesquisa – ProAPP/IFG, de 06 de Março de 2013. Disponível em: <http://www.ifg.edu.br/dppg/images/Editais/edital_008_2013_proppg-ifg_ProAPP.pdf>. Acesso em: 21 maio 2013c.

IFG. **Instituto Federal de Educação, Ciência e Tecnologia de Goiás**. Edital Nº 001/2013-PROPPG. Programa Institucional de incentivo à participação em eventos científicos e tecnológicos para servidores do Instituto Federal de Educação, Ciência e Tecnologia de Goiás (PIPECT/IFG), de 22 de Janeiro de 2013. Disponível em: < http://www.ifg.edu.br/dppg/images/Editais/edital_001-2013-proppg-ifg_pipect.pdf>. Acesso em: 26 maio 2013d.

INMON, W. H.; HACKATHORN, R. D. **Como usar o Data Warehouse**. Rio de Janeiro, Editora Infobook, 1997.

ISBN**.** ***International Standard Book Number***. Disponível em: <http://www.isbn.bn.br/>. Acesso em: 06 jul. 2013.

IYODA, E. M. **Inteligência Computacional no projeto automático de Redes Neurais Híbridas e Redes NeuroFuzzy heterogêneas.** Dissertação de M. Sc, Faculdade de Engenharia Elétrica e da Computação, UNICAMP, Campinas-SP, 2000. Disponível em: <ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/theses/emi_mest/capa.pdf> Acesso em: 07 maio 2013.

JAIN, S.; AALAM, A.; DOJA, M.N. **K-Means Clustering using WEKA Interface**. Proceedings of the 4th National Conference. Jamia Hamdard University. New Delhi, India, 2010. Disponível em: <http://www.bvicam.ac.in/news/INDIACom%202010%20Proceedings/papers/Group3/INDIACom10_388_Paper.pdf> Acesso em: 26 set. 2013.

KAMPFF, A. J. C. **Mineração de Dados Educacionais para Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à prática docente**. Tese de D.

Sc., Informática na Educação. Universidade Federal do Rio Grande do Sul. Porto Alegre, RJ, Brasil, 2009. Disponível em: < http://www.lume.ufrgs.br/bitstream/handle/10183/19032/000734349.pdf?sequence=1> Acesso em: 26 jan. 2012.

LAROSE, D. T. **Discovering Knowledge in Data. An Introducing to DATA MNING,** New Jersey, John Wiley & Sons, 2006. Disponível em: <http://books.google.com.br/books?id=JbPMdPWQIOwC&printsec=frontcover&hl=pt-BR#v=onepage&q=medicine&f=false> Acesso em: 08 jan. 2014.

LEITE FILHO, G. A. **Perfil da produção científica dos docentes e programas de pós-graduação em ciências contábeis no Brasil**. Revista de Contabilidade e Controladoria, ISSN 1984-6266. Universidade Federal do Paraná, v.2, n.2, p. 1-13, Curitiba, PR, Brasil, 2010. Disponível em: <http://ojs.c3sl.ufpr.br/ojs2/index.php/rcc/article/viewFile/19370/13279> Acesso em: 20 mar. 2013.

LEMOS, E. P. **Análise de crédito bancário com o uso de Data Mining: Redes Neurais e Árvores de Decisão**. Dissertação de M. Sc, Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, Curitiba, PR, Brasil, 2003. Disponível em: <http://www.ppgmne.ufpr.br/arquivos/diss/70.pdf> Acesso em: 30 abr. 2013.

LINDEN, R. **Técnicas de Agrupamento.** Revista de Sistemas de Informação da FSMA (Faculdade Salesiana Maria Auxiliadora), nº 4, Macaé, RJ, Brasil, 2009. Disponível em: <http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf> Acesso em: 29 de set. 2013.

MARCONDES, C. H. **Representação e economia da Informação**. Ciência da Informação. V. 30, n.1, p. 61-70, Brasília, DF, Brasil, 2001. Disponível em: < http://www.scielo.br/pdf/ci/v30n1/a08v30n1.pdf> Acesso em: 09 jan. 2012.

MONTEIRO, M. J. F. **Extracção de Conhecimento de Dados I – Classificação**. Dissertação de M. Sc, Faculdade de Economia, Universidade do Porto, Porto, Portugal, 2005. Disponível em: < http://manuelmonteiro.eu/conteudos/docs/ECD1_MM_JG_Trab2.pdf> Acesso em: 26 nov. 2013.

MORAIS, B. C. S. **Extração de Conhecimento da Plataforma Lattes utilizando Técnicas de Mineração de Dados: Estudo de Caso POLI/UPE**. Trabalho de Conclusão de Curso em Engenharia de Computação - Universidade de Pernambuco, Recife, PE, Brasil, 2010. Disponível em: <http://tcc.ecomp.poli.br/20102/Monografia%20VF%20-%20Bruno%20Carlos.pdf> Acesso em: 20 mar. 2013.

MORAIS, A. C.; MACHADO, R. N. **Abordagem Métrica da Produção Científica dos Docentes do ICI/UFBA**. XIV Encontro Regional de Estudantes de Biblioteconomia, Documentação, Ciência da Informação e Gestão da informação. Universidade Federal do Maranhão, São Luís, MA, Brasil, 2011. Disponível em: < http://rabci.org/rabci/sites/default/files/ABORDAGEM%20M%C3%89TRICA%20DA%20PRODU%C3%87%C3%83O%20CIENT%C3%8DFICA%20DOS%20DOCENTES%20DO%20ICIUFBA%20%282006-2009%29.pdf> Acesso em: 20 mar. 2013.

MOTA, L. M.; CARDOSO, E. A.; SANTOS, L. S. **Uma imagem atual da atividade de pesquisa na Rede Federal de Educação Profissional Científica e Tecnológica**.

Instituto Federal da Bahia, Barbalho, BA, Brasil, 2010. Disponível em: <http://connepi.ifal.edu.br/ocs/index.php/connepi/CONNEPI2010/paper/viewFile/870/603>. Acesso em: 14 maio 2013.

NEGREIROS, A. V.; LIMA, F. C. **Data Mining**. Instituto Federal da Paraíba, João Pessoa, PB, Brasil, 2009. Disponível em: <http://flaviocorreialima.com/wp-content/uploads/2009/11/Data_Mining.pdf> Acesso em: 01 maio 2013.

NONAKA, I.; TAKEUCHI, H.. **Criação do conhecimento na empresa.** Rio de Janeiro: Elsevier, 1997.

PAULA, M. V. **Explorando o Potencial da Plataforma Lattes como fonte de conhecimento organizacional em Ciência e Tecnologia**. Dissertação de M. Sc, Programa de Pós-Graduação em Gestão do Conhecimento e da Tecnologia da Informação, Universidade Católica de Brasília, Brasília, DF, Brasil, 2004. Disponível em: <http://www.bdtd.ucb.br/tede/tde_arquivos/3/TDE-2004-12-03T093632Z-156/Publico/Dissertacao%20Marcelo.pdf>. Acesso em: 30 abr. 2013.

PASTA, A. **Aplicação da Técnica de Data Mining na base de dados do ambiente de Gestão Educacional: Um estudo de caso de uma Instituição de Ensino Superior de Blumenau-SC.** Dissertação M. Sc, Mestrado Acadêmico em Computação Aplicada, Universidade do Vale do Itajaí – UNIVALI. São José, SC, Brasil, 2011. Disponível em: <http://www6.univali.br/tede/tde_arquivos/10/TDE-2011-07-04T161405Z-766/Publico/Arquelau%20Pasta.pdf> Acesso em: 07 maio 2013.

PIZZIRANI, F. **Otimização Topológica de Estruturas utilizando algoritmos Genéticos.** Dissertação M. Sc, Faculdade de Engenharia Mecânica, UNICAMP, Campinas, SP, Brasil, 2003. Disponível em: <http://www.bibliotecadigital.unicamp.br/document/?down=vtls000295246> Acesso em: 08 maio 2013.

POSTGRESQLBRASIL. **Comunidade Brasileira de PostgreSQL.** Disponível em: <http://www.postgresql.org.br>. Acesso em: 07 jul. 2013.

PRASS, F. S. **Estudo comparativo entre algoritmos de análise de agrupamentos em Data Mining**. Dissertação M. Sc, Mestrado em Ciência da Computação, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil, 2004. Disponível em: <https://repositorio.ufsc.br/xmlui/bitstream/handle/123456789/87267/210022.pdf?sequence=1> Acesso em: 04 out. 2013.

PRETO, D.; SILVEIRA, S. R. **Um Estudo de Caso da Aplicação de Mineração de Dados em uma Instituição de Ensino Superior**. Centro Universitário Ritter dos Reis. Porto Alegre, RS, Brasil, 2010. Disponível em: <http://ensino.univates.br/~cetec/wet/anais2010/C08-wet2010.pdf> Acesso em: 26 jan. 2012.

RABELO, E. **Avaliação de Técnicas de Visualização para Mineração de Dados.** Dissertação de M. Sc, Programa de Pós-Graduação em Ciência da Computação, Universidade Estadual de Maringá, Maringá, PR, Brasil, 2007. Disponível em: <http://www.din.uem.br/~mestrado/diss/2007/rabelo.pdf> Acesso em: 07 maio 2013.

REZENDE, S. O. **Sistemas Inteligentes: Fundamentos e Aplicações**. Editora Manole Ltda, Barueri, SP, Brasil, 2005.

RIBEIRO, N. M.; SANTOS, L. L.; MOTA, L. M. **Indicadores de produção científica da Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT): Uma análise comparativa entre o IFBA e algumas Instituições desta rede.** V CONNEPI – Congresso Norte-Nordeste de Pesquisa e Inovação. Maceió, AL, Brasil, 2010. Disponível em: <http://connepi.ifal.edu.br/ocs/index.php/connepi/CONNEPI2010/paper/viewFile/1391/616> Acesso em: 27 set. 2013.

ROMÃO, W.; NIEDERAUER, C.; MARTINS, A.; TCHOLAKIAN, A.; PACHECO, R.; BARCIA, R. **Extração de regras de associação em C&T: O algoritmo Apriori**. In: ENEGEP, UFSC, Florianópolis, SC, Brasil, 1999. Disponível em: <http://www.abepro.org.br/biblioteca/ENEGEP1999_A0901.PDF> Acesso em: 08 out. 2013.

ROMÃO, W. **Descoberta de Conhecimento relevante em Banco de Dados sobre Ciência e Tecnologia**. Tese de D. Sc., Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina – UFSC, Florianópolis, SC, Brasil, 2002. Disponível em: <http://www.din.uem.br/~intersul/intersul_arquivos/documentos/Tese%20Wesley.pdf> Acesso em: 30 abr. 2013.

SANTANA, G. A.; SOBRAL, N. V.; FERREIRA, M. H. W.; SILVA, F. M. **Indicadores Científicos: Uma Análise da Produção do Programa de Pós-Graduação em Sociologia (PPGS) da UFPE a partir dos currículos da Plataforma Lattes (PL)**. XIV Encontro Regional de Estudantes de Biblioteconomia, Documentação, Ciência da Informação e Gestão da informação. Universidade Federal do Maranhão, São Luís, MA, Brasil, 2011. Disponível em: <http://www.academia.edu/1025364/INDICADORES_CIENTIFICOS_Uma_Analise_da_Producao_do_Programa_de_Pos-Graduacao_em_Sociologia_PPGS_da_UFPE_a_partir_dos_curriculos_da_Plataforma_Lattes_PL_> Acesso em: 20 mar. 2013.

SANTOS, L. D. M.; MIKAMI, R.; VENDRAMIN, A. C.; KAESTNER C. A. **Procedimentos de Validação Cruzada em Mineração de Dados para ambiente de Computação Paralela.** ERAD 2009. Caxias do Sul, RS, Brasil, 2009. Disponível em: < http://www.lbd.dcc.ufmg.br/colecoes/erad/2009/047.pdf > Acesso em: 27 out. 2013.

SEIDEL, E. J.; MOREIRA JÚNIOR, F. J.; ANSUJ, A. P.; NOAL, M. R. C. **Comparação entre o método Ward e o método K-médias no agrupamento de produtores de leite**. Universidade Federal de Santa Maria. Ciência e Natura – Revista do Centro de Ciências Naturais e Exatas, Vol. 30, nº 1, Santa Maria, RS, Brasil, 2008. Disponível em: <http://cascavel.ufsm.br/revistas/ojs-2.2.2/index.php/cienciaenatura/article/view/9737/5830> Acesso em: 29 set. 2013.

SIAPEnet. **Sistema Integrado de Administração de Recursos Humanos**. Disponível em: <http://www.siapenet.gov.br/oque.htm> Acesso em: 28 set. 2013.

SILVA, P. H. O. **Sistema hipermídia adaptativo baseado no descobrimento e análise de padrões de uso utilizando Mineração da Web.** Dissertação M. Sc, Mestrado em Modelagem Matemática e Computacional, CEFET-MG, Belo Horizonte, MG, Brasil, 2010. Disponivel em: < http://www.files.scire.net.br/atrio/cefet-mg-ppgmmc_upl/THESIS/18/pedrohenriqueoliveiraesilva.pdf> Acesso em: 07 out. 2013.

SOARES, F. **WebServices - EJB 3.0.** Especialização em Desenvolvimento de Aplicações Web com Interfaces Ricas, Instituto de Informática da Universidade Federal de Goiás, Goiânia, GO, Brasil, 2012. Disponível em: <http://www.inf.ufg.br/~fabrizzio/web/ejb/aula13.pdf> Acesso em: 08 jul. 2013.

SOUZA FILHO, M. C. **Classificação Automática de Gêneros de Áudio Digital.** Trabalho de Conclusão de Curso de Engenharia da Computação, Escola Politécnica de Pernambuco, Recife, PE, Brasil, 2006. Disponível em: <http://tcc.ecomp.poli.br/20062/MoacirFilho.pdf> Acesso em: 21 jun. 2013.

SUMATHI, S; SIVANANDAM, S.N. **Introduction to Data Mining and its Applications**, Berlim, Spinger, 2006. Disponível em: <http://books.google.com.br/books?hl=pt-BR&id=QwveL0jFk_gC&q=genetic+algorithms#v=onepage&q=evolutionary%20process&f=false >Acesso em: 08 de maio 2013.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introdução ao DATA MINING Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.

TARAPANOFF, K. **Inteligência Organizacional e Competitiva**. Editora Universidade de Brasília, Brasília, DF, Brasil, 2001. 344 p.

TERRA, J. C. C.; ALMEIDA, C. **Gestão do Conhecimento e Inteligência Competitiva: duas faces da mesma moeda.** Terra Fórum Consultores. 2003. Disponível em: < http://biblioteca.terraforum.com.br/BibliotecaArtigo/Duas%20faces%20da%20mesma%20moeda.pdf> Acesso em: 09 jan. 2012.

THOMPSON. J. R. **Estimation equations for kappa statistics - Statistics in Medicine**, Vol. 20, Edição 19, 2895 - 2906, Outubro 2001.

TURBAN, E.; SHARDA, R.; DELEN, D. **Decision Support and Business Intelligence Systems.** Pearson/Prentice Hall, 9th Edition, 696p. 2010.

VIEIRA, A. M.; ENSSLIN, S. R.; SILVA, H. A. S. **Perfil da produção científica dos docentes dos departamentos de contabilidade de três Universidades Federais do sul do Brasil**. DOI 10.4025/enfoque. vol. 30i3. 13255, p. 44-59. 2011. Disponível em: <http://periodicos.uem.br/ojs/index.php/Enfoque/article/view/13255/8327> Acesso em: 12 mar. 2013.

WEKA. **Data Mining Software in Java**, Disponível em: <http://www.cs.waikato.ac.nz/ml/WEKA/>. Acesso em: 30 set. 2013.

WITTEN, I. H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. Elsevier, Second Edition, 2005. Disponível em: <http://books.google.com/books?id=QTnOcZJzlUoC&printsec=frontcover&dq=data+mining&hl=pt-BR#v=onepage&q=&f=false>. Acesso em: 29 abr. 2013.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. Elsevier, Third Edition, 2011. Disponível em: <http://books.google.com.br/books?id=bDtLM8CODsQC&printsec=frontcover&dq=WITTEN,+Ian.+H.;+FRANK,+Eibe;+HALL,+Mark+A.+Data+Mining:+Practical+Machine+Learning+Tools+and+Techniques&hl=pt-BR&sa=X&ei=2ELDUcnLK9i54AOo2oCwBw&ved=0CDEQ6AEwAA>. Acesso em: 20 jun. 2013.

ANEXOS

ANEXO A – Formulário do ProAPP

INSTITUTO FEDERAL
GOIÁS

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇAO, CIENCIA E TECNOLOGIA DE GOIAS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE APOIO À PRODUTIVIDADE EM PESQUISA DO IFG (ProAPP/2013)
ANALISE TECNICA (PRODUTIVIDADE DO PESQUISADOR)

TÍTULO DO PROJETO:			
AUTOR:		**TÍTULO:**	**REGIME TRABALHO:**
PARTICIPOU PROAPP/2012/13?	**SUBMETEU ARTIGO PARA PUBLICAÇÃO?**	**QUANTOS? (ANEXAR COMPROVAÇÃO)**	

PRODUÇÃO CIENTÍFICA	PONTUAÇÃO
Artigo publicado em periódicos – Qualis A (10 pontos)	
Artigo publicado em periódicos – Quais B (08 pontos)	
Livro publicado, com ISBN (10 pontos)	
Capítulo de livro publicado, com ISBN (05 pontos)	
Organização, adaptação ou tradução publicada de livro ou cap. Livro (03 pontos)	
Trabalho completo publicado em anais de eventos (03 pontos)	
Resumo estendido publicado em anais de eventos (02 pontos)	
Resumo publicado em anais de eventos (01 ponto)	
Registro de obras artísticas (filmes, textos, partituras de música, fotos, outros) (02 pontos)	
Registro de plantas arquitetônicas/projetos (02 pontos)	
Registro de programas de computador (03 pontos)	
Patente concedida (10 pontos)	
Patente depositada (05 pontos)	
Registro de desenho industrial (06 pontos)	
Registro de cultivar (06 pontos)	
Registro de topografia de circuito integrado (06 pontos)	
Orientações concluídas de Doutorado (03 pontos)	
Orientações concluídas de Mestrado (02 pontos)	
Orientações concluídas de Iniciação Científica e Tecnológica e de Especialização no IFG (1,5 pontos)	
Orientações concluídas Monografia de Graduação no IFG (01 ponto)	
Organização de eventos científicos e tecnológicos (01 ponto)	
Obs: Considerar somente as produções dos últimos 3 (três) anos, com limites previstos no item 6.3.1 do Edital.	
SUB TOTAL:	
Titulação (Doutor = 15 pontos; Mestre = 10 pontos)	
Regime de DE (05 pontos)	
TOTAL:	

Fonte: IFG (2013c)

ANEXO B - Formulário do PIPECT

INSTITUTO FEDERAL GOIÁS	**Programa Institucional de Incentivo à Participação em Eventos Científicos e Tecnológicos para servidores do IFG (PIPECT/IFG)** **FORMULÁRIO PARA ANÁLISE DO CURRÍCULO DO SERVIDOR**	
Solicitante:	Processo Número	Demanda /
Câmpus:		
Link do Currículo Lattes:		

CRITÉRIOS DE ANÁLISE E JULGAMENTO DE MÉRITO E RELEVÂNCIA	*NOTA*
A. Produção Científica: (60 pontos no máximo) Atribuir pontos para produções declaradas como "Produção científica, tecnológica e artística/cultural" do currículo Lattes, de acordo com a seguinte tabela:	
- livro produzido na área de conhecimento do projeto apresentado (autor ou organizador) (4,0 pontos); - capítulo de livro (3,0 pontos); - artigo completo em periódico arbitrado internacional (3,5 pontos); - artigo completo em periódico arbitrado nacional (3,0 pontos); - projeto realizado em colaboração com outras instituições ou financiado por órgãos de fomento (2,5 pontos); - trabalho completo em anais de congressos (2,0 pontos); - resumo em anais de congressos (1,0 ponto); - artigo em jornais noticiosos ou revistas (0,5 ponto); - trabalho técnico (0,5 ponto).	
Obs: Considerar somente as produções dos últimos 5 anos, limitadas a 3 em cada categoria.	
C. Orientação: (30 pontos no máximo) Atribuir pontos para orientações, de acordo com a seguinte tabela:	
- trabalho de pós-graduação *stricto senso* (4,0 pontos); - trabalho de pós-graduação *lato senso* (3,0 pontos); - Trabalho de Conclusão de Curso de graduação (2,0 pontos); - programa institucional de Iniciação Científica/Tecnológica (2,5 pontos);	
Obs: Considerar somente as produções dos últimos 5 anos, limitadas a 3 em cada categoria.	
D. Regime de Trabalho (10 pontos para Dedicação Exclusiva; 05 pontos para 40 h)	
Nota Total (pontos)	

Obs: a) A pontuação final de cada projeto será dada pelo somatório das notas atribuídas aos três (3) itens acima

Observações:
Data:
Assinatura:

Fonte: IFG (2013d)

ANEXO C - Código do *script* PHP desenvolvido

- Arquivo **index.php**

```
<?php

//-----Insere bibliotecas usadas
require_once 'database.php';
require_once 'funcoes.php';

//-----Define variáveis
global $cv_id;
global $array_tabelas;
global $array_tabelas_id;
$lista_de_arquivos = new DirectoryIterator('xml');

//-----Lista de forma sequencial os arquivos xml da pasta e já insere a tag "curriculo_vitae" no banco
//-----E para cada XML acessado, roda a função "acessa_xml" que grava as outras TAGs no banco.
foreach ($lista_de_arquivos as $arquivo) {
    $nome_arquivo = $arquivo->getFilename();
    if($nome_arquivo != '.' && $nome_arquivo != '..'){
        $array_tabelas = array();
        $array_tabelas_id = array();
        $GLOBALS['cv_id'] = '"'.trim($nome_arquivo, '.xml').'"';
        $url = "xml/".$nome_arquivo;
        $root = simplexml_load_file($url);
        $tag_xml = $root->getName();
        $chave_pai_nome = "xml";
        $chave_pai_valor = (int)$GLOBALS["cv_id"];
        $sql_create_colunas = "";
        $tabela_colunas = array();
        $vetor_atributos = array();
        foreach ($root->attributes() as $rotulo=>$informacao) {
            $vetor_atributos[$rotulo] = substr($informacao,0, 999);
            array_push($tabela_colunas, strtolower(str_replace('-', '_', $rotulo)));
            $sql_create_colunas = $sql_create_colunas.", ".strtolower(str_replace('-', '_', $rotulo))."
varchar(1000)";
        }
        $tabela_nome = strtolower(str_replace('-', '_', $tag_xml));
        $tabela_nome = gerencia_tabela_db($tabela_nome, $tabela_colunas, $sql_create_colunas, "cv_id",
$chave_pai_nome);
        $id_insercao =
insert_banco($tabela_nome,$vetor_atributos,"cv_id",$GLOBALS["cv_id"],$chave_pai_nome
,$chave_pai_valor);
        list_node($root, 1, "curriculo_vitae_id", 0);
    }
}

function list_node($node, $nivel, $chave_pai_nome, $chave_pai_valor) {
    foreach ($node as $element) {
        $tag_xml = $element->getName();
        $sql_create_colunas = "";
        $tabela_colunas = array();
        $vetor_atributos = array();
        foreach ($element->attributes() as $rotulo=>$informacao) {
            $vetor_atributos[$rotulo] = substr($informacao,0, 999);
            array_push($tabela_colunas, strtolower(str_replace('-', '_', $rotulo)));
            $sql_create_colunas = $sql_create_colunas.", ".strtolower(str_replace('-', '_', $rotulo))."
varchar(1000)";
```

```
        }
        $tabela_nome = strtolower(str_replace('-', '_', $tag_xml));
        $tabela_nome = gerencia_tabela_db($tabela_nome, $tabela_colunas, $sql_create_colunas, "cv_id",
$chave_pai_nome);
        $id_insercao =
insert_banco($tabela_nome,$vetor_atributos,"cv_id",$GLOBALS["cv_id"],$chave_pai_nome
,$chave_pai_valor);
        if ($element->count() <> 0) {
            list_node($element, $nivel+1, $tabela_nome."_id", $id_insercao);
        }
    }
}
pg_close($dbcon);
?>
```

- Arquivo database.php

```
<?php

//----- Coneta a um banco de dados chamado "cliente" na máquina "localhost" com um usuário e senha

$con_string = "host=localhost port=5432 dbname=lattes user=xxxxxx password=xxxxxx";
if(!$dbcon = pg_connect($con_string)) die ("Erro ao conectar ao banco<br>".pg_last_error($dbcon));

?>
```

- Arquivo funcoes.php

```
<?php

//-----Mantém tabela no banco. Se ela não existir, será criada. Se faltar campos, será adicionado.-----//
function gerencia_tabela_db($tabela_nome, $tabela_colunas, $sql_create_colunas, $chave_geral_nome,
$chave_pai_nome){
    $array_tabelas_banco = verifica_tabelas_db();
    $existe_tabela = array_search($tabela_nome, $array_tabelas_banco);
    if ($existe_tabela){
        $colunas_atuais = verifica_colunas_tabelas($tabela_nome);
    }
    else{
        $colunas_atuais = array();
        $colunas_atuais[3] = "";
    }
    if (count($colunas_atuais)>=3){
        $nome_coluna_pai_inserida = $colunas_atuais[3];
    }else{
        $nome_coluna_pai_inserida = "-";
    }
    if ($existe_tabela and ($nome_coluna_pai_inserida == $chave_pai_nome || $chave_pai_nome ==
"root")){
        $colunas_diferentes = array_diff($tabela_colunas, $colunas_atuais);
        if($colunas_diferentes){
            $sql_ater_table = '';
            foreach ($colunas_diferentes as $coluna) {
                $sql_ater_table = "ALTER TABLE $tabela_nome ADD $coluna VARCHAR(1000);";
                $result = pg_query($sql_ater_table);
             }
        }
    }else{
        if ($existe_tabela){ /*
            $tabela_pai = str_replace('_id', '', $chave_pai_nome);
            $tabela_nome = $tabela_nome."_".$tabela_pai;
            if (strlen($tabela_nome)> 61){

                $tabela_nome = substr($tabela_nome,0, 61);
                $incremento_tabela = 1;
                $tabela_nome_incrementada = $tabela_nome."_".$incremento_tabela;
                $existe_tabela = array_search($tabela_nome_incrementada, $array_tabelas_banco);
```

```
            while ($existe_tabela){
               $incremento_tabela = $incremento_tabela + 1;
               $tabela_nome_incrementada = $tabela_nome."_".$incremento_tabela;
               $existe_tabela = array_search($tabela_nome, $array_tabelas_banco);
            }
            $tabela_nome = $tabela_nome_incrementada;
          }*/
      }
      $chave_pai_nome_trecho2 = ", $chave_pai_nome";
      $chave_pai_nome_trecho1 = ", $chave_pai_nome INTEGER";
      $sql_chaves = "id SERIAL NOT NULL , $chave_geral_nome VARCHAR(20) NOT NULL
$chave_pai_nome_trecho1";
      $sql_create_colunas = $sql_chaves.$sql_create_colunas;
      $sql = "CREATE TABLE $tabela_nome ($sql_create_colunas ,PRIMARY KEY (id,
$chave_geral_nome $chave_pai_nome_trecho2));";
      pg_query($sql) or FALSE;
   }
   return $tabela_nome;
}

//-----Verifica tabelas no banco de dados e monta um vetor com o resultado-----//

function verifica_tabelas_db(){
   $result = pg_query("SELECT tablename AS tabela FROM pg_catalog.pg_tables WHERE schemaname
NOT IN ('pg_catalog', 'information_schema', 'pg_toast') ORDER BY tablename");
   $tabelas = array();
   $cont = 1;
   while ($row = pg_fetch_row($result)) {
      $tabelas[$cont] = $row[0];
      $cont = $cont+1;
   }
   return $tabelas;
}

//-----Verifica os campos de uma tabela e monta um vetor com o resultado-----//

function verifica_colunas_tabelas($tabela_nome){
   $result = pg_query("select column_name from INFORMATION_SCHEMA.COLUMNS where
table_name = '$tabela_nome';");
   $colunas = array();
   $cont = 1;
   while ($row = pg_fetch_row($result)) {
      $colunas[$cont] = $row[0];
      $cont = $cont+1;
   }
   return $colunas;
}

function insert_banco($tabela, $vetor_atributos, $chave_geral_nome, $chave_geral_valor,
$chave_pai_nome , $chave_pai_valor){
      $colunas = $chave_geral_nome;
      $valores = $chave_geral_valor;
      $colunas = $colunas.", $chave_pai_nome";
      $valores = $valores.", $chave_pai_valor";
      foreach ($vetor_atributos as $coluna => $valor){
         $colunas = $colunas.",".$coluna;
         $valores = $valores.",'".pg_escape_string($valor)."'";
      }
      $colunas = strtolower(str_replace('-', '_', $colunas));
      $colunas = preg_replace('/^\,/', '', $colunas);
      $valores= preg_replace('/^\,/', '', $valores);
      $sql = "INSERT INTO $tabela ($colunas)
            VALUES ($valores) RETURNING id";
      $result = pg_query($sql) or FALSE;
      $obj = pg_fetch_object($result);
      $ultimo_id = $obj->id;
      return $ultimo_id;  }?>
```

ANEXO D - *Views* criadas no banco de dados

- *View* **v_apresentacaotrabalho**

```
CREATE OR REPLACE VIEW v_apresentacaotrabalho AS
 SELECT dados_basicos_da_apresentacao_de_trabalho.id,
dados_basicos_da_apresentacao_de_trabalho.cv_id, dados_basicos_da_apresentacao_de_trabalho.ano,
dados_basicos_da_apresentacao_de_trabalho.natureza
   FROM dados_basicos_da_apresentacao_de_trabalho
  WHERE dados_basicos_da_apresentacao_de_trabalho.ano::text >= '2011'::text
  ORDER BY dados_basicos_da_apresentacao_de_trabalho.cv_id;
```

- *View* **v_areadeatuacao**

```
CREATE OR REPLACE VIEW v_areadeatuacao AS
 SELECT aa.id, aa.cv_id, aa.sequencia_area_de_atuacao AS sequencia,
        CASE
            WHEN aa.nome_grande_area_do_conhecimento::text = 'ENGENHARIAS'::text THEN 'ENG'::text
            WHEN aa.nome_grande_area_do_conhecimento::text =
'CIENCIAS_EXATAS_E_DA_TERRA'::text THEN 'CET'::text
            WHEN aa.nome_grande_area_do_conhecimento::text = 'CIENCIAS_HUMANAS'::text THEN
'CH'::text
            WHEN aa.nome_grande_area_do_conhecimento::text = 'CIENCIAS_SOCIAIS_APLICADAS'::text
THEN 'CSA'::text
            WHEN aa.nome_grande_area_do_conhecimento::text = 'CIENCIAS_BIOLOGICAS'::text THEN
'CB'::text
            WHEN aa.nome_grande_area_do_conhecimento::text = 'LINGUISTICA_LETRAS_E_ARTES'::text
THEN 'LLA'::text
            WHEN aa.nome_grande_area_do_conhecimento::text = 'CIENCIAS_DA_SAUDE'::text THEN
'CS'::text
            WHEN aa.nome_grande_area_do_conhecimento::text = 'CIENCIAS_AGRARIAS'::text THEN
'CA'::text
            ELSE '?'::text
        END AS grandearea,
        CASE
            WHEN aa.nome_da_area_do_conhecimento::text = ANY (ARRAY['Saúde Coletiva'::character
varying::text, 'Parasitologia'::character varying::text, 'Engenharia Agrícola'::character varying::text,
'Zootecnia'::character varying::text, 'Ecologia'::character varying::text, 'Botânica'::character varying::text,
'Genética'::character varying::text, 'Bioquímica'::character varying::text, 'Imunologia'::character
varying::text, 'Microbiologia'::character varying::text, 'Farmacologia'::character varying::text,
'Biofísica'::character varying::text, 'Morfologia'::character varying::text, 'Zoologia'::character varying::text,
'Farmácia'::character varying::text, 'Fonoaudiologia'::character varying::text, 'Probabilidade e
Estatística'::character varying::text, 'Ciência Política'::character varying::text, 'Psicologia'::character
varying::text, 'Economia'::character varying::text, 'Engenharia de Minas'::character varying::text,
'Engenharia de Produção'::character varying::text, 'Engenharia de Materiais e Metalúrgica'::character
varying::text, 'Engenharia Química'::character varying::text, 'Ciências Ambientais'::character varying::text])
THEN 'Outras'::character varying
            WHEN aa.nome_da_area_do_conhecimento::text = ''::text THEN '?'::character varying
            ELSE sem_acentos(aa.nome_da_area_do_conhecimento)
        END AS area
   FROM area_de_atuacao aa
  WHERE aa.sequencia_area_de_atuacao::text = '1'::text
  ORDER BY aa.cv_id;
```

- *View* **v_artigo**

```
CREATE OR REPLACE VIEW v_artigo AS
```

```
 SELECT ap.id, ap.cv_id, dba.natureza, dba.ano_do_artigo, dda.issn
   FROM artigo_publicado ap
   JOIN dados_basicos_do_artigo dba ON ap.id = dba.artigo_publicado_id
   JOIN detalhamento_do_artigo dda ON dba.id = dda.artigo_publicado_id
  WHERE dba.ano_do_artigo::text >= '2011'::text AND dda.issn::text <> ''::text;
```

- *View* **v_capitulo**

```
CREATE OR REPLACE VIEW v_capitulo AS
 SELECT c.id, c.cv_id, dbc.ano, dc.isbn
   FROM capitulo_de_livro_publicado c
   JOIN dados_basicos_do_capitulo dbc ON c.id = dbc.capitulo_de_livro_publicado_id
   JOIN detalhamento_do_capitulo dc ON dbc.id = dc.capitulo_de_livro_publicado_id AND dbc.ano::text >=
'2011'::text AND dc.isbn::text <> ''::text;
```

- *View* **v_idioma**

```
CREATE OR REPLACE VIEW v_idioma AS
 SELECT idioma.id, idioma.cv_id, idioma.idioma, idioma.descricao_do_idioma,
        CASE
            WHEN idioma.proficiencia_de_escrita::text = 'BEM'::text THEN 'B'::text
            WHEN idioma.proficiencia_de_escrita::text = 'RAZOAVELMENTE'::text THEN 'R'::text
            WHEN idioma.proficiencia_de_escrita::text = 'POUCO'::text THEN 'P'::text
            ELSE NULL::text
        END AS escrita_ingles
   FROM idioma
  WHERE idioma.idioma::text = 'EN'::text AND idioma.proficiencia_de_escrita::text <>
'NAO_INFORMADO'::text
  ORDER BY idioma.cv_id; _id;
```

- *View* **v_livro**

```
CREATE OR REPLACE VIEW v_livro AS
 SELECT l.id, l.cv_id, dbl.ano, dl.isbn
   FROM livro_publicado_ou_organizado l
   JOIN dados_basicos_do_livro dbl ON l.id = dbl.livro_publicado_ou_organizado_id
   JOIN detalhamento_do_livro dl ON dl.id = dbl.livro_publicado_ou_organizado_id AND dbl.ano::text >=
'2011'::text AND dl.isbn::text <> ''::text;
```

- *View* **v_materialdidatico**

```
CREATE OR REPLACE VIEW v_materialdidatico AS
 SELECT dados_basicos_do_material_didatico_ou_instrucional.id,
dados_basicos_do_material_didatico_ou_instrucional.cv_id,
dados_basicos_do_material_didatico_ou_instrucional.ano,
dados_basicos_do_material_didatico_ou_instrucional.natureza
   FROM dados_basicos_do_material_didatico_ou_instrucional
  WHERE dados_basicos_do_material_didatico_ou_instrucional.ano::text >= '2011'::text
  ORDER BY dados_basicos_do_material_didatico_ou_instrucional.cv_id;
```

- *View* **v_orientacoes**

```
CREATE OR REPLACE VIEW v_orientacoes AS
        SELECT dados_basicos_de_outras_orientacoes_concluidas.id,
dados_basicos_de_outras_orientacoes_concluidas.cv_id,
dados_basicos_de_outras_orientacoes_concluidas.natureza,
dados_basicos_de_outras_orientacoes_concluidas.ano
          FROM dados_basicos_de_outras_orientacoes_concluidas
         WHERE (dados_basicos_de_outras_orientacoes_concluidas.natureza::text =
'TRABALHO_DE_CONCLUSAO_DE_CURSO_GRADUACAO'::text OR
dados_basicos_de_outras_orientacoes_concluidas.natureza::text = 'INICIACAO_CIENTIFICA'::text OR
dados_basicos_de_outras_orientacoes_concluidas.natureza::text =
```

```
'MONOGRAFIA_DE_CONCLUSAO_DE_CURSO_APERFEICOAMENTO_E_ESPECIALIZACAO'::text)
AND dados_basicos_de_outras_orientacoes_concluidas.ano::text >= '2011'::text
UNION
        SELECT dados_basicos_de_orientacoes_concluidas_para_mestrado.id,
dados_basicos_de_orientacoes_concluidas_para_mestrado.cv_id,
dados_basicos_de_orientacoes_concluidas_para_mestrado.natureza,
dados_basicos_de_orientacoes_concluidas_para_mestrado.ano
          FROM dados_basicos_de_orientacoes_concluidas_para_mestrado
         WHERE dados_basicos_de_orientacoes_concluidas_para_mestrado.ano::text >= '2011'::text;
```

- *View* **v_processosetecnicas**

```
CREATE OR REPLACE VIEW v_processosetecnicas AS
 SELECT dados_basicos_do_processos_ou_tecnicas.id,
dados_basicos_do_processos_ou_tecnicas.cv_id, dados_basicos_do_processos_ou_tecnicas.ano,
dados_basicos_do_processos_ou_tecnicas.natureza
   FROM dados_basicos_do_processos_ou_tecnicas
  WHERE dados_basicos_do_processos_ou_tecnicas.ano::text >= '2011'::text
  ORDER BY dados_basicos_do_processos_ou_tecnicas.cv_id;
```

- *View* **v_projetos**

```
CREATE OR REPLACE VIEW v_projetos AS
 SELECT projeto_de_pesquisa.id, projeto_de_pesquisa.cv_id, projeto_de_pesquisa.natureza,
projeto_de_pesquisa.ano_fim
   FROM projeto_de_pesquisa
  WHERE projeto_de_pesquisa.ano_fim::text >= '2011'::text OR projeto_de_pesquisa.ano_fim::text =
''::text
  ORDER BY projeto_de_pesquisa.cv_id;
```

- *View* **v_relatoriotecnico**

```
CREATE OR REPLACE VIEW v_relatoriotecnico AS
 SELECT rp.id, rp.cv_id, rp.ano
   FROM dados_basicos_do_relatorio_de_pesquisa rp
  WHERE rp.ano::text >= '2011'::text
  ORDER BY rp.cv_id;
```

- *View* **v_resumo**

```
CREATE OR REPLACE VIEW v_resumo AS
 SELECT t.id, t.cv_id, dbt.natureza, dbt.ano_do_trabalho
   FROM trabalho_em_eventos t
   JOIN dados_basicos_do_trabalho dbt ON t.id = dbt.trabalho_em_eventos_id
   JOIN detalhamento_do_trabalho dt ON dbt.id = dt.trabalho_em_eventos_id AND
dbt.ano_do_trabalho::text >= '2011'::text AND dbt.natureza::text = 'RESUMO'::text
  ORDER BY t.cv_id, dbt.natureza;
```

- *View* **v_resumoexpandido**

```
CREATE OR REPLACE VIEW v_resumoexpandido AS
 SELECT t.id, t.cv_id, dbt.natureza, dbt.ano_do_trabalho
   FROM trabalho_em_eventos t
   JOIN dados_basicos_do_trabalho dbt ON t.id = dbt.trabalho_em_eventos_id
   JOIN detalhamento_do_trabalho dt ON dbt.id = dt.trabalho_em_eventos_id AND
dbt.ano_do_trabalho::text >= '2011'::text AND dbt.natureza::text = 'RESUMO_EXPANDIDO'::text
  ORDER BY t.cv_id, dbt.natureza;
```

- *View* **v_software**

```
CREATE OR REPLACE VIEW v_software AS
```

```
 SELECT dados_basicos_do_software.id, dados_basicos_do_software.cv_id,
dados_basicos_do_software.ano, dados_basicos_do_software.natureza
   FROM dados_basicos_do_software
  WHERE dados_basicos_do_software.ano::text >= '2011'::text
  ORDER BY dados_basicos_do_software.cv_id;
```

- *View* **v_suap_docente**

```
CREATE OR REPLACE VIEW v_suap_docente AS
 SELECT suap_pessoa.id, suap_pessoa.nome, suap_cnpq_curriculovittaelattes.numero_identificador AS
cv_id, suap_servidor.matricula, suap_pessoa_fisica.cpf, suap_pessoa_fisica.sexo, date_part('year'::text,
age(suap_pessoa_fisica.nascimento_data::timestamp with time zone)) AS idade,
suap_pessoa_fisica.estado_civil_id,
        CASE
            WHEN suap_jornada_trabalho.nome::text = 'DEDICACAO EXCLUSIVA'::text THEN 'DE'::text
            WHEN suap_jornada_trabalho.nome::text = '40 HORAS SEMANAIS'::text THEN '40H'::text
            WHEN suap_jornada_trabalho.nome::text = '20 HORAS SEMANAIS'::text THEN '20H'::text
            ELSE NULL::text
        END AS regime_trabalho,
        CASE
            WHEN suap_unidadeorganizacional.nome::text = 'CAMPUS APARECIDA DE GOIÂNIA'::text
THEN 'APA'::text
            WHEN suap_unidadeorganizacional.nome::text = 'CAMPUS GOIÂNIA'::text THEN 'GYN'::text
            WHEN suap_unidadeorganizacional.nome::text = 'CAMPUS URUACU'::text THEN 'URU'::text
            WHEN suap_unidadeorganizacional.nome::text = 'CAMPUS JATAÍ'::text THEN 'JAT'::text
            WHEN suap_unidadeorganizacional.nome::text = 'REITORIA'::text THEN 'REI'::text
            WHEN suap_unidadeorganizacional.nome::text = 'CAMPUS ITUMBIARA'::text THEN 'ITU'::text
            WHEN suap_unidadeorganizacional.nome::text = 'CAMPUS FORMOSA'::text THEN 'FOR'::text
            WHEN suap_unidadeorganizacional.nome::text = 'CAMPUS LUZÂNIA'::text THEN 'LUZ'::text
            WHEN suap_unidadeorganizacional.nome::text = 'CAMPUS ANÁPOLIS'::text THEN 'ANA'::text
            WHEN suap_unidadeorganizacional.nome::text = 'CAMPUS GOIÁS'::text THEN 'GOI'::text
            WHEN suap_unidadeorganizacional.nome::text = 'CAMPUS INHUMAS'::text THEN 'INH'::text
            ELSE NULL::text
        END AS campus,
        CASE
            WHEN date_part('year'::text, age(suap_pessoa_fisica.nascimento_data::timestamp with time
zone)) >= 50::double precision THEN '50oumais'::text
            WHEN date_part('year'::text, age(suap_pessoa_fisica.nascimento_data::timestamp with time
zone)) >= 40::double precision AND date_part('year'::text,
age(suap_pessoa_fisica.nascimento_data::timestamp with time zone)) <= 49::double precision THEN
'40a49'::text
            WHEN date_part('year'::text, age(suap_pessoa_fisica.nascimento_data::timestamp with time
zone)) >= 30::double precision AND date_part('year'::text,
age(suap_pessoa_fisica.nascimento_data::timestamp with time zone)) <= 39::double precision THEN
'30a39'::text
            WHEN date_part('year'::text, age(suap_pessoa_fisica.nascimento_data::timestamp with time
zone)) >= 20::double precision AND date_part('year'::text,
age(suap_pessoa_fisica.nascimento_data::timestamp with time zone)) <= 29::double precision THEN
'20a29'::text
            ELSE NULL::text
        END AS faixaidade
   FROM suap_servidor
   JOIN suap_pessoa_fisica ON suap_pessoa_fisica.username::text = suap_servidor.matricula::text
   JOIN suap_setor ON suap_setor.id = suap_servidor.setor_exercicio_id
   JOIN suap_unidadeorganizacional ON suap_unidadeorganizacional.id = suap_setor.uo_id
   JOIN suap_situacao ON suap_situacao.id = suap_servidor.situacao_id
   JOIN suap_cargo_emprego ON suap_cargo_emprego.id = suap_servidor.cargo_emprego_id
   JOIN suap_jornada_trabalho ON suap_jornada_trabalho.id = suap_servidor.jornada_trabalho_id
   JOIN suap_grupo_cargo_emprego ON suap_grupo_cargo_emprego.id =
suap_cargo_emprego.grupo_cargo_emprego_id
   JOIN suap_pessoa ON suap_pessoa.id = suap_pessoa_fisica.pessoa_ptr_id
   JOIN suap_cnpq_curriculovittaelattes ON suap_cnpq_curriculovittaelattes.pessoa_fisica_id =
suap_pessoa.id
  WHERE suap_servidor.excluido = false AND suap_situacao.id = 1 AND
suap_grupo_cargo_emprego.categoria::text = 'docente'::text AND (suap_jornada_trabalho.id = ANY
(ARRAY[1, 3, 4])) AND suap_cnpq_curriculovittaelattes.numero_identificador <> ''::text
  ORDER BY suap_pessoa.nome;
```

- *View* **v_textojornalrevista**

```
CREATE OR REPLACE VIEW v_textojornalrevista AS
 SELECT t.id, t.cv_id, dbt.ano_do_texto
   FROM texto_em_jornal_ou_revista t
   JOIN dados_basicos_do_texto dbt ON t.id = dbt.texto_em_jornal_ou_revista_id
   JOIN detalhamento_do_texto dt ON dbt.id = dt.texto_em_jornal_ou_revista_id AND
dbt.ano_do_texto::text >= '2011'::text ORDER BY t.cv_id;
```

- *View* **v_titulacao**

```
CREATE OR REPLACE VIEW v_titulacao AS
 SELECT dados_gerais.cv_id, dados_gerais.nome_completo, titulacao.id, titulacao.nome_curso,
titulacao.nome_instituicao, titulacao.status_do_curso, titulacao.ano_de_inicio,
       CASE
          WHEN titulacao.ano_de_conclusao::text = ''::text THEN '3000'::character varying
          ELSE titulacao.ano_de_conclusao
       END AS ano_de_conclusao, titulacao.nivel, titulacao.id_nivel
   FROM dados_gerais
   LEFT JOIN (       (       (       (         SELECT pos_doutorado.id, pos_doutorado.cv_id,
pos_doutorado.nome_instituicao, '-' AS nome_curso, pos_doutorado.status_do_curso,
pos_doutorado.ano_de_inicio, pos_doutorado.ano_de_conclusao, 'POS-DOUTORADO' AS nivel, 5 AS
id_nivel
                                    FROM pos_doutorado
                          UNION
                                  SELECT doutorado.id, doutorado.cv_id, doutorado.nome_instituicao,
doutorado.nome_curso, doutorado.status_do_curso, doutorado.ano_de_inicio,
doutorado.ano_de_conclusao, 'DOUTORADO' AS nivel, 4 AS id_nivel
                                    FROM doutorado)
                    UNION
                            SELECT mestrado.id, mestrado.cv_id, mestrado.nome_instituicao,
mestrado.nome_curso, mestrado.status_do_curso, mestrado.ano_de_inicio, mestrado.ano_de_conclusao,
'MESTRADO' AS nivel, 3 AS id_nivel
                              FROM mestrado)
              UNION
                      SELECT especializacao.id, especializacao.cv_id, especializacao.nome_instituicao,
especializacao.nome_curso, especializacao.status_do_curso, especializacao.ano_de_inicio,
especializacao.ano_de_conclusao, 'ESPECIALIZACAO' AS nivel, 2 AS id_nivel
                        FROM especializacao)
       UNION
               SELECT graduacao.id, graduacao.cv_id, graduacao.nome_instituicao, graduacao.nome_curso,
graduacao.status_do_curso, graduacao.ano_de_inicio, graduacao.ano_de_conclusao, 'GRADUACAO' AS
nivel, 1 AS id_nivel
                 FROM graduacao) titulacao ON dados_gerais.cv_id::text = titulacao.cv_id::text
  WHERE titulacao.status_do_curso::text <> 'INCOMPLETO'::text
  ORDER BY dados_gerais.nome_completo, titulacao.id_nivel;
```

- *View* **v_titulacaomax** (Derivada de v_titulacao)

```
CREATE OR REPLACE VIEW v_titulacaomax AS
 SELECT DISTINCT dados_gerais.cv_id, dados_gerais.nome_completo AS nome,
       CASE
          WHEN (2013 - c.ano_de_conclusao::integer) = (-987) THEN 0
          ELSE 2013 - c.ano_de_conclusao::integer
       END AS anosformacaonum,
       CASE
          WHEN (2013 - c.ano_de_conclusao::integer) = (-987) THEN 'Cursando'::text
          WHEN (2013 - c.ano_de_conclusao::integer) = ANY (ARRAY[0]) THEN '0anos'::text
          WHEN (2013 - c.ano_de_conclusao::integer) = ANY (ARRAY[1, 2, 3]) THEN '1a3anos'::text
          WHEN (2013 - c.ano_de_conclusao::integer) >= 4 THEN '4oumais'::text
          ELSE NULL::text
       END AS anosformacao,
       CASE
          WHEN c.nivel = 'GRADUACAO'::text THEN 'G'::text
```

```
            WHEN c.nivel = 'ESPECIALIZACAO'::text THEN 'E'::text
            WHEN c.nivel = 'MESTRADO'::text THEN 'M'::text
            WHEN c.nivel = 'DOUTORADO'::text THEN 'D'::text
            WHEN c.nivel = 'POS-DOUTORADO'::text THEN 'P'::text
            ELSE '?'::text
        END AS titulacaomax,
        CASE
            WHEN vta.nivel = 'GRADUACAO'::text THEN 'G'::text
            WHEN vta.nivel = 'ESPECIALIZACAO'::text THEN 'E'::text
            WHEN vta.nivel = 'MESTRADO'::text THEN 'M'::text
            WHEN vta.nivel = 'DOUTORADO'::text THEN 'D'::text
            WHEN vta.nivel = 'POS-DOUTORADO'::text THEN 'P'::text
            ELSE 'Nenhum'::text
        END AS cursando
   FROM dados_gerais
   LEFT JOIN ( SELECT n.cv_id, n.nome_completo, n.nome_instituicao, n.nome_curso,
n.status_do_curso, n.ano_de_inicio, n.ano_de_conclusao, n.nivel, n.id_nivel
          FROM v_titulacao n
         WHERE n.id_nivel = (( SELECT max(v_titulacao.id_nivel) AS max
                  FROM v_titulacao
                 WHERE v_titulacao.status_do_curso::text = 'CONCLUIDO'::text
                 GROUP BY v_titulacao.cv_id
                HAVING v_titulacao.cv_id::text = n.cv_id::text))) c ON c.cv_id::text = dados_gerais.cv_id::text
   LEFT JOIN ( SELECT DISTINCT vt.cv_id, vt.nome_completo, vt.status_do_curso, vt.nivel, vt.id_nivel
      FROM v_titulacao vt
     WHERE vt.id_nivel = (( SELECT max(v_titulacao.id_nivel) AS max
              FROM v_titulacao
             GROUP BY v_titulacao.cv_id, v_titulacao.status_do_curso
            HAVING v_titulacao.cv_id::text = vt.cv_id::text AND v_titulacao.status_do_curso::text =
'EM_ANDAMENTO'::text
            LIMIT 1)) AND vt.status_do_curso::text = 'EM_ANDAMENTO'::text
     ORDER BY vt.cv_id) vta ON vta.cv_id::text = dados_gerais.cv_id::text
  ORDER BY dados_gerais.nome_completo;
```

- *View* **v_trabalhocompleto**

```
CREATE OR REPLACE VIEW v_trabalhocompleto AS
 SELECT t.id, t.cv_id, dbt.natureza, dbt.ano_do_trabalho
   FROM trabalho_em_eventos t
   JOIN dados_basicos_do_trabalho dbt ON t.id = dbt.trabalho_em_eventos_id
   JOIN detalhamento_do_trabalho dt ON dbt.id = dt.trabalho_em_eventos_id AND
dbt.ano_do_trabalho::text >= '2011'::text AND dbt.natureza::text = 'COMPLETO'::text
  ORDER BY t.cv_id, dbt.natureza;
```

- *View* **v_trabalhotecnico**

```
CREATE OR REPLACE VIEW v_trabalhotecnico AS
 SELECT t.id, t.cv_id, dbt.natureza, dbt.ano
   FROM trabalho_tecnico t
   JOIN dados_basicos_do_trabalho_tecnico dbt ON t.id = dbt.trabalho_tecnico_id
   JOIN detalhamento_do_trabalho_tecnico dt ON dt.id = dbt.trabalho_tecnico_id AND dbt.ano::text >=
'2011'::text
  ORDER BY t.cv_id;
```

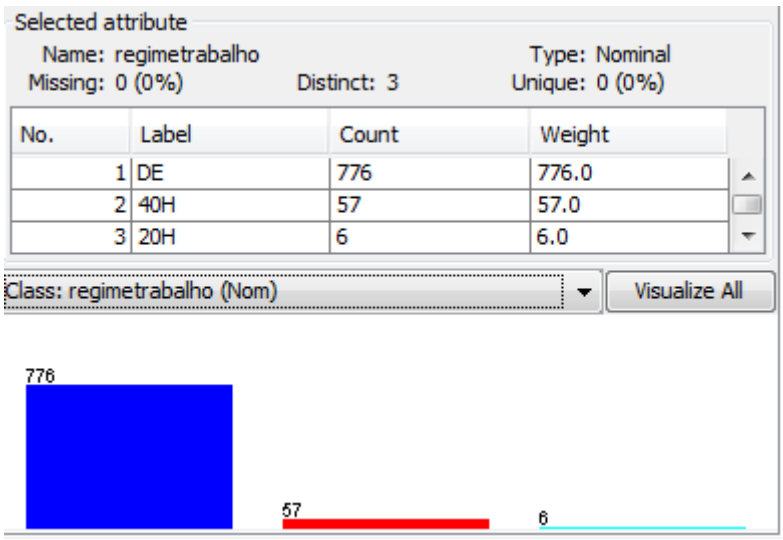
ANEXO E – *View* v_arff (retorna todos os dados do arquivo lattes.ARFF)

```
CREATE OR REPLACE VIEW v_arff AS
 SELECT v_suap_docente.nome, v_suap_docente.regime_trabalho, v_suap_docente.sexo,
v_suap_docente.faixaidade, v_titulacaomax.titulacaomax, v_titulacaomax.anosformacao,
v_titulacaomax.cursando,
        CASE
            WHEN v_idioma.escrita_ingles IS NULL THEN '?'::text
            ELSE v_idioma.escrita_ingles
        END AS escritaingles, v_suap_docente.campus,
        CASE
            WHEN v_areadeatuacao.grandearea IS NULL THEN '?'::text
            ELSE v_areadeatuacao.grandearea
        END AS grandearea,
        CASE
            WHEN v_areadeatuacao.area IS NULL THEN '?'::character varying
            ELSE v_areadeatuacao.area
        END AS area, dados_adicionais.atua_pos, dados_adicionais.nucleo_pesquisa,
        CASE
            WHEN projetos.qtdprojeto > 0 THEN projetos.qtdprojeto
            ELSE 0::bigint
        END AS qtdprojeto,
        CASE
            WHEN projetos.qtdprojeto IS NULL THEN 'Nenhum'::text
            WHEN projetos.qtdprojeto = 1 THEN 'Um'::text
            WHEN projetos.qtdprojeto = 2 THEN 'Dois'::text
            WHEN projetos.qtdprojeto = 3 THEN 'Tres'::text
            WHEN projetos.qtdprojeto = 4 THEN 'Quatro'::text
            WHEN projetos.qtdprojeto >= 5 THEN 'Cinco_oumais'::text
            ELSE projetos.qtdprojeto::text
        END AS projeto,
        CASE
            WHEN v_publicacoes.artigo > 0 THEN v_publicacoes.artigo
            ELSE 0::bigint
        END AS artigos,
        CASE
            WHEN v_publicacoes.trabalhocompleto > 0 THEN v_publicacoes.trabalhocompleto
            ELSE 0::bigint
        END AS trabalho_completo,
        CASE
            WHEN v_publicacoes.resumoexpandido > 0 THEN v_publicacoes.resumoexpandido
            ELSE 0::bigint
        END AS resumo_expandido,
        CASE
            WHEN v_publicacoes.resumo > 0 THEN v_publicacoes.resumo
            ELSE 0::bigint
        END AS resumo,
        CASE
            WHEN v_publicacoes.apresentacaotrabalho > 0 THEN v_publicacoes.apresentacaotrabalho
            ELSE 0::bigint
        END AS apresentacao_trabalho,
        CASE
            WHEN v_publicacoes.capitulo > 0 THEN v_publicacoes.capitulo
            ELSE 0::bigint
        END AS capitulo,
        CASE
            WHEN v_publicacoes.livro > 0 THEN v_publicacoes.livro
            ELSE 0::bigint
        END AS livro,
```

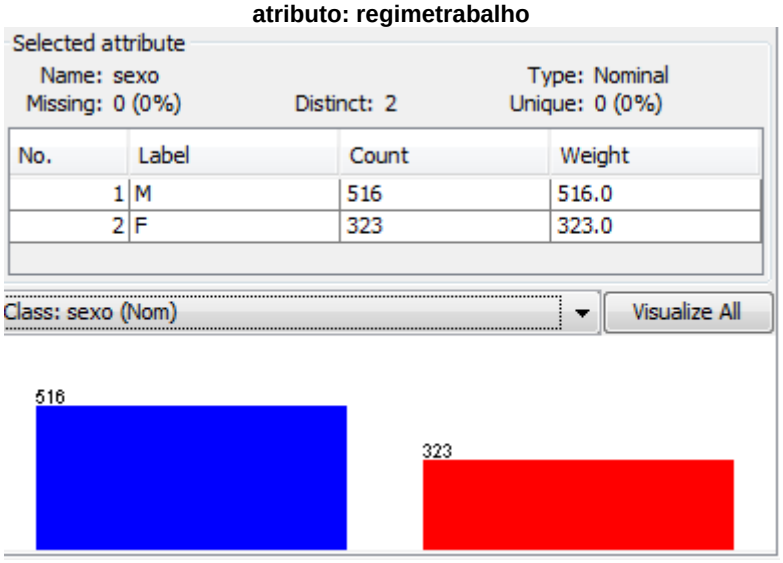
```
    CASE
      WHEN v_publicacoes.materialdidatico > 0 THEN v_publicacoes.materialdidatico
      ELSE 0::bigint
    END AS material_didatico,
    CASE
      WHEN v_publicacoes.processosetecnicas > 0 THEN v_publicacoes.processosetecnicas
      ELSE 0::bigint
    END AS processos_tecnicas,
    CASE
      WHEN v_publicacoes.relatoriotecnico > 0 THEN v_publicacoes.relatoriotecnico
      ELSE 0::bigint
    END AS relatorio_tecnico,
    CASE
      WHEN v_publicacoes.software > 0 THEN v_publicacoes.software
      ELSE 0::bigint
    END AS software,
    CASE
      WHEN v_publicacoes.textojornalrevista > 0 THEN v_publicacoes.textojornalrevista
      ELSE 0::bigint
    END AS texto_jornal_revista,
    CASE
      WHEN v_publicacoes.trabalhotecnico > 0 THEN v_publicacoes.trabalhotecnico
      ELSE 0::bigint
    END AS trabalho_tecnico,
    CASE
      WHEN orientacaomestrado.qtdorientacao > 0 THEN orientacaomestrado.qtdorientacao
      ELSE 0::bigint
    END AS orientacaomes_num,
    CASE
      WHEN orientacaomestrado.qtdorientacao > 0 THEN 'S'::text
      ELSE 'N'::text
    END AS orientacaomes,
    CASE
      WHEN orientacaoespec.qtdorientacao > 0 THEN orientacaoespec.qtdorientacao
      ELSE 0::bigint
    END AS orientacaoespec_num,
    CASE
      WHEN orientacaoespec.qtdorientacao IS NULL THEN 'Nenhuma'::text
      WHEN orientacaoespec.qtdorientacao = 1 THEN 'Uma'::text
      WHEN orientacaoespec.qtdorientacao >= 2 THEN 'Duas_oumais'::text
      ELSE orientacaoespec.qtdorientacao::text
    END AS orientacaoespec,
    CASE
      WHEN orientacaoic.qtdorientacao > 0 THEN orientacaoic.qtdorientacao
      ELSE 0::bigint
    END AS orientacaoic_num,
    CASE
      WHEN orientacaoic.qtdorientacao IS NULL THEN 'Nenhuma'::text
      WHEN orientacaoic.qtdorientacao = 1 THEN 'Uma'::text
      WHEN orientacaoic.qtdorientacao = 2 THEN 'Duas'::text
      WHEN orientacaoic.qtdorientacao >= 3 THEN 'Tres_oumais'::text
      ELSE orientacaoic.qtdorientacao::text
    END AS orientacaoic,
    CASE
      WHEN orientacaograduacao.qtdorientacao > 0 THEN orientacaograduacao.qtdorientacao
      ELSE 0::bigint
    END AS orientacaograd_num,
    CASE
      WHEN orientacaograduacao.qtdorientacao IS NULL THEN 'Nenhuma'::text
      WHEN orientacaograduacao.qtdorientacao = 1 THEN 'Uma'::text
      WHEN orientacaograduacao.qtdorientacao = 2 THEN 'Duas'::text
      WHEN orientacaograduacao.qtdorientacao = 3 THEN 'Tres'::text
      WHEN orientacaograduacao.qtdorientacao >= 4 THEN 'Quatro_oumais'::text
      ELSE orientacaograduacao.qtdorientacao::text
    END AS orientacaograd
  FROM v_suap_docente
  JOIN dados_adicionais ON v_suap_docente.cv_id = dados_adicionais.cv_id::text
```

```
   JOIN v_titulacaomax ON v_suap_docente.cv_id = v_titulacaomax.cv_id::text
   LEFT JOIN v_areadeatuacao ON v_suap_docente.cv_id = v_areadeatuacao.cv_id::text
   LEFT JOIN ( SELECT v_projetos.cv_id, count(v_projetos.cv_id) AS qtdprojeto
   FROM v_projetos
  GROUP BY v_projetos.cv_id) projetos ON v_suap_docente.cv_id = projetos.cv_id::text
   LEFT JOIN ( SELECT v_orientacoes.cv_id, count(v_orientacoes.cv_id) AS qtdorientacao
   FROM v_orientacoes
  GROUP BY v_orientacoes.cv_id) orientacoes ON v_suap_docente.cv_id = orientacoes.cv_id::text
   LEFT JOIN ( SELECT v_orientacoes.cv_id, count(v_orientacoes.cv_id) AS qtdorientacao
   FROM v_orientacoes
  WHERE v_orientacoes.natureza::text = 'Dissertação de mestrado'::text
  GROUP BY v_orientacoes.cv_id) orientacaomestrado ON v_suap_docente.cv_id =
orientacaomestrado.cv_id::text
   LEFT JOIN ( SELECT v_orientacoes.cv_id, count(v_orientacoes.cv_id) AS qtdorientacao
   FROM v_orientacoes
  WHERE v_orientacoes.natureza::text =
'MONOGRAFIA_DE_CONCLUSAO_DE_CURSO_APERFEICOAMENTO_E_ESPECIALIZACAO'::text
  GROUP BY v_orientacoes.cv_id) orientacaoespec ON v_suap_docente.cv_id =
orientacaoespec.cv_id::text
   LEFT JOIN ( SELECT v_orientacoes.cv_id, count(v_orientacoes.cv_id) AS qtdorientacao
   FROM v_orientacoes
  WHERE v_orientacoes.natureza::text = 'INICIACAO_CIENTIFICA'::text
  GROUP BY v_orientacoes.cv_id) orientacaoic ON v_suap_docente.cv_id = orientacaoic.cv_id::text
   LEFT JOIN ( SELECT v_orientacoes.cv_id, count(v_orientacoes.cv_id) AS qtdorientacao
   FROM v_orientacoes
  WHERE v_orientacoes.natureza::text =
'TRABALHO_DE_CONCLUSAO_DE_CURSO_GRADUACAO'::text
  GROUP BY v_orientacoes.cv_id) orientacaograduacao ON v_suap_docente.cv_id =
orientacaograduacao.cv_id::text
   LEFT JOIN v_idioma ON v_suap_docente.cv_id = v_idioma.cv_id::text
   LEFT JOIN v_publicacoes ON v_suap_docente.cv_id = v_publicacoes.cv_id
  ORDER BY v_suap_docente.nome;
```

ANEXO F – Histogramas dos dados da pesquisa distribuídos por atributos



atributo: regimetrabalho



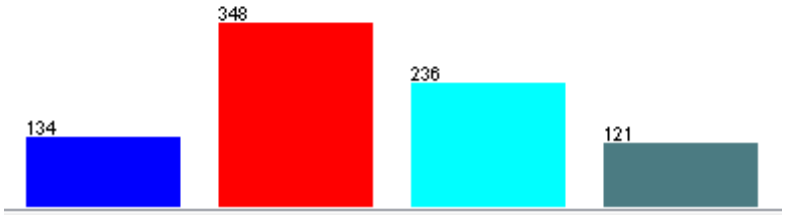
atributo: sexo

Selected attribute

Name: idade Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	20a29	134	134.0
2	30a39	348	348.0
3	40a49	236	236.0
4	50oumais	121	121.0

Class: idade (Nom) Visualize All



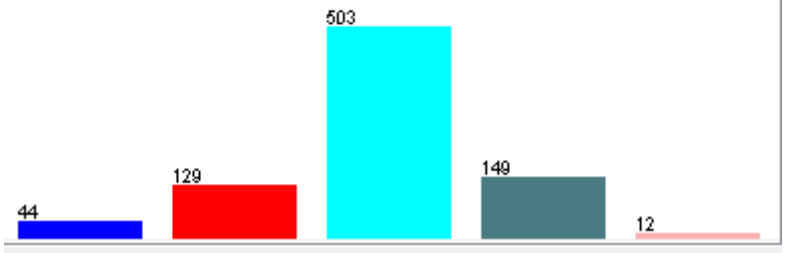
atributo: idade

Selected attribute

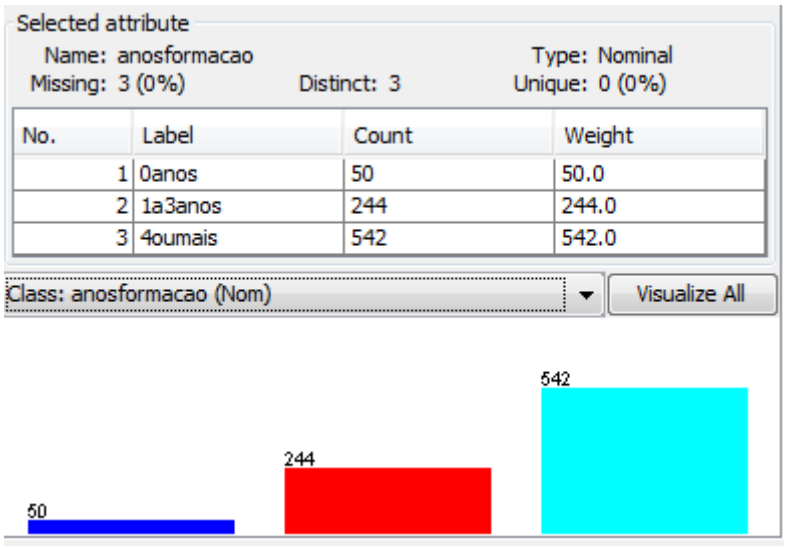
Name: titulacaomax Type: Nominal
Missing: 2 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	G	44	44.0
2	E	129	129.0
3	M	503	503.0
4	D	149	149.0
5	P	12	12.0

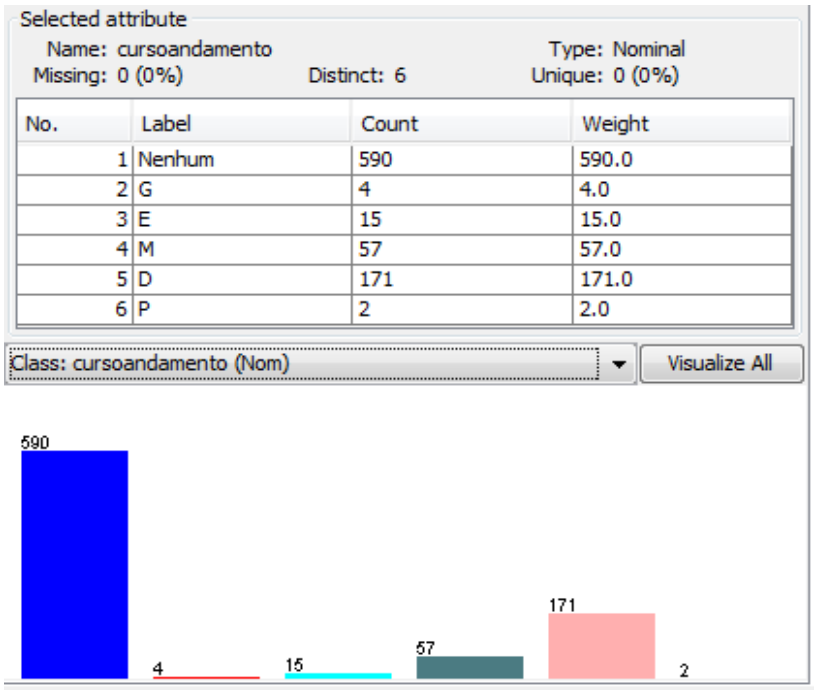
Class: titulacaomax (Nom) Visualize All



Atributo: titulacaomax



Atributo: anosformacao



Atributo: cursoandamento

Selected attribute

Name: escritaingles Type: Nominal
Missing: 159 (19%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	P	192	192.0
2	B	167	167.0
3	R	321	321.0

Class: escritaingles (Nom) Visualize All

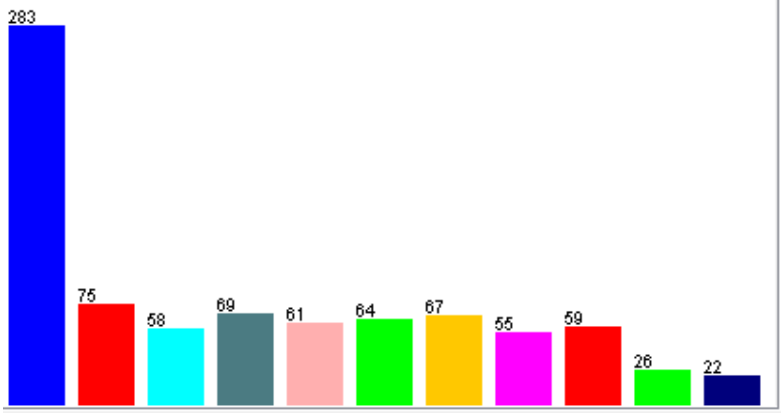
Atributo: escritaingles

Selected attribute

Name: campus Type: Nominal
Missing: 0 (0%) Distinct: 11 Unique: 0 (0%)

No.	Label	Count	Weight
1	GYN	283	283.0
2	JAT	75	75.0
3	LUZ	58	58.0
4	ANA	69	69.0
5	APA	61	61.0
6	ITU	64	64.0
7	INH	67	67.0
8	FOR	55	55.0
9	URU	59	59.0
10	REI	26	26.0
11	GOI	22	22.0

Class: campus (Nom) Visualize All



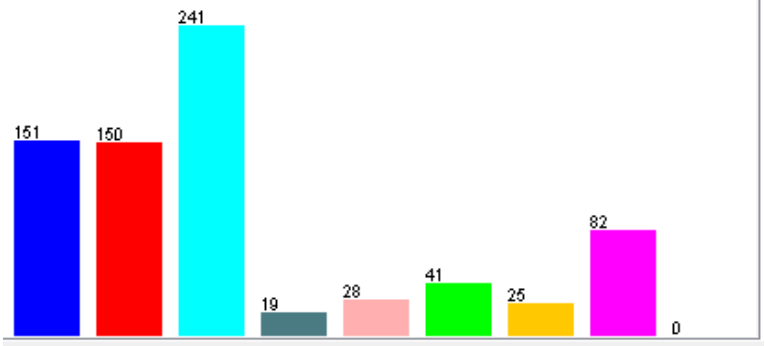
Atributo: campus

Selected attribute

Name: grandearea Type: Nominal
Missing: 102 (12%) Distinct: 8 Unique: 0 (0%)

No.	Label	Count	Weight
1	ENG	151	151.0
2	CH	150	150.0
3	CET	241	241.0
4	CA	19	19.0
5	CB	28	28.0
6	CSA	41	41.0
7	CS	25	25.0
8	LLA	82	82.0
9	Outros	0	0.0

Class: grandearea (Nom) Visualize All



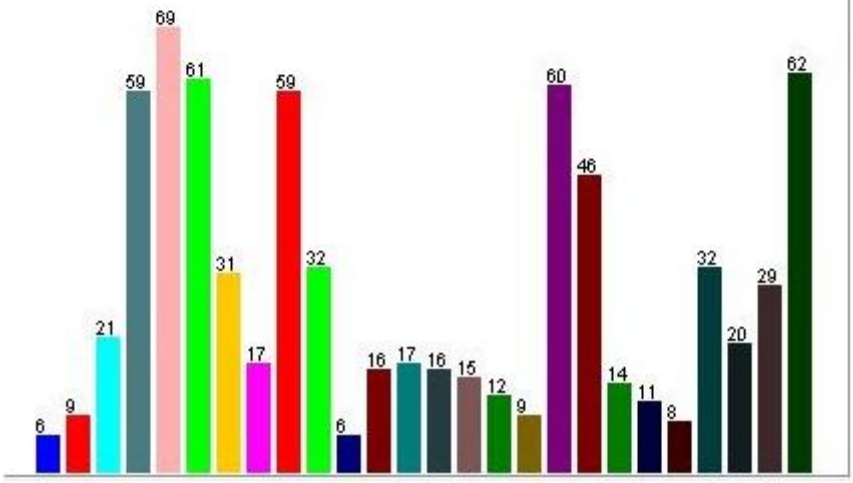
Atributo: grandearea

Selected attribute

Name: area Type: Nominal
Missing: 102 (12%) Distinct: 26 Unique: 0 (0%)

No.	Label	Count	Weight
1	Agronomia	6	6.0
2	Ciencia e Tecnologia de Alimentos	9	9.0
3	Educacao Fisica	21	21.0
4	Quimica	59	59.0
5	Ciencia da Computacao	69	69.0
6	Matematica	61	61.0
7	Fisica	31	31.0
8	Geociencias	17	17.0
9	Educacao	59	59.0
10	Historia	32	32.0
11	Biologia Geral	6	6.0
12	Filosofia	16	16.0
13	Geografia	17	17.0
14	Sociologia	16	16.0
15	Administracao	15	15.0
16	Arquitetura e Urbanismo	12	12.0
17	Turismo	9	9.0
18	Engenharia Eletrica	60	60.0
19	Engenharia Civil	46	46.0
20	Engenharia Mecanica	14	14.0
21	Engenharia de Transportes	11	11.0
22	Engenharia Sanitaria	8	8.0
23	Letras	32	32.0
24	Linguistica	20	20.0
25	Artes	29	29.0
26	Outras	62	62.0

Class: area (Nom) Visualize All



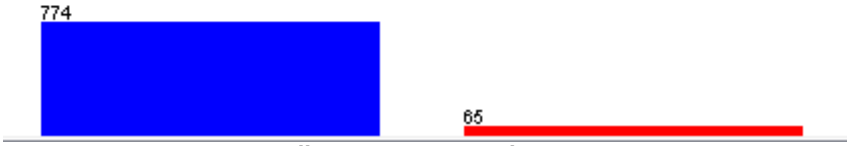
Atributo: area

Selected attribute

Name: atuaposgraduacao Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	N	774	774.0
2	S	65	65.0

Class: atuaposgraduacao (Nom) Visualize All



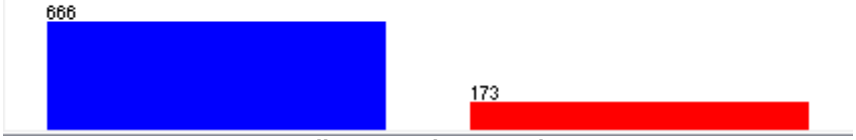
Atributo: atuaposgraduacao

Selected attribute

Name: nucleopesquisa Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	N	666	666.0
2	S	173	173.0

Class: nucleopesquisa (Nom) Visualize All



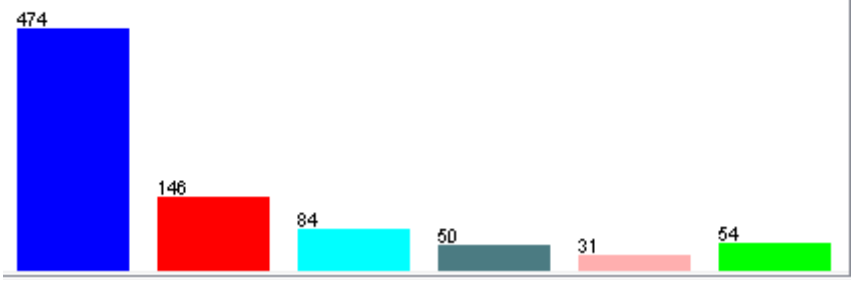
Atributo: nucleopesquisa

Selected attribute

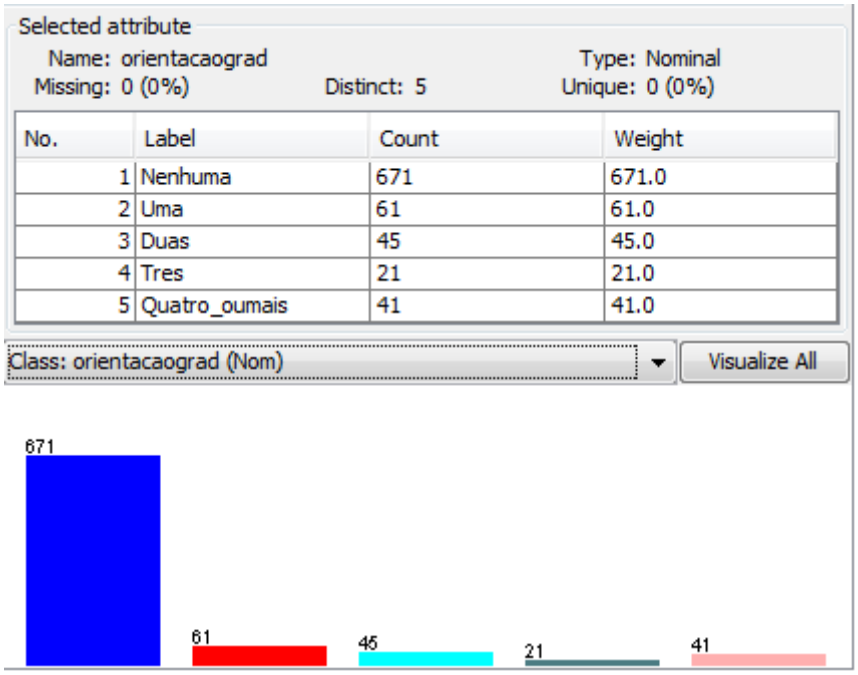
Name: projeto Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

No.	Label	Count	Weight
1	Nenhum	474	474.0
2	Um	146	146.0
3	Dois	84	84.0
4	Tres	50	50.0
5	Quatro	31	31.0
6	Cinco_oumais	54	54.0

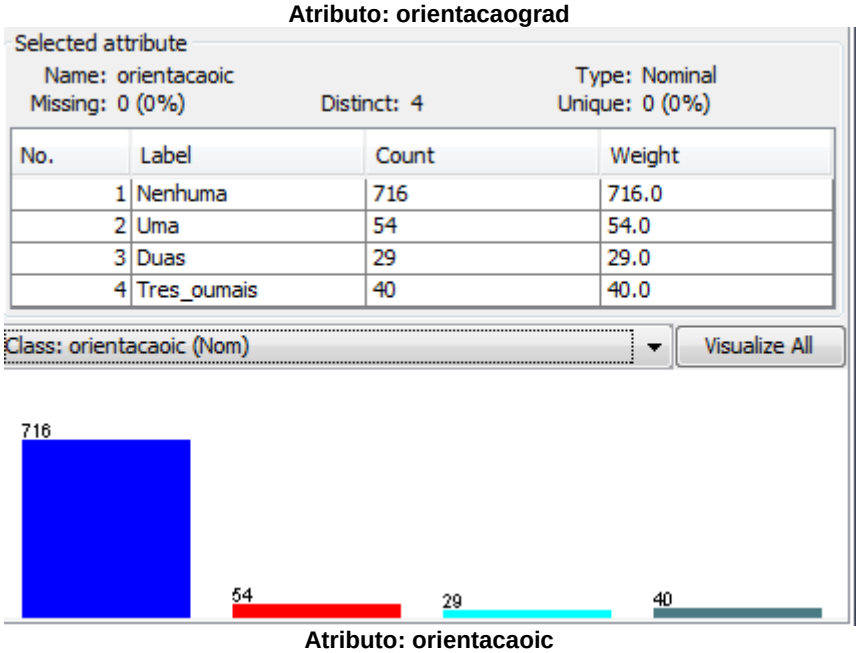
Class: projeto (Nom) Visualize All



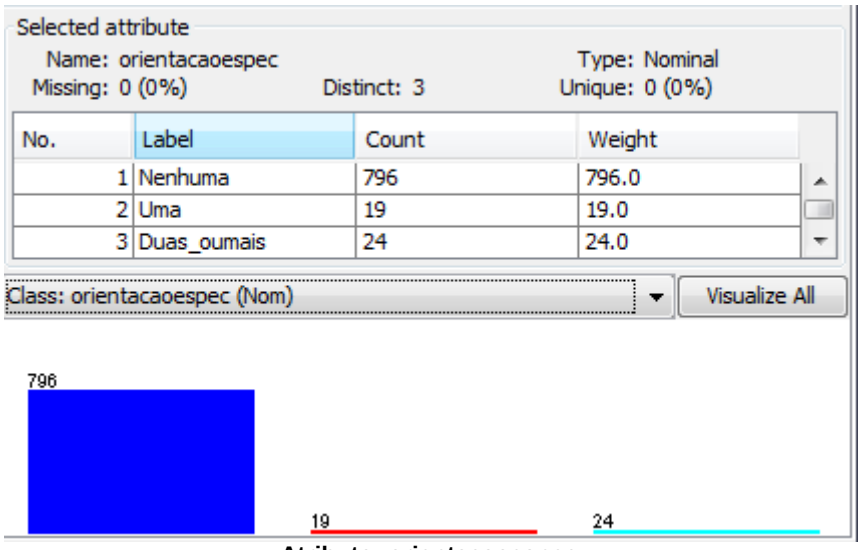
Atributo: projeto



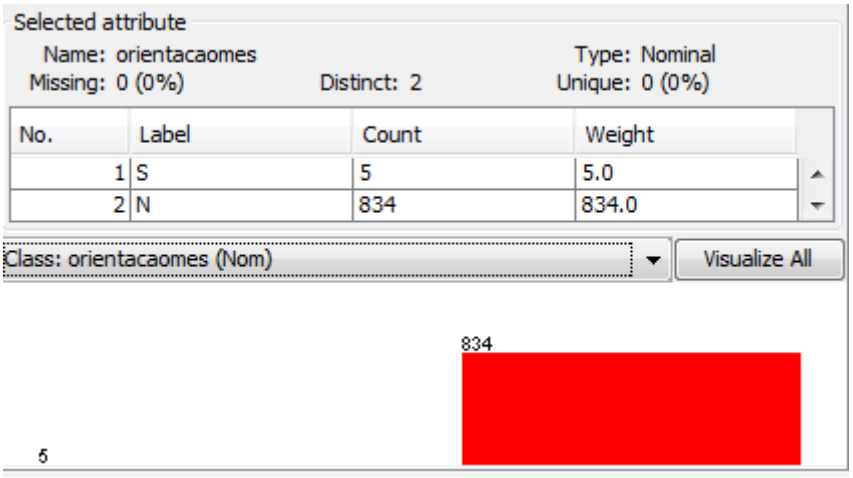
Atributo: orientacaograd



Atributo: orientacaoic



Atributo: orientacaoespec



Atributo: orientacaomes

ANEXO G – *View* v_publicacoes (soma os itens de publicações para atribuição do atributo classe)

```
CREATE OR REPLACE VIEW v_publicacoes AS
 SELECT v_suap_docente.cv_id, v_suap_docente.nome, v_apresentacaotrabalho.apresentacaotrabalho,
v_artigo.artigo, v_capitulo.capitulo, v_livro.livro, v_materialdidatico.materialdidatico,
v_processosetecnicas.processosetecnicas, v_relatoriotecnico.relatoriotecnico, v_software.software,
v_textojornalrevista.textojornalrevista, v_trabalhocompleto.trabalhocompleto,
v_resumoexpandido.resumoexpandido, v_resumo.resumo, v_trabalhotecnico.trabalhotecnico,
COALESCE(v_apresentacaotrabalho.apresentacaotrabalho, 0::bigint) + COALESCE(v_artigo.artigo,
0::bigint) + COALESCE(v_capitulo.capitulo, 0::bigint) + COALESCE(v_livro.livro, 0::bigint) +
COALESCE(v_materialdidatico.materialdidatico, 0::bigint) +
COALESCE(v_processosetecnicas.processosetecnicas, 0::bigint) +
COALESCE(v_relatoriotecnico.relatoriotecnico, 0::bigint) + COALESCE(v_software.software, 0::bigint) +
COALESCE(v_textojornalrevista.textojornalrevista, 0::bigint) +
COALESCE(v_trabalhocompleto.trabalhocompleto, 0::bigint) +
COALESCE(v_resumoexpandido.resumoexpandido, 0::bigint) + COALESCE(v_resumo.resumo, 0::bigint)
+ COALESCE(v_trabalhotecnico.trabalhotecnico, 0::bigint) AS total
   FROM v_suap_docente
   LEFT JOIN ( SELECT v_apresentacaotrabalho.cv_id, count(v_apresentacaotrabalho.cv_id) AS
apresentacaotrabalho
          FROM v_apresentacaotrabalho
         GROUP BY v_apresentacaotrabalho.cv_id) v_apresentacaotrabalho ON v_suap_docente.cv_id =
v_apresentacaotrabalho.cv_id::text
   LEFT JOIN ( SELECT v_artigo.cv_id, count(v_artigo.cv_id) AS artigo
      FROM v_artigo
     GROUP BY v_artigo.cv_id) v_artigo ON v_suap_docente.cv_id = v_artigo.cv_id::text
   LEFT JOIN ( SELECT v_capitulo.cv_id, count(v_capitulo.cv_id) AS capitulo
   FROM v_capitulo
  GROUP BY v_capitulo.cv_id) v_capitulo ON v_suap_docente.cv_id = v_capitulo.cv_id::text
   LEFT JOIN ( SELECT v_livro.cv_id, count(v_livro.cv_id) AS livro
   FROM v_livro
  GROUP BY v_livro.cv_id) v_livro ON v_suap_docente.cv_id = v_livro.cv_id::text
   LEFT JOIN ( SELECT v_materialdidatico.cv_id, count(v_materialdidatico.cv_id) AS materialdidatico
   FROM v_materialdidatico
  GROUP BY v_materialdidatico.cv_id) v_materialdidatico ON v_suap_docente.cv_id =
v_materialdidatico.cv_id::text
   LEFT JOIN ( SELECT v_processosetecnicas.cv_id, count(v_processosetecnicas.cv_id) AS
processosetecnicas
   FROM v_processosetecnicas
  GROUP BY v_processosetecnicas.cv_id) v_processosetecnicas ON v_suap_docente.cv_id =
v_processosetecnicas.cv_id::text
   LEFT JOIN ( SELECT v_relatoriotecnico.cv_id, count(v_relatoriotecnico.cv_id) AS relatoriotecnico
   FROM v_relatoriotecnico
  GROUP BY v_relatoriotecnico.cv_id) v_relatoriotecnico ON v_suap_docente.cv_id =
v_relatoriotecnico.cv_id::text
   LEFT JOIN ( SELECT v_software.cv_id, count(v_software.cv_id) AS software
   FROM v_software
  GROUP BY v_software.cv_id) v_software ON v_suap_docente.cv_id = v_software.cv_id::text
   LEFT JOIN ( SELECT v_textojornalrevista.cv_id, count(v_textojornalrevista.cv_id) AS textojornalrevista
   FROM v_textojornalrevista
  GROUP BY v_textojornalrevista.cv_id) v_textojornalrevista ON v_suap_docente.cv_id =
v_textojornalrevista.cv_id::text
   LEFT JOIN ( SELECT v_trabalhocompleto.cv_id, count(v_trabalhocompleto.cv_id) AS trabalhocompleto
   FROM v_trabalhocompleto
```

```
  GROUP BY v_trabalhocompleto.cv_id) v_trabalhocompleto ON v_suap_docente.cv_id =
v_trabalhocompleto.cv_id::text
   LEFT JOIN ( SELECT v_resumoexpandido.cv_id, count(v_resumoexpandido.cv_id) AS
resumoexpandido
   FROM v_resumoexpandido
  GROUP BY v_resumoexpandido.cv_id) v_resumoexpandido ON v_suap_docente.cv_id =
v_resumoexpandido.cv_id::text
   LEFT JOIN ( SELECT v_resumo.cv_id, count(v_resumo.cv_id) AS resumo
   FROM v_resumo
  GROUP BY v_resumo.cv_id) v_resumo ON v_suap_docente.cv_id = v_resumo.cv_id::text
   LEFT JOIN ( SELECT v_trabalhotecnico.cv_id, count(v_trabalhotecnico.cv_id) AS trabalhotecnico
   FROM v_trabalhotecnico
  GROUP BY v_trabalhotecnico.cv_id) v_trabalhotecnico ON v_suap_docente.cv_id =
v_trabalhotecnico.cv_id::text  ORDER BY v_suap_docente.nome;
```