INSTITUTO FEDERAL DO PARANÁ

KEWERSON HUGO PEREIRA DE MELO

PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS: EXTRAÇÃO DE INFORMAÇÕES ACERCA DA REALIDADE DOS ESTUDANTES BRASILEIROS

LONDRINA

2015

INSTITUTO FEDERAL DO PARANÁ

KEWERSON HUGO PEREIRA DE MELO

PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS: EXTRAÇÃO DE INFORMAÇÕES ACERCA DA REALIDADE DOS ESTUDANTES BRASILEIROS

Trabalho de Conclusão de Curso apresentado ao Curso Superior em Análise e Desenvolvimento de Sistemas do Instituto Federal do Paraná – Campus Londrina.

Orientadores: Adriana Carniello e Gilson Doi Junior

LONDRINA

2015

SUMÁRIO

RESUMO

Informação é um recurso importante para apoiar o processo de tomada de decisão. Ela consiste em um subsídio para os gestores poderem avaliar melhor os riscos e os impactos das alternativas disponíveis e optar pela alternativa que satisfaça aos critérios de decisão com maior aderência. No entanto, a informação é um produto da interpretação dos dados disponíveis e nem sempre ela é disponibilizada de forma acessível e processada para facilitar sua interpretação. Quando a situação envolve avaliar um pequeno conjunto de dados, a análise dos dados pode ser um processo trivial. Apesar disso, à medida que o tamanho da massa de dados aumenta, a complexidade de análise e interpretação das informações aumenta proporcionalmente. O desenvolvimento deste trabalho segue de encontro a esta necessidade de prover subsídios mais eficazes para a melhoria da qualidade da educação por meio da extração de conhecimentos específicos pois existem muitos dados educacionais disponibilizados pelo governo onde se pode aplicar o processo de Descoberta de Conhecimento em Base de Dados. Por meio da realização de um levantamento bibliográfico, se levantou a hipótese de que o mau desempenho escolar de estudantes está ligado à escolaridade dos pais. Para testar essa hipótese foi utilizado a base de dados do Exame Nacional do Ensino Médio que possui diversas informações sobre os candidatos, seu desempenho e suas condições sociais. Os resultados obtidos indicam que os candidatos com pais de maior grau de escolaridade possuem melhor desempenho na redação e na nota média do exame. Esse resultado possui caráter quantitativo, sendo que a realização futura de uma pesquisa qualitativa complementar se faz necessária para investigar quais as causas que tornam os estudantes de pais com o nível de escolaridade referido terem desempenho maior.

Palavras-chave: Mineração de dados, Descoberta de Conhecimento em Base de Dados, Pentaho, Kettle, Weka, Escolaridade, Educação, Fatores de Risco.

ABSTRACT

Information is an important resource to support the decision making process. It consists as an allowance for managers to better evaluate the risks and impacts of the available choices and choose the one that satisfy the choice criteria with more adhesion. However, information is a result of available interpretation data and it is not always reachable in accessible and processed way to facilitate understanding. When the situation involves evaluating a small data group, the analisys may be an easy process. Despite this, as the data size grows, the complexity of analisys and understanding of the information grows proportionally. The development of this paper aims to supply this necessity of more efficient supports to improve the education quality through especific knowledge, because there are lots of educational data available by the government that the application of Knowledge Discovering in Databases is suitable. Through bibliographic review, it was raised a hypothesis that the bad school performance is connected to parents' education. To test this hypothesis it was used the database of Exame Nacional do Ensino Médio that contains several information about its candidates, their performance and social condition. The obtained results points that the candidates with parents of a higher level of education has better scores in the dissertation and the average score of the Exame. This result has a quantitative nature, and a future additional qualitative research becomes necessary to investigate what causes makes students with parents of higher education reach a better performance.

Keywords: Data mining, Knowledge Discovering in Databases, Pentaho, Kettle, Weka, Schooling, Education, Risk factors.

1 INTRODUÇÃO

Com os grandes avanços da tecnologia, passou-se a armazenar cada vez mais dados, tornando o trabalho de analisar esses dados cada vez mais difícil chegando a um nível no qual é impossível um ser humano realizar essas análises em tempo satisfatório de forma manual. A Mineração de dados, parte do processo de Descoberta de Conhecimento em Base de Dados, surgiu justamente para auxiliar na extração de informação de base de dados grandes que devido ao grande acúmulo de dados tornou-se uma tarefa muito complexa para humanos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O Brasil, por meio do Portal Dados Abertos, coleta diversos tipos de dados na área da saúde, de auxílios sociais, de educação, entre outros. Esses dados estão disponíveis para qualquer pessoa e é um direito garantido pela lei de acesso à informação pública (Lei 12.527/2011) (BRASIL A, 2007).

Por meio deste domínio são oferecidas bases de dados que encontram-se organizados e estruturados, na maioria dos casos. Todavia, devido a massiva quantidade de dados disponibilizados, a análise das bases e a descoberta de informações relevantes para a tomada de decisão resulta em um processo que demanda um esforço excessivo por parte do responsável da pesquisa.

Com esses dados é possível buscar informações na tentativa de tomar decisões, mas esse processo de tomada de decisões requer tempo e preparo. É necessário analisar, manipular e tratar diversos dados para se obter uma informação que pode ser utilizada como conhecimento na tomada de uma decisão. Esse processo pode ser realizado de forma trivial, no entanto, fatores como quantidade de dados, complexidade de relações ou a importância da decisão final podem exigir um maior tempo e preparo.

A Mineração de Dados é um processo computacional para descoberta de padrões interessantes e inéditos, como também modelos descritivos, entendíveis e preditivos em dados de larga escala (ZAKI; MEIRA JR., 2014). Esse processo pode ser utilizado em conjunto de dados para obtenção de informações. As informações obtidas pela mineração de dados são utilizadas para tomadas de decisão, muitas

vezes na administração de negócios, mas isso não impede de se utilizá-las em outras aplicações e setores.

Este trabalho visa descobrir características que influenciam no aprendizado dos alunos pois pode ser uma das maneiras mais eficazes para realizar uma intervenção por parte do educador, tanto no contexto de uma sala de aula quanto no contexto mais amplo da educação brasileira.

PROBLEMA

A área de gestão educacional sofre com uma carência de análises aprofundadas nas bases de dados educacionais públicas. Esta carência leva a informações incompletas e/ou pouco relevantes para o processo de tomada de decisão na gestão educacional, o que pode comprometer o alcance da meta de melhoria da qualidade do ensino brasileiro.

Apesar de existirem trabalhos que analisam dados educacionais levantados em campo pelos pesquisadores, os dados levantados não são disponibilizados ao público e também baseiam-se em amostras pouco representativas como foi levantado na revisão bibliográfica na seção Error: Reference source not found. Diante deste problema, este trabalho visa responder a seguinte pergunta: será que a aplicação do processo de Descoberta de Conhecimento em Base de Dados em uma base de dados pública com informações dos estudantes brasileiros consegue extrair informações que refletem com maior precisão a realidade educacional brasileira?

OBJETIVO GERAL

Aplicar o processo de Descoberta de Conhecimento em Base de Dados para análise e extração de conhecimento em uma base de dados educacional pública em busca de informações que auxiliem no processo de tomada de decisão na gestão educacional.

OBJETIVOS ESPECÍFICOS

- Compreender os fatores que influenciam no rendimento escolar dos alunos em base de dados educacionais e públicas;
- Realizar a preparação dos dados por meio das atividades de seleção, pré-processamento e transformação em uma base de dados educacional do Exame Nacional do Ensino Médio;
- Aplicar técnicas e algoritmos de mineração de dados na base pré-processada;
- Analisar as informações obtidas pela mineração de dados, visando reconhecer padrões nos dados;
- Apresentação e análise de conhecimentos obtidos por meio dos padrões encontrados.

2 REVISÃO BIBLIOGRÁFICA

2.1 RENDIMENTO ESCOLAR DE ALUNOS

O desempenho do aluno também deve ser pensado como resultado do desempenho dos pais. Uma das hipóteses mais difundidas sobre o baixo desempenho escolar estão atribuídos a famílias de baixo nível socioeconômico, baixa escolaridade e pouco empenhada na educação formal do filho (DAZZANI; FARIA, 2009).

Um estudo apresentado pelo Instituto Glia de Neurociência e Desenvolvimento aponta como um dos fatores de risco que mais prejudicam o desempenho dos alunos é o grau de escolaridade do chefe da família, aumentando as chances de baixo rendimento escolar em 5,8 vezes. (GLIA, 2010).

Mazzetto; Bravo; Carneiro (2002) apontam indícios de que a concorrência pelo curso de licenciatura de química tenha ficado mais concorrido e que uma das características similares entre os candidatos é a alta escolaridade dos pais e também notaram uma queda no número de reprovações no primeiro ano do curso.

Sampaio et al (2011) concluiu que os alunos que desistem dos cursos da Universidade Federal de Pernambuco para tentar outros cursos da mesma instituição possuem melhores notas no vestibular, renda familiar elevada e pais com mais anos de escolaridade se comparado com os alunos que evadem o curso e não são observados repetindo o concurso.

Esses estudos levantam a hipótese de que a educação escolar dos pais afeta no rendimento escolar dos alunos se limitados à amostragens locais. Os dados utilizados nessas pesquisas também não são divulgados ao público para que sejam reproduzidos os resultados aplicando outras técnicas para obtenção de conhecimento. A partir deste levantamento, este trabalho busca a veracidade da hipótese levantada em um contexto de nível nacional por meio dos resultados dos alunos em exames nacionais, como o Exame Nacional do Ensino Médio.

2.2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

No nosso cotidiano, tomar decisões é algo que requer tempo e preparo. É necessário analisar, manipular e trtar diversos dados para se obter uma informação que pode ser utilizada como conhecimento na tomada de uma decisão. Esse processo pode ser realizado de forma trivial, no entanto, fatores como quantidade de dados, complexidade de relações ou a importância da decisão final podem exigir um maior tempo e preparo.

Com os grandes avanços da tecnologia, passou-se a armazenar cada vez mais dados, tornando o trabalho de analisar esses dados cada vez mais difícil chegando a um nível no qual é impossível um ser humano realizar essas análises em tempo satisfatório de forma manual. A Mineração de dados, parte do processo de Descoberta de Conhecimento em Base de Dados, surgiu justamente para auxiliar na extração de informação de base de dados grandes que devido ao grande acúmulo de dados tornou-se uma tarefa muito complexa para humanos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A Descoberta de Conhecimento em Base de Dados estrutura uma sequência de etapas para obtenção de conhecimento em base de dados por meio das etapas de seleção, pré-processamento, transformação, mineração de dados e avaliação. A 1 apresenta um diagrama representando o processo de extração de conhecimento. Apesar de ser estruturado de forma sequencial, podem ocorrer ciclos internos caso uma das etapas não alcance os resultados desejados e seja necessário voltar o processo para etapas anteriores (SFERRA; CORRÊA, 2003).

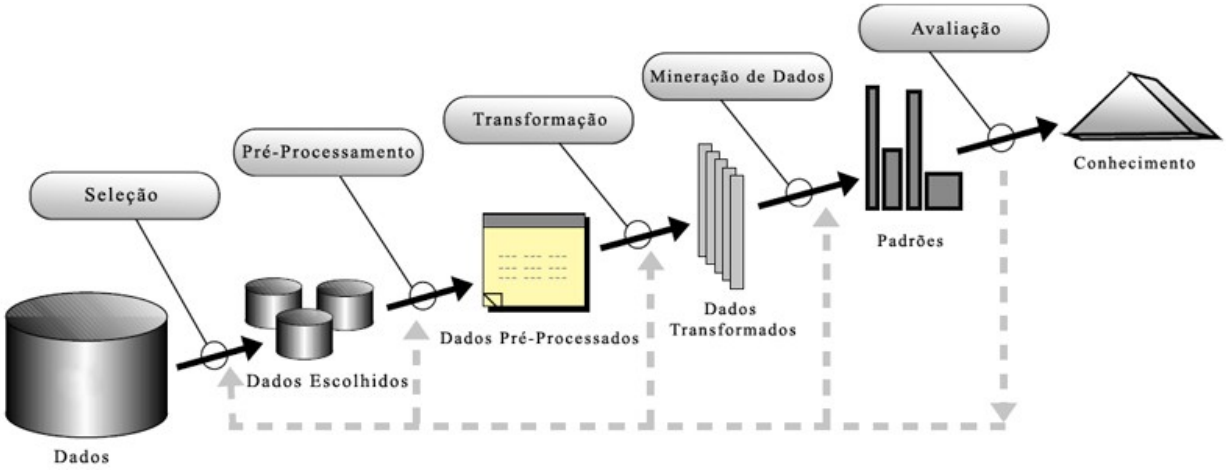

Figura 1: Etapas do Processo de Descoberta de Conhecimento em Base de Dados.

Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

2.2.1 Seleção

Na etapa de seleção, a partir de um conjunto de dados maior, busca-se selecionar apenas os dados que possuem relevância e que foram utilizados para a pesquisa. Sendo necessário conhecer bem os objetivos da pesquisa (CÔRTES; PORCARO; LIFSCHITZ, 2002).

Após a etapa de Pré-processamento é natural voltar para a etapa de Seleção para selecionar apenas os dados que não estão ausentes, fora do padrão ou inconsistentes e que já foram integrados com outros conjuntos de dados coerentes.

2.2.2 Pré-processamento

A etapa de pré-processamento possui duas subetapas dentro dela, sendo elas a de Limpeza dos Dados e Integração dos dados.

A subetapa de Limpeza dos Dados, é utilizada para identificar dados que estão ausentes, que são inconsistentes ou fora do padrão.

Os dados ausentes são dados de uma determinada tupla (célula de uma planilha) que não foram preenchidos. Esse problema acontece com frequência em formulários no qual é necessário o preenchimento manual ou um campo opcional do formulário.

A inconsistência trata-se de um valor que não faz sentido em uma determinada tupla como por exemplo, um nome em um campo no qual deveria existir apenas números ou o nome de uma cidade incompleto.

Para contornar o problema de ausência dos dados pode-se utilizar as seguintes soluções: preenchimento manual do dado, valor fixo para dados ausentes, valor médio, valor mais comum ou a Integração com outros conjuntos de dados que possam fornecer os dados ausentes com maior precisão.

O problema de inconsistência pode ser corrigido alterando o valor para o correto, caso seja possível, ou ignorando a linha, o que nem sempre pode ser uma opção viável pois pode afetar diretamente nos resultado (CÔRTES; PORCARO; LIFSCHITZ, 2002).

A subetapa de Integração dos Dados, é utilizada para combinar diversas fontes de dados em uma única base coerente. Nessa etapa é preciso tomar cuidado com: valores redundantes como a idade e ano de nascimento; valores que dependem de outros valores para realizar a junção dos conjuntos de dados; e valores conflitantes, que seriam dados de categorias diferentes mas com o mesmo valor (CAMILO; SILVA, 2009).

2.2.3 Transformação

A transformação é a etapa na qual os dados são preparados para atender restrições de formatos para que sejam utilizados em determinado algoritmo da etapa subsequente. Alguns algoritmos de mineração de dados trabalham apenas com valores nominais e outros com valores numéricos. Caso os dados selecionados não sejam adequados pode ser necessário adaptá-los para o tipo de algoritmo que será utilizado na etapa de mineração de dados visando facilitar o processo (SFERRA; CORRÊA, 2003).

Um exemplo de transformação seria alterar faixas de notas de 0 até 100 em categorias com as letras de A até D, no qual cada letra receberia uma faixa de nota como mostrado na 2.

A	76 até 100
B	51 até 75
C	26 até 50
D	0 até 25

Figura 2: Exemplo de transformação

Na transformação também podem ser utilizadas técnicas de redução para obter uma representação reduzida dos dados sem comprometer a integridade. Um exemplo prático seria o de reduzir as vendas trimestrais de cada ano em uma única tabela menor apresentando apenas as vendas anuais como mostrado na 3.

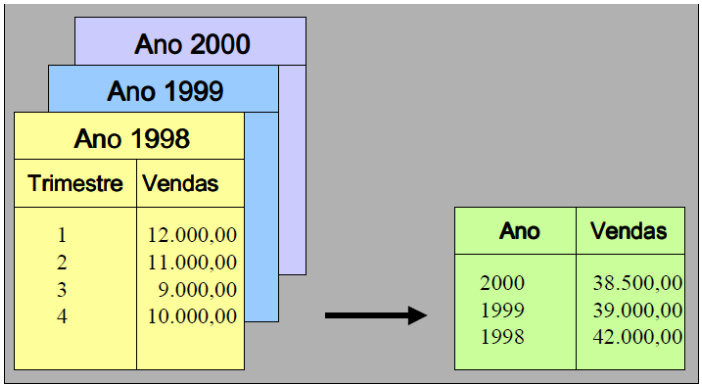

Figura 3: Transformação de tabelas.

Fonte: (CÔRTES; PORCARO; LIFSCHITZ, 2002)

2.2.4 Mineração de dados

É a etapa na qual são utilizados algoritmos para identificar padrões entre os dados. Para isso são necessários grandes conjuntos de dados, de forma que, esses padrões possam ser confirmados por meio do número de ocorrências e repetições.

Essas técnicas são divididas de acordo com o objetivo e o tipo de padrão que podem identificar. Elas são divididas entre: Associação, Classificação, Predição Numérica e Agrupamento (CAMILO; SILVA, 2009). Cada uma dessas técnicas possui um objetivo específico:

- Associação: Técnicas de associação são utilizadas para identificar relações entre itens mais frequentes de um determinado conjunto. Os resultados obtidos ao aplicar algoritmos que buscam associações são chamados de Regras de Associação e são visualizados da seguinte forma: SE cliente compra leite E pão ENTÃO compra manteiga. A 4 apresenta regras detectadas pela aplicação da técnica associação no qual associa características de uma pessoa como idade, escolaridade e avaliação de crédito como fatores que influenciam na compra de um computador. No exemplo da Regra 1, se a pessoa for um jovem e não for um estudante, então ele não compra computadores.

Regra 1: SE *idade* = *jovem* AND *estudante* = *não* ENTÃO *compra computadores* = *não*
Regra 2: SE *idade* = *jovem* AND *estudante* = *sim* ENTÃO *compra computadores* = *sim*
Regra 3: SE *idade* = *média* ENTÃO *compra computadores* = *sim*
Regra 4: SE *idade* = *adulto* AND *avaliação de crédito* = *excelente* ENTÃO *compra computadores* = *sim*
Regra 5: SE *idade* = *adulto* AND *avaliuação de crédito* = *ruim* ENTÃO *compra computadores* = *não*

Figura 4: Exemplo da técnica de Associação.

- Classificação: A classificação é utilizada para separar em grupos ou categorias que possuem algum tipo de relação, por exemplo: um banco pode utilizar modelos de classificação para separar os seus clientes em categorias de clientes especiais ou de risco, com isso é possível aplicar diferentes estratégias de negócios para ambas categorias. Um dos algoritmos mais conhecidos de classificação são as Árvores de decisão, as quais a partir de um ponto inicial classificam quais os pontos finais de cada um dos valores classificados. No exemplo da 5 o ponto do topo, nomeado de 'Idade?', representa a raiz da árvore. Neste caso realiza-se o teste de idade, sendo que cada resultado leva o valor para um teste diferente como o ponto 'Garantias?' ou um dos nós finais como 'Classe A', 'Classe B' ou 'Classe C', por exemplo: uma pessoa jovem e com garantia alta seria classificado no grupo com nome 'Classe A'.

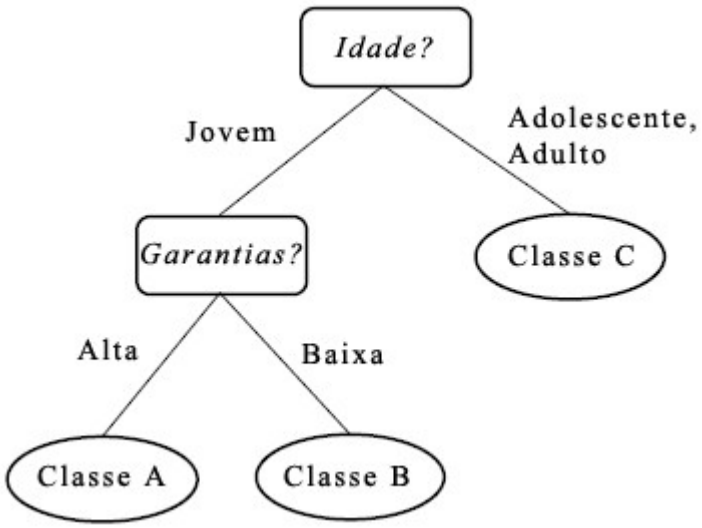

Figura 5: Exemplo da técnica de Classificação.

- Predição Numérica: Utilizada para prever ou estimar valores por meio da correlação entre valores de um determinado grupo. No exemplo na 6 foi traçado uma linha mostrando o crescimento dos valores em um determinado período de tempo. A técnica estatística para realizar uma predição é conhecido como regressão numérica. Alguns dos algoritmos que podem ser utilizadas são os de Regressão Linear e de Regressão Não-Linear.

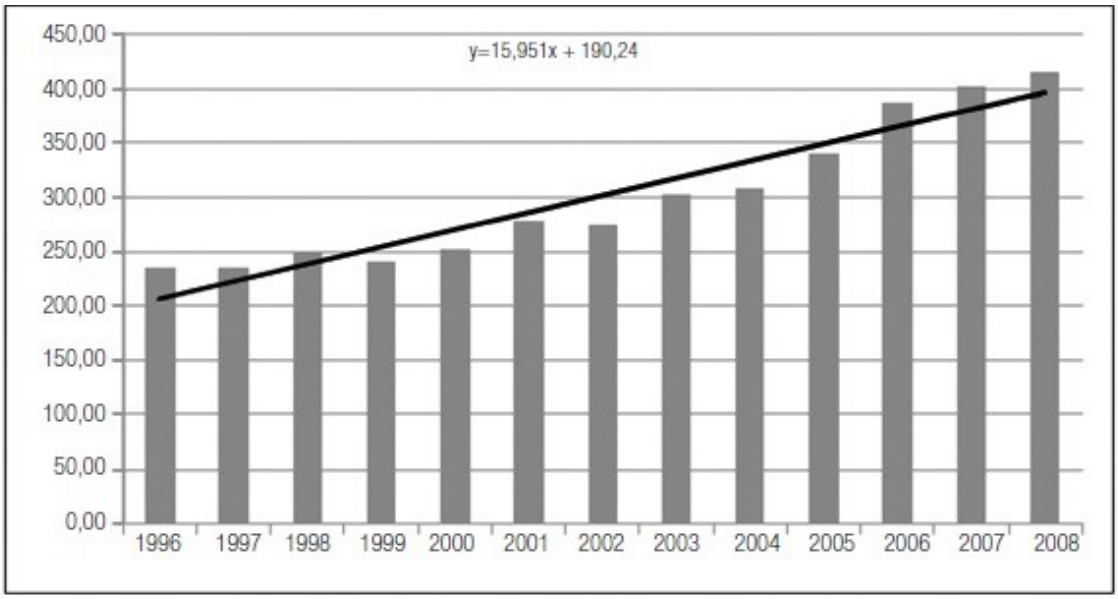

Figura 6: Exemplo da técnica de Regressão Linear.

- Agrupamento: As técnicas de agrupamento separam um conjunto de registro em uma matriz, na qual os valores mais semelhantes são mais próximos. Os elementos dentro de um agrupamento são considerados semelhantes entre si e dissimilares com elementos de outros agrupamentos como na 7.

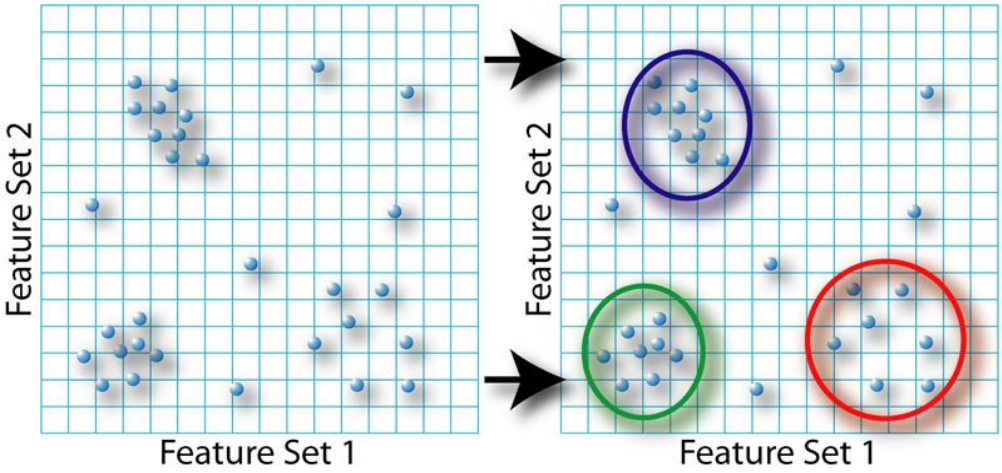

Figura 7: Exemplo da Técnica de Agrupamento.

2.2.5 Avaliação

O resultado da aplicação de cada uma das técnicas de mineração de dados resulta em uma nova organização dos dados, os padrões. Os padrões obtidos devem então passar por uma avaliação para verificar se os padrões encontrados têm relevância para a pesquisa. Nem todos os resultados obtidos no processo de mineração podem ser aplicados para a pesquisa pois o algoritmo faz uma análise baseada em padrões sem nenhuma análise de contexto no qual pode ocorrer falsos positivos, por isso é necessário cientista de dados para trabalho de identificar e interpretar quais padrões apresentados pelos algoritmos podem ser utilizados e apresentados como conhecimento.

2.2.6 Apresentação e assimilação de conhecimento

Nesta etapa os padrões obtidos no processo de mineração dos dados são aplicados para tomada de decisão (CÔRTES; PORCARO; LIFSCHITZ, 2002).

Os objetivos principais desta etapa são:

- Apresentar os conhecimentos obtidos;
- Determinar a melhor forma para aplicar o conhecimento obtido na tomada de decisão;
- Definir as vantagens e desvantagens do projeto;
- Reavaliar o projeto;
- Criar outros projetos.

2.3 EXTRAÇÃO DE CONHECIMENTO EM BASE DE DADOS EDUCACIONAIS

O processo de mineração de dados é muito utilizado para identificação de certos padrões para tomada de decisão e pode ser aplicado de forma satisfatória em diversas áreas como diagnóstico médico, previsão financeira, previsão do tempo, marketing, identificação de fraudes em cartões de crédito, entre outros (BRAMER, 2007).

Existe um campo de pesquisa chamado Mineração de Dados Educacionais que busca resultados para o melhoramento do ensino desde os primeiros anos na escola até abordagens como o EAD (Educação a Distancia). As pesquisas são focadas em resolver problemas com evasão, métodos de aprendizagem, qualidade de ensino e prever fatores de risco. Trata-se de uma área relativamente recente no qual a maioria das publicações tratam de problemas relacionados a uma instituição de ensino em específico.

Existem diversas pesquisas na área de educação demonstrando ótimos resultados que podem servir como exemplo para pesquisas utilizando os dados disponibilizados do ENEM.

Manhães; Santos Filho; Rodrigues (2012) trazem uma abordagem quantitativa buscando padrões para o elevado número de evasões em cursos superiores e seus resultados obtidos revelam características similares entre os alunos que cancelaram a matrícula. Com os resultados obtidos o autor propõe a criação de um sistema acadêmico para alertar os educadores quais alunos estão mais propensos à evasão para que seja possível tomar medidas prévias.

Rigo et al.(2012) apresenta em sua publicação uma proposta de melhorias possíveis na aplicação de mineração de dados para detecção de perfil de alunos com risco de evasão propondo uma solução interativa para obtenção dos resultados para um diagnóstico precoce. Os autores destacam a importância de ter dados relevantes sobre os estudantes para que o perfil previsto possua maior precisão. Em experimentos preliminares utilizando algoritmos de redes neurais para classificação foi possível traçar o perfil de estudantes com risco de evasão com 90% de precisão.

Adeodato et al.(2014) buscaram demonstrar padrões em escolas privadas utilizando a base de dados do ENEM 2011 que demonstram um desempenho bom ou ruim entre elas utilizando algoritmos de classificação, como árvore de decisão. O resultado final deixa claro que o fator que mais influencia na qualidade de ensino das escolas privadas é o econômico-financeiro dos pais, mas é necessário uma pesquisa mais aprofundada para avaliar como o fator econômico pode influenciar na qualidade de ensino dos alunos.

Morais e Fechine (2012) utilizam algoritmos de classificação de Árvore de Decisão e Redes Bayesianas para identificar fatores relevantes no desempenho dos alunos em ambiente virtual de ensino a distância (EAD). Os resultados revelaram que os alunos com maior contato com o ambiente virtual e possuem um desempenho maior que outros alunos da turma, isto permite que o professor tome abordagens diferentes focando em utilizar mais atividades nos ambientes virtuais. Os autores destacam o baixo poder estatístico da pesquisa por ela trabalhar com um grupo pequeno de pessoas, limitado apenas à plataforma de ensino a distância Moodle.

Kampff; Reategui; de Lima (2008) propõem um sistema utilizando técnicas de agrupamento para identificar alunos em risco de evasão ou reprovação gerando alertas aos docentes para que tomem as medidas específicas para cada aluno, focando no dialogo e em suas dificuldades. Além do sistema de alertas baseado em agrupamentos, o artigo também propõe entrevistas com os docentes para buscar as melhores abordagens pedagógicas.

3 METODOLOGIA

Para realizar o processo de descoberta de conhecimento foi necessário realizar os seguintes passos:

- Foi utilizado o processo de Descoberta de Conhecimento em Base de Dados para a extração de conhecimento passando por todas etapas citadas na seção 2.2.
- Foi definido que a escolaridade dos pais é um grande fator de risco para o rendimento escolar dos alunos por meio do levantamento bibliográfico realizado na seção 2.1. Com base na hipótese de que a escolaridade dos pais influencia no rendimento escolar dos alunos foi selecionado a base de dados do Exame Nacional do Ensino Médio para
- As etapas de Seleção, Pré-processamento e Transformação foram realizadas utilizando o software gratuito e de código aberto chamado Kettle, da empresa Pentaho na versão 6.0.0.0-353. Este software permite abrir a base de dados de diversos formatos de arquivos ou bancos de dados como CSV, XLS e XML, permitindo também realizar as etapas de seleção, transformação e integração utilizando uma interface gráfica com os processos em forma de diagrama e exportação do resultado do processo para quaisquer formatos suportados pelo software.
- Realizar a etapa de seleção para selecionar apenas os dados com relevância para a pesquisa.
- Realizar a etapa de pré-processamento, aplicando as subetapas de limpeza dos dados e integração de dados caso necessário.
- Realizar a etapa de transformação para utilização do conjunto no software de mineração de dados.

- A etapa de Mineração de Dados foi realizada utilizando o software gratuito e de código aberto chamado Weka, também da empresa Pentaho, na versão 3.6.12. O Weka possui diversas ferramentas para visualização dos dados e aplicação de algoritmos de técnicas de mineração de dados como citado na seção 2.2.4. A aplicação das técnicas de mineração de dados não altera o conjunto de dados, pois elas só servem para verificação de padrões para obtenção de informações sobre o conjunto de dados.
- Utilizar a técnica de classificação, pois esta técnica mostra-se adequada para responder à perguntas como: “Qual a influência que a escolaridade dos pais possuem sobre o desempenho escolar de seus filhos?”.
- Aplicar o algoritmo de classificação nomeado de J48, por gerar uma visualização de fácil interpretação para a etapa de avaliação, em seus parâmetros padrão.
- Utilizar o Weka no modo de treinamento de conjunto com a escolaridade dos pais como classe principal para a ser classificada, no qual a partir das notas serão classificadas as escolaridades.
- O Treinamento de Conjunto utiliza todo o conjunto de dados para criar a regra do algoritmo e também para testar essa regra.
- O resultado da técnica de classificação apresenta dois resultados importantes para serem interpretados, o modelo classificador e a matriz de confusão.
- O Modelo Classificador apresenta a regra que o algoritmo criou para posteriormente realizar os testes. No caso do algoritmo J48, este modelo pode ser apresentado em formato de árvore.
- A Matriz de Confusão apresenta como dos valores foram classificados durante o teste utilizando o Modelo Classificador. Essa informação é exibida em forma de tabela, na qual as suas colunas representam como o valor foi classificado e as linhas apresentam qual o valor que foi classificado. Visualizando as informações obtidas pela Matriz de Confusão é possível verificar os problemas de imprecisão do conjunto de dados e voltar para alguma das etapas anteriores do Processo de Descoberta de Conhecimento em Base de Dados para realizar as alterações necessárias com o objetivo de melhorar os resultados.

- Realizar a avaliação dos resultados da etapa de mineração de dados apontando os padrões obtidos.
- Apresentar os conhecimentos extraídos dos padrões obtidos e apresentar de forma simples a visualização desse conhecimento extraído.

4 RESULTADOS

Os dados do Exame Nacional do Ensino Médio de 2013 foram obtidos no site Dados Abertos (BRASIL A, 2007) que possui diversos dados governamentais de saúde, educação e segurança, mantido pelo próprio governo para manter o direito de todo cidadão ao acesso as informações governamentais garantido pela lei de acesso à informação pública (Lei 12.527/2011) (BRASIL B, 2007). O arquivo disponibilizado vem compactado e contem as pastas Dicionário, Inputs, Planilhas, Leia-me e Documentação, Provas e Gabaritos, e Dados. Os arquivos que utilizaremos encontram-se nas pastas Dicionário e Dados.

A pasta dicionário possui o arquivo de planilha nomeado de Dicionário_Microdados_Enem_2013 que contem as informações de cada uma das colunas existentes no conjunto de dados, mostrando qual o nome da variável, a descrição, os valores que a variável pode assumir, o tamanho e o tipo da variável.

A pasta Dados possui dois arquivos no formato CSV com os nomes CONSISTENCIA_ENEM_ESCOLA_2013 e MICRODADOS_ENEM_2013, que é o conjunto de dados utilizados nesta pesquisa.

Com os dados em mãos, foi possível aplicar as etapas do processo de Descoberta de Conhecimento em Base de Dados começando pela etapa de Seleção. Nas próximas seções apresentam-se os resultados obtidos em cada etapa deste processo

4.1 SELEÇÃO

Nesta primeira etapa foram selecionados quais dados serão utilizados para alcançar o objetivo. Para isso foi necessário definir exatamente quais candidatos farão parte do conjunto de dados. Neste trabalho, foram selecionados apenas os dados dos candidatos devidamente matriculados em escolas, que participaram de todas etapas da prova sem serem desclassificados e que estão cursando o ensino médio. Diante desta seleção, foram utilizadas apenas as colunas NOTA_CN, NOTA_CH, NOTA_LC, NOTA_MT, NU_NOTA_REDACAO, Q001, Q002 e

ST_CONCLUSAO.

As colunas NOTA_CN, NOTA_CH, NOTA_LC, NOTA_MT, NU_NOTA_REDACAO contem as notas dos candidatos e são do tipo numérico. Elas suportam valores de 0 até 1000 que variam de acordo com o desempenho do candidato.

As colunas Q001 e Q002 possuem as respostas dos candidatos referentes a escolaridade de seus pais. Q001 é a coluna referente ao pai e Q002 é a coluna referente a mãe. Os valores suportados por essas colunas podem ser visualizados na 1.

VALOR	DESCRIÇÃO
A	Não estudou
B	Da 1ª à 4ª série
C	Da 5ª à 8ª série
D	Ensino Médio
E	Ensino Médio
F	Ensino Superior incompleto
G	Ensino Superior
H	Pós-graduação
I	Não sei

VALOR	DESCRIÇÃO
1	Já concluí o Ensino Médio
2	Estou cursando e concluirei o Ensino Médio em 2013
3	Estou cursando e concluirei o Ensino Médio após 2013
4	Não concluí e não estou cursando o Ensino Médio

Após concluir a etapa de seleção gerou-se um novo conjunto de dados com as informações apresentadas na 3.

COLUNAS	DESCRIÇÃO	VALOR	SIGNIFICADO
NOTA_CN	Nota da prova de Ciências da Natureza	0 à 1000	Nota do Aluno
NOTA_CH	Nota da prova de Ciências Humanas	0 à 1000	Nota do Aluno
NOTA_LC	Nota da prova de Linguagens e Códigos	0 à 1000	Nota do Aluno
NOTA_MT	Nota da prova de Matemática	0 à 1000	Nota do Aluno
NU_NOTA_REDACAO	Nota da prova de redação	0 à 1000	Nota do Aluno
ST_CONCLUSAO	Situação de conclusão do Ensino Médio	1	Já concluí o Ensino Médio
		2	Estou cursando e concluirei o Ensino Médio em 2013
		3	Estou cursando e concluirei o Ensino Médio após 2013
		4	Não concluí e não estou cursando o Ensino Médio
Q001	Até quando seu pai estudou?	A	Não estudou
		B	Da 1ª à 4ª série
		C	Da 5ª à 8ª série
		D	Ensino Médio incompleto
		E	Ensino Médio
		F	Ensino Superior incompleto
		G	Ensino Superior
		H	Pós-graduação
		I	Não sei
Q002	Até quando sua mãe estudou?	A	Não estudou
		B	Da 1ª à 4ª série
		C	Da 5ª à 8ª série
		D	Ensino Médio incompleto
		E	Ensino Médio
		F	Ensino Superior incompleto
		G	Ensino Superior
		H	Pós-graduação
		I	Não sei

4.2 PRÉ-PROCESSAMENTO

Nessa etapa foi verificado que não existem valores que precisavam ser tratados no conjunto de dados utilizado, sendo assim não foi necessário aplicar nenhuma das subetapas existentes.

4.3 TRANSFORMAÇÃO

Na etapa de transformação alterou-se os valores das colunas para posterior aplicação dos algoritmos na etapa de Mineração de Dados.

Para que o conjunto de dados seja lido pelo software Weka, alterou-se o delimitador de colunas utilizado pelo arquivo CSV, que por padrão é um ponto e vírgula (;), para uma vírgula (,).

Todos os candidatos do ENEM que possuíam os valores 2 e 3 na coluna de ST_CONCLUSAO foram mantidos, no qual o valor 2 representa os candidatos que concluem o ensino médio no mesmo ano da prova e o valor 3 representa os candidatos que concluirão o ensino médio após o ano da prova. Os candidatos que possuem os valores 1 e 4, que representam respectivamente os candidatos que já concluíram o ensino médio e que não estou cursando o ensino médio, foram removidos. Apenas os candidatos que estão cursando o ensino médio foram mantidos no conjunto de dados.

O nome das colunas Q001 e Q002 para PAI e MAE. Também foram alterados os valores das colunas PAI e MAE de letras de A, B, C, D, E, F, G, H e I para as palavras NAO_ESTUDOU, PRIMARIO, GINASIO, MEDIO_INC, MEDIO_COMP, SUPERIOR_INC, SUPERIOR_COMP, POS_GRADUCAO e NAO_SABE. Essas alterações foram necessárias para que seja mais fácil visualizar os dados durante a etapa de mineração de dados. As alterações podem ser visualizadas na 4.

VALOR	DESCRIÇÃO
NAO_ESTUDOU	Não estudou
PRIMARIO	Da 1ª à 4ª série
GINASIO	Da 5ª à 8ª série
MEDIO_INC	Ensino Médio incompleto
MEDIO_COMP	Ensino Médio
SUPERIOR_INC	Ensino Superior incompleto
SUPERIOR_COMP	Ensino Superior
POS_GRADUCAO	Pós-graduação
NAO_SABE	Não sei

Tabela 4: Valores suportados pela coluna PAI e MAE após renomeação dos valores.

Foram criadas duas novas colunas com os nomes REDACAO_NOMINAL e NOTA_MEDIA_NOMINAL para a posterior aplicação do algoritmo de classificação na etapa de mineração. A coluna REDACAO_NOMINAL possui apenas a nota da coluna NU_NOTA_REDACAO enquanto a coluna NOTA_MEDIA_NOMINAL possui a média das colunas NOTA_CN, NOTA_CH, NOTA_LC, NOTA_MT, e NU_NOTA_REDACAO. Essas colunas suportarão valores nominais que vão indicar a nota do candidato de acordo com letras de A à J em múltiplos de 100 como pode ser visualizado na 5.

VALOR	DESCRIÇÃO
A	Nota de 901 até 1000
B	Nota de 801 até 900
C	Nota de 701 até 800
D	Nota de 601 até 700
E	Nota de 501 até 600
F	Nota de 401 até 500
G	Nota de 301 até 400
H	Nota de 201 até 300
I	Nota de 101 até 200
J	Nota de 0 até 100

Tabela 5: Valores possíveis para as colunas REDACAO_NOMINAL e NOTA_MEDIA_NOMINAL.

As colunas NOTA_CN, NOTA_CH, NOTA_LC, NOTA_MT, e NU_NOTA_REDACAO foram removidas, pois não serão utilizadas na etapa de mineração. Ao final da etapa de transformação foi gerado um novo conjunto de dados contendo apenas as informações apresentadas pela 6.

COLUNAS	DESCRIÇÃO	VALOR	SIGNIFICADO
NOTA_MEDIA_NOMINAL	Nota média	A	Nota de 901 até 1000
		B	Nota de 801 até 900
		C	Nota de 701 até 800
		D	Nota de 601 até 700
		E	Nota de 501 até 600
		F	Nota de 401 até 500
		G	Nota de 301 até 400
		H	Nota de 201 até 300
		I	Nota de 101 até 200
		J	Nota de 0 até 100
REDACAO_NOMINAL	Nota da redação	A	Nota de 901 até 1000
		B	Nota de 801 até 900
		C	Nota de 701 até 800
		D	Nota de 601 até 700
		E	Nota de 501 até 600
		F	Nota de 401 até 500
		G	Nota de 301 até 400
		H	Nota de 201 até 300
		I	Nota de 101 até 200
		J	Nota de 0 até 100
PAI	Até quando seu pai estudou?	NAO_ESTUDOU	Não estudou
		PRIMARIO	Da 1ª à 4ª série
		GINASIO	Da 5ª à 8ª série
		MEDIO_INC	Ensino Médio incompleto
		MEDIO_COMP	Ensino Médio
		SUPERIOR_INC	Ensino Superior incompleto
		SUPERIOR_COMP	Ensino Superior
		POS_GRADUCAO	Pós-graduação
		NAO_SABE	Não sei
MAE	Até quando sua mãe estudou?	NAO_ESTUDOU	Não estudou
		PRIMARIO	Da 1ª à 4ª série
		GINASIO	Da 5ª à 8ª série
		MEDIO_INC	Ensino Médio incompleto
		MEDIO_COMP	Ensino Médio
		SUPERIOR_INC	Ensino Superior incompleto
		SUPERIOR_COMP	Ensino Superior
		POS_GRADUCAO	Pós-graduação
		NAO_SABE	Não sei

Tabela 6: Informações das colunas do conjunto de dados após a etapa de transformação.

Na 7 é possível observar uma prévia do conjunto de dados depois de aplicar todas as alterações.

REDACAO_NOMINAL	NOTA_MEDIA_NOMINA	PAI	MAE
G	F	MEDIO_COMP	MEDIO_COMP
G	F	NAO_SABE	GINASIO
G	G	NAO_SABE	NAO_SABE
H	G	NAO_ESTUDOU	NAO_ESTUDOU
C	E	SUPERIOR_COMP	SUPERIOR_COMP
C	D	MEDIO_COMP	SUPERIOR_COMP
F	F	PRIMARIO	GINASIO
F	F	PRIMARIO	GINASIO
E	E	GINASIO	MEDIO_INC
F	F	SUPERIOR_COMP	POS_GRADUCAO

Tabela 7: Prévia do conjunto de dados após etapa de transformação

4.4 MINERAÇÃO DE DADOS

O algoritmo de classificação escolhido foi o J48, pois gera uma visualização de fácil interpretação para a etapa de avaliação. O algoritmo J48 foi aplicado utilizando os parâmetros padrão do algoritmo, sendo necessário definir apenas qual a coluna que será testada pelo algoritmo, que no caso deste trabalho são as colunas referentes à escolaridade dos pais.

Um dos parâmetros do Weka é o modo de teste, que define como será divido o conjunto de dados para o algoritmo criar a regra e testá-la depois. Foi escolhido o Treinamento de Conjunto que utiliza todo o conjunto de dados para gerar a regra e também para testá-la.

O algoritmo J48 foi aplicado somente para duas colunas por vez, SENDO que uma das colunas representa a nota da redação ou da nota média enquanto a outra coluna utilizada representa a escolaridade da mãe ou do pai. Os algoritmos foram aplicados entre as colunas na seguinte ordem, REDACAO_NOMINAL e MAE, REDACAO_NOMINAL e PAI, NOTA_MEDIA_NOMINAL e MAE, e NOTA_MEDIA_NOMINAL e PAI. Os resultados reproduzidos pela técnica de classificação exibe diversas informações como os parâmetros que foram utilizados,

a modelo classificado, avaliação do algoritmo, precisão da coluna que foi testada e matriz de confusão. Atentando-se apenas ao modelo classificado e matriz de confusão pois nas informações exibidas por eles que pode-se visualizar os padrões.

4.5 AVALIAÇÃO

Ao aplicar o algoritmo para as colunas REDACAO_NOMINAL e MAE o Modelo Classificador apresentou o resultado exibido na 8.

Para cada valor da coluna REDACAO_NOMINAL, o algoritmo correlaciona um valor da coluna MAE, destacando ao final entre parênteses a quantidade de valores classificados com esta regra e em seguida, a quantidade de valores erroneamente classificados.

Na primeira linha da 8 a REDACAO_NOMINAL com o valor G teve 384732 classificações como MEDIO_COMP sendo que 293732 foram classificações incorretas. Essas classificações incorretas são todos os valores que não são do valor MEDIO_COMP mas que foram classificadas nesse grupo.

```
REDACAO_NOMINAL = G: MEDIO_COMP (384732.0/293732.0)
REDACAO_NOMINAL = H: PRIMARIO (107929.0/78040.0)
REDACAO_NOMINAL = F: MEDIO_COMP (564539.0/414580.0)
REDACAO_NOMINAL = D: MEDIO_COMP (282940.0/201026.0)
REDACAO_NOMINAL = E: MEDIO_COMP (602124.0/432102.0)
REDACAO_NOMINAL = C: MEDIO_COMP (140722.0/101308.0)
REDACAO_NOMINAL = B: SUPERIOR_COMP (43181.0/31591.0)
REDACAO_NOMINAL = J: PRIMARIO (54629.0/38920.0)
REDACAO_NOMINAL = I: PRIMARIO (23413.0/15920.0)
REDACAO_NOMINAL = A: SUPERIOR_COMP (15671.0/10902.0)
```

Figura 8: Resultado do Modelo Classificador da escolaridade da mãe e nota da redação.

No resultado do Modelo Classificador é possível notar um padrão, no qual os candidatos com valores A e B na coluna REDACAO_NOMINAL foram classificados como SUPERIOR_COMP. O mesmo se repete para os valores C, D, E, F, e G, que foram classificados como MEDIO_COMP e os valores H, I e J que foram

classificados como PRIMARIO. Os outros valores da coluna MAE não aparecem, pois nem todos valores são considerados relevantes para a regra criada pelo algoritmo por conter valores que não podem ser classificados nas regras criadas, como por exemplo: um candidato que tem a mãe que estudou até o primário pode ter tirado uma nota alta que se enquadra no valor A da coluna REDACAO_NOMINAL, dessa forma ele será classificado como SUPERIOR_COMP devido à sua nota alta. Essas classificações erradas podem ser visualizadas nas Matriz de Confusão da 9.

A Matriz de Confusão gerada pelo algoritmo demonstra a disposição da classificação dos registros da base de dados em uma matriz. A quantidade de linhas e colunas é equivalente a quantidade de valores da coluna classificada definida no parâmetro de aplicação do algoritmo do Weka, que neste caso é a escolaridade da MAE. As linhas e colunas representam o mesmo valor da coluna MAE e os valores de cada letra pode ser verificado no canto direito da 9. As colunas indicam no qual os valores foram classificados e as linhas representam qual o valor foi classificado. Os valores contidos na diagonal principal, colunas e linhas equivalentes, representam os dados corretamente classificados, enquanto que os valores além desta diagonal representam os dados classificados incorretamente pelo algoritmo.

Observa-se que várias colunas tiveram nenhum registro classificado (valores zero na 9), indicando uma baixa representatividade destes valores na análise. Por essa pouca relevância dos dados, o algoritmo optou por classificar em outras colunas com maior similaridade de valores, o que acabou causando uma maior taxa de erro.

```
     a      b      c      d      e      f      g      h      i   <-- classified as
532309  35071  14623      0      0      0      0      0      0 |      a = MEDIO_COMP
346494  53091   3106      0      0      0      0      0      0 |      b = PRIMARIO
239698   9275  16359      0      0      0      0      0      0 |      c = SUPERIOR_COMP
348743  40471   4164      0      0      0      0      0      0 |      d = GINASIO
 49355  13109    250      0      0      0      0      0      0 |      e = NAO_ESTUDOU
144519  12306   2716      0      0      0      0      0      0 |      f = MEDIO_INC
 71587  13411    628      0      0      0      0      0      0 |      g = NAO_SABE
148681   4794  12747      0      0      0      0      0      0 |      h = POS_GRADUCAO
 93671   4443   4259      0      0      0      0      0      0 |      i = SUPERIOR_INC
```

Figura 9: Matriz de Confusão da escolaridade da mãe e nota da redação.

As colunas da 9 representam no qual os valores foram classificados enquanto as linhas representam quais valores foram classificados. Pode-se visualizar que a maioria dos valores foram classificados como MEDIO_COMP na coluna 'a'. Ainda assim é possível notar um padrão, no qualos valores de NAO_ESTUDOU, PRIMARIO, GINASIO, MEDIO_INC, MEDIO_COMP e NAO_SABE tiveram muitos valores classificados como PRIMARIO na coluna 'b' se comparado com os valores classificados como SUPERIOR_COMP na coluna 'c' enquanto os valores de SUPERIOR_COMP e POS_GRADUACAO tiveram mais valores classificados como SUPERIOR_COMP na coluna 'c' se comparado com a quantidade classificada como PRIMARIO da coluna 'b'.

Aplicou-se então o algoritmo para as colunas REDACAO_NOMINAL e PAI para verificar se o padrão encontrado no resultado anterior também se aplica a essas colunas. O resultado do Modelo Classificador pode ser visto na 10.

```
REDACAO_NOMINAL = G: PRIMARIO (384732.0/274046.0)
REDACAO_NOMINAL = H: PRIMARIO (107929.0/72317.0)
REDACAO_NOMINAL = F: PRIMARIO (564539.0/424445.0)
REDACAO_NOMINAL = D: MEDIO_COMP (282940.0/209523.0)
REDACAO_NOMINAL = E: MEDIO_COMP (602124.0/456374.0)
REDACAO_NOMINAL = C: MEDIO_COMP (140722.0/104301.0)
REDACAO_NOMINAL = B: SUPERIOR_COMP (43181.0/32573.0)
REDACAO_NOMINAL = J: PRIMARIO (54629.0/36684.0)
REDACAO_NOMINAL = I: PRIMARIO (23413.0/14851.0)
REDACAO_NOMINAL = A: SUPERIOR_COMP (15671.0/11227.0)
```

Figura 10: Resultado do Modelo Classificador da escolaridade do pai e nota da redação.

É possível notar que o resultado da classificação entre a escolaridade do pai e a nota da redação apresentou um padrão semelhante ao da classificação entre a escolaridade da mãe e nota da redação, sendo que as maiores notas, de valores A e B, foram classificadas como SUPERIOR_COMP. Nota-se também que apenas três valores, C, D e E, foram classificados como MEDIO_COMP enquanto os cinco valores restantes, F, G, H, I e J, foram classificados como PRIMARIO, apresentando assim um padrão um pouco diferente do exibido na 8.

A Matriz de Confusão da classificação entre REDACAO_NOMINAL e PAI pode ser visualizada na 11.

```
     a      b      c      d      e      f      g      h      i   <-- classified as
255588 217844  14056      0      0      0      0      0      0 |      a = MEDIO_COMP
189489 312899   4692      0      0      0      0      0      0 |      b = PRIMARIO
131207  65937  15052      0      0      0      0      0      0 |      c = SUPERIOR_COMP
160938 214284   5183      0      0      0      0      0      0 |      d = GINASIO
 78979 129696   2194      0      0      0      0      0      0 |      e = NAO_SABE
 27654  69500    428      0      0      0      0      0      0 |      f = NAO_ESTUDOU
 67581  70076   3083      0      0      0      0      0      0 |      g = MEDIO_INC
 63150  23968   9733      0      0      0      0      0      0 |      h = POS_GRADUCAO
 51200  31038   4431      0      0      0      0      0      0 |      i = SUPERIOR_INC
```

Figura 11: Matriz de Confusão da escolaridade do pai e nota da redação.

A Matriz de Confusão da classificação entre a escolaridade do pai e nota da redação da 11 apresenta uma classificação muito diferente da que foi visualizada na Matriz de Confusão da classificação entre a escolaridade da mãe e nota da redação da 9. A quantidade de classificações como PRIMARIO, representado na 11 pela letra 'b', possuem valores altos para todas escolaridades se comparado com a 9. Também pode-se verificar que os valores classificados como MEDIO_COMP, representado pela letra 'a', e PRIMARIO, representado pela letra 'b', são muito próximos e em várias linhas da 11 a quantidade de classificações como PRIMARIO, representado por 'b', superam a quantidade de classificações como MEDIO_COMP, representado por 'a', sendo o caso das linhas 'b', 'd', 'e', 'f' e 'g'. Mas ainda assim é possível verificar um desempenho superior para escolaridades mais elevadas como POS_GRADUACAO, SUPERIOR_INC, SUPERIOR_COMP e MEDIO_COMP se comparado com as escolaridades NAO_ESTUDOU, PRIMARIO, GINASIO, MEDIO_INC e NAO_SABE.

O algoritmo foi aplicado dessa vez para as colunas NOTA_MEDIA_NOMINAL e MAE para verificar se existem padrões relacionados à nota média do candidato com a escolaridade da mãe. O resultado do Modelo Classificador pode ser visualizado na 12.

```
NOTA_MEDIA_NOMINAL = F: MEDIO_COMP (1022845.0/763659.0)
NOTA_MEDIA_NOMINAL = E: MEDIO_COMP (735221.0/510157.0)
NOTA_MEDIA_NOMINAL = G: PRIMARIO (218735.0/152041.0)
NOTA_MEDIA_NOMINAL = D: SUPERIOR_COMP (210976.0/150316.0)
NOTA_MEDIA_NOMINAL = H: PRIMARIO (4951.0/3271.0)
NOTA_MEDIA_NOMINAL = C: SUPERIOR_COMP (26770.0/16781.0)
NOTA_MEDIA_NOMINAL = B: POS_GRADUCAO (382.0/227.0)
```

Figura 12: Resultado do Modelo Classificador da escolaridade da mãe e nota média.

Ao visualizar o resultado da 12 é possível notar que não existem os valores A, I e J da coluna NOTA_MEDIA_NOMINAL e isso ocorre devido à dificuldade de se tirar uma nota média acima de 900 e abaixo de 200. Para atingir uma nota média de 900 o candidato precisa ter um desempenho muito alto em todas áreas de conhecimento e para atingir uma nota média abaixo de 200 é necessário ter um desempenho muito abaixo do esperado pela média dos candidatos em todas as áreas de conhecimento.

Ao verificar as notas mais elevadas é possível notar que segue o mesmo padrão visto na 8 e na 10, sendo que as melhores notas foram classificadas com escolaridades maiores enquanto as menores estão relacionadas às menores escolaridades.

Na 13 a maioria dos valores foram classificados como MEDIO_COMP na coluna 'a'. As linhas que representam as escolaridades mais altas como POS_GRADUACAO, SUPERIOR_COMP, SUPERIOR_INC e MEDIO_COMP tiveram muitos valores classificados como SUPERIOR_COMP na coluna 'c' enquanto as escolaridades restantes tiveram muito mais valores classificados como PRIMARIO na coluna 'b'.

```
     a      b      c      d      e      f      g      h      i   <-- classified as
484250  39125  58588      0      0      0      0     40      0 |      a = MEDIO_COMP
324550  68374   9767      0      0      0      0      0      0 |      b = PRIMARIO
185609   8923  70649      0      0      0      0    151      0 |      c = SUPERIOR_COMP
328760  50143  14470      0      0      0      0      5      0 |      d = GINASIO
 45269  16866    579      0      0      0      0      0      0 |      e = NAO_ESTUDOU
135007  14220  10309      0      0      0      0      5      0 |      f = MEDIO_INC
 65870  17290   2465      0      0      0      0      1      0 |      g = NAO_SABE
108982   4332  52753      0      0      0      0    155      0 |      h = POS_GRADUCAO
 79769   4413  18166      0      0      0      0     25      0 |      i = SUPERIOR_INC
```

Figura 13: Matriz de Confusão da escolaridade da mãe e nota média.

O último teste realizado reproduziu o Modelo Classificador que pode ser visto na 14. Assim como na 12, esse Modelo Classificador também não apresenta os valores A, I e J para a NOTA_MEDIA_NOMINAL. Nesse modelo as notas B, C e D foram classificadas como SUPERIOR_COMP enquanto apenas a nota E foi classificada como MEDIO_COMP. As notas F, G e H foram classificados como PRIMARIO.

```
NOTA_MEDIA_NOMINAL = F: PRIMARIO (1022845.0/735776.0)
NOTA_MEDIA_NOMINAL = E: MEDIO_COMP (735221.0/534786.0)
NOTA_MEDIA_NOMINAL = G: PRIMARIO (218735.0/141829.0)
NOTA_MEDIA_NOMINAL = D: SUPERIOR_COMP (210976.0/155451.0)
NOTA_MEDIA_NOMINAL = H: PRIMARIO (4951.0/3036.0)
NOTA_MEDIA_NOMINAL = C: SUPERIOR_COMP (26770.0/17166.0)
NOTA_MEDIA_NOMINAL = B: SUPERIOR_COMP (382.0/219.0)
```

Figura 14: Resultado do Modelo Classificador da escolaridade do pai e nota média.

Na Matriz de Confusão da 15 a maioria dos valores foram classificados como MEDIO_COMP na coluna 'a' e PRIMARIO na coluna 'b'. Os valores de SUPERIOR_COMP, representado pela linha 'c', obteve mais valores classificados como SUPERIOR_COMP na coluna 'c' do que como PRIMARIO na coluna 'b'. Os valores de POS_GRADUACAO, representado pela linha 'h', foram em sua maioria classificados como SUPERIOR_COMP na linha 'c', superando os valores classificados como MEDIO_COMP na coluna 'a'.

```
     a      b      c      d      e      f      g      h      i   <-- classified as
200435 229774  57279      0      0      0      0      0      0 |      a = MEDIO_COMP
125800 365890  15390      0      0      0      0      0      0 |      b = PRIMARIO
 92937  53967  65292      0      0      0      0      0      0 |      c = SUPERIOR_COMP
117774 244226  18405      0      0      0      0      0      0 |      d = GINASIO
 52522 150364   7983      0      0      0      0      0      0 |      e = NAO_SABE
 14604  81783   1195      0      0      0      0      0      0 |      f = NAO_ESTUDOU
 52162  76771  11807      0      0      0      0      0      0 |      g = MEDIO_INC
 39519  16240  41092      0      0      0      0      0      0 |      h = POS_GRADUCAO
 39468  27516  19685      0      0      0      0      0      0 |      i = SUPERIOR_INC
```

Figura 15: Matriz de Confusão da escolaridade do pai e nota média.

4.6 APRESENTAÇÃO E ASSIMILAÇÃO DE CONHECIMENTO

Na etapa anterior foi possível identificar um padrão no qual quanto maior a escolaridade do pai ou da mãe maior é a nota dos candidatos. Para verificar esse resultado foram criados gráficos com a representação em porcentagem de cada uma das escolaridades dos pais e as notas da redação e média das notas gerais utilizando o conjunto de dados. É importante ressaltar que as técnicas de mineração de dados não alteram o conjunto de dados, sendo as técnicas utilizadas apenas como fonte de informação para a extração de conhecimento.

Nos gráficos da 16, 17, 18 e 19, o eixo vertical representa a porcentagem de notas dos estudantes por faixa de notas enquanto o eixo horizontal representa o grau de escolaridade dos pais.

A 16 apresenta um gráfico no qual é possível observar que a quantidade de notas na redação de 600 à 800 e acima de 800 crescem conforme a escolaridade da mãe aumenta. Quanto menor a escolaridade, maior é a presença de notas baixas de 200 à 400 na nota da redação e menor quantidade de notas altas de 600 ou mais.

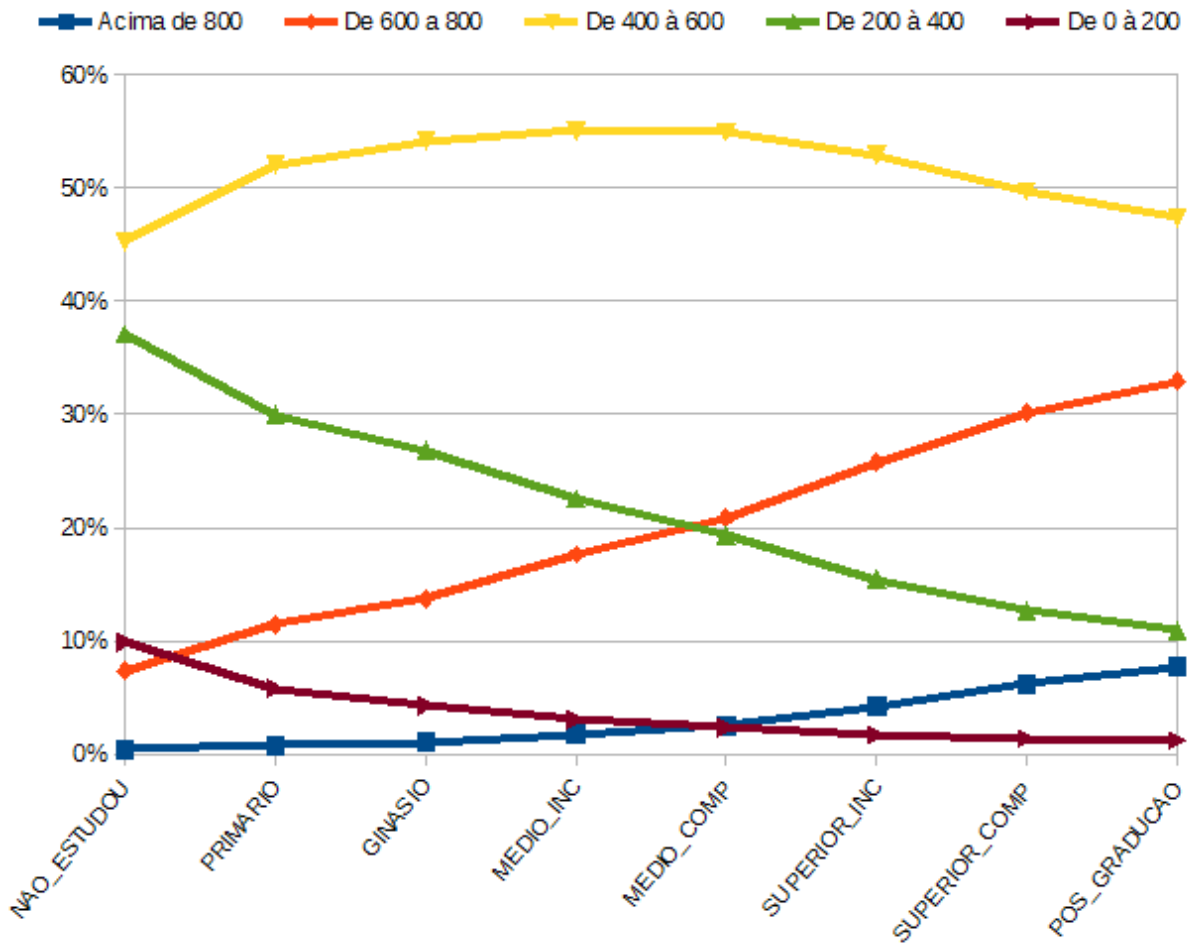

Figura 16: Gráfico com a porcentagem de notas da redação por escolaridade da mãe.

A 17 apresenta um gráfico semelhante no qual a quantidade de notas na redação de 600 à 800 e acima de 800 aumentam de acordo com o aumento da escolaridade do pai. Quanto menor a escolaridade do pai, menor quantidade de notas altas de 600 para cima e maior a quantidade de notas baixas de 200 à 400 e de 0 à 200.

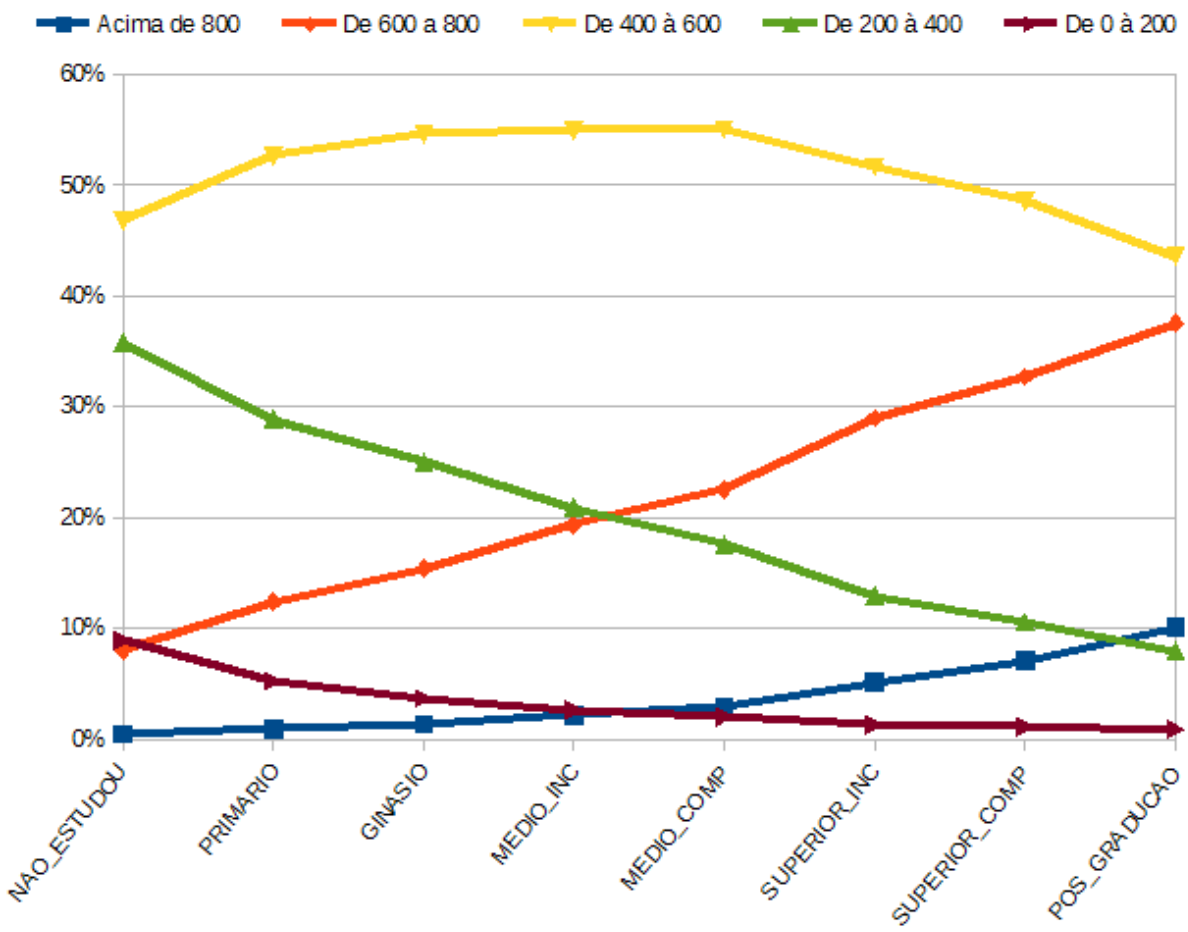

Figura 17: Gráfico com a porcentagem de notas da redação por escolaridade do pai.

A 18 apresenta muito poucas notas média acima de 800 e abaixo de 200, mas ainda é possível observar um aumento na quantidade de notas média de 600 à 800 e uma diminuição nas notas baixas de 200 à 400 conforme a escolaridade da mãe aumenta.

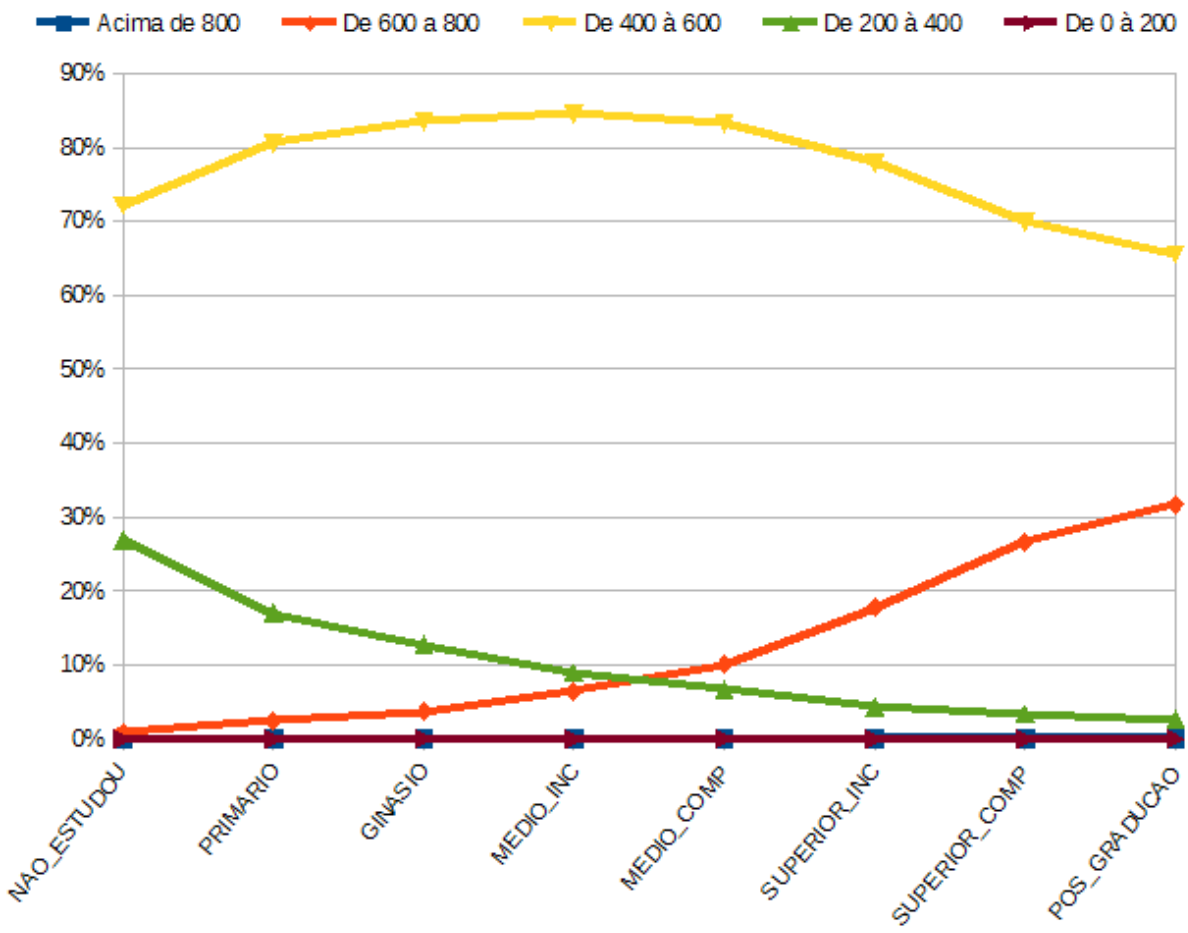

Figura 18: Gráfico com a porcentagem de notas média por escolaridade da mãe.

A 19 apresenta um gráfico semelhante ao anterior no qual existe um aumento na quantidade de notas média de 600 à 800 e uma diminuição nas notas baixas de 200 à 400 conforme a escolaridade do pai aumenta.

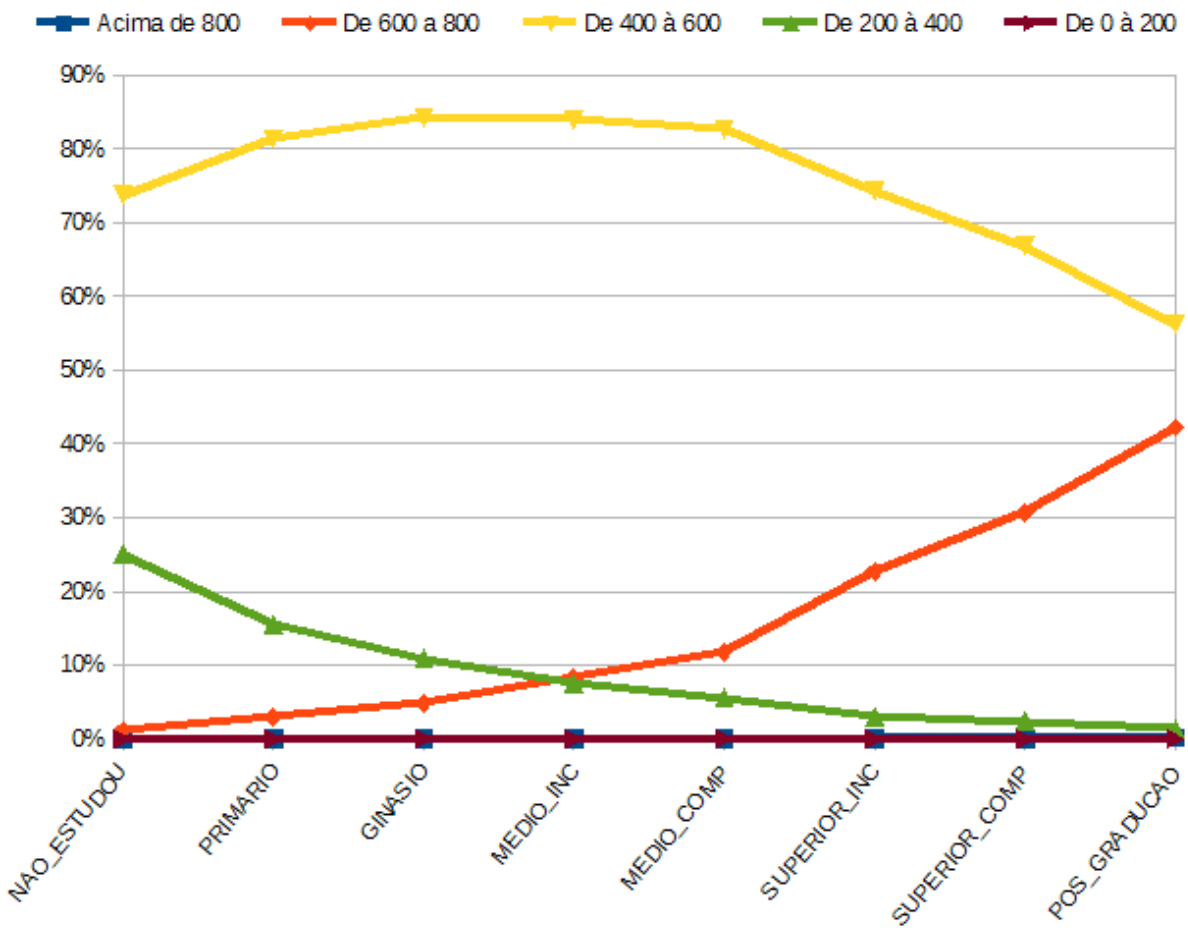

Figura 19: Gráfico com a porcentagem de notas média por escolaridade do pai.

5 CONSIDERAÇÕES FINAIS

O objetivo geral de aplicar o processo de Descoberta de Conhecimento em Base de Dados para análise e extração de conhecimento em uma base de dados educacional pública em busca de informações sobre a realidade dos estudantes brasileiros foi concluído neste trabalho por meio do alcance dos seguintes objetivos específicos:

Na seção 2.1 foi concluído o objetivo de compreender os fatores que influenciam no rendimento escolar dos alunos fazendo um levantamento de alguns trabalhos que citavam fatores de riscos que mais afetavam no rendimento escolar dos alunos, sendo possível compreender que um dos grandes fatores de risco é a escolaridade dos pais.

Nas seções 4.1, 4.2 e 4.3 foi concluído objetivo de realizar a preparação dos dados por meio das atividades de seleção, pré-processamento e transformação em uma base de dados educacional realizando toda a preparação dos dados para que fossem utilizados os algoritmos de mineração de dados. Como citado na metodologia na seção 3, foi utilizado o software Kettle para realizar todo o processo de preparação dos dados.

Na seção 4.4 foi concluído o objetivo de aplicar técnicas e algoritmos de mineração de dados na base pré-processada utilizando o software Weka, como citado na metodologia na seção 3, para aplicação do algoritmo de classificação J48. A escolha da técnica foi feita observando as hipóteses que foram levantadas e como seria possível respondê-las, que como foi levantado na revisão bibliográfica era a hipótese era: “Qual a influência que a escolaridade dos pais possui sobre o desempenho escolar de seus filhos?”.

Na seção 4.5 foi concluído o objetivo de analisar as informações obtidas pela mineração de dados, visando reconhecer padrões nos dados ao realizar a leitura dos modelos classificadores e matrizes de confusão obtidos pode-se verificar um padrão em comum com todos os resultados obtidos na etapa de mineração de dados. O padrão encontrado relacionava a escolaridade dos pais com as maiores notas da redação e nota média do ENEM.

A assimilação e apresentação de conhecimento na seção 4.6, concluiu o último objetivo de apresentar e analisar os conhecimentos obtidos por meio dos padrões encontrados e também conclui última etapa do processo de Descoberta de Conhecimento em Base de Dados. Nesta seção foi utilizado gráficos em linha onde o eixo horizontal apresenta o nível de escolaridade dos pais e o eixo vertical apresenta a porcentagem de alunos que tirou determinada nota em cada uma das escolaridades sendo possível observar que quanto maior a escolaridade dos pais, maiores são a quantidade de notas melhores e menores a quantidade de notas piores.

O resultado desta pesquisa valida o fator de risco escolaridade dos pais levantado na pesquisa do Instituto Glia de neurologia e desenvolvimento (GLIA, 2010) por meio da extração do mesmo conhecimento em uma base de dados mais representativa. O fato de ambas pesquisas identificarem o fator de risco baixa escolaridade dos pais como sendo um fator de risco para o baixo rendimento escolar dos estudantes representa uma contribuição deste trabalho à medida que aponta esta característica como forte candidata à investigação no sentido de contribuir com a meta de melhoria da qualidade do ensino.

Outra contribuição deste trabalho refere-se à aplicação em si do processo de Descoberta de Conhecimento em Base de Dados na base de dados do ENEM, que pode ser utilizado como referência para interessados em consultar os procedimentos executados.

Este trabalho pode ser complementado pelas seguintes atividades:

- Análise de outros fatores de risco para o baixo rendimento escolar dos estudantes;
- Aplicação dos mesmos procedimentos de pesquisa para as bases de dados do ENEM de outros anos;
- Investigação da relação causa e efeito da escolaridade dos pais no rendimento escolar do estudante por meio de uma pesquisa interdisciplinar.

REFERÊNCIAS

ADEODATO, Paulo JL; SANTOS FILHO, Maílson M.; RODRIGUES, Rodrigo L. Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar. In: **Anais do Simpósio Brasileiro de Informática na Educação**. 2014. p. 891-895.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, [S.l.], v. 19, n. 02, p. 03, ago. 2011. ISSN 1414-5685. Disponível em: <http://www.br-ie.org/pub/index.php/rbie/article/view/1301>. Acessado em 29 de Novembro de 2015.

BRAMER, Max. *Principles of data mining*. Vol. 131. London: Springer, 2007.
SFERRA, Heloisa Helena; CORRÊA, A. M. C. J. Conceitos e aplicações de data mining. **Revista de ciência & tecnologia**, v. 11, n. 22, 2003.

BRASIL A, 2007. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **INEP.** Disponível em: <http://dados.gov.br/dataset/microdados-do-exame-nacional-do-ensino-medio-enem>. Acessado em 29 de Novembro de 2015.

BRASIL B, 2007. Ministério do Planejamento, Orçamento e Gestão. **Dados Abertos.** Disponível em: <http://dados.gov.br/sobre>. Acessado em 29 de Novembro de 2015.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Goiânia: Universidade Federal de Goiás**, 2009.

CORTES, Sérgio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. Mineração de dados–Funcionalidades, técnicas e abordagens. **PUC-RioInf, Rio de Janeiro. IN THE BRAZILIAN GOVERNMENT USING CREDIT SCORING Leonardo sales**, 2002.

DAZZANI, Maria Virgínia Machado; FARIA, Marcelo O. Família, escola e desempenho acadêmico. **Avaliação Educacional: desatando e reatando nó**, *2009*, 249-264.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

GLIA, Instituto Glia de neurologia e desenvolvimento. Educando com a ajuda das Neurociências. **Projeto Atenção Brasil**, 2010.

KAMPFF, Adriana Justin Cerveira; REATEGUI, Eliseo Berni; DE LIMA, José Valdeni. Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. **RENOTE**, v. 6, n. 1, 2008.

MANHÃES, Laci Mary Barbosa et al. Identificaçao dos fatores que influenciam a evasao em cursos de graduaçao por meio de sistemas baseados em mineraçao de dados: Uma abordagem quantitativa. **Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo**, 2012.

MAZZETTO, Selma Elaine; BRAVO, Claudia Christina; CARNEIRO, Sá. Licenciatura em química da UFC: perfil sócio-econômico, evasão e desempenho dos alunos. ***Química Nova* 25.6/B** (2002): 1204-1210.

MORAIS, Alana M.; FECHINE, Joseana. Mineração de Dados Educacionais no Apoio ao Processo de Tomada de Decisão do Docente. **Universidade Federal de Campina Grande.** 2012 . Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/wei/2013/0017.pdf> Acessado em 29 de Novembro de 2015.

RIGO, Sandro J.; CAZELLA, Silvio C.; CAMBRUZZI, Wagner. Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In: **Anais do Workshop de Desafios da Computação Aplicada à Educação**. 2012. p. 168-177.

SAMPAIO, Breno; SAMPAIO, Yoni; de MELLO, Euler P. G.; MELO, Andrea. S. Desempenho no vestibular, background familiar e evasão: evidências da UFPE. ***Economia Aplicada*****,** 2011, *15*(2), 287-309.

ZAKI, Mohammed J.; MEIRA, Wagner, Jr.Data Mining and Analysis: Fundamental Concepts and Algorithms, 2014. **Cambridge University Press.** Disponivel em: <http://www.dataminingbook.info/pmwiki.php/Main/BookDownload>. Acessado em 29 de Novembro de 2015.