UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS AGRÁRIAS
DEPARTAMENTO DE COMPUTAÇÃO

AMANDA FONSECA CAMPANHARO

ANÁLISE DO PERFIL DOS ALUNOS DE ESCOLAS PÚBLICAS DE ALEGRE USANDO MINERAÇÃO DE DADOS

ALEGRE - ES
2014

AMANDA FONSECA CAMPANHARO

ANÁLISE DO PERFIL DOS ALUNOS DE ESCOLAS PÚBLICAS DE ALEGRE USANDO MINERAÇÃO DE DADOS

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Bruno Vilela Oliveira

ALEGRE – ES

2014

ANÁLISE DO PERFIL DOS ALUNOS DE ESCOLAS PÚBLICAS DE ALEGRE USANDO MINERAÇÃO DE DADOS

Aprovado em 9 de dezembro de 2014.

BANCA EXAMINADORA

Prof. Dr. Bruno Vilela Oliveira

Prof. Bruno Vilela Oliveira
Orientador

Jacson R. C. Silva
Prof. MSc. Jacson Rodrigues Correia da Silva

Prof. MSc. Valéria Alves da Silva

RESUMO

Um dos grandes desafios da área de educação, a se destacar em instituições públicas no Brasil, é o desempenho dos alunos, que frequentemente está abaixo dos padrões dos países desenvolvidos. Trata-se de um problema de análise complexa e que pode ser influenciada por diversos fatores como características cognitivas e socioeconômicas dos alunos. Esboçar um perfil dos alunos pode representar uma abordagem estratégica bastante importante para a escola, tendo em vista que essa informação pode auxiliar na identificação dos fatores que influenciam o desempenho dos mesmos. O processo de descoberta de conhecimento é um processo que permite extrair conhecimento previamente desconhecido e potencialmente útil, armazenadas em grandes bases de dados. Ele é composto por diversas etapas destacando-se a Mineração de Dados. Este trabalho visa aplicar uma das principais tarefas da mineração de dados, as regras de associação, com o propósito de se identificar os fatores que poderiam influenciar o desempenho acadêmico dos alunos das escolas públicas de Alegre.

Palavras-chave: Educação pública; Desempenho escolar; Descoberta de conhecimento em base de dados; Mineração de dados; Regras de associação.

ABSTRACT

A major challenge in the field of education, especially in public institutions in Brazil, is the students' performance, which is often below the standards of developed countries. It is a problem of complex analysis which can be influenced by several factors such as cognitive and socioeconomic characteristics of the students. Sketch a profile of the students can be a very important to the school strategic approach, given that this information can assist in identifying the factors that influence their performance. The process of knowledge discovery is a process that allows extracting of previously unknown and potentially useful knowledge stored in large databases. It consists of several stages highlighting Data Mining. This paper aims to apply one of the major tasks of data mining, association rules, in order to identify factors that could influence the academic performance of students in public schools of Alegre.

Keywords: Public Education; School performance; Knowledge discovery in databases; Data mining; Association rules.

SUMÁRIO

1 INTRODUÇÃO.....11

1.1 O PROBLEMA E SUA IMPORTÂNCIA.....12

1.2 OBJETIVOS.....13

1.2.1 Objetivo Geral 13

1.2.2 Objetivos Específicos 14

2 REVISÃO BIBLIOGRÁFICA..... 15

2.1 TRABALHOS CORRELATADOS.....15

2.2 CONCEITOS E DEFINIÇÕES.....17

2.2.1 Processo de KDD..... 17

2.2.1.1 Pré Processamento 18

2.2.1.2 Seleção 19

2.2.1.3 Transformação..... 19

2.2.1.4 Mineração de Dados 20

2.2.1.5 Interpretação 20

2.2.1.6 Conhecimento 20

2.2.2 Mineração de Dados..... 20

2.2.3 Regras de Associação 22

2.2.3.1 Definição..... 22

2.2.3.2 Algoritmo Apriori 26

2.1.4 Ferramenta de mineração de dados 28

3 METODOLOGIA 30

3.1 AQUISIÇÃO DOS DADOS ESCOLARES.....30

3.2 SELEÇÃO DOS DADOS.....31

3.3 PRÉ-PROCESSAMENTO.....32

3.4 TRANSFORMAÇÃO DOS DADOS.....41

3.5 MINERAÇÃO DE DADOS.....42

3.5.1 Suporte/Confiança.....43

3.1.2 LIFT.....44

4 RESULTADOS 47

5 CONCLUSÃO 50

6 REFERÊNCIAS 52

APÊNDICE A – MODELO DE ARQUIVO NO FORMATO ARFF **..... 57**

LISTA DE TABELAS

Tabela 1: exemplos dos índices de suporte e confiança 25

Tabela 2: Regras geradas com a medida LIFT 26

Tabela 3: Exemplo da planilha fornecida pelas escolas 31

Tabela 4: Medidas estatísticas da base de dados dos alunos 33

Tabela 5: Medidas estatísticas da base de dados dos alunos da escola A.....34

Tabela 6: Medidas estatísticas da base de dados dos alunos da escola B 34

Tabela 7: Medidas estatísticas da base de dados dos alunos da escola C 34

Tabela 8: exemplos de regras que evidenciam a relação entre os atributos Notas e Faltas 47

Tabela 9: exemplos de regras encontradas com o atributo Bolsa Família 48

Tabela 10: exemplos de regras que evidenciam a associação entre os atributos notas de português e notas de matemática 49

Tabela 11: exemplos de regras não interessante para a pesquisa 49

LISTA DE FIGURAS

Figura 1 – Processo de KDD 18

Figura 2 – Representação da base de dados de produtos 22

Figura 3 – base de dados com transações de compras 23

Figura 4 - Algoritmo Apriori 27

Figura 5- Tela de pré-processamento da ferramenta WEKA. 28

Figura 6 – Tela de Definição dos Parâmetros do Apriori 29

Figura 7- Distribuição da idade dos alunos das três escolas 35

Figura 8 - Distribuição das notas de matemática dos alunos das três escolas 35

Figura 9- Distribuição das notas de português dos alunos das três escolas 36

Figura 10- Distribuição da idade dos alunos das três escolas 36

Figura 11 – Gráfico de dispersão das notas de português e porcentagem de faltas 37

Figura 12 – Gráfico de dispersão das Notas de português e Porcentagem de Faltas 38

Figura 13 – Gráfico de dispersão das notas de português e notas de matemática 38

Figura 14 – Gráfico de dispersão das idade e notas de português 39

Figura 15 – Gráfico de dispersão das idade e notas de matemática 39

Figura 16 – Reprovação dos alunos em anos anteriores 40

Figura 17- Cadastro dos alunos no programa Bolsa família. 40

Figura 18- turno que os dos alunos estudam 41

Figura 19- Tela de pré-processamento mostrando o caminho pra discretização de atributos 42

Figura 20- Exemplos regras geradas 43

Figura 22 – Regras geradas relacionadas aos alunos que não possuem bolsa família..... 44

Figura 23 – Regras geradas relacionada aos alunos que possuem bolsa família 44

Figura 24 – Regras geradas que relacionam notas e faltas dos alunos..... 45

Figura 25 – Regras geradas que relacionam notas e faltas dos alunos..... 45

Figura 26 – Regras geradas que sobre os alunos que não possui bolsa fámilia 45

Figura 27 – Regras geradas que sobre os alunos que possui bolsa fámilia 46

1 INTRODUÇÃO

A educação é considerada uma variável determinante para desenvolvimento social, político e econômico de um país. Por isso é importante identificar os obstáculos dominantes do sistema de ensino público, de forma a torná-lo apto para uma transformação sociocultural (GOMES, 1964).

Diversos são os problemas presentes na educação brasileira, especialmente no que diz respeito à educação pública. Segundo Castro (1999), o desempenho dos alunos brasileiros permanece muito distante dos padrões internacionais e daqueles necessários ao desenvolvimento nacional.

Para atingir a qualidade do ensino das escolas públicas no país é preciso explorar os fatores que influenciam o desempenho escolar dos alunos. O conhecimento do perfil dos alunos é uma poderosa ferramenta para a educação. Com um maior conhecimento e entendimento do comportamento dos alunos, podem-se desenvolver melhorias no atendimento e no ensino, de acordo com as necessidades de diferentes grupos (MIYAMOTO *et al*.,2011).

Para Menezes Filho (2007), os componentes que mais influenciam o desempenho escolar são características familiares e do aluno como educação da mãe, atraso escolar, reprovação prévia, condição socioeconômica e trabalho fora de casa.

Além disso, é importante notar que não basta que os dados dos alunos estejam disponíveis, é necessário que existam meios de se trabalhar com eles e de se fazer diferentes cruzamentos para gerar resultados que comprovem determinadas hipóteses sugeridas (VIECELLI *et al*.,2008).

Para essa finalidade, pode-se aplicar mecanismos modernos de tratamento de informações da área de mineração de dados. Como uma disciplina, a mineração de dados estuda técnicas que têm por objetivo extrair padrões de informações a partir de conjuntos de dados. Certos padrões extraídos, quando devidamente interpretados, podem constituir conhecimentos relevantes para a tomada de decisões estratégicas.

Quando utilizamos as técnicas de mineração de dados no contexto educacional, temos a Mineração de Dados Educacionais (*Educational Data Mining* ou EDM). A EDM é uma área de pesquisa que tem como principal objetivo o desenvolvimento de técnicas para explorar conjuntos de dados coletados em ambientes educacionais. Assim, é possível compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o papel do contexto em que a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem (BAKER; ISOTANI; CARVALHO, 2011).

1.1 O PROBLEMA E SUA IMPORTÂNCIA

Segundo Cavalcante e Santos Junior (2013), vários são os fatores que podem influenciar o desempenho escolar de alunos das escolas públicas. Esses fatores podem ser internos ao funcionamento da instituição de ensino ou externos, relacionados ao ambiente de convívio do aluno. Entre os fatores internos se destacam aqueles que compõem a estrutura da escola como laboratórios, bibliotecas, sistema de monitoria e a capacitação dos professores. Os fatores externos estão associados ao perfil socioeconômico do aluno.

Mesmo que a escola tenha uma ótima estrutura e corpo docente qualificado, não se pode garantir bom desempenho se o aluno não estiver motivado na busca pelo conhecimento. Para Gil (2011), além dos hábitos de estudos, a motivação do aluno é uma importante variável na determinação do sucesso na aprendizagem. Assim, uma conjugação entre fatores relacionados à escola e ao aluno poderia trazer um melhor desempenho escolar.

Como fatores externos associados ao desempenho escolar, Barros e Mendonça (2000) destacam aqueles relacionados ao ambiente familiar como a escolaridade do pai, a escolaridade da mãe e a renda da família.

Para Biondi e Felício (2007), a determinação dos fatores que afetam o desempenho acadêmico dos alunos é fundamental para rever e avaliar as práticas educacionais e nortear as estratégias educacionais para a melhoria da qualidade da educação pública brasileira principalmente no que diz respeito ao desempenho escolar.

O uso de técnicas de mineração de dados é uma interessante abordagem para se extrair e analisar os fatores que contribuem para o desempenho escolar dos alunos.

Esse problema foi abordado no trabalho de Namem e Soares (2011) onde demostrou com o uso de mineração de dados a identificação de associações entre variáveis relacionadas ao ensino de Língua Portuguesa para alunos da quarta série do ensino fundamental do Estado do Rio de Janeiro, que alguns fatores, como falta de incentivo dos pais, reprovação prévia do aluno e atuação do aluno em trabalho doméstico e/ou em trabalho fora de casa, entre outros, exercem influência sobre o aprendizado e o desempenho do estudante.

No presente trabalho, o problema do desempenho acadêmico em escolas públicas da cidade de Alegre-ES será abordado por meio do emprego de uma estratégia de mineração de dados como ferramenta para extração de informações potencialmente úteis, implicitamente existentes nas bases de dados dos alunos dessas escolas.

1.2 OBJETIVOS

Nessa seção serão descritos o objetivo geral do trabalho e os objetivos específicos.

1.2.1 Objetivo Geral

O objetivo deste trabalho é estudar os conceitos e técnicas de mineração de dados e conduzir um processo de descoberta de conhecimento em base de dados (KDD, do inglês *knowledge-discovery in databases*) na tentativa de extrair padrões a partir de dados de alunos de escolas públicas da cidade de Alegre – ES que possam ser utilizados para se compreender quais fatores poderiam ter relação com o desempenho escolar desses estudantes. Será utilizada a ferramenta de mineração de dados WEKA (acrônimo para a expressão em inglês: *Waikato Environment for Knowledge Analysis*) (UNIVERSITY OF WAIKATO, 2004)

1.2.2 Objetivos Específicos

- Realizar um estudo sobre mineração de dados e suas técnicas e definir a tarefa de mineração de dados a ser aplicada ao problema abordado nesse trabalho.
- Obter e selecionar os dados educacionais relevantes para a pesquisa.
- Analisar os dados e representa-los em formato aceito pela ferramenta WEKA.
- Aplicar o algoritmo APRIORI para a extração de regras de associação.
- Analisar e interpretar os resultados obtidos

2 REVISÃO BIBLIOGRÁFICA

Esse capítulo está organizado em duas partes. Na primeira parte são destacados trabalhos e relatos de pesquisas referentes ao processo de busca de informações em bases de dados educacionais, na sequência, são apresentados conceitos e definições necessários à compreensão deste trabalho, relacionadas com a área de mineração de dados.

2.1 TRABALHOS CORRELATADOS

Merceron e Yacef (2005) propuseram um estudo demonstrando como o uso da técnica Regras de Associação (Witten et al., 2011) pode ajudar a descobrir conhecimentos pedagógicos relevantes extraídos do banco de dados do ITS LogicITA utilizado na Universidade de Sydney. Buscou-se descobrir quais são os erros que ocorrem com frequência durante a resolução dos exercícios. Estes resultados puderam ser utilizados por professores para revisão de materiais do curso ou para enfatizar detalhes durante a explicação dos conceitos aos estudantes.

Gottardo (2012) aplicou técnicas de mineração de dados sobre um um conjunto de atributos para representar estudantes em um curso a distância (EAD) utilizando um AVA, foi possível então a obter estimativas sobre o desempenho futuro de estudantes. Esse conjunto de atributos escolhido considera três dimensões: perfil de uso do AVA, interação estudante-estudante e interação bidirecional estudante-professor. Neste trabalho, um estudo com sete experimentos foram realizados e apresentam diferentes cenários em que as estimativas sobre o desempenho foram obtidas. Os resultados desses experimentos apontam para a viabilidade desta proposta, tendo em vista os índices promissores de acurácia obtidos na classificação de estudantes quanto ao seu desempenho final nos cursos.

Carniel (2013) abordou o tema evasão escolar em instituições de ensino superior e utilizou dados dos alunos do curso de Ciência da Computação da

UNIVALI (Universidade do Vale do Itajaí), Campus Itajaí/SC, identificando grupos de risco, faixa etária de alunos mais propensos a evasão e classificando alunos por tipo de afastamento. Aplicou três tarefas de mineração de dados.

A primeira foi a Clusterização, em três experimentos distintos apresentados com a finalidade de agrupar os alunos pelas suas semelhanças sendo que cada cluster agrupou alunos com mais características comuns, uma constatação importante gerada através do uso de cluster foi que a evasão independe das notas dos alunos, sendo que alunos com notas para aprovação evadem na mesma proporção de alunos com notas para evasão

Depois aplicou a regra associação com o objetivo de encontrar regras frequentes que identificam relações entre a evasão e algum comportamento ou característica frequente entre os alunos evadidos, também nesta linha tentou-se encontrar estas relações para os alunos não evadidos, foi realizado em quatro experimentos distintos. Com uso desse tarefa pode fazer um apontamento de disciplinas e áreas específicas as quais os alunos evadidos cursaram no semestre em que evadiram, em determinados anos da pesquisa. Na maioria dos casos são disciplinas dos primeiros períodos acadêmicos.

Por fim se aplicou a Classificação na tentativa de agrupar os alunos em classes pré-definidas através dos atributos referentes a evasão e ao desempenho escolar como por exemplo Tipo de Afastamento e as Médias Calculadas. Para esta tarefa foram trabalhadas três pesquisas no qual indicam que os resultados gerados através dessa regra pode ser passíveis de novos estudos, pois com uma melhor análise dos atributos dos alunos, selecionando atributos disponíveis já no início da vida acadêmica do aluno pode ser possível classificar novos alunos e tentar prever possíveis alunos com possíveis dificuldades de aprendizado no curso e ou consequentemente passiveis de evasão.

Pimentel e Omar (2006) descrevem um estudo de caso e apresentam um modelo para a aplicação das tarefas de regra de associação e classificação a fim de descobrir conhecimentos a respeito de um aluno ou de um grupo de alunos, em dados coletados por meio de avaliações. Espera-se com isto criar mecanismos capazes de apoiar a avaliação formativa com feedback adequado para professor e aluno.

Em uma pesquisa semelhante realizado por Trindade e Costa (2012), com o objetivo de aplicar uma técnica de Mineração de Dados para avaliar o nível de gestão acadêmica de alunos pré-vestibulandos, comparando a localidade com o tipo de escola (administração pública ou privada), estudou resultados apresentados por uma análise de atributos, utilizando-se 392 alunos, em Minas Gerais, com o critério de seleção do suporte mínimo de 1%, confiança mínima de 50% e lift mínimo de 1.01. Como resultados, os autores encontraram que 302 alunos que cursaram o ensino médio em Escola Estadual, apresentaram confiança (conf) de 77%, onde o valor do lift destacou que a chance de um candidato de Diamantina ter cursado o ensino médio em escola estadual era 1.61 vezes maior que o suporte esperado. Como conclusões, a partir das regras geradas com essas medidas de interesse, foi possível concluir que há uma diferença no perfil dos candidatos quanto à localização e o tipo de escola onde estudaram: onde os alunos do interior de Minas Gerais concluíram o ensino médio em escola estadual. Já os candidatos da região metropolitana, concluíram o ensino médio em escola particular.

2.2 CONCEITOS E DEFINIÇÕES

2.2.1 Processo de KDD

A descoberta de conhecimento em base de dados consiste do processo de descobrir conhecimento útil em grandes bases de dados. Segundo Fayyad, *et al* (1996), KDD pode ser caracterizado como um processo não trivial de identificação de padrões válidos, inovadores, potencialmente úteis e compreensíveis em bases de dados. Na Figura 1 é apresentado como esse processo é divido em várias etapas que serão descritos nas seguintes subseções.

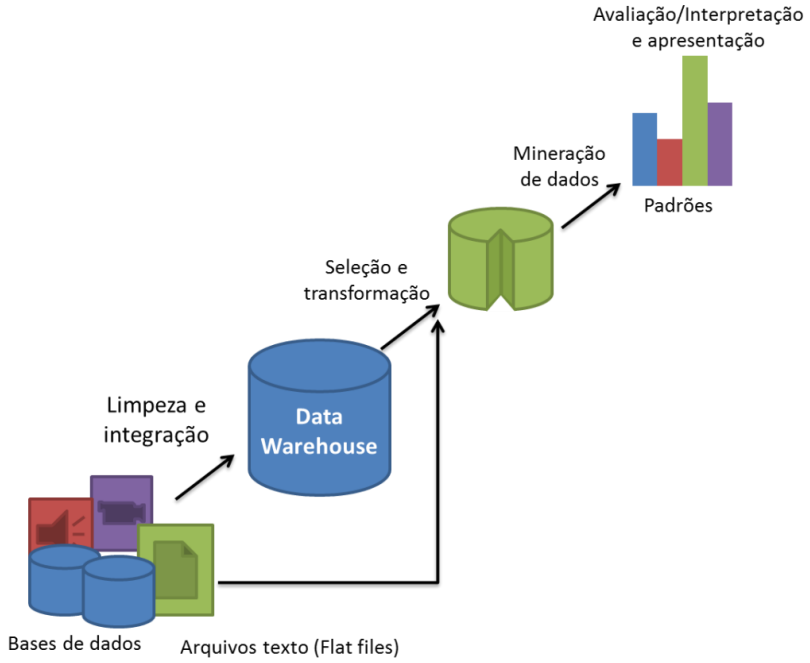

Figura 1 – Processo de KDD

Fonte: Elaborada pelo autor

2.2.1.1 Pré Processamento

Os dados muitas vezes precisam de preparação antes de passar para a etapa de mineração. Nessa etapa, ocorre a preparação das informações para a geração de uma base de dados consistente que possa ser utilizada de forma adequada. A integração de dados de diferentes bases de dados que muitas vezes são obtidos em diferentes formatos, a remoção de informações consideradas desnecessárias e a adequação dos dados para assegurar formatos consistentes são tarefas comuns no pré-processamento.

Abaixo são resumidos detalhes de algumas das atividades que podem ser realizadas nessa etapa preliminar de preparação dos dados (HAN; KAMBER, 2006):

- **Valores ausentes**: os valores de alguns atributos podem não estar presentes nas bases. Para resolver esse problema pode-se: ignorar os registros que possuem valores ausentes; preencher os valores ausentes manualmente; usar um valor constante ou usar o valor médio do atributo.
- **Valores fora do padrão (valores aberrantes ou *outliers*)**: Para lidar com esse tipo de problema pode-se utilizar análise de agrupamentos, que consiste no agrupamento dos valores similares, facilitando a identificação e

exclusão de valores fora do padrão; inspeção humana e computadorizada para identificar e excluir os valores fora do padrão; e regressão, na qual os dados podem ser ajustados por meio de funções de regressão.

- **Dados inconsistentes**: erros que ocorrem normalmente porque o usuário entra com um dado incorreto. Outro tipo de inconsistência nos dados é a redundância, que ocorre quando são registrados dados que possuem o mesmo valor semântico, mas que foram inseridos com valores diferentes. Uma forma de resolver esse problema é por meio da análise de correlação, que consiste na medida de relacionamento entre dois atributos.

2.2.1.2 Seleção

Geralmente as bases de dados armazenam vários atributos de uma entidade, mas nem sempre todos estes atributos são relevantes para a análise em questão. Atributos são as informações que referenciam a entidade, a entidade é uma representação de um conjunto de informações sobre determinado conceito, por exemplo aluno seria uma entidade e idade, matricula e CPF seriam atributos dessa entidade.

Faz-se necessário uma criteriosa seleção de quais atributos no conjunto de dados realmente representam informações relevantes para uso no processo de KDD (MORAIS, 2010).

2.2.1.3 Transformação

A etapa de transformação consiste na transformação ou consolidação dos dados em formatos utilizáveis, de acordo com a técnica de mineração de dados utilizada na próxima etapa. Entre as transformações tipicamente utilizadas tem-se a normalização de atributos, a padronização de atributos, a redução de dados e a criação de novos atributos (BÜRKLE, 2006)

2.2.1.4 Mineração de Dados

Essa etapa é a mais importante do processo KDD, na qual são aplicadas tarefas de mineração para a extração de padrões de interesse a partir dos dados. Essa etapa será descrita detalhadamente na seção 2.2.2.

2.2.1.5 Interpretação

Nem todos os padrões descobertos têm importância para o problema em questão. Para Morais (2010) os resultados obtidos devem ser avaliados por analistas para que seja verificada a utilidade dos mesmos. Os padrões que forem julgados úteis devem ser interpretados de forma que se tornem compreensíveis para os usuários.

2.2.1.6 Conhecimento

Na última etapa do processo de KDD ocorre a extração do conhecimento que é incorporado a um sistema, podendo ser utilizado diretamente pelo usuário final ou, simplesmente, documentado e relatado às pessoas interessadas. Os resultados dos passos anteriores ainda devem ser analisados para que possíveis conflitos entre o conhecimento existente e o conhecimento adquirido sejam resolvidos (FAYYAD, *et al*., 1996).

2.2.2 Mineração de Dados

Segundo Hand *et al*. (2001), a mineração de dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam úteis para a tomada de decisão.

Essa é uma área de pesquisa multidisciplinar, que inclui inteligência computacional, matemática, estatística e banco de dados. Um dos grandes desafios da mineração de dados é propor algoritmos que sejam capazes de trabalhar com grandes quantidades de dados. (MORAIS, 2010).

Para Fayyad, *et al* (1996), o processo de mineração de dados possui dois objetivos principais: predição, em que a ideia é prever o comportamento futuro de algumas variáveis da base de dados e a descrição, em que a ideia é

identificar padrões existentes nas bases de dados que sejam possível interpretá-los.

As tarefas em mineração de dados estão relacionadas ao objetivo que se deseja buscar, isto é, ao tipo de padrão que se deseja encontrar nas bases de dados enquanto que uma técnica está relacionada a como encontrar os padrões de interesse. Podem ser citadas as seguintes tarefas de mineração de dados (FAYYAD, *et al* 1996):

- **Classificação**: consiste em determinar uma função que associe cada item de uma base dados a uma das classes previamente definidas de acordo com a similaridade de alguma característica dos dados. Por exemplo, em um supermercado podemos classificar os produtos em algumas classes predefinidas: frios, laticínios, produtos de higiene, etc.
- **Associação**: identifica grupos de dados que ocorrem em conjunto ou de forma associada e usa essa informação para a tomada de decisões. Exemplo de seu uso é se 50% dos consumidores que compram o produto A comprarem também o produto B, colocar esses produtos em prateleiras próximas ou estimular a compra em conjunto desses itens pode aumentar as vendas. Essa tarefa será mais detalhada na seção 2.2.3.
- **Regressão**: consiste em prever o valor de uma variável com base nos valores de outras variáveis. Algumas aplicações dessa tarefa incluem: estimar o PIB de um país; estimar a probabilidade de um paciente sobreviver e prever séries temporais.
- **Análise de Séries Temporais**: evidencia sequências de ações ou acontecimentos presentes em uma base de dados organizada de forma temporal.
- **Sumarização**: consiste em definir um conjunto mínimo de características que seja capaz de identificar um subconjunto de objetos. São geralmente aplicadas em análise exploratória de dados e geração de relatórios automatizados.
- **Agrupamento (*Clustering*)**: visa identificar e agrupar dados baseando-se em algum critério de similaridade. Os dados são agrupados de maneira que aqueles que pertencem ao mesmo grupo possuem alta similaridade entre si, mas são pouco similares aos dados associados aos

demais grupos. Uma possível aplicação dessa tarefa pode ser a separação de comportamentos suspeitos em base de dados em um processo de auditoria (CAMILO; SILVA, 2009).

2.2.3 Regras de Associação

2.2.3.1 Definição

A regra de associação é uma das principais tarefas da mineração de dados que consiste em encontrar relacionamentos entre dados armazenados. A ideia básica da extração de regras de associação está na análise da ocorrência conjunta de itens em uma mesma transação (VASCONCELOS. CARVALHO 2004). Uma transação consiste por uma ou mais operações de acesso à base de dados, onde cada transação é composta por um conjunto de itens, presentes na base de dados transacional. As bases de dados transacionais são todas as informações históricas armazenadas ao longo do tempo de vida de uma organização, como por exemplo cadastros de clientes, produtos, fornecedores, entradas e saídas de estoque.

As regras de associação são representadas na forma X => Y, que se lê X implica em Y, sendo X denominado o antecessor e Y o consequente, e X e Y são dois *itemsets* distintos em uma base de dados.

Supondo um exemplo prático, ilustra-se uma pequena base de dados de produtos de uma empresa fictícia como na Figura 2. A cada grupo de itens que é vendido pela empresa dá-se o nome de *itemset*. Um *itemse*t com k elementos é chamado de k-*itemse*t. A Figura 3 mostra esses itemsets em transações comerciais.

Produto (Item)
biscoito
manteiga
arroz
feijão
ovos
café
leite
pão
queijo

Figura 2 – Representação da base de dados de produtos

Fonte: Morais, 2010

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

Figura 3 – base de dados com transações de compras

Fonte: Morais, 2010

Um exemplo de regra de associação gerada a partir das transações de compra da Figura 3, seria {queijo} => {pão}. A interpretação desta regra é que clientes que compram queijo e leite tendem a comprar pão também. Porém, a recíproca desta relação não necessariamente é verdadeira, isto é, nem sempre pode-se afirmar que clientes que compram pão tendem a comprar queijo.

Medidas de interesse são utilizadas para avaliar o quão interessante é uma regra de associação ou sua validade em relação à base de dados da qual essa regra foi extraída, suporte e confiança são as principais medidas de interesse aplicadas às regras de associação (GONÇALVES, 2005). Essas métricas podem ser utilizadas para limitar a quantidade de regras que serão extraídas e para descrever o quão interessante é uma regra.

O suporte do *itemset* A, denotado por sup(A), indica a porcentagem de transações da base de dados que contém os itens de A, ou seja:

$$Sup\ (A) = \frac{n(A)}{N}.100$$

Em que n(A) é número de transações nas quais A ocorre e N é o número total de transações da base de dados. Na Figura 3, por exemplo, o suporte do *itemset* café é 80%.

Já o suporte da regra de associação é a frequência com a qual um *itemset* ocorre simultaneamente com outro *itemset* considerando-se todas as transações da base, ou seja:

$$Sup\ (X => Y) = \frac{FrequênciaXeY}{TotaldeTransações}.100$$

Como exemplo, o suporte da regra {café, leite} =>{manteiga} considerando os dados presentes na tabela 2 é calculado como:

$$Sup\ (\{café\} => \{pao\}) := \frac{3}{5}.100 = 60$$

A confiança de uma regra de associação é a frequência com que X e Y ocorrem juntos em relação a todas as transações de que contêm X (MOTTA, 2010). A confiança é obtida pela fórmula:

$$Confiança(X, Y = X) = \frac{FrequênciadeXeY}{FrequênciadeX}.100$$

Considerando os dados da Figura 3, a confiança da regra {café, leite} => {pão} é calculado como:

$$Confiança(\{café, leite\} => \{pão\}) = \frac{3}{3}.100 = 100$$

Segundo Berry e Linoff (1997) a confiança é a frequência com que o relacionamento mantém-se verdadeiro na amostra e o suporte como a frequência com que a combinação acontece.

Segundo Gonçalves (2005) o modelo mais usado em regras de associação, costuma ser referenciado na literatura como modelo Suporte/Confiança que consiste em encontrar todas as regras que possuam suporte e confiança maiores ou iguais, a um suporte mínimo (SupMin) e uma confiança mínima (ConfMin), estabelecido pelo usuário. No modelo Suporte/Confiança os dados são minerados em duas fases:

Na primeira fase encontra-se todos os conjuntos de itens que possuem suporte maior ou igual ao SupMin. Estes conjuntos são chamados de conjuntos frequentes. Depois para cada conjunto frequente encontrado na primeira fase, gera-se as regras de associação que possuem confiança maior ou igual a ConfMin.

Ao usar o modelo Suporte/Confiança com os dados da Figura 3, é possível obter várias regras de associação, algumas das quais estão indicadas na Tabela 1. Os valores mostrados na segunda (SUP_A) e terceira coluna (SUP_B) são os suporte do item antecedente e o suporte do item consequente na base de dados respectivamente, já a quarta e a quinta coluna são o suporte e a confiança da regra. Considerando SupMin de 30% e ConfMin de 60%.

Tabela 1: exemplos dos índices de suporte e confiança

Regra (A=>B)	SUP_A	SUP_B	SUB	CONF
Leite=>Café	60%	80%	60%	100%
Queijo => Pão	40%	80%	40%	100%
Ovos => Pão	40%	80%	40%	100%

Experimentos apresentados em Zheng, *et al* (2001) demonstraram que a mineração de bases de dados reais usando o modelo Suporte/Confiança pode levar à geração de centenas de milhares de regras. Além disso, grande parte destes resultados minerados costuma ser composto por regras óbvias, redundantes ou, até mesmo, contraditórias.

Outras medidas de interesse diferentes do suporte e da confiança foram propostas com o intuito de identificar as regras que são, de fato, relevantes e úteis dentre as muitas que podem ser mineradas.

O LIFT é um exemplo de outra medida de interesse usada em regras de associação. Também conhecida como *Interest*, foi proposta por BRIN *et al*.(1997) e é utilizada para avaliar as dependências entre o conjunto de itens do antecedente e do consequente de uma regra de associação. Esta medida indica o quanto mais frequente torna-se Y quando X ocorre.

$$LIFT(X => Y) = \frac{Confiança(X => Y)}{Suporte(Y)}$$

De acordo com Gonçalves (2005) a variação do LIFT mostra que:

- LIFT(X => Y) = 1: independência entre os conjuntos de itens antecedente e consequente.
- LIFT(X => Y) < 1:os conjuntos de itens X e Y são negativamente dependentes.
- LIFT(X => Y) > 1:os conjuntos de itens X e Y são positivamente dependentes.

Na Tabela 2 são apresentadas as medidas de suporte, confiança e LIFT das regras geradas na aplicação do modelo suporte-confiança, utilizando SupMin de 20% e ConfMin de 60%, da base de dados ilustrada na Figura 2.

Tabela 2: Regras geradas com a medida LIFT

Regra (A=>B)	SUP_A	SUP_B	SUB	CONF	LIFT
Leite=>Café	60%	80%	60%	100%	1.25
Queijo => Pão	40%	80%	40%	100%	1.25
Ovos => Pão	40%	80%	40%	100%	1.25
Arroz => Feijão	40%	40%	20%	50%	1.25

É importante observar que o LIFT consegue destacar com maior facilidade a dependência positiva entre conjuntos de itens que possuem uma confiança baixa. Por exemplo {Arroz => Feijão} não satisfazem a confiança mínima, porém essa regra possui um LIFT positivo o que mostra dependência positiva entre o antecedente e o consequente da regra, o suporte real da regra é 1,25 maior que o suporte esperado. O suporte real é a probabilidade conjunta de A e B do conjunto de itens ($A \cup B$), enquanto a probabilidade conjunta sob independência representa o suporte esperado do mesmo conjunto (Bürkle, 2006). Ou seja, a probabilidade conjunta de arroz e feijão ocorrerem juntos é maior que probabilidade deles ocorrerem de forma independentes na base de dados.

2.2.3.2 Algoritmo Apriori

O algoritmo Apriori é considerado um dos mais importantes algoritmos para extração de Regras de Associação. Desenvolvido por Agrawal e Srikant (1994), esse algoritmo é capaz de encontrar todos os *itemsets* frequentes em uma base de dados. Os *itemsets* frequentes são aqueles que apresentam suporte maior que o suporte mínimo estabelecido pelo usuário.

Esse algoritmo possui a propriedade antimonotônica que diz que um conjunto de itens não é frequente se possuir pelo menos um subconjunto de itens que não é frequente, ou seja, para que um *itemset* seja frequente é necessário que todos os *itemsets* contidos nele sejam também frequentes. Caso um único *itemset* contido nele não seja frequente, seu suporte nem precisa ser calculado (o banco de dados não precisa ser varrido), pois sabemos previamente que ele nunca poderá ser frequente (AMO, 2004).

Segundo Gillmeister e Cazella (2009), primeiramente o algoritmo faz várias passagens sobre a base de dados com o intuito de selecionar todos os conjuntos de itens frequentes, sendo que, em cada um destes passos, primeiro gera um conjunto de itens candidatos e então percorre a base de dados para determinar se os candidatos satisfazem o suporte mínimo estabelecido por parâmetro. Na Figura 4 é apresentado o funcionamento do algoritmo.

```
• APRIORI(D, freq_min) {
• L1 = {todos os itens freqüentes em D de tamanho 1};
• Para (k = 2 ; Lk-1 <> vazio; k++) {
  – Ck = {candidatos gerados pela combinação dos elementos de Lk-1};
  – Para cada elemento E de Ck {
      • Para cada subconjunto S de E {
          – Se S <> Lk-1 então Ck = Ck – E;
      • };
  – };
  – Para cada transação T de D {
      • Para cada elemento E de Ck {
          – Se E está contido em T então freq(E) = freq(E) + 1;
      • };
  – };
  – Para cada elemento E de Ck {
      • Se freq(E) > freq_min então Lk = Lk + E;
  – };
• };
• retornar UkLk;
• };
```

Figura 4 - Algoritmo Apriori

De acordo com a figura 4, primeiramente, L1 que é o conjunto de itens com somente um elemento, é gerado, na sequência tem-se um laço com k passos, basicamente é a geração do grupo de itens candidatos Ck, através dos itens gerados no passo anterior (conjunto de Lk-1). Depois é realizado a poda que é quando são descartados para cada subconjunto de itens de Ck, os que não estiverem em Lk-1 , pois não terá a menor chance de ser frequente. Assim, nesta fase é calculado o conjunto $(C_k) = (C'_k)$ – {conjunto de itemsets podados}.

Depois é executado a tarefa no laço k, que consisti um outro laço para contagem do suporte dos itens (E) do grupo candidato Ck, onde cada transação da base de dados é analisada. Neste momento, cada candidato c terá ao final o seu suporte computado, e no próximo passo k, os itens que não obtiveram o suporte mínimo estabelecido são excluídos.

2.1.4 Ferramenta de mineração de dados

Nesse trabalho, para realizar a extração de regras de associação a partir da base de dados acadêmica analisada, foi utilizada a ferramenta WEKA(*Waikato Environment for Knowledge Analysis*), por ser *um framework* de código aberto e por prover na mesma ferramenta o pré-processamento de dados, agrupamento ou clustering, classificação, regressão, visualização e regras de associação de dados, o WEKA permite que os algoritmos possam ser aplicados diretamente a determinado conjunto de dados através da interface gráfica de usuário, que foi desenvolvido pela Universidade de Waikato, na Nova Zelândia (UNIVERSITY OF WAIKATO, 2004).

Na Figura 5 é apresentada a tela de pré-processamento da ferramenta.

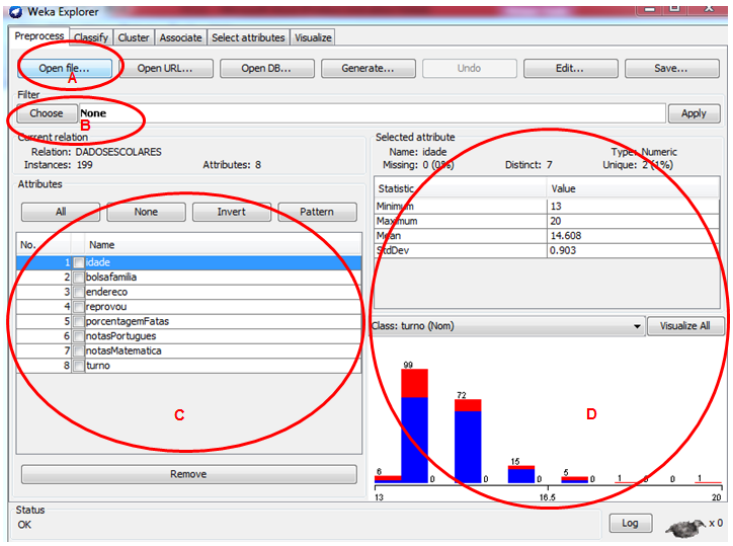

Figura 5- Tela de pré-processamento da ferramenta WEKA.

Fonte: Costa, 2007

As partes destacadas na figuram são descritas a seguir:

A - Botão que permite a seleção de bases de dados no formato ARFF (*Attribute-Relation File Format*);

B- Nessa área podem ser selecionados algoritmos de pré-processamento dos dados, como por exemplo, discretização de atributos;

C- Listagem dos atributos da base de dados;

D- Área reservada para exibição de estatísticas dos atributos da base de dados.

A ferramenta WEKA recebe como entrada arquivos no formato ARFF, que é descrito no apêndice A, composto por 3 elementos: *relation*, que define um nome para a base de dados; *atribute*, onde são especificados os atributos que compõem a base de dados e *data*, que são os dados separados por vírgulas (MORAIS, 2010).

A ferramenta possui implementações de vários algoritmos de mineração de dados, incluindo o Apriori. Nesse ambiente, o usuário pode escolher os algoritmos de regras de associação pelo botão "*Choose*".

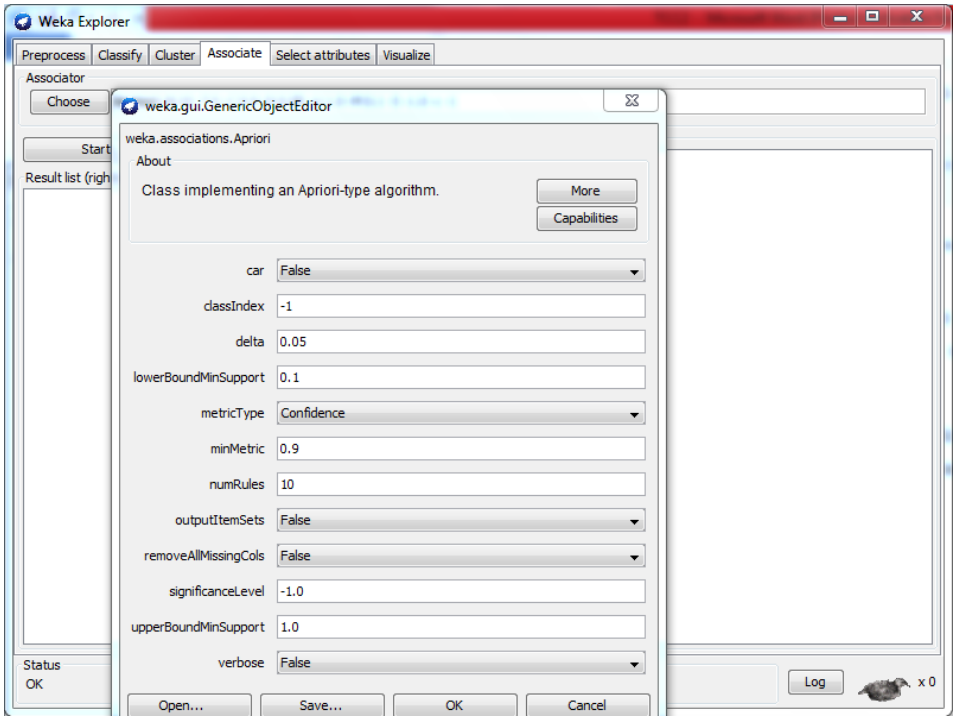

Figura 6 – Tela de Definição dos Parâmetros do Apriori

No WEKA, as regras geradas são listadas de forma ordenada pelo valor da medida de interesse e os valores numéricos à direita dos itens antecedentes e consequentes das regras representam o número de suas ocorrências, isto é, o suporte do *itemset*. Por exemplo a regra:

porcentagemFaltas='(9.47-12.26]'31==>notasMatematica='(59.5-67.6]'22 conf:(0.71).

PorcentagemFaltas='(9.47-12.26]'31 representa o antecedente da regra, sendo o valor 31 o número de registros na base de dados em que o atributo porcentagemFaltas obteve valores entre 9.47 e 12.26 e consequente notasMatematica='(59.5-67.6]' 22 diz e sendo que 22 vezes o atributo nota de matemática obteve valores entre 59,5 e 67,6.

3 METODOLOGIA

Esta seção descreve a metodologia de mineração de dados desenvolvida no presente trabalho.

A metodologia proposta nesse trabalho consiste de se aplicar o algoritmo Apriori, por meio da ferramenta WEKA, para a extração de regras de associação a partir do dados de alunos de três escolas públicas do município de Alegre. O objetivo é buscar associações que indiquem fatores que possam influenciar o desempenho escolar desses alunos. Os dados utilizados foram dos alunos do nono ano (antiga oitava série) obtidos em três escolas do Município de Alegre. Duas dessas escolas situam-se em bairros de Alegre, sendo uma no centro e outra no bairro Vila do Sul. A outra escola é localizada no distrito de Anutiba.

3.1 AQUISIÇÃO DOS DADOS ESCOLARES

Ponto inicial do projeto foi a aquisição dos dados escolares junto as escolas. Os dados foram obtidos nas próprias escolas de forma manual por meio do preenchimento de uma planilha eletrônica com os dados dos alunos.

A base de dados em questão é formada por dados de 199 alunos do nono ano em 2013 de três escolas do município de Alegre. Em que 120 alunos são da escola localizada no centro de Alegre, 58 alunos pertencem a outra escola localizada em Alegre e 21 alunos pertencem a escola localizada em um distrito do município. Como forma de identificar as escolas ao longo do trabalho as escolas serão associadas como escola A, escola B e escola C respectivamente

Um dos problemas foi coletar os dados com informações associados à origem socioeconômica do alunos como escolaridade dos pais e renda familiar do aluno. As escolas não possuem esse tipo de informação e a diretoia não autorizou a aplicação de um questionário para que fossem obtidos mais informações dos alunos. A única informação relacionada a situação

socioeconômica dos alunos fornecida pela escola foi se os alunos possui ou não cadastro no programa bolsa família.

Outro problema encontrado na obtenção dos dados foi que as escolas não quiseram ceder os dados de mais de uma série, e como nono ano era a série com maior número de alunos, optou-se por solicitar os dados dessa turma.

As informações dos alunos fornecidos pelas escolas foram: data de nascimento, endereço, se possui bolsa família, se possui reprovação, porcentagem de faltas, nota de português e matemática, como mostrado da tabela 3.

Tabela 3: Exemplo da planilha fornecida pelas escolas

Nota		**Data de Nascimento**	**Possui bolsa Família**	**Endereço**	**Possui reprovação**	**Faltas(%)**
port	**Mat**					
71	81	24/08/1999	Sim	Rua Iberaldo Anchieta (Vila do Sul)	Sim	2%
95	84	26/05/1998	Não	Rua Maurício Lacerda (Vila do Sul)	Não	1%
73	70	04/11/1999	Sim	Rua Vereador José Corrente (Vila do Sul)	Não	1%
88	82	01/02/1997	Não	Rua Porfirio de Sousa Freire (Vila do Sul)	Não	2%

3.2 SELEÇÃO DOS DADOS

Nessa etapa ocorreu a seleção os dados para o processo de mineração de dados. Durante a aquisição dos dados nas escolas, foram escolhidas as informações dos alunos que poderiam ter alguma relevância para a pesquisa, os dados selecionados foram:

- Idade: Idade no aluno no ano 2013, ano que este cursava o nono ano.
- Bolsa família: Se o aluno participa(S) ou não(N) do programa do governo federal Bolsa Família que consiste na ajuda financeira às famílias pobres (definidas como aquelas que possuem renda per capita de 70 até 140 reais) e extremamente pobres (com renda per capita menor que 70 reais) para que as famílias beneficiárias mantenham seus filhos e/ou dependentes com frequência na escola.

- Endereço: Bairro onde o aluno reside.
- Reprovação: Se o aluno possui(S) ou não(N) reprovação nos anos escolares anteriores.
- Porcentagem de faltas: porcentual da quantidade de aulas que os alunos faltaram no nono ano em 2013, a porcentagem de faltas varia de 0 a 100%. O aluno tem o direito de faltar até o limite de 25% do total de aulas dadas. Se ultrapassar este limite estará reprovado no período letivo correspondente.
- Notas português: Nota de português que aluno obteve no nono referente a avaliação do professor para o esforço do aluno nos diversos graus na disciplina de português, com provas , trabalhos, pontos extras por comportamento, etc. O nota varia de 0-100. o aluno precisa ter média final 60 para não repetir o ano.
- Notas matemática: Nota de matemática que aluno obteve no nono referente a avaliação do professor para o esforço do aluno nos diversos graus na disciplina de matemática, com provas , trabalhos, pontos extras por comportamento, etc. O nota varia de 0-100. o aluno também precisa ter média final 60 para não repetir o ano.
- Turno: Turno em que o aluno estuda, a nono ano das três escolas são oferecidos no turno matutino(M) e no vespertino (V).

Informações de matrícula, número de documentos como o CPF e o nome dos alunos não foram incluídos na base gerada, já que os valores desses atributos são únicos para cada dado, não permitindo descrever nenhum tipo de padrão genérico a partir deles.

3.2 PRÉ-PROCESSAMENTO

Na fase de pré-processamento foram eliminadas as inconsistências, como mesmo endereço registrados de diferentes formas, registros redundantes. Pela análise dos dados dos alunos obtidos nas escolas de forma manual, detectou-se registros redundantes e muitos erros de digitação.

Para o atributo endereço foi retirado o nome da rua, mantendo apenas o nome do bairro. No caso da escola situada no distrito foi utilizado o nome do

mesmo já que o distrito não possui bairros. Dados com erros de tipografia , foram padronizados com o valor correto para todos eles. Como exemplo, o bairro São Manoel ocorria com duas grafias em diferentes registros: “São Manoel” e “São Manuel”. Os registros cujos bairros estavam cadastrados com a grafia errada, isto é, “São Manuel”, tiveram o valor do atributo bairro corrigido.

Também na fase de pré-processamento foi possível obter uma visualização dos atributos em caráter estatístico que busca representar a realidade dos alunos. A análise exploratória dos dados é importante para se compreender os dados e suas características.

Os gráficos estatísticos mostrados nessa análise foi gerada pela ferramenta R. Uma ferramenta livre e de código aberto criada em1996 por Ross Ihaka e Robert Gentleman, muito utilizada para análise e manipulação de dados, com testes paramétricos e não paramétricos, modelagem linear e não linear, análise de séries temporais, análise de sobrevivência, simulação e estatística espacial, além de apresentar facilidade na elaboração de diversos tipos de gráficos. (SOUZA; PETERNELLI, 2009)

Foi feita uma análise usando a estatística descritiva sobre os dados quantitativos dos alunos (Idade,notas de português, notas de matemática e porcentagem de faltas) como forma de descrever como os dados estão distribuídos na base de dados. Na tabela 4 é apresentada algumas medidas estatísticas dos atributos quantitativos da base de dados de todos os alunos, essas medidas são normalmente usadas para descrever o conjunto de dados, como a média, a mediada e o desvio padrão.

Tabela 4: Medidas estatísticas da base de dados dos alunos

Atributo	Média	Mediana	Desvio Padrão
Idade	14,46	14	0,77
Notasde Português	72,79	74,5	11,68
Notas de Matemática	71,45	72	11,6
Porcentagem de Faltas	8,63	5,5	5,42

Nas tabelas 5, 6 e 7 são apresentadas a média, a mediana e o desvio padrão das escolas A, B e C respectivamente. Pode se observar com essas

medidas estatísticas que os dados das três escolas separadas são muito semelhantes.

Tabela 5: Medidas estatísticas da base de dados dos alunos da escola A

Atributo	Média	Mediana	Desvio Padrão
Idade	14,46	14	0,77
Notas de Português	72,79	74,5	11,68
Notas de Matemática	71,45	72	11,6
Porcentagem de Faltas	8,63	7,5	5,42

Tabela 6: Medidas estatísticas da base de dados dos alunos da escola B

Atributo	Média	Mediana	Desvio Padrão
Idade	14,95	14	0,74
Notas de Português	71,31	69	9,25
Notas de Matemática	74,31	71	11,5
Porcentagem de Faltas	9,07	8,5	5,64

Tabela 7: Medidas estatísticas da base de dados dos alunos da escola C

Atributo	Média	Mediana	Desvio Padrão
Idade	14,63	15	0,71
Notas de Português	74,59	74	11,05
Notas de Matemática	71,02	68	12,69
Porcentagem de Faltas	8,52	7,9	5,3

Também foram apresentados de forma gráfica como esses atributos estão dispostos na base dados. Um histograma é uma visão geral da distribuição dos valores dos atributos em relação a todo o conjunto dos dados Na Figura 7 é apresentada um gráfico de pizza onde mostra como o atributo idade está distribuído, é possível notar que idade que maioria dos alunos possui 14 e 15 anos. A pequena variação desse atributo pode ser explicado por baixo desvio padrão como foi observado na Tabela 4.

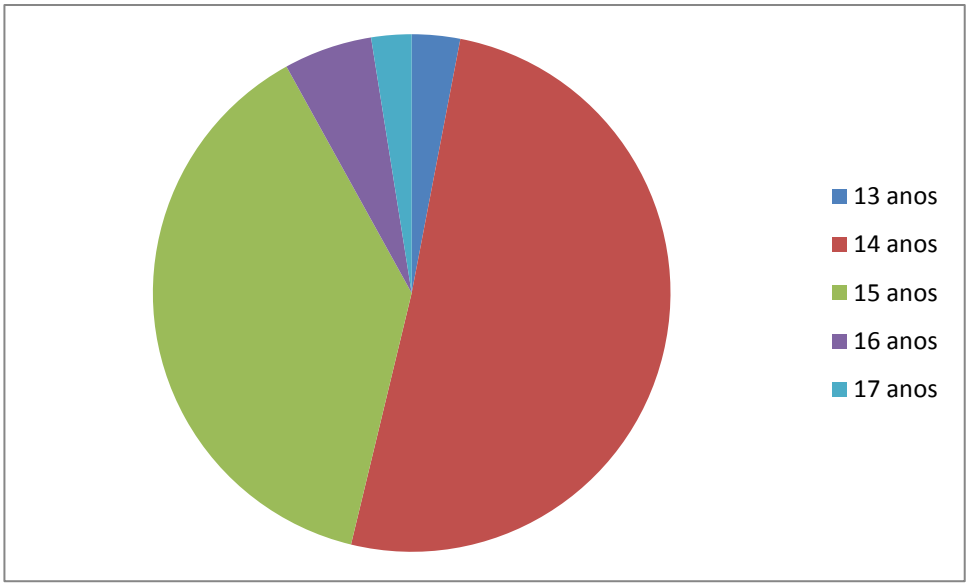

Figura 7- Distribuição da idade dos alunos das três escolas

Na Figura 8 é apresenta como as notas de português dos alunos do nono estão distribuídas, é possível notar que existe a maior frequência de alunos com notas entre 60 a 80. Pode-se notar também que o número de alunos com notas menores de 60 é quase inexpressivo.

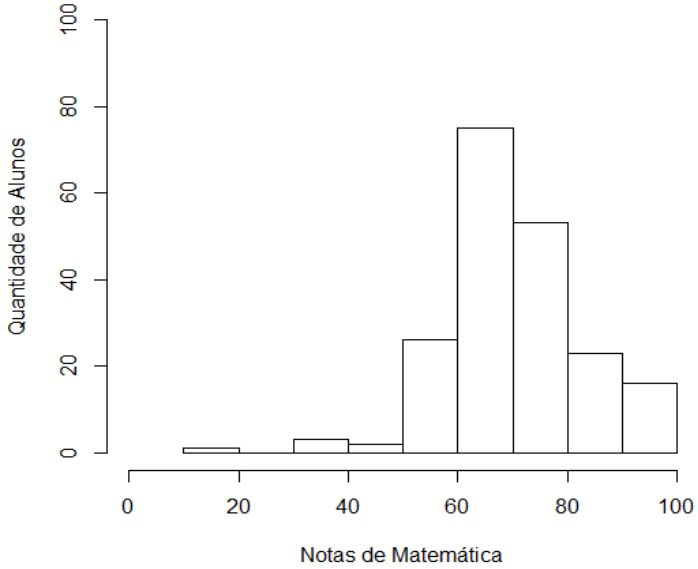

Figura 8 - Distribuição das notas de matemática dos alunos das três escolas

Conforme pode ser observado no histograma da Figura 9, a distribuição de notas de português é bem parecida com a distribuição das notas de matemática, onde as notas entre 60 a 80 aparecem com mais frequência. Nota-se também que o número de alunos com notas menores de 60 é praticamente inexpressivo.

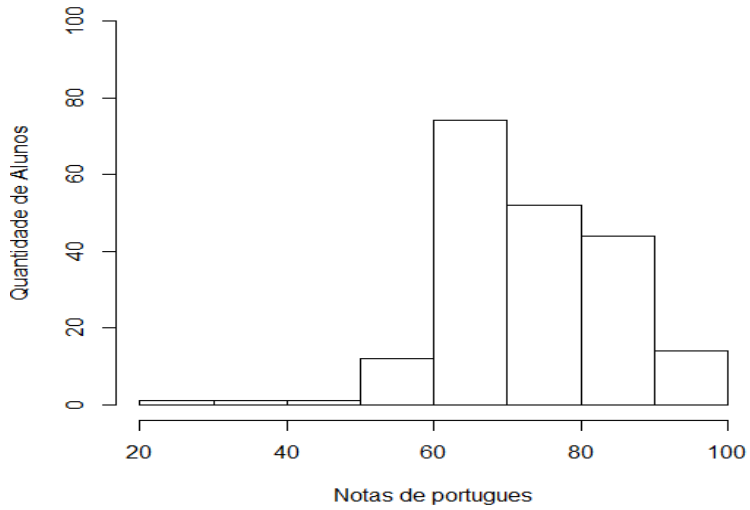

Figura 9- Distribuição das notas de português dos alunos das três escolas

Na Figura 10 é apresentado o histograma do atributo porcentagem de faltas, onde a maioria das faltas se concentra entre 0 a 15 %, e dessas existe mais alunos com percentual de faltas entre 0 e 5%, que é positivo, pois caracteriza uma alta frequência dos alunos nas aulas.

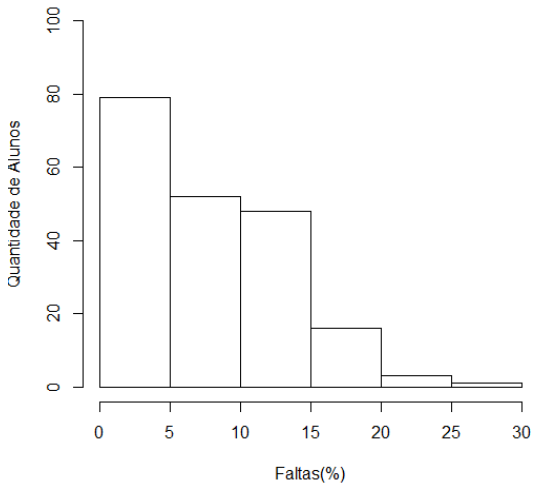

Figura 10- Distribuição da idade dos alunos das três escolas

O gráfico de dispersão nos fornece uma ideia do tipo de relacionamento entre os dois atributos, e ainda existe uma medida que explica de forma mais precisa o grau de relacionamento entre duas variáveis, que é o coeficiente de

correlação. O coeficiente de correlação pode variar de –1,00 a + 1,00 e será tanto mais fraca quanto mais próximo o coeficiente estiver de zero.

A partir disso, foi possível obter uma ideia geral das relações entre alguns atributos. Na Figura 11 é apresentado um gráfico de dispersão no qual se pode verificar a existência de correlação negativa entre porcentagem de faltas e notas de português. O coeficiente de correlação desses atributos é de -0,67, mostra que valores decrescentes do atributo porcentagem de faltas estão associados a valores crescentes do atributo notas de português.

O mesmo pode ser dito com relação a porcentagem de faltas e notas de matemática como apresentado na Figura 12, esses atributos possuem uma correlação negativa, com o coeficiente de correlação de -0,68, no qual observa que valores decrescentes do atributo porcentagem de faltas estão associados a valores crescentes do atributo notas de matemática.

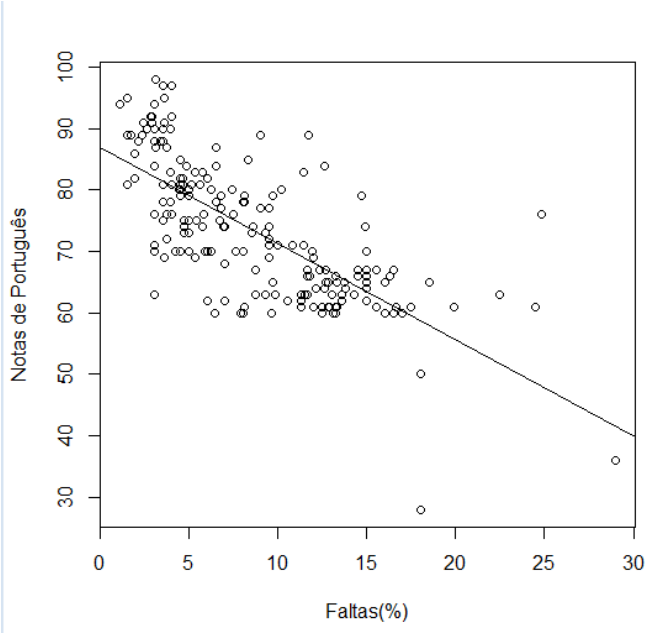

Figura 11 – Gráfico de dispersão das notas de português e porcentagem de faltas

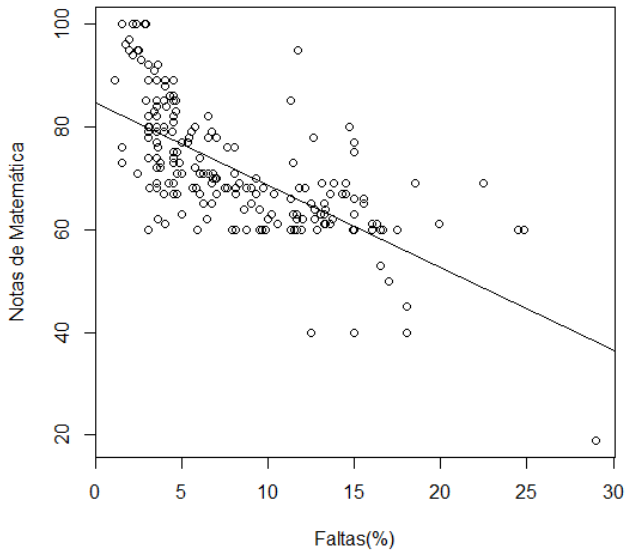

Figura 12 – Gráfico de dispersão das Notas de português e Porcentagem de Faltas

Outra observação diz respeito à relação notas de português e notas de matemática. Na Figura 13 é apresentado o gráfico de dispersão desses atributos, onde percebe-se uma correlação positiva entre os mesmos, com coeficiente de correlação de 0,74, mostra que os valores crescentes de um atributo estão associados a valores crescentes do outro atributo, ou seja, à medida que as notas de português aumentam, também aumentam a nota de matemática.

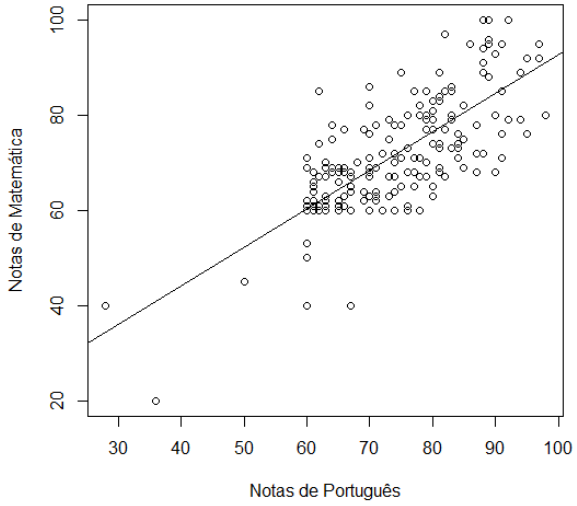

Figura 13 – Gráfico de dispersão das notas de português e notas de matemática

No que diz respeito a idade e nota dos alunos, não é possível perceber uma relação entre os atributos. Na Figura 14 é apresentado o gráfico de dispersão que mostra como as notas de português estão dispersa ao longo das

idade dos alunos. A não correlação entre idade dos alunos e nota de português pode ser explicado pelo coeficiente de correlação de -0,08, no qual está próximo a 0, mostrando assim a inpendencia entre os atributos.

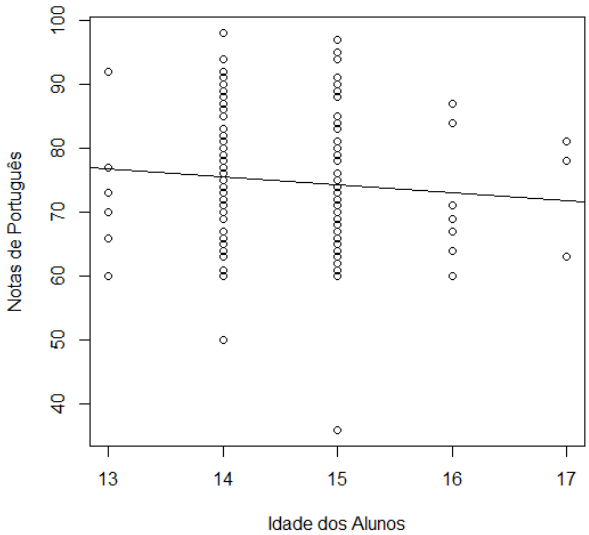

Figura 14 – Gráfico de dispersão das idade e notas de português

O mesmo pode ser observado com a relação aos atributos idade e notas de matemática como apresentado na Figura 15, não é possível perceber uma relação entre esses dois atributos, seu coeficiente de correlação é de -0,05 que é próximo a zero, o que explica a não correlação entre eles.

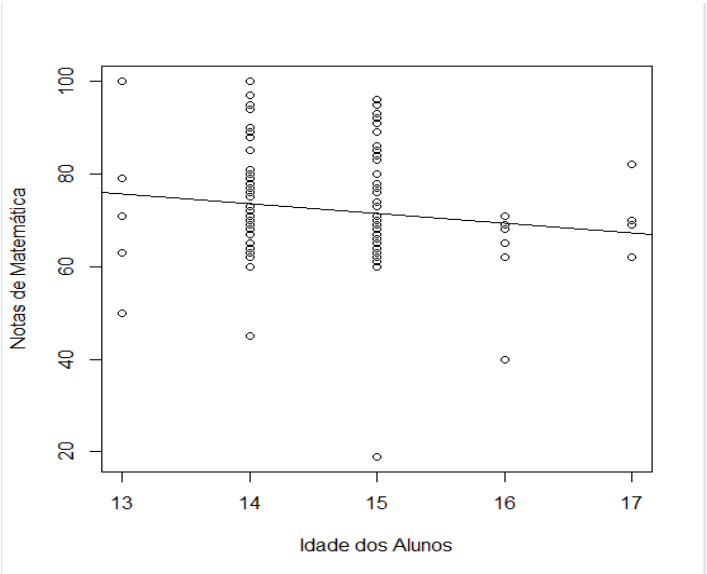

Figura 15 – Gráfico de dispersão das idade e notas de matemática

Pode observar também que poucos os alunos do nono possuem reprovações nas séries anteriores conforme é apresentado na Figura 16. Na Figura 17 mostra a quantidade de alunos que possui cadastro no programa do governo Bolsa Família.

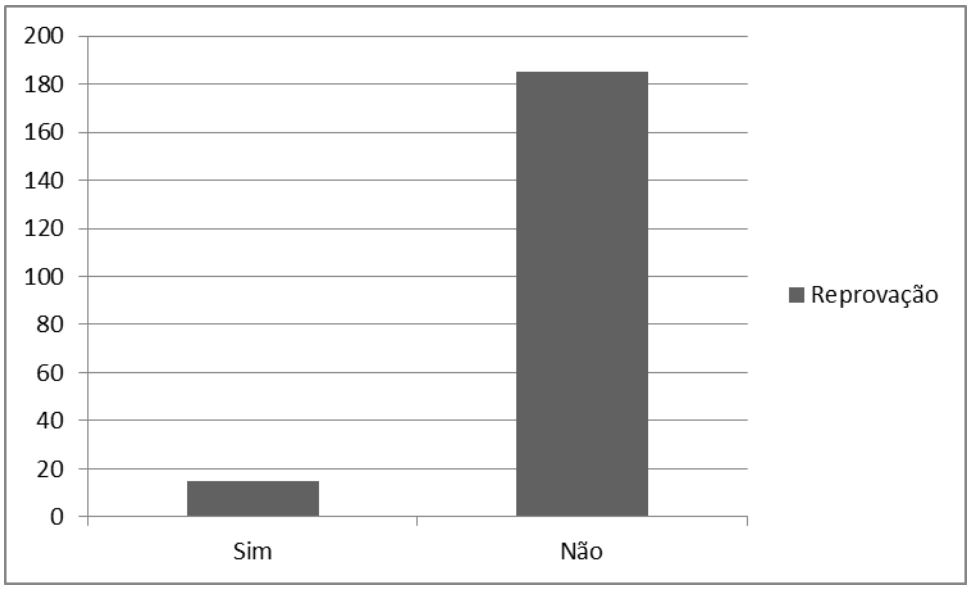

Figura 16 – Reprovação dos alunos em anos anteriores

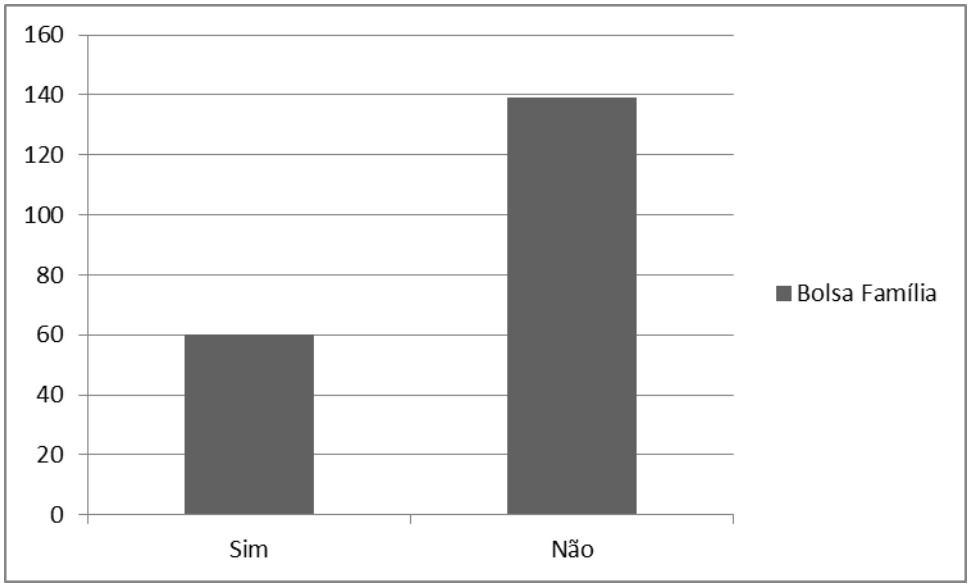

Figura 17- Cadastro dos alunos no programa Bolsa família.

Por fim é apresentado os turnos que os alunos que compõem a base da pesquisa estudam, conforme a Figura 18, a maioria dos alunos estudam no turno matutino.

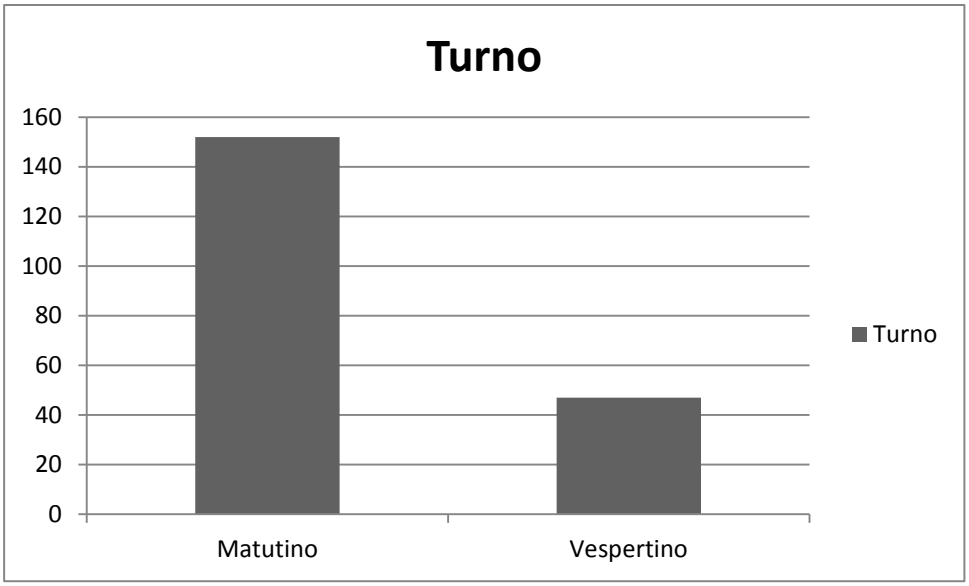

Figura 18- turno que os dos alunos estudam

3.4 TRANSFORMAÇÃO DOS DADOS

Após o pré-processamento e análise preliminar dos dados, verificou-se a necessidade de transformar alguns dados com o objetivo de atender alguns requisitos do algoritmo Apriori.

O atributo data de nascimento foi convertido para idade tomando como base a data de 2013 para gerar essa informação.

Atributos quantitativos como notas e faltas recebem uma grande quantidade de valores diferentes, isto significa que se forem utilizados os valores reais destes atributos para a geração de regras de associação, existe uma grande chance de que nenhuma, ou pouquíssimas, regra contendo esses atributos sejam geradas. Isto ocorre devido à frequência muito baixa que cada valor tende a apresentar na base de dados.

Dessa forma, foi necessário discretizar os atributos que eram numéricos, usando intervalos iguais, e trocando a categoria numérica por uma nominal correspondente ao intervalo. Esse processo foi realizado por meio de filtros utilitários disponibilizados juntamente com a ferramenta WEKA.

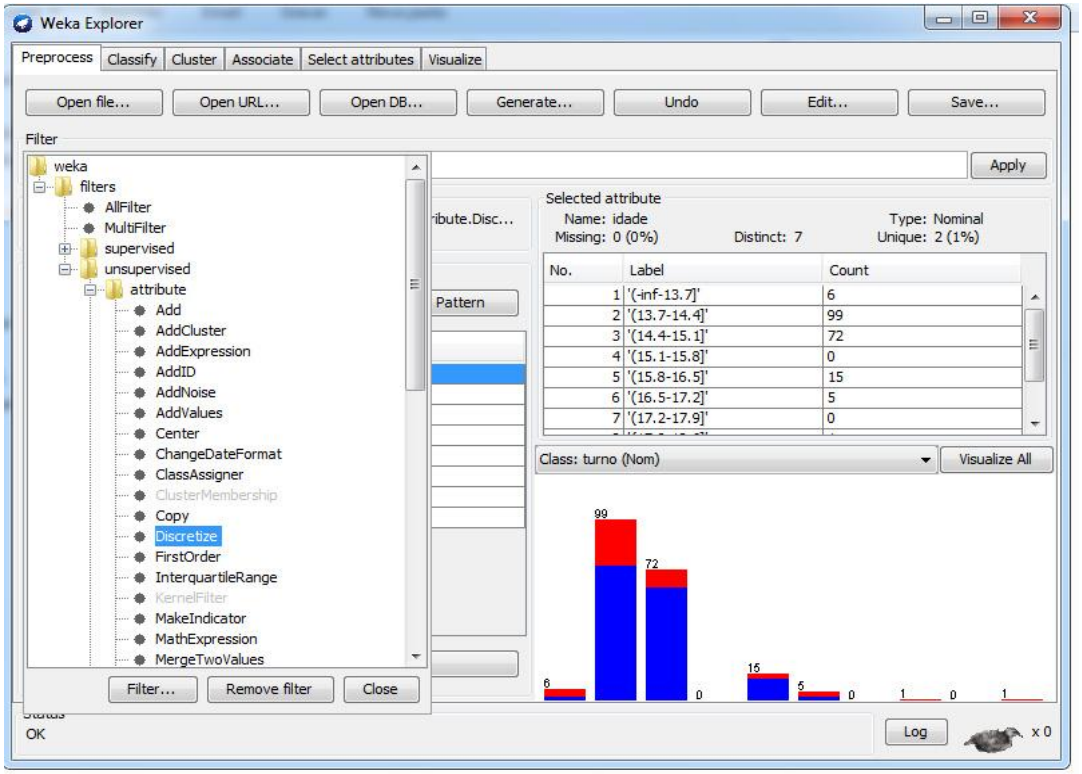

Figura 19- Tela de pré-processamento mostrando o caminho pra discretização de atributos

Fonte:UNIVERSITY OF WAIKATO, 2004

3.5 MINERAÇÃO DE DADOS

Na etapa de mineração de dados foi aplicada a tarefa de regras de associação por meio do algoritmo Apriori disponibilizada pela ferramenta WEKA, na tentativa de identificar associações entre atributos da base de dados que pudessem explicar de alguma forma o desempenho escolar dos estudantes.

Por a base de dados ser pequenas (199 alunos) e as medidas estatísticas das três escolas serem parecidas, o que mostra que os dados das três escolas estão distribuídos de forma semelhante como foi apresentado no pre processamento, optou por aplicar a regra de associação na base de dados completa.

O WEKA permite que as regras resultantes ser classificadas de acordo com diferentes medidas de interesse. Neste trabalho, foi usado o suporte/confiança e o LIFT como os critérios.

3.5.1 Suporte/Confiança

Nesta subseção serão apresentadas as principais regras extraídas com o Modelo Suporte/Confiança aplicados com todos os atributos da base de dados os alunos (data de nascimento, endereço, bolsa família, reprovação, porcentagem de faltas, notas de português e notas de matemática). Foram definidos parâmetros de entrada relacionados ao suporte e confiança mínimos desejados para obtenção de regras, com os seguintes valores:

- Suporte Mínimo = 5%;
- Confiança Mínima = 60%.

Cabe ressaltar que a escolha de um valor pequeno para o suporte mínimo visou possibilitar a identificação de combinação de atributos que, apesar de não tão frequentes, pudessem ter grande relevância , ou seja, confiança alta. Para esse experimento foram solicitadas ao programa 200 regras de associação.

Grande parte das regras não geraram nenhuma informação que possa justificar o desempenho dos alunos. Apesar do alto suporte e da alta confiança nem o termo antecedente nem consequente estavam relacionados com o desempenho atual do aluno. Na figura 20 apresenta um exemplo das regras com alta confiança .

```
idade='(13.8-14.2]' 101 ==> reprovacao=N 101    conf:(1)
idade='(13.8-14.2]' turno=M 76 ==> reprovacao=N 76    conf:(1)
idade='(13.8-14.2]' bolsafamilia=N 68 ==> reprovacao=N 68    conf:(1)
idade='(14.6-15]' turno=M 63 ==> reprovacao=N 63    conf:(1)
idade='(13.8-14.2]' bolsafamilia=N turno=M 55 ==> reprovacao=N 55   conf:(1)
```

Figura 20- Exemplos regras geradas

Além das regras geradas na figura 20, o WEKA gerou regras que sugerem a associação entre atributos que podem descrever algo sobre o desempenho escolar dos alunos de Alegre.

Algumas regras extraídas sugerem a existência de uma associação entre frequência e desempenho acadêmico nas disciplinas de português e matemática como é apresentado na Figura 21.

notasPortugues='(91-inf)' 11 ==> porcentagemFaltas='(-inf-3.89]' 11 conf:(1)

bolsafamilia=N porcentagemFaltas='(3.89-6.68]' notasPortugues='(77-84]' 13 ==> notasPortugues='(77-84]' 12 conf:(0.9)

porcentagemFaltas='(9.47-12.26]' 24 ==> notasMatematica='(59.5-67.6]' 18 conf:(0.75)

Figura 21 – Regras geradas que relacionam os atributos Notas e porcentagem de falta

Outras regras extraídas também sugere um relacionamento entre os atributos notas dos alunos (português e matemática), porcentagem de faltas e bolsa família. Pode se observar na Figura 22 as regras relacionados aos alunos que não possuem cadastro no programa bolsa família. Na Figura 23 apresenta as regras relacionadas aos alunos que possuem o cadastro no programa.

porcentagemFaltas='(12.26-15.05]' notasMatematica='(59.5-67.6]' turno=M 13 ==> bolsafamilia=N 13 conf:(1)

notasPortugues='(56-63]' 30 ==> bolsafamilia=N 29 conf:(0.97)

notasPortugues='(70-77]' turno=M 30 ==> bolsafamilia=N 29 conf:(0.97)

porcentagemFaltas='(12.26-15.05]' 29 ==> bolsafamilia=N 28 conf:(0.97)

notasMatematica='(59.5-67.6]' 27 ==> bolsafamilia=N 26 conf:(0.96)

Figura 22 – Regras geradas relacionadas aos alunos que não possuem bolsa família

notasMatematica='(83.8-91.9]' 17 ==> bolsafamilia=S 16 conf:(0.94)

porcentagemFaltas='(-inf-3.89]' 47 ==> bolsafamilia=S 33 conf:(0.7)

notasPortugues='(84-91]' 25 ==> bolsafamilia=S 17 conf:(0.68)

notasMatematica='(75.7-83.8]' 34 ==> bolsafamilia=S 21 conf:(0.62)

Figura 23 – Regras geradas relacionada aos alunos que possuem bolsa família

3.1.2 LIFT

Nesta subseção serão apresentadas as regras extraídas usando a medida de interesse LIFT. A medida de interesse LIFT que é uma das medidas de interesse mais utilizadas para avaliar dependências entre atributos, ou seja, quanto maior o valor do LIFT mais interessante à regra torna.

Foi realizado um experimento, as regras foram geradas a partir da base de dados completa com LIFT mínimo de 1,1 para eliminar as dependências negativas.

As regras geradas com o maior LIFT dizem respeito à correlação de alguns atributos. Na Figura 24 são apresentadas as regras que confirmam a correlação entre os atributos relacionados a notas e faltas.

```
porcentagemFaltas='(-inf-3.89]' 50 ==> notasPortugues='(84-91]' 18    conf:(0.36) < lift:(2.99)>
notasMatematica='(59.5-67.6]' 56 ==>  porcentagemFaltas='(9.47-12.26]' 18    conf:(0.32) < lift:(2.67)>
notasPortugues='(63-70]' 46 ==> porcentagemFaltas='(12.26-15.05]' 16    conf:(0.35) < lift:(2.39)>
porcentagemfaltas='(3.89-6.68]' 55 ==> notasPortugues='(77-84]' 24    conf:(0.44) < lift:(1.93)>
```

Figura 24 – Regras geradas que relacionam notas e faltas dos alunos.

Na Figura 25 são apresentadas regras geradas que evidência a associação entre os atributos notas de português e notas de matemática.

```
notasPortugues='(56-63]' 30 ==> notasMatematica='(59.5-67.6]' 21    conf:(0.7) < lift:(2.63)>
notasPortugues='(77-84]' 45 ==> notasMatematica='(75.7-83.8]' 16    conf:(0.36) < lift:(1.77)>
notasMatematica='(67.6-75.7]' 64 ==> notasPortugues='(77-84]' 18    conf:(0.28) < lift:(1.24)>
```

Figura 25 – Regras geradas que relacionam notas e faltas dos alunos

Outras regras geradas nesse experimento usando LIFT diz respeito associação entre o atributo bolsa família com os atributos relacionados a notas e faltas. Na figura 26 é apresentado as regras relacionados aos alunos que não possuem cadastro no programa bolsa família.

```
bolsafamilia=N notasPortugues='(63-70]' 42 ==> porcentagemFaltas='(12.26-15.05]' 16    conf:(0.38) < lift:(2.61)>
bolsafamilia=N notasPortugues='(56-63]' 29 ==> notasMatematica='(59.5-67.6]' 21    conf:(0.72) < lift:(2.57)>
bolsafamilia=N 139 ==> notasPortugues='(63-70]' notasMatematica='(67.6-75.7]' 19    conf:(0.14) < lift:(1.43)>
bolsafamilia=N 139 ==> porcentagemFaltas='(12.26-15.05]' 28    conf:(0.2) < lift:(1.38)>
porcentagemFaltas='(9.47-12.26]' notasMatematica='(59.5-67.6]' 20 ==>bolsafamilia=N 18    conf:(0.9) < lift:(1.29)>
```

Figura 26 – Regras geradas que sobre os alunos que não possui bolsa fámilia

Na figura 27 é apresentado as regras relacionados aos alunos que não possuem cadastro no programa bolsa família.

```
bolsafamilia=S 60 ==> notasMatematica='(83.8-91.9]' 16    conf:(0.27) < lift:(2.95)> lev:(0.05) [10] conv:(1.21)
bolsafamilia=S 60 ==> porcentagemFaltas='(-inf-3.89]' 33    conf:(0.55) < lift:(2.19)> lev:(0.09) [17] conv:(1.6)
bolsafamilia=S 60 ==> notasPortugues='(77-84]' 25    conf:(0.42) < lift:(1.84)> lev:(0.06) [11] conv:(1.29)
notasMatematica='(75.7-83.8]' 40 ==> bolsafamilia=S 21    conf:(0.53) < lift:(1.74)> lev:(0.04) [8] conv:(1.4)
```

Figura 27 – Regras geradas que sobre os alunos que possui bolsa fámilia

4 RESULTADOS

Esse capítulo apresenta os resultados obtidos com a mineração de regras de associação a partir da base de dados dos alunos e suas possíveis interpretações. Para tanto, utilizou-se as medidas de interesse Suporte/Confiança e LIFT com suporte mínimo de 5%, confiança mínima de 60% e lift mínimo de 1.1 Em razão do número de regras obtidas, somente aquelas julgadas mais importantes foram selecionadas.

É importante observar que as informações extraídas com essas regras não pode refletir sobre todos os alunos das escolas e sim de uma série.

Os experimentos realizados na seção de mineração de dados na metodologia do presente trabalho usando a medida suporte/confiança e a medida LIFT extraíram regras semelhantes que evidenciaram relações entre alguns atributos que podem caracterizar os alunos quanto ao desempenho escolar. Como resultados para melhor compreensão, pode-se destacar algumas regras, tais como é apresentada ma Tabela 8:

Tabela 8: exemplos de regras que evidenciam a relação entre os atributos Notas e Faltas

Regra (A=> B)	Conf	LIFT
notasMatematica='(84-91)' ==> porcentagemFaltas='(-inf-3.89]'	**1**	**3,98**
porcentagemFaltas='(9.47-12.26]' ==> notasMatematica='(59.5-67.6]'	**0,79**	**2,79**

Conforme a primeira regra mostrada na Tabela 8, os alunos que possuíam a média de 84 entre 91, na disciplina de português, apresentaram-se com um percentual de faltas menor ou igual a 3,89%, com uma confiança (conf.) de 100%, onde o LIFT apresentou o suporte real da regra de 3,98 vezes o valor do seu suporte esperado. Já a segunda regra sugere que alunos com um percentual de faltas de 9,47% entre 12,26% tendem a obter de 59,5 a 67,5 de média em matemática, com uma conf. de 79% e LIFT de 2,79.

Com essas regras observou-se um apontamento com relação aos atributos notas e porcentagem de faltas (quanto maior a nota do aluno menor a porcentagem de faltas). A frequência do aluno influenciou no seu rendimento escolar, que quanto mais assíduos foram às aulas, maior eram as possibilidades de conseguirem melhores notas. Essa relação – confirmada

usando a regra de associação – já havia sido identificada no pré processamento, onde notou-se a correlação entre esses atributos.

Algumas regras encontradas diz respeito ao atributo Bolsa Familia. A Tabela 9 apresenta alguma dessas regras.

Tabela 9: exemplos de regras encontradas com o atributo Bolsa Família

Regra (A=> B)	Conf	LIFT
notasPortugues='[84-91]' ==> bolsafamilia=S	0,75	2,5
bolsafamilia=S ==> porcentagemFaltas='(-inf-3.89]'	0,69	3,7
bolsafamilia=N ==> notasMatematica='(59.5-67.6]	0,87	3,1
porcentagemFaltas='[9.47-12.26]' ==> bolsafamilia=N	0,71	2,7

A a primeira regra da Tabela 9 sugere que alunos que tem de 84 a 91 de média em português tendem a ser cadastradas no programa bolsa família e, a segunda regra, pode sugerir que que alunos que são cadastradas no programa bolsa família tendem a ter um percentual de faltas inferior ou igual a 3,89%.

As terceira e a quarta regra apresentada na Tabela 9 dizem respeito aos alunos que não possuem o cadastro no programa. Especificamente, a regra 3 indica que alunos que não são cadastradas no bolsa família tendem a obter de 59,5 a 67,6 de média em português. E a regra 4 sugere que alunos que obtêm um percentual de faltas de 12.26 tendem a não serem cadastrados no bolsa família.

Com essas regras pode observar que os alunos que possuem cadastro no programa governo bolsa família possuem um rendimento e uma frequência melhor que os alunos que não possui o auxílio. Para a Felicetti e Trevisol (2012) isso pode ser comprovado pelo fato de muitas famílias terem receio de perder o benefício e isso faz com que a frequência desses alunos seja alta. O programa exige o mínimo de 85% de frequência mensal às aulas para estudantes de 6 a 15 anos e 75% para aqueles com idades entre 16 e 17 anos. Com essa exigência do programa, a maior frequência do aluno acaba influenciando no seu desempenho, como sugerido pelas regras já citadas, que mostram a associação entre os atributos relacionados com as notas e faltas.

Alguns regras extraídas confirmam a associação entre o atributo nota de português e nota de matemática identificada na etapa de pre processamento que foi obtida uma análise exploratória dos dados. Como resultados para melhor compreensão, pode-se destacar algumas regras, a Tabela 10 apresentada essas regras:

Tabela 10: exemplos de regras que evidenciam a associação entre os atributos notas de português e notas de matemática

Regra (A=> B)	Conf	LIFT
notasPortugues[56-63]==> notasMatematica='[59.5-67.6]'	**0,7**	**2,6**
notasPortugues[77-84]==> notasMatematica='[75,7-83,8]'	**0,36**	**2,63**

Com essas regras é possível observar que os alunos tedem a obter desempenhos semelhantes nas disciplinas de português e matemática.

Outra observação a ser feita é sobre os atributos frequentes na base de dados como reprovou=N e turno=M e idade=(13.7-14.4], por serem frequentes acabam compondo conseqüente da maioria das regras obtidas. Assim grande parte das as regras mineradas não parecem expressar relacionamentos interessantes para a pesquisa, mesmo quando os valores da medida de confiança são muito altos, como apresentado na Tabela 11:

Tabela 11: exemplos de regras não interessante para a pesquisa

Regra (A=> B)	Conf	LIFT
idade='(13.7-14.4]' turno=M 74 ==> reprovou=N 74	1	1,31
idade='(13.7-14.4]' bolsafamilia=N ==> reprovou=N	1	1,31

Nas regras apresentadas na Tabela 11, dizem que alunos com idade em torno de 14 anos não possuem reprovação, o que é obvio pois a idade escolar dos alunos cursam o nono ano é 14 anos de idade. A ***idade escolar*** é o período regular frequentado por alunos em ambiente escolar.

5 CONCLUSÃO

A conclusão das análises dos resultados obtidos, considerando todos os testes e possibilidades de associação é que desempenho escolar parece estar associado à frequência do aluno na escola, o que também pode justificar um melhor desempenho dos alunos que participam do programa bolsa família.

Foi possível perceber também, com a aplicação do algoritmo Apriori, que a utilização das medidas de interesse suporte/confiança e LIFT, apesar de extrair regras semelhantes, as mesmas se comportavam de maneiras diferentes. Os resultados obtidos através da confiança, demonstram em que as regras geradas são confiáveis mediante a base de dados de onde ela foi extraída, podendo-se perceber através do estudo, que alguns atributos foram mais frequentes que outros, como por exemplo, o turno e o fator não reprovação, que são qualitativos (onde o aluno poderia ter reprovado ou não, além de poder estudar no turno matutino ou vespertino). Outro exemplo a ser destacado, foi o atributo idade dos alunos, que em sua maioria eram de 14 a 15 anos, aparecendo com mais frequência, por possuir o valor de suporte alto.

No entanto, os atributos relacionados as notas e faltas recebem uma grande quantidade de valores diferentes, possuindo assim um suporte baixo. Portanto, os itens mais frequentes acabam compondo a maioria das regras obtidas, tornando muito trabalhosa a sua avaliação.

Com a utilização da medida LIFT a identificação das regras que caracterizam o aluno quanto ao desempenho escolar foram identificadas com maior facilidade, pois o LIFT é capaz de analisar a dependência entre o antecedente e o consequente de uma regra de associação, ou seja, com o LIFT as regras, mesmo com a confiança baixa, podem sugerir uma certa dependência entre os itens consequentes e os antecedentes das regras.

Apesar do processo de aquisição de dados ter sido obtido por um número muito pequeno de dados sobre os alunos, os experimentos foram úteis para demonstrar que os algoritmos de mineração de regras de associações, realmente são úteis nesse tipo de pesquisa, confirmando algumas expectativas

a respeito de associações entre determinados fatores que influenciam o desempenho acadêmico dos estudantes.

A mineração de dados descreve tarefas poderosas para se explorar dados acadêmicos a fim de resolver problemas desse tipo de domínio. Também é importante se ressaltar o aprendizado adquirido com a aplicação do processo de KDD, o uso de uma ferramenta para pesquisas em mineração de dados e a tarefa específica de regras de associação.

Entretanto, os resultados desse trabalho foram influenciados por fatores como: a dificuldade na obtenção de informações sobre os alunos das escolas envolvidas na pesquisa. As direções das escolas só disponibilizaram dados de uma turma, o que resultou uma base de dados pequena para uma análise com suporte adequado de dados. A pouca quantidade de dados pode reduzir significativamente a quantidade e as características dos padrões que podem ser extraídos da base. Certos padrões só podem ser estabelecidos a partir da análise de uma combinação específica de atributos dos dados e, portanto, a escassez de informações disponíveis na base utilizada no presente trabalho contribuiu para um cenário não adequado para tarefas de mineração de dados.

Uma nova pesquisa poderia utilizar uma base com mais detalhes sobre os alunos, como características socioeconômicas, de modo a permitir a exploração de um domínio mais abrangente de informações.

6 REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. **Fast algorithms for mining association rules**. Proceedings 20th International Conference Very Large Data Bases, VLDB, p. 487_499, 1994.

AMO, S. D. **Técnicas de Mineração de Dados**. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 14., 2004, Uberlândia. Disponível em: <www.inca.gov.br/.../14_artigo_aplicacao_metodos_computacionais_min... de FA Rodrigues - 2012 - Artigos relacionados>. Acesso em: 14 maio 2014.

BARROS, R. P.; MENDONÇA, R. **Uma análise dos determinantes do desempenho educacional no Brasil**. Brasília – Projeto Nordeste, 2000. (Série Estudos, n. 8). Disponível em: <http://www.dominiopublico.gov. br/download/texto/me000557.pdf>. Acesso em: 03 jan. 2014.

BAKER, Ryan Shaun Joazeiro de; ISOTANI, Seiji; CARVALHO, Adriana Maria Joazeiro Baker de. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. Worcester, 2011. Disponível em: <http://users.wpi.edu/~rsbaker/BD-RBIE-pt-v22.pdf>. Acesso em: 02 fev. 2014.

BERRY, M. J. A.; LINOFF, G. **Data Mining techniques – for marketing, sales, andcustomer support.** United States: Wiley Computer Publishing, 1997.

BIONDI, R. L.; FELÍCIO, F. **Atributos escolares e o desempenho dos estudantes:** uma análise em painel dos dados do Saeb. Brasília: INEP, 2007. (Série Documental. Textos para Discussão, n. 28). Disponível em: <http://www.publicacoes.inep.gov.br/detalhes. asp?pub=4321>. Acesso em: 28 nov. 2013.

BRIN, S., MOTWANI, R., ULLMAN, J. D. e TSURM S., Dynamic itemset counting and implication rules for market basket data, Proceedings of the ACM SIGMOD International Conference on Management of Data, Arizona, USA, 1997.

BÜRKLE, Paula Yamada. **Um Método de Pós Processamento de Regras de Associação com Base nas Relações de Dependências entre Atributos.** 2006. 107 f. Dissertação (Mestrado) - Curso de Ciências da Computação, Universidade Federal Fluminense, Niterói, 2006. Disponível em: <www.bdtd.ndc.uff.br/tde_arquivos/41/.../Paula Burkle-dissert.pdf>. Acesso em: 10 out. 2014.

CAMILO, C. O.; SILVA, J. C. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.** Goiás: Instituto de Informática Universidade Federal de Goiás, 2009.

CAVALCANTE, Carmem Haab Lutte; SANTOS JUNIOR, Pedro Aureliano dos. Fatores que influenciam o desempenho escolar: a percepção dos estudantes do curso Técnico em Contabilidade do IFRS – Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul, Campus Porto Alegre. **Revista Liberato,** Novo Hamburgo, v. 14, n. 21, p.30-49, jun. 2013. Disponível em: <http://www.liberato.com.br/SITES/default/files/arquivos/Revista_SIER/v. 14, n. 21 (2013)/03Desempenho.pdf>. Acesso em: 22 fev. 2014.

CARNIEL, Mauricio. **APLICAÇÃO DE ALGORITMOS DE MINERAÇÃO DE DADOS PARA IDENTIFICAÇÃO DE FATORES QUE INFLUENCIAM A EVASÃO DE ALUNOS DO CURSO DE CIÊNCIA DA COMPUTAÇÃO DA UNIVALI.** 2013. 138 f. TCC (Graduação) - Curso de CiÊncia da ComputaÇÃo, Departamento de Centro de CiÊncias TecnolÓgicas da Terra e do Mar, Universidade do Vale do ItajaÍ, Itajaí, 2013. Disponível em: <http://siaibib01.univali.br/pdf/Mauricio Carniel.pdf>. Acesso em: 22 out. 2014.

CASTRO, M. H. **A educação para o século XXI: o desafio da qualidade e da eqüidade**. Rio de Janeiro: Academia Brasileira de Ciências (ABC), 1999. Disponível em: <http://www.inep.gov.br.> Acesso em: 02 dez. 2013.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining toknowledge discovery in databases**. AI Magazine, v. 17, p. 37_54, 1996.

FELICETTI, Adriana Aparecida; TREVISOL, Joviles Vitório. BOLSA FAMÍLIA E DESEMPENHO ESCOLAR: AVALIAÇÃO DE UMA POLÍTICA PÚBLICA DE INCLUSÃO SOCIOEDUCACIONAL. In: IX ANPED SUL, 9., 2012, Caxias do Sul. **Anais.** Caxias do Sul,: Aaa, 2012. v. 11, p. 1 - 19. Disponível em:<http://www.portalanpedsul.com.br/admin/uploads/2012/Sociologia_da_Educacao/Trabalho/06_53_58_2913-7360-1-PB.pdf>. Acesso em: 25 jun. 2014.

FERREIRA NETO, P. N.; BARROSO**,** B. C. **DESCOBERTA DE CONHECIMENTO NA BASE DE DADOS DE UMA LOCADORA DE FILMES.** 2006. 72 f. Monografia (Especialização) - Curso de Ciência da Computação, Departamento de Centro de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, 2006. Disponível em: <www.portaltcc.icen.ufpa.br>. Acesso em: 13 jan. 2014.

GIL, A. C. **Didática do ensino superior**. São Paulo: Atlas, 2011.

GILLMEISTER, Paulo Ricardo Guglieri; CAZELLA, Sílvio César. Uma análise comparativa de algoritmos de regras de associação: minerando dados da indústria automotiva.**Escola Regional de Banco de Dados,** Caxias do Sul,

p.10, out. 2007. Disponível em: <http://www.ais.com.br/pos/>. Acesso em: 10 out. 2014.

GOMES, S. ANTÓNIO. **O desenvolvimento socioeconômico e a educação**. Lisboa: Iscul, 1964. v. 2, 652-670 p. Disponível em: <http://analisesocial.ics.ul.>. Acessoem: 09 nov. 2013.

GONÇALVES, Eduardo Corrêa. Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas. **Infocomp: Journal of Computer Science,** Niterói, v. 4, n. 1, p.26-35, 29 mar. 2005. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v4.1/art04.pdf?origin=publication_detail>. Acesso em: 17 out. 2013.

GOTTARDO, Ernani. **ESTIMATIVA DE DESEMPENHO ACADÊMICO DE ESTUDANTES EM UM AVA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS.** 2012. 84 f. Dissertação (Mestrado) - Curso de Sistemas de Informação, Universidade Tecnológica Federal do Paraná, Curitiba, 2012. Disponível em: <http://repositorio.utfpr.edu.br/jspui/bitstream/1/439/1/CT_PPGCA_M_Gottardo, Ernani_2012.pdf>. Acesso em: 15 out. 2014.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques.** [S.l.]: Morgan Kaufmann, 2006.

HAND, D; MANNILA, H; SMYTH, P. **Principles of Data Mining**. MIT Press, 2001.

MENEZES FILHO, N.A. **Os Determinantes do Desempenho Escolar do Brasil**. Instituto Futuro Brasil, IBMEC-SP, 2007. Disponível em < http://veja.abril.com.br/gustavo_ioschpe/arquivos_270908> Acesso em: 28 nov. 2013.

MERCERON, A.; YACEF, K. **Educational data mining: a case study. Artificial Intelligence in Education**. Proceedings of the 12th International Conference on Artificial Intelligence, p.467-474, 2005.

MIYAMOTO, Anderson P. et al. ANÁLISE DOS PERFIS DOS ESTUDANTES DO INSTITUTO SUPERIOR DE TECNOLOGIA DA FAETEC. **Revista Eletrônica de Computação,** Rio de Janeiro, v. 1, n. 1, p.1-41, mar. 2011. Disponível em: <http://rec.univercidade.edu.br/artigos/artigoE01_06.pdf>. Acesso em: 12 dez. 2013.

MORAIS, B. C. S. **Extração de conhecimento da plataforma lattes utilizando técnicas de mineração de dados: estudo de caso poli/upe**. Recife. Departamento de Sistemas e Computação Escola Politécnica de Pernambuco Universidade de Pernambuco, 2010.

MOTTA, C. G. L. **METODOLOGIA PARA MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO MULTINÍVEIS INCLUINDO PRÉ E PÓS-PROCESSAMENTO.** 2010. 103 f. Tese (Doutorado) - Curso de Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010. Disponível

em:<http://fenix3.ufrj.br/60/teses/coppe_d/CustodioGouveaLopesDaMotta.pdf>. Acesso em: 13 jan. 2013.

NAMEN, Anderson Amendoeira; SOARES, Ana Carolina Soares e. **MINERAÇÃO DE DADOS RELACIONADOS AO APRENDIZADO DE LÍNGUA PORTUGUESA – UM ESTUDO EXPLORATÓRIO**. In: ENCONTRO DE MODELAGEM COMPUTACIONAL, 14., 2011, Rio de Janeiro**.** Nova Friburgo , 2011. Disponível em: <http://www.uva.br/observatorioeducacao/File/Artigo14EMCNamenSoares.pdf> . Acesso em: 20 out. 2014.

PIMENTEL, E.P., Omar, N. **Descobrindo Conhecimentos em Dados de Avaliacao Aprendizagem com Tecnicas de Mineracao de Dado**. Workshop sobre Informática na Escola. Anais do Congresso da Sociedade Brasileira de Computação, 147-155, 2006. Disponível em:< http://www.br-ie.org/pub/index.php/wie/article/view/885/871>. Acesso em: 20 out. 2014.

SANTOS, Rafael.**Weka na Munheca:**Um guia para uso do Weka em scripts e integração com aplicações em Java. São Paulo, 2005. 20 p. Disponível em: <http://www.dcc.ufrj.br/~valeriab/DTM-weka-na-munheca.pdf>. Acesso em: 25 nov. 2013.

SOUZA, Emanuel Fernando Maia de; PETERNELLI, Luiz Alexandre. **INTRODUÇÃO AO SISTEMA LIVRE R EM ANÁLISES ESTATÍSTCAS.** Viçosa: Editora Ufv, 2009. 13 p. Disponível em: <http://www.dpi.ufv.br/~peternelli/tutoriaisR/>. Acesso em: 10 dez. 2014.

TRINDADE, Áthila R.; COSTA, Marcus Vinícius B.. USO DE MINERAÇÃO DE DADOS PARA DESOBERTA DE CONHECIMENTO: ESTUDO DE CASO DO VESTIBULAR DA UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI (UFVJM). **Revista de Ciencia e Tecnologia do Vale do Mucuri,**Mucuri, v. 4, p.62-74, nov. 2012. Disponível em: <http://www.rctvm.com.br/revista/index.php/rctvm/article/view/47/27>.Acessoem: 11 nov. 2014.

UNIVERSITY OF WAIKATO (Nova Zelândia). **Weka 3 – Machine Learning Software in Java**. NZ. 2004. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka>. Acesso em: 06 jan. 2014.

VASCONCELOS L. M. R.; CARVALHO C. L. **Aplicação de Regras de Associação para Mineração de Dados na Web.** (TechnicalReport - RT-INF_004-04 - Relatório Técnico). GO: 2004. Disponível em: <http://www.inf.ufg.br/virtualbib/RT-INF_004-04.pdf.> Acesso em: 05 jan. 2014.

VIECELLI, E.;TREVISOL, J. V.; MORAES, D. C. F. **O desenvolvimento de um banco de dados para cálculo de indicadores educacionais baseados nos micro dados do Inep**. 2008. Universidade do Oeste de Santa Catarina. Joaçaba, 2008.Disponívelem:<http://srvwt.unoesc.edu.br/sites/default/files/Artigo_Minerador_de_dados_educacionais_revisao.pdf>. Acesso em: 13 dez. 2013

WITTEN, I.H.; FRANK E.; HALL, M.A**. Data mining: Practical machine learning tools and techniques**. San Francisco: Morgan Kaufmann, 3 ed., 2011.

Z. Zheng, R. Kohavi e L. Mason, Real World Performance of Association Rule Algorithms, Proc.of the 7thACM SIGKDD **Conf. on Knowledg e Discovery and Data Mining**, São Francisco, Estados Unidos, 2001, 401–406.

APÊNDICE A – MODELO DE ARQUIVO NO FORMATO ARFF

```
@RELATION DADOSESCOLARES
@attribute idade real
@attribute bolsafamilia {S, N}
@attribute endereco {'VILA DO SUL','SAO MANOEL','EMILIO MARINS','COLINA','BILAU','CHARQUEADO','CONCEICAO','ZONA RURAL','ANUTIBA','CENTRO','GUARAREMA','VILA ALTA','NOVA ALEGRE','COBRINHA','CASAS POPULARES','CAMPO DE VIACAO'}
@attribute reprovacao {S, N}
@attribute porcentagemFaltas real
@attribute notasPortugues real
@attribute notasMatematica real
@attribute turno {M,V}

@data
14,S,'VILA DO SUL',N,3.5,78,80,M
14,S,'VILA DO SUL',N,3,84,74,M
14,N,'SAO MANOEL',N,5,73,73,M
14,N,'VILA DO SUL',N,17.5,61,60,M
15,N,'EMILIO MARINS',N,15.5,67,65,M
14,N,'VILA DO SUL',N,4.5,77,80,M
15,N,'EMILIO MARINS',N,6,62,74,M
15,S,'BILAU',N,4.6,81,83,M
15,N,'SAO MANOEL',N,15,64,75,M
14,N,'VILA DO SUL',N,6,75,80,M
```