UNIVERSIDADE FEDERAL DE SANTA CATARINA

Rafael Damiani Alves

PREDIÇÃO DO DESEMPENHO DA REDAÇÃO DO ENEM UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Araranguá

2018

Rafael Damiani Alves

PREDIÇÃO DO DESEMPENHO DA REDAÇÃO DO ENEM UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Trabalho de Conclusão do Curso de Graduação do Centro de Ciência, Tecnologia e Saúde da Universidade Federal de Santa Catarina como requisito para a obtenção do Título de Bacharel em Tecnologias da Informação e Comunicação.
Orientador: Prof. Dr. Cristian Cechinel
Coorientador: Me. Emanuel Marques Queiroga

Araranguá
2018

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Alves, Rafael Damiani
PREDIÇÃO DO DESEMPENHO DA REDAÇÃO DO ENEM UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS / Rafael Damiani Alves ; orientador, Cristian Cechinel, coorientador, Emanuel Marques Queiroga, 2018.
67 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2018.

Inclui referências.

1. Tecnologias da Informação e Comunicação. 2. Mineração de dados. 3. Predição. 4. Redação . 5. ENEM. I. Cechinel, Cristian . II. Queiroga, Emanuel Marques . III. Universidade Federal de Santa Catarina. Graduação em Tecnologias da Informação e Comunicação. IV. Título.

Rafael Damiani Alves

PREDIÇÃO DO DESEMPENHO DA REDAÇÃO DO ENEM UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de "Bacharel em Tecnologia da Informação e Comunicação".

Araranguá, 21 de junho de 2018.

Prof. Patricia Jantsch Fiuza, Dr.
Coordenadora do Curso

Prof. Vinicius Faria Culmant Ramos, D.Sc.
Coordenador do Bacharelado em Tecnologias da Informação e Comunicação
Centro de Ciências, Tecnologias e Saúde
Port. Nº 1681/2017 - SIAPE 1814701

Banca Examinadora:

Prof. Cristian Cechinel, Dr.
Orientador
Universidade Federal de Santa Catarina

Prof. Juarez Bento da Silva, Dr.
Universidade Federal de Santa Catarina

Prof. Robson Rodrigues Lemos, Dr.
Universidade Federal de Santa Catarina

Dedico esse trabalho primeiramente a Deus, mas também a meus pais, familiares, amigos e a todos aqueles que de alguma forma me deram forças nessa caminhada.

AGRADECIMENTOS

Primeiramente agradeço a Deus por me dar força, fé e sabedoria para elaboração desse trabalho e por proporcionar que aos poucos eu atinja minhas metas, como é o caso da obtenção do Grau de Bacharel em Tecnologias da Informação e Comunicação. Agradeço imensamente também aos meus pais Sander e Luciani, por todo amor, confiança e atenção passados para mim. Por terem me proporcionado todas as oportunidades possíveis para a realização da graduação e consequentemente deste trabalho, deixo claro que sem os mesmos nada disso seria possível. Devo ressaltar minha gratidão pelos meus amigos, que nos momentos de cansaço e esgotamento me proporcionaram alegria e descontração, contribuindo então para a retomada do foco na conquista dos meus objetivos.

No meio acadêmico deixo meus sinceros agradecimentos ao professor e meu orientador Dr. Cristian Cechinel e também ao meu coorientador Me. Emanuel Marques Queiroga, por sanarem todas as minhas dúvidas e também por em nenhum momento desistirem desse projeto, ou seja, pelo apoio incondicional prestado pelos mesmos. Registro aqui minha enorme gratidão pela Universidade Federal de Santa Catarina – UFSC por todo o suporte prestado pela mesma, e pela grande oportunidade de realizar essa graduação de qualidade. Agradeço também a todos os professores que tanto me ensinaram e me proporcionaram conhecimento nessa jornada. Por fim agradeço a todos aqueles que de alguma forma direta ou indireta contribuíram para este projeto e para a minha formação acadêmica.

Escreva algo que valha a pena ler ou faça algo que valha a pena escrever.

(Benjamin Franklin)

RESUMO

O avanço das tecnologias da informação e comunicação tem proporcionado o armazenamento de bases de dados cada vez maiores. Devido a isso, as diversas técnicas de mineração de dados vêm sendo utilizadas para realizar descoberta de padrões que permitem a melhoria em muitas áreas, bem como vantagens competitivas, quando a mesma é aplicada sobre bases de dados comerciais. As técnicas de mineração de dados podem ser empregadas em quaisquer áreas, porém em algumas áreas como Marketing, Detecção de fraude, Investimento financeiros, essas técnicas vêm sendo utilizadas com mais regularidade. A mineração de dados educacionais (EDM) é uma das diversas linhas de pesquisas da mineração, sendo a responsável pela descoberta de informações (KDD) em bases de dados que contém informações acadêmicas. Desta forma esse trabalho propõe-se a encontrar padrões e gerar um modelo preditivo do indicador de desempenho das notas da prova de redação referentes aos dados educacionais do Exame Nacional do Ensino Médio (ENEM) de 2016. Neste contexto, foram feitos experimentos onde os dados foram categorizados, para obter um melhor resultado na aplicação dos algoritmos. A classe predita foi a nota da redação que foi categorizada como: baixo, médio, alto e nulo. Os modelos finais foram treinados e testados por meio dos algoritmos: Naive Bayes e J48. Esses algoritmos foram utilizados através do pacote de software WEKA. Por meio da utilização dessas técnicas de mineração de dados, o modelo com o maior desempenho conseguiu prever 61.7464% das amostras presentes na base de dados do ENEM 2016.

Palavras-chave: Mineração de Dados, Dados educacionais, ENEM, Modelo preditivo.

ABSTRACT

The advancement of information and communication technologies has provided an increasing capacity of storage databases. Due to this, several data mining techniques have been used to discover patterns that allow improvement in many areas, as well as competitive advantages, when applied on commercial databases. Data mining techniques can be used in any area; however, areas as marketing, fraud detection and financial investments have been employing these techniques more frequently. Educational data mining (EDM) is one of many data mining research lines, and it is responsible for knowledge-discovery in databases (KDD) that contain academic information. Therefore, the present study aims to detect patterns and creat a predictive model of the performance indicator of writing test scores for the National High School Exam (ENEM) - 2016 data. In this context, experiments were made where data were categorized, in order to obtain better results in the application of algorithms. The predicted class was the writing test score, which was categorized as: low, medium, high and null. Final models were trained and tested through the algorithms: Naive Bayes and J48. These algorithms were used through the WEKA software package. The use of these data mining techniques allowed the model with the best performance to predict correctly 61.7464% of the samples found in the ENEM 2016 databases.

Keywords: data mining, educational data, ENEM, predictive model.

LISTA DE FIGURAS

Figura 1: O ciclo do processo de KDD. 19
Figura 2: Ferramenta Weka Explorer. 28
Figura 3: Processos utilizados na metodologia. 40
Figura 4: Processo de configuração para remoção das acentuações. 42
Figura 5: Processo de Table output Mapping 42
Figura 6:Fluxo percorrido pelos dados no Pentaho 43
Figura 7: Função para importar Instâncias do banco de dados. 44
Figura 8: Função para descobrir pontos de corte. 44
Figura 9: Efetuando a categorização da variável "NU_NOTA_REDACAO". 45
Figura 10: Seleção dos dados do estado de Santa Catarina. 48
Figura 11: Seleção dos dados da cidade de Araranguá. 48
Figura 12: Resumo dos resultados obtidos por meio do algoritmo J48 relacionados aos dados de Santa Catarina.. 53
Figura 13: Matriz de Confusão gerada por meido do algoritmo J48 relacionados aos dados de Santa Catarina. 53
Figura 14: Resumo dos resultados obtidos por meio do algoritmo Naive Baye e relacionados aos dados de Santa Catarina. 54
Figura 15: Matriz de Confusão gerada por meio do algoritmo Naive Baye e relacionada aos dados de Santa Catarina. 54
Figura 16: Resumo dos resultados obtidos por meio do algoritmo J48 e relacionados aos dados de Araranguá. 55
Figura 17: Matriz de Confusão gerada por meio do algoritmo J48 e relacionada aos dados de Araranguá. 56
Figura 18: Resumo dos resultados obtidos por meio do algoritmo Naive Baye relacionado aos dados de Araranguá. 56
Figura 19: Matriz de Confusão gerada por meio do algoritmo Naive Baye e relacionado aos dados de Araranguá. 57
Figura 20: Visualização parcial da árvore de decisão da cidade de Araranguá. 60

LISTA DE QUADROS

Quadro 1: Modelo de ranqueamento. 31
Quadro 2: História e eventos do ENEM. 34
Quadro 3: Quantitativos dos totais de instâncias. 39
Quadro 4: Quantitativos dos totais de instâncias em relação aos tipos de escolas. ..39
Quadro 5: Regras para a classificação das notas da redação. 45
Quadro 6: Regras para a classificação da variável Q005. 46
Quadro 7: Regras para a classificação da variável Q006. 46
Quadro 8: Regras para a classificação da idade. 47
Quadro 9: Acurácia detalhada do primeiro experimento, utilizando o algoritmo J48 e apenas os dados de Santa Catarina. 53
Quadro 10: Acurácia detalhada do primeiro experimento, utilizando o algoritmo Naive Bayes e apenas os dados de Santa Catarina. 54
Quadro 11: Acurácia detalhada do primeiro experimento, utilizando o algoritmo J48 e apenas os dados de Araranguá. 56
Quadro 12: Acurácia detalhada do primeiro experimento, utilizando o algoritmo Naive Bayes e apenas os dados de Araranguá. 57

LISTA DE ABREVIATURAS E SIGLAS

APIs - Application Programming Interface

ARFF - Attribute-Relation File Format

BI - Business Intelligence

CSV - Comma-separated values

DDL - Linguagem de definição de dados

EDM - Mineração de Dados Educacionais

ENEM- Exame Nacional do Ensino Médio

ENCCEJA - Exame para Certificação de Competências de Jovens e Adultos

ETL - Extração, Transformação e Carregamento

FIES - Fundo de Financiamento ao Estudante do Ensino Superior

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

ISO - International Standardization Organization

KDD - Processo de Descoberta do Conhecimento

MDE – Mineração de Dados Educacionais

MD – Mineração de Dados

OLAP - Online Analytical Processing

PROUNI - Programa Universidade para Todos

RDF - Resource Description Framework

SISU - Sistema de Seleção Unificada

SPARQL - Protocol and RDF Query Language

SQL - Linguagem de Consulta Estruturada

TCC - Trabalho de Conclusão de Curso

URI - Identificador Uniforme de Recursos

URLs- Uniform Resource Locator

WEKA- Waikato Environment for Knowledge Analysis

W3C - World Wide Web Consortium

SUMÁRIO

1 INTRODUÇÃO 15

1.1 Contextualização Do Problema e Justificativa 15

1.2 Objetivos 16

1.2.1 Objetivo Geral 16

1.2.2 Objetivos Específicos 17

2 REFERENCIAL TEÓRICO 18

2.1 Mineração de Dados 18

2.2 Mineração de dados Educacionais 20

2.2.1 Métodos para EDM 21

2.3 Algoritmos de Aprendizagem de Máquina 24

2.3.1 Algoritmo J48 25

2.3.2 Algoritmo Naive Bayes 25

2.4 Árvores de Decisão 26

2.5 Matriz de Confusão 26

2.6 Ferramentas Computacionais que auxiliam o processo de Mineração de Dados 27

2.6.1 Weka 27

2.6.2 Pentaho 28

2.7 Dados abertos 28

2.7.1 Dados abertos governamentais 29

2.7.2 Modelo de ranqueamento 31

2.8 ENEM 32

2.9 Trabalhos relacionados 35

3 METODOLOGIA E EXPERIMENTOS 38

3.1 Contexto 38

3.2 Metodologia 40

3.3 Seleção e Pré-processamento dos dados 40
3.3.1 Transformação dos dados 43
3.3.2 Seleção dos dados de Santa Catarina e Araranguá 47
3.4 Geração e avaliação¸ dos modelos de predição 48
4 RESULTADOS E DISCUSSÕES 52
4.1 Resultados do primeiro experimento 52
4.2 Resultados do segundo experimento 55
4.3 Comparativo dos resultados do primeiro experimento com os do segundo experimento 57
4.4 Comparação com os Trabalhos Relacionados 58
4.5 Árvores de Decisão 59
4.5.1 Árvore de Decisão gerada no primeiro experimento 59
4.5.2 Árvore de Decisão gerada no segundo experimento 59
5 CONCLUSÃO 61

1 INTRODUÇÃO

O Exame Nacional do Ensino Médio (ENEM), segundo INEP (2018), se trata de uma avaliação criada e aplicada pelo Ministério da Educação para mensurar a performance dos estudantes que completaram o ensino médio. A prova do ENEM é subdividida em 4 áreas, são elas: Ciências Humanas e suas Tecnologias, Ciências da Natureza e suas Tecnologias, Linguagens, Códigos e suas Tecnologias, Matemática e suas Tecnologias. Conforme INEP (2018) além destas 4 áreas também é realizada uma prova para avaliação da competência de redação. É nessa prova que esse trabalho de conclusão de curso (TCC) tem o intuito de prever o desempenho dos alunos, através da mineração de dados.

Segundo Fayyad et al. (1996), as áreas científicas, governamentais, corporativas vem gerando um grande aumento nos seus bancos de dados, devido a isso se torna difícil analisar estes dados, necessitando da utilização de novas ferramentas e técnicas para análise automática e inteligente de bancos de dados. Devido a esse contexto, conforme Luan (2007), a mineração de dados pode ser utilizada para descoberta de tendências e padrões ocultos, com base nesses resultados gerados pela mineração de dados, os analistas e estudiosos da área podem analisar os mesmos e tomar decisões mais precisas, podendo focar a atenção nos pontos ou casos mais críticos e específicos.

Além das áreas já descritas nesse trabalho de conclusão de curso, segundo Romero e Ventura (2010) outra área que vem produzindo um grande volume de dados é a educação, devido a isso surgiu uma nova área de estudo da mineração de dados chamada de mineração de dados educacionais (MDE). De acordo com Romero e Ventura (2010) mineração de dados educacionais é um segmento de pesquisa interdisciplinar que possui métodos para estudar dados originados a partir do cenário educacional.

Tendo em vista que o ENEM é um dos exames que mais mobilizam os estudantes, e que vem ocorrendo um crescimento em relação a MDE e a evolução de suas técnicas, identificou-se uma oportunidade para uma pesquisa nessa área.

1.1 Contextualização Do Problema e Justificativa

De acordo com a Universidade Metodista de São Paulo (2011), a redação vem sendo uma das maiores dificuldades dos alunos, quando realizam algum vestibular ou exame. Nos vestibulares de diversas universidades essa nota vem sendo determinante para aprovação dos

candidatos, em relação ao ENEM não é diferente, a nota da redação é decisiva para o aluno conquistar uma boa média final. Além disso segundo Brasil (2018), a redação é critério de desempate no Programa Universidade Para Todos (Prouni).

Segundo Globo (2016), os dados do ENEM demonstram que houve um crescimento no número total de alunos que tiveram suas redações zeradas. Em 2015, 53 mil participantes obtiveram nota zero, já na edição de 2016 foram 84.236 . De acordo com Globo (2016), em 2016, só 77 inscritos alcançaram a nota mil na redação (maior nota possível), enquanto que no ano anterior foram 104 e no ano de 2014, o total era de 250 redação com notas mil.

Conforme a Universidade Metodista de São Paulo (2011), os professores assim como os alunos vem dando uma atenção especial a redação, pois diversos fatores podem fazer com que o aluno obtenha um resultado positivo ou negativo nessa avaliação.

Sendo assim qualquer descoberta em relação a esse tipo de prova, pode ajudar os professores, gestores e especialistas da área educacional a melhorar o ensino, tendo potencial para contribuir não apenas para um aluno em especifico, mas para toda uma coletividade, que vem sofrendo com as baixas notas nas provas de redação. Baseado nisso acentua-se a hipótese de que modelos preditivos e a identificação de variáveis influenciadoras no desempenho dos alunos e a mineração de dados de um modo geral poderiam auxiliar na melhoria significativa das notas de redação.

Diante deste contexto, surge a seguinte pergunta de pesquisa: É possível prever com um mínimo de exatidão o desempenho dos alunos na redação do ENEM?

1.2 Objetivos

Esse trabalho de conclusão de curso busca gerar modelos para predição do desempenho da redação do ENEM. Por meio dos mesmos pode-se ajudar os gestores e especialistas da área da educação a entender o desempenho dos alunos na prova de redação do ENEM. Dessa forma, por intermédio dos modelos de predição e consequentemente da descoberta de conhecimento, busca-se ofertar contribuições para a criação ou enrobustecimento de iniciativas que visem melhorias no desempenho dos inscritos do ENEM.

1.2.1 **Objetivo Geral**

Gerar modelos para predição do desempenho da redação do ENEM, por meio dos microdados do ENEM 2016 e algoritmos de classificação.

1.2.2 **Objetivos Específicos**

- Utilizar os microdados do ENEM 2016;
- Implementar *scripts* para a fase de transformação;
- Utilizar os algoritmos de classificação J48 e Naive Bayes;
- Gerar árvores de decisão baseada nos microdados do ENEM 2016;
- Implementar técnicas de mineração de dados em geral.

2 REFERENCIAL TEÓRICO

2.1 Mineração de Dados

Através do processo de globalização e crescimento da economia diversos novos segmentos de mercado surgiram e outros sofreram grandes modificações. Um fator muito impactado pela globalização foi o aumento da competitividade do mercado, devido a isso as empresas buscam de todo modo por vantagens que as ajudem a superar seus concorrentes.

Baseado a esse contexto de aquecimento do mercado, a cada dia que passa aumenta o número de transações e de dados, criando bases de dados cada vez maiores. Então esse grande volume de dados, se torna uma das possibilidades das empresas de se diferenciarem no mercado, mas para isso é necessário a análise desses dados para extração de informações uteis. Realizar manualmente essa etapa de análise de dados em uma base de dados de grande volume é praticamente inviável.

Com base em Berry e Linoff (2004), para obter informações a partir dessas grandes bases de dados surgiu a mineração de dados, que pode ser compreendida como a exploração e análise de grandes quantidades de dados, de maneira automática ou semiautomática, com o propósito de descobrir padrões e regras relevantes.

Existe diversas perspectivas sobre o conceito de mineração de dados, com base nisso são apresentadas duas definições consideradas importantes:

Segundo Zaki e Meira Junior (2014) a mineração de dados (Data Mining) é o processo de descoberta de padrões significativos e desconhecidos (novos), bem como modelos descritivos, compreensíveis e preditivos a partir de um grande volume de dados.

De acordo com Han, Kamber e Pei (2011) o objetivo de um processo de Mineração de dados (MD) é extrair conhecimento, que sejam importantes para tomadas de decisões, que podem se encontrar em informações que estão “ocultas” no grande volume de dados.

Conforme Han, Kamber e Pei (2011) a mineração de dados utiliza métodos de diferentes áreas científicas, como Estatística, Banco de Dados, Inteligência Artificial, Aprendizado de Máquina e Reconhecimento de Padrões, definindo e indicando que a MD é um processo interdisciplinar.

De acordo com Gomes (2015) a descoberta de padrões é um processo que se inicia pela escolha da base de dados, a mesma deve ser escolhida com base na possibilidade de responder as indagações dos especialistas da área que se deseja descobrir padrões. Após a escolha da base de dados a ser explorada, a próxima etapa a ser realizada é o pré-processamento,

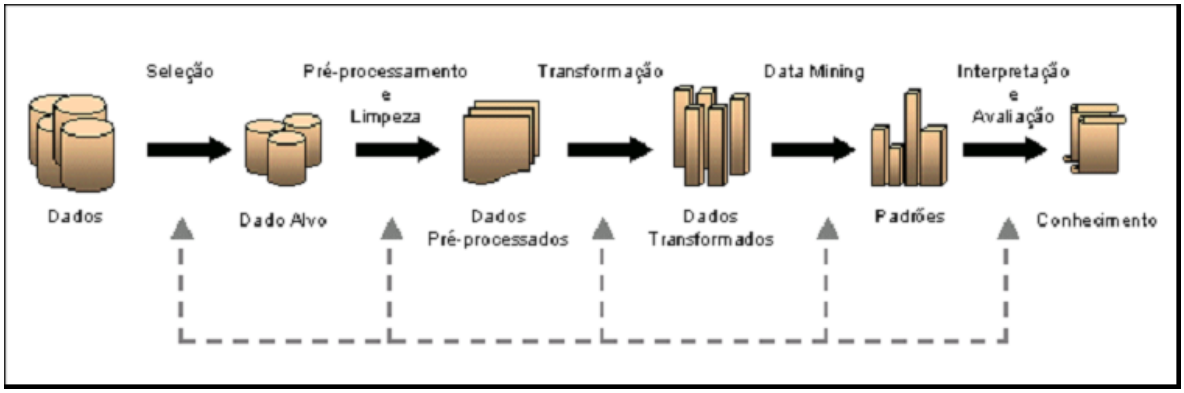
pois segundo Gomes (2015) os dados inclusos na base de dados, carecem de pré-processamento para que fiquem estruturados, limpos, selecionados e padronizados ou seja prontos para que o seguinte processo de mineração de dados seja executado.

No processo de extração de informações através da mineração de dados, pode ser utilizado diversos softwares, um exemplo é o WEKA (Waikato Environment for Knowledge Analysis). Nesta etapa de mineração de dados diversos métodos inteligentes podem ser aplicados para descoberta dos padrões.

Segundo Han, Kamber e Pei (2011) alguns veem a mineração de dados como uma etapa fundamental no processo de descoberta do conhecimento (do inglês Knowledge Discovery in Databases - KDD).

Existe diversas definições para KDD, porém para esse trabalho destaca-se FAYYAD et al. "o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados". Para Gomes (2015) o KDD contém diversas fases que possivelmente são o caminho que os dados percorrem até virarem conhecimento.

Figura 1: O ciclo do processo de KDD.



Fonte: Adaptação de FAYYAD et al. (1996).

O primeiro passo para a descoberta de conhecimento é a compreensão do problema e definição do objetivo do processo de KDD que se deseja atingir. Em um segundo momento a etapa que deve ser realizada é a seleção. A seleção trata-se de um processo onde é selecionado um conjunto de dados (ou concentrar-se em um subconjunto), os quais serão expostos as próximas etapas de KDD para concretizar o objetivo da mesma.

A terceira etapa se trata do pré-processamento de dados, etapa está que é responsável pela remoção de ruídos, coletando as informações necessárias para modelar ou explicar o ruído,

seleção de atributos importantes, formatação dos dados, tratamento de campos ausentes, ou seja, uma limpeza de modo geral.

A quarta etapa se trata da transformação, a mesma tem o objetivo de redução da dimensionalidade dos dados ou até mesmo complementar os dados. O número final de variáveis pode ser reduzido, ou permanecer constante.

A quinta etapa se trata da mineração de dados, onde os dados já foram selecionados, pré-processados e transformados utilizando os algoritmos específicos de aprendizagem, como por exemplo, algoritmos para efetuar: associação, classificação, clusterização e etc.

Através do uso desses algoritmos busca-se encontrar padrões onde posteriormente os mesmos serão analisados e interpretados pelos especialistas.

A sexta e última fase consiste na documentação do conhecimento ou na inclusão do mesmo e feedback para os responsáveis. Nesta fase deve ser realizado também a interpretação dos padrões pelos especialistas.

Os conceitos sobre KDD e os processos e etapas de KDD presentes nesse capitulo são em sua grande parte, baseados na obra de FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996).

2.2 Mineração de dados Educacionais

Segundo Queiroga (2017) muitos trabalhos buscam modelar comportamentos de alunos com o objetivo de realizar previsões sobre os mesmos, recorrendo a diversas técnicas e bases de dados, na sua maioria com resultados satisfatórios.

Estas pesquisas normalmente se distinguem no conjunto de técnicas utilizadas e na sua meta de previsão, sendo assim algumas destas pesquisas são focalizadas para a predição das notas dos alunos em uma ou diversas disciplinas, enquanto em outros trabalhos o principal objetivo se trata da predição de evasão dos alunos, apresentando se o mesmo está em situação de risco de evasão ou não. Segundo Detoni, Cechinel e Araujo (2015), a identificação com antecedência de estudantes que sofrem algum tipo de risco de evasão pode ajudar de maneira decisiva o trabalho de professores e tutores.

Baseado no que já foi dito, percebe-se que é viável minerar dados de instituições de ensino, alunos e diversos outros dados ligados a educação, com objetivos normalmente de colher informações que possam levar a melhoria do ensino nas instituições ou aumento do desempenho do aluno.

Por meio dessas circunstâncias, ergueu-se uma nova área de pesquisa chamada de "Mineração de Dados Educacionais" (EDM), que se trata de uma Subárea da mineração de dados. Conforme Baker, Carvalho e Isotani (2011), a EDM é definida como a área de pesquisa que tem como objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais.

Segundo Kampff (2009) para estabelecer os dados a serem minerados, é preciso saber claramente o problema que se deseja resolver e também ter conhecimento sobre os dados que se tem à disposição. De forma iterativa, é necessário aplicar as etapas de KDD até que os resultados possam auxiliar na resolução dos problemas que originaram a busca de conhecimento.

Devido a esse contexto, conforme Baker e Yacef (2009), pesquisadores da área de mineração de dados educacionais, estudam não somente essa área, mas também diversas outras, incluindo aprender com software educacional, aprendizagem colaborativa apoiada por computador, testes adaptativos de computador (e testes mais abrangentes) e os fatores associados com falha do aluno ou não retenção nos cursos.

De acordo com Baker e Yacef (2009), em todos esses segmentos, uma das principais áreas de atuação da EDM tem sido a melhoria dos modelos de estudantes, esses modelos apresentam diversas informações sobre as características de um aluno ou estado, como o conhecimento atual do aluno, motivação, metacognição e atitudes. Esses autores ainda ressaltam que pesquisadores também conseguiram ampliar a modelagem do estudante, de uma forma que vai além do software educacional, no sentido de descobrir quais fatores são preditivos de falha do aluno ou não-retenção em cursos universitários ou na faculdade.

Para que quaisquer dos possíveis objetivos da EDM já falados antes, possa ser atingindo é preciso a utilização de métodos e técnicas, a seguir, encontra-se as principais técnicas e métodos utilizadas na EDM que segundo Baker, Carvalho e Isotani (2011) são: predição, agrupamento, mineração de relações, destilação de dados para facilitar decisões humanas, descobertas com modelos.

2.2.1 Métodos para EDM

As técnicas e métodos propostos por BAKER; ISOTANI; CARVALHO (2011), são aceitas por um vasto número de pesquisadores, e seguem a seguinte taxonomia:

- Predição
  - Classificação
  - Regressão
  - Estimação de Densidade
- Agrupamento
- Mineração de relações
  - Mineração de Regras de associação
  - Mineração de Correlações
  - Mineração de Padrões Sequenciais
  - Mineração de Causas
- Destilação de dados para facilitar decisões humanas
- Descobertas com modelos

No que tange a **predição**, de acordo com Baker, Carvalho e Isotani (2011) seu objetivo é desenvolver modelos que possam deduzir alguns aspectos específicos dos dados, também chamados de variáveis preditivas, por meio de análise e fusão dos variados aspectos descoberto nos dados, denominados de variáveis preditoras. Para a realização da predição é necessário que exista uma quantidade relativa de dados, e os mesmo devem passar por uma codificação manual para que seja possível a identificação de uma ou diversas variáveis preditoras previamente conhecidas.

Segundo Baker, Carvalho e Isotani (2011) dos três tipos de predição os mais utilizados são: classificação, regressão, tendo em vista que a estimação de densidade é raramente utilizada na EDM devido à falta de independência estatística dos dados. Em relação a classificação e a regressão, as variáveis preditoras podem ser de dois tipos: categóricas ou numéricas. Quando a variável preditora é numérica os algoritmos mais utilizados são os de regressão linear e redes neurais. Já quando essas variáveis são binarias ou categóricas os algoritmos de maior destaque são os de classificação como por exemplo arvores de decisão.

Na área de **agrupamento**, o objetivo predominante é classificar os dados em diversos grupos e/ou categorias, de forma que os mesmos se agrupem naturalmente. É significativo salientar que estes grupos e categorias inicialmente são desconhecidos. Através de processos de descoberta de conhecimento, técnicas de agrupamento, os grupos/categorias são automaticamente identificados. É possível criar esses grupos e/ou categorias utilizando diferentes tipos de análise, como por exemplo é possível identificar grupos de escolas (para

investigar as diferenças e similaridades entre escolas), ou achar grupos de alunos (para investigar as diferenças e similaridades entre alunos).

Mineração de relações, tem como objetivo descobrir possíveis relações entre variáveis que estão presentes em bancos de dados. Um dos processos é identificar quais variáveis são mais fortemente relacionadas com uma variável especifica. Também pode-se descobrir as relações entre quaisquer variáveis presentes nos dados. Existem quatro formas para identificar essas relações são elas: regras de associação, correlações, sequências e causas.

A **mineração de regras de associação** tem como principal meta gerar/identificar regras do tipo se-então (if-then) que possibilitem associar o valor analisado de uma variável ao valor de uma outra variável, isso quer dizer que procura-se por descoberta de relações. Por exemplo, ao analisar um suposto banco de dados educacionais seria possível encontrar uma regra que faz a associação entre a variável "objetivo do aluno", onde a mesma poderia ser uma variável binária podendo ter os valores alcançado ou não alcançado, e uma outra variável também do tipo binária "pedir ajuda ao professor" que poderia ter os valores sim ou não. Neste sentido se um determinado aluno tem o objetivo de aprender uma determinada disciplina, porém tem dificuldade em entender a mesma, ele deve pedir ajuda ao professor, criando então a regra "pedir ajuda ao professor".

Quando se fala em **mineração de correlações**, deve se ter em mente que a mesma tem o objetivo de achar entre as variáveis correlações positivas ou negativos. Dessa forma, ao analisar uma base de dados seria viável identificar ligação entre as variáveis que revelam o comportamento atual do aluno e seu desempenho(nota) nas atividades seguintes.

A **mineração de sequências** tem como meta achar associação temporal entre eventos e o impacto dos mesmos no valor de uma variável. Nesse contexto se torna possível identificar qual percurso de ações podem levar o aluno a atingir uma aprendizagem efetiva. Sabendo dessas informações se torna viável criar diversas atividades institucionais que poderiam melhorar a qualidade do ensino, e ajudar o aluno a atingir o objetivo final que é o aprendizado efetivo do conteúdo.

A **mineração de causas**, dispõem de algoritmos e técnicas para averiguar se um evento causa outro evento através da análise dos padrões de covariância. Baseado nisso seria possível saber por exemplo, quais as causas de um aluno ter ido mal na prova (que poderia ser por eventos comportamentais), existiria a possibilidade de caso ocorresse a reprovação do aluno, saber quais eventos levaram a acontecer a reprovação.

Na área de **destilação de dados para facilitar decisões humanas**, a meta desta área é apresentar dados que são complexos de uma maneira que facilite a compreensão do mesmo e evidenciar de forma clara suas características mais importantes. Os métodos dessa subárea simplificam e descomplicam a visualização da informação contida nos dados educacionais coletados por softwares educacionais.

Através dos métodos apresentados, se torna possível obter informações que ajudem a melhorar o setor da educação e do ensino. Segundo Baker, Carvalho e Isotani (2011) assim, é viável entender de forma mais precisa e adequada os alunos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de fatores diversos que influenciam a aprendizagem.

De acordo com Luan (2007), um exemplo da utilização das técnicas e métodos de EDM serias as instituições de ensino usar a classificação, para uma análise abrangente das características dos estudantes, ou usar estimativa para prever a probabilidade de uma variedade de resultados, como transferência, persistência, retenção e também o tão desejado sucesso. Este tópico foi na sua maioria, alicerçado e baseado no trabalho de BAKER; ISOTANI; CARVALHO (2011).

2.3 Algoritmos de Aprendizagem de Máquina

Algoritmos de Aprendizagem de Máquina são os responsáveis por possibilitar que os computadores aprendam, ou seja, os algoritmos fazem os computadores tomarem decisões baseados em tentativas de resoluções de problemas anteriores bem-sucedidas. Dessa forma, pode-se dizer que os algoritmos criam um padrão o qual o computador irá se basear para resolver os problemas de mesma espécie, a terminologia da realização desse processo se chama aprendizado de máquina (subárea da inteligência artificial).

De acordo com Gomes (2015), em relação aos padrões que se deseja que as máquinas aprendam e a disponibilidade de dados para treinamento, normalmente separa-se em dois grupos de aprendizado, os quais são conhecidos como paradigmas de aprendizado de máquina: aprendizado supervisionado e não supervisionado. Destaca-se que os algoritmo J48, utilizados nesse trabalho de conclusão de curso é um algoritmo de aprendizagem supervisionado. Já em relação aos algoritmos não supervisionados um exemplo deles é o Apriori.

2.3.1 Algoritmo J48

Segundo Witten e Frank (2005) O algoritmo J48 surgiu devido a necessidade de recodificar o algoritmo C4.5, que inicialmente era um código da linguagem C, para a linguagem Java.

Conforme Librelotto e Mozzaquatro (2014), o J48 tem o objetivo de gerar árvores de decisão embasada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias presentes no conjunto de teste. Um dos motivos pelos quais os especialistas em data mining fazem a utilização frequente do algoritmo J48 é que o mesmo mostra-se apropriado para as tarefas, relacionadas as variáveis (dados) qualitativas contínuas e discretas presentes nas bases de dados. Ainda de acordo com Librelotto e Mozzaquatro (2014), o algoritmo J48, é um algoritmo de classificação e é apontado como o algoritmo que apresenta o melhor resultado na produção de árvores de decisão (na maioria das vezes), baseado em um conjunto de dados de treinamento.

Conforme Witten e Frank (2005), para que a árvore seja gerada, o J48 usa o método de abordagem de dividir-para-conquistar, dessa forma um problema de difícil resolução é dividido em diversos pequenos (relativamente) e mais simples problemas onde é utilizado da recursividade para que cada pequeno problema seja resolvido. Dessa forma, o problema complexo tende a ser resolvido. Segundo Gomes (2015), as árvores de decisão são geradas do topo para a base, a partir da seleção o atributo mais indicado para cada situação. O J48 é preparado para trabalhar com classes (classe a ser predita) binárias, nominais e valores faltantes de classe. Em relação aos atributos binários, de data, nominais, numéricos e valores faltantes.

2.3.2 Algoritmo Naive Bayes

Segundo Garcia et al. (2013), o classificador bayesiano, como o algoritmo Naive Bayes, é um simples classificador probabilístico fundamentado na aplicação do Teorema de Bayes com um certo grau de independência. Ou seja, um classificador que presume que o aparecimento de uma característica particular de uma classe, não está ligada com o aparecimento de qualquer outra característica. De acordo com Chimieski e Fagundes (2001), o classificador Naive Bayes gera uma estimativa de probabilidade, em vez de classificações maciças. Para cada valor de classe, o algoritmo Naive Bayes estima a probabilidade de uma suposta tupla pertencer a uma determinada classe.

Para Chimieski e Fagundes (2001), o algoritmo Naive Bayes prevê a probabilidade de uma tupla pertencer a uma classe especifica. Similarmente às árvores de decisão e classificadores de redes neurais, o algoritmo Naïve Bayesian demonstra alta precisão e velocidade, quando utilizados em bancos de dados.

Conforme Garcia et al. (2013), esse algoritmo é de grande eficiente em diferentes aplicações, pois é muito robusto em relação aos atributos irrelevantes. Além de que, requer um pequeno conjunto de dados de treino para poder estimar os critérios necessários para a classificação.

2.4 Árvores de Decisão

Segundo Queiroga (2017), as arvores de decisão se tratam de algoritmos de classificação supervisionada, pois nos mesmos é preciso saber quais são as classes de cada registro do conjunto de treinamento. Nesse trabalho de conclusão de curso, o algoritmo utilizado para gerar essas árvores foi o J48.

De acordo Queiroga (2017), algoritmos desta categoria geram uma estrutura de árvore que classificam as amostras desconhecidas. Ainda conforme Queiroga (2017), para esse grupo de algoritmo é preciso que ele sempre defina quais são os elementos da árvore, pois dessa maneira se torna possível a analogia com uma árvore real (presente na natureza), e observar seus nós conectados às ramificações.

Conforme Queiroga (2017), existem três tipos de nós: o nó raiz, que inicia a árvore, nós comuns que dividem uma determinada variável e geram algumas ramificações e também os nós folha que contém as informações de classificação do algoritmo. Segundo Quilan (1986) as ramificações contém todos os valores possíveis das variáveis indicadas no nó, para facilitar a interpretação e entendimento.

Conforme Queiroga (2017) com a arvore estruturada, cada nó recebe a tarefa de testar um atributo dos novos nós. Segundo Pichiliani (2008), pode-se dizer que a variável que melhor classificar os dados deve ser escolhido como um nó da arvore, para simplificar o entendimento é habitual colocar os valores das probabilidades de cada classe dentro do nó.

2.5 Matriz de Confusão

Conforme Monard e Baranauskas (2003), a matriz de confusão de uma hipótese h fornece uma medida efetiva do modelo de classificação, ao apresentar o número de

classificações corretas versus as classificações preditas para cada classe, sobre um conjunto de exemplos T.

2.6 Ferramentas Computacionais que auxiliam o processo de Mineração de Dados

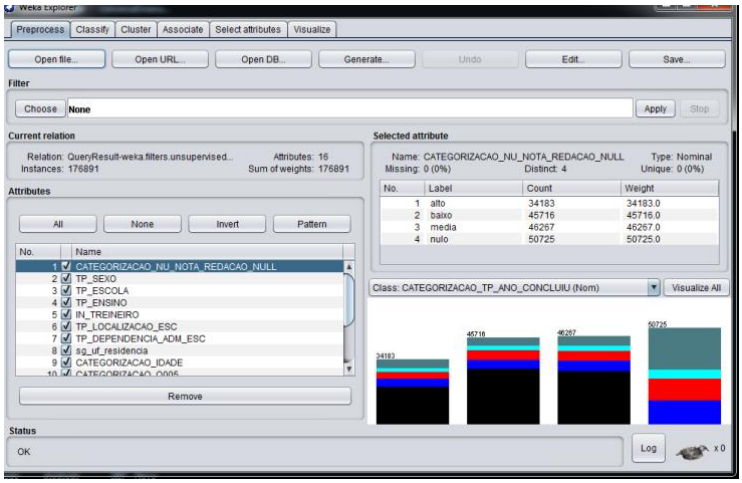
Nos dias atuais, encontram-se várias ferramentas computacionais que prestam um certo tipo de ajuda no processo de mineração de dados. Diversas delas são direcionadas a usuários finais. No decorrer desse referencial teórico destaca-se duas ferramentas desse tipo: Weka e Pentaho, que foram utilizadas nesse trabalho de conclusão de curso.

2.6.1 Weka

De acordo com Weka (2018), Weka se trata de uma coleção de algoritmos de aprendizado de máquina para realizar tarefas de mineração de dados. Esses algoritmos podem ser utilizados diretamente a um conjunto de dados, por meio da interface do Weka, ou podem serem chamados a partir de um código Java. O Weka conta também com ferramentas para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização. Também é utilizado para o desenvolvimento de novos esquemas de aprendizado de máquina.

Segundo Frank et al. (2010), para carregar os dados no Weka é possível utilizar várias fontes, incluindo arquivos, URLs (Uniform Resource Locator) e bancos de dados. Os formatos de arquivo suportados incluem o formato ARFF (Attribute-Relation File Format) do próprio Weka, formato CSV (Comma-separated values) e entre outros. Também é possível gerar dados usando uma fonte de dados artificial e editar os mesmo de uma forma manual usando um conjunto de dados editora.

Figura 2: Ferramenta Weka Explorer.



Fonte: Elaborado pelo autor.

2.6.2 Pentaho

Conforme Oliveira, Jesus e Braz (2015), o Pentaho é um software que simplifica a preparação, análise e visualização dos dados. Pentaho é uma plataforma completa de BI (Business Intelligence), criado em linguagem Java com vantagens que vão desde uma distribuição gratuita, até uma simples integração com muitas fontes de dados e aplicativos que utilizam padrões abertos. A ferramenta permite também o uso de APIs (Application Programming Interface) e tem flexibilidade nas opções de saída podendo ser em diversos formatos. Além disso, é uma ferramenta que pode ser usada em diversos sistemas operacionais.

Segundo Oliveira (2018), o Pentaho conta como uma ferramenta chamada Pentaho Data Integration, que é uma ferramenta ETL (Extração, Transformação e Carregamento), que permite acessar e preparar fontes de dados para análise, mineração e geração de relatórios OLAP (Online Analytical Processing).

2.7 Dados abertos

De acordo com a Open Knowledge International (2018), dados são abertos quando qualquer pessoa pode acessar livremente, usá-los e redistribuí-los, estando sujeito a obrigação de referenciar a sua autoria e compartilhar pela mesma licença. Os dados abertos podem ser de diversos tipos como: cultura, finanças, estatísticas, clima, ambiente, educação e entre outros.

Existem diversos motivos para a abertura de dados, na gestão pública por exemplo. Segundo Abertos (2018), existem 5 motivos que levam a abertura dos dados nesse setor, são eles: transparência na gestão pública, contribuição da sociedade com serviços inovadores ao cidadão, aprimoramento na qualidade dos dados governamentais, viabilização de novos negócios e obrigatoriedade por lei. A partir desse contexto, pode se dizer que os dados abertos são de grande importância para sociedade, pois através deles é possível verificar a atuação do governo em várias áreas, como transporte, saúde e educação. Além de poderem ser objeto de estudo de diversas pessoas podendo então surgir inovações baseadas nesses dados ou até mesmo a analise desses dados para resolução de um problema da área de onde é oriunda a base de dados.

De acordo com Open Knowledge Brasil (2014), em 2014 a Fundação do Conhecimento Aberto realizou um estudo com o objetivo de saber o Índice Global de Dados Abertos. O mesmo relata que apesar de ter ocorrido um certo avanço, a maioria dos governos ainda não disponibiliza informações chave e em formato acessível para seus cidadãos e empresas. A pesquisa traz um ranking de países fundamentado na disponibilidade e acessibilidade de informações em dez áreas, que incluem gastos governamentais, resultados eleitorais, horários dos meios de transportes e níveis de poluição. Na pesquisa publicada em 2014, o Brasil obteve uma pequena melhora, 54% contra 48% na edição anterior, mas mesmo assim sofreu uma queda no ranking geral já que na pesquisa anterior alcançou a 24º posição, enquanto na de 2014 ficou em 26º lugar.

2.7.1 Dados abertos governamentais

Para Ribeiro e Almeida (2011), os Dados Abertos Governamentais são compreendidos como o esforço para a publicação e divulgação das informações do setor público na Web, permitindo a reutilização e a integração destes dados.

Segundo Ribeiro e Almeida (2011) dados do governo estão sendo publicados na web para ampliar a responsabilidade, fornecer informações valiosas sobre o mundo e permitir que o governo, o país e o mundo funcionem com mais eficiência. Berners-lee (2009) complementa que está divulgação de dados abertos permite que a sociedade tenha uma visão mais clara sobre o desempenho do governo em relação aos objetivos acordadas, bem como sobre o desenvolvimento de políticas públicas.

Conforme Diniz (2010), publicar dados na web vem sendo cada vez mais comum, seja em uma página web, ou até mesmo em um arquivo para "*download*". No entanto a publicação de dados abertos presume que algumas características sejam respeitadas, de modo que garanta que os mesmos possam ser acessados e reutilizados por máquinas.

De acordo com Diniz (2010) os Dados governamentais abertos deverão garantir as seguintes características:

- Possuir independência de plataformas tecnológicas.

- Ser baseado em formatos padronizados. A garantia de progresso e melhoria contínua da representação dos dados está nas tecnologias amparadas por organismos internacionais de padrões como W3C (World Wide Web Consortium) e ISO (International Standardization Organization).

- Os dados devem estar desvinculados das páginas web, relatórios ou ferramentas que os conceberam.
- O formato utilizado para representação dos dados, deve permitir o manuseio dos mesmos através de máquinas. Além disso os dados deverão estar estruturados, pois uma estruturação adequada permite que terceiros possam fazer uso automatizado dos dados. Formatos que somente podem ser visualizados, e não manuseados e extraídos, não são bons e devem ser evitados.

- Cada conjunto de dados terá de disponibilizar uma "descrição externa de si próprio (metadados) de tal forma que seja identificada a sua natureza, conheça se a sua origem e qualidade e seja possível um estudo dos dados através de um conjunto de instruções de máquina que descreve os dados e suas relações."

- Sempre que for viável, deve ser inserido conteúdos semânticos no código da página web onde os dados estão disponíveis. Pois assim facilita a leitura dos dados por outras máquinas e dessa forma os mecanismos de buscas como Google ou Yahoo encontrarão os dados mais facilmente.

- Quando a disponibilização de dados for por meio de uma interface de aplicativos, os dados devem ser separados da interface.

- Sempre que for viável, deve ser criado URIs (Identificador Uniforme de Recursos) para cada objeto dos seus dados. A URI se trata de um padrão de codificação para fornecer uma representação numérica universal e sem ambiguidade para cada objeto de maneira independente da plataforma de software e do idioma.

Ainda baseado em Diniz (2010), a publicação dos dados e disseminação deverá ser feita no ambiente da rede mundial de computadores, expondo claramente em catálogo construído especificamente para tal o caminho para encontrá-los, bem como os seus respectivos metadados. Quanto maior for a quantidade de formatos que os conjuntos de dados estiverem publicados, maiores serão as chances dos usuários terem acesso aos dados.

2.7.2 **Modelo de ranqueamento**

Segundo Berners-lee (2006) o mesmo elaborou um modelo de ranqueamento de 1 a 5 estrelas (similar ao método de ranqueamento dos hotéis), que tem o objetivo de medir a qualidade dos dados abertos, ou seja, quanto mais estrelas o dado receber significa que mais poderoso ele é e que possui maior facilidade para utilização. No Quadro 1 são apresentados os graus e regras desse ranqueamento:

Quadro 1: Modelo de ranqueamento.

Número de estrelas:	Regra:
1	Os dados devem estar, disponível na Web em qualquer formato (pdf, imagens), mas com uma licença aberta, para ser Open Data.
2	Os dados precisam ser disponibilizados como dados estruturados que sejam legíveis por máquina (por exemplo, excel em vez de digitalização de imagem de uma tabela).

3	Similar ao ranqueamento 2 porém com formato não proprietário (por exemplo, CSV em vez de excel).
4	Deve seguir todas as regras acima e utilizar de Padrões W3C (RDF e SPARQL).
5	Deve seguir todas as regras acima e deve vincular seus dados aos dados de outras fontes para fornecer contexto.

Fonte: Adaptado de Berners-lee (2006).

Os dados abertos utilizados neste trabalho de conclusão de curso, são os dados do (ENEM 2016), considerando o ranqueamento elaborado por Berners-lee (2006), é constatado que os mesmos estão na classificação de 3 estrelas, pois estão presentes na web perante a licença aberta, são estruturados e legíveis por maquinas e por fim utilizam formato CSV. Sendo assim, classificados como 3 estrelas, os dados utilizados nesse TCC, não utilizam Padrões W3C (RDF e SPARQL) e também não são integrados aos dados de outras fontes.

2.8 ENEM

De acordo com Andriola (2011), o ENEM foi criado em 1998, com o objetivo de avaliar o desempenho do estudante ao fim da escolaridade básica, pretendendo aferir o desenvolvimento das competências e habilidades fundamentais ao exercício pleno da cidadania. Segundo Travitzki (2013), inicialmente tímido, o Exame Nacional do Ensino Médio foi crescendo ano após ano, realizando diversas mudanças, aumentando os objetivos, e tornou-se uma avaliação consolidada no sistema educacional brasileiro, cada vez mais procurado pelos estudantes.

Segundo Andriola (2011), o Ministério da Educação apresentou uma proposta de reformulação do ENEM, no ano de 2010 , onde ficou decretado que o mesmo passaria a ter 180 questões ao invés de 63 como era até 2008, além disso o ENEM passou a ser aplicado em dois dias. Além de que começou a explorar quatro áreas do conhecimento humano, dividindo as questões igualmente (45 questões) para cada um das seguintes áreas do conhecimento:

- Linguagens, códigos e suas tecnologias –Literatura, Educação Física, Língua Estrangeira, Língua Portuguesa, (Inglês ou Espanhol), Artes e Tecnologias da Informação e Comunicação.

- Matemática e suas tecnologias.
- Ciências da Natureza e suas tecnologias – Biologia, Física e Química.
- Ciências Humanas e suas tecnologias – Filosofia, História, Geografia e Sociologia

Podemos notar que com o passar do tempo o ENEM, foi aos poucos se organizando e se ligando a outras políticas federais, como por exemplo o Programa Universidade para Todos (PROUNI) e também o Fundo de Financiamento Estudantil (FIES). Embora um pequeno grupo de universidades, já usassem os resultados do ENEM como parte do processo seletivo desde o ano de 2000, foi há pouco tempo que o mesmo se consolidou como um tipo de vestibular nacional, sendo utilizado por praticamente todas as instituições federais através de um sistema próprio chamado de Sistema de Seleção Unificada (SISU). Outra mudança impactante feita no ENEM, foi a possibilidade do estudante conseguir a isenção da taxa de inscrição concedida pela primeira vez em 2001. Conforme INEP (2018), atualmente quatro perfis têm direito à isenção, sendo eles:

- O pretendente que estiver cursando a última série do Ensino Médio no ano atual do exame, em qualquer modalidade de ensino, em escola da rede pública declarada ao Censo Escolar.
- O pretendente que obteve a Certificação de Conclusão do Ensino Médio, por meio Exame Nacional de Certificação de Competências de Jovens e Adultos (ENCCEJA) ano anterior ao exame.
- O pretendente que concluiu todo o Ensino Médio em escola da rede pública ou como bolsista integral na rede privada e tenha renda per capita igual ou inferior a um salário mínimo e meio.
- O pretendente que declare estar em situação de vulnerabilidade socioeconômica por ser membro de família inscrita no Cadastro Único para Programas Sociais (CadÚnico).

Conforme INEP (2018), nos dias atuais o principal objetivo do ENEM é a análise do desempenho escolar e acadêmico ao final do Ensino Médio. Os resultados podem:

- Proporcionar o estabelecimento de parâmetros para a autoavaliação do participante, buscando a continuidade de sua formação e a sua introdução no mercado de trabalho;
- Proporcionar a criação de referência nacional para a melhoria dos currículos do Ensino Médio;
- Ser utilizados como mecanismo único, alternativo ou complementar para o ingresso na Educação Superior, especialmente, a disponibilizada pelas instituições federais de educação superior;
- Proporcionar a entrada do participante em programas governamentais de financiamento ou apoio ao estudante da Educação Superior;
- Ser usado como instrumento de seleção para ingresso nos diferentes setores do mundo do trabalho;
- Possibilitar o desenvolvimento de estudos e indicadores sobre a educação brasileira.

O Quadro 2 apresenta de forma mais objetiva diversos eventos e mudanças realizadas no ENEM:

Quadro 2: História e eventos do ENEM.

Ano:	Eventos e mudanças importantes:
1998	Criação do ENEM.
2000	Pequeno grupo de universidades começam a usar o ENEM como parte do critério de seleção.
2001	Criação da isenção de taxa de inscrição para alunos desfavorecidos.
2004	É criado o PROUNI.
2005	PROUNI é vinculado ao ENEM.
2006	Começam a ser divulgadas as médias do ENEM por escola.
2007	É criado o REUNI.

2009	É criado o SISU.
2010	ENEM começa a valer como certificação do ensino médio.
2011	ENEM passa a ser obrigatório para participantes requisitarem o FIES.

Fonte: Adaptado de Travitzki (2013).

2.9 Trabalhos relacionados

Nos últimos anos o cenário de Mineração de dados educacionais tem crescido muito, dessa forma esta importante área de pesquisa que possibilita a análise de um grande conjunto de dados, vem ajudando na solução de problemas voltados a educação.

Segundo Rodrigues et al. (2014), na área de mineração de dados educacionais, do ano de 2006 há 2010 o número de publicações não chegou há 20. Porém, com a crescente evolução da área, do ano de 2011 até o mês de maio de 2014 foram publicados em torno de 47 artigos, mostrando uma grande evolução em função do tempo.

No trabalho de Gomes (2015), foram utilizados os dados do ENEM de 2013 e 2014, organizando os mesmos em quatro bases diferentes levando em conta apenas a região do Nordeste e do estado de Pernambuco. A base de dados do ano de 2014 que possuía os dados dos inscritos do Nordeste, conta com um total de 2.444.754 instâncias, já a base de 2013 possui 2.358.851 instâncias.

O estudo em questão buscou incialmente encontrar algumas regras de associação através do algoritmo Apriori. Para encontrar essas regras Gomes (2015), manteve os padrões da ferramenta WEKA, porém o mesmo utilizou um parâmetro de confiança de 80% e alterou a quantidade de regras a encontrar-se para 30. Em um segundo momento Gomes (2015), realizou a análise dos dados para gerar algumas estatísticas, um exemplo foi a descoberta de que aproximadamente 53% dos inscritos que declararam renda familiar de até 2 salários mínimos tiveram um desempenho máximo de 500 pontos na prova de matemática.

Em um terceiro momento Gomes (2015), realizou análise de desempenho utilizando o algoritmo J48 com o objetivo de analisar o desempenho dos inscritos do cenário regional e também do cenário local, nas quatro áreas de conhecimento. O atributo classificador que foi utilizado foi o gênero, e os parâmetros foram mantidos os padrões do J48, porém foram utilizados 60% dos dados para criação do modelo. No cenário estadual, os modelos que foram

criados, apontaram que candidatos do sexo feminino normalmente tem notas inferiores ou iguais a Nota Média.

Ainda sobre trabalho de Gomes (2015), o mesmo realizou a análise de redações com notas zero com o objetivo de encontrar possíveis padrões. Nessa análise foi utilizado o cenário nacional que conta com um total de 328.440 instâncias. Para tentar atingir esse objetivo foi usado o algoritmo J48 onde o status da redação é o atributo classificador.

Simon e Cazella (2017) realizaram um estudo que possuía o objetivo de gerar um modelo preditivo do indicador de desempenho médio na área de ciências da natureza e suas tecnologias dos alunos de escolas do ensino médio através dos dados abertos do ENEM 2015.

Para esse estudo ser realizado Simon e Cazella (2017), converteram o arquivo ".XLSX", que possuía os dados, para ".CSV" devido a limitações do software WEKA. Desta forma a análise foi feita a partir de um arquivo com as nove colunas selecionadas e 15599 instâncias, com a identificação das colunas e as escolas as quais os resultados são informados.

Simon e Cazella (2017) utilizaram a variável Média Escola para indicar o desempenho médio dos estudantes da escola, na área de ciências da natureza e suas tecnologias, devido a forma como o algoritmo utilizado trabalha (J48) foi necessário realizar a categorização da mesma. Através do software WEKA os mesmos realizaram a análise por árvore de decisão usando o algoritmo j48, que foi feita a partir da opção de cross-validation, com o valor para fold igual a dez e com a variável a ser predita Média Escola. A árvore de decisão gerada por Simon e Cazella (2017) conseguiu acertar 77,02% instâncias das 15998 inseridas.

No trabalho de Adeodato, Santos Filho e Rodrigues (2014), foram usados os microdados do ENEM 2011 e também os dados do censo escolar 2011, que detalha as condições das escolas secundarias e infraestrutura do corpo docente, somente as escolas privadas com mais de 15 alunos foram selecionadas. Como o objetivo do estudo se tratava de dizer se a escola era boa ou não.

Para poder classificar as escolas como boas, Adeodato, Santos Filho e Rodrigues (2014) precisaram se preocupar com duas questões: que métrica seria usada para avaliar a qualidade da escola e que limiar seria adotado como critério para definir o que seria uma escola "boa". A partir dessas questões Adeodato, Santos Filho e Rodrigues (2014) consideraram a média aritmética das notas dos alunos como medida de qualidade das escolas (processo similar ao que é utilizado no ranking das escolas). A seguir os mesmos, definiram o quartil superior como limiar de binarização da nota para caracterizar o objetivo como escola forte ou fraca.

Ainda sobre o estudo de Adeodato, Santos Filho e Rodrigues (2014), a regressão logística gerou um classificador capaz de definir uma pontuação de tendência ao sucesso da

escola, por meio das suas características e também dos seus docentes e discentes e famílias. Além disso, no trabalho dos mesmos ele utilizaram árvore de decisão para extrair o conhecimento explicitando.

Outras técnicas também utilizadas, criaram regras similares que ajudaram os resultados e evidenciaram que os principais fatores que influenciam a boa qualidade das escolas estão ligados a situação econômica e financeira, seja de maneira direta (renda familiar) ou indiretamente, ou em aspectos culturais (nível de educação da mãe ou do pai) da família.

3 METODOLOGIA E EXPERIMENTOS

Esta seção descreve em detalhes o contexto em que os dados foram usados, bem como os processos utilizados na metodologia, do desenvolvimento desse trabalho de conclusão de curso, que foram: Seleção e Pré-processamento dos dados, Transformação dos dados, Seleção dos dados de Santa Catarina e Araranguá, Geração e avaliação dos modelos de predição. Além disso essa seção apresenta também a descrição dos experimentos realizados.

3.1 Contexto

Para a elaboração deste trabalho foram utilizados e analisados os dados abertos do Exame Nacional do Ensino Médio 2016 (ENEM), sendo estes obtidos na página web do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), onde existe a disponibilidade de diversas bases de dados para download. Conforme INEP (2007), nos Microdados do ENEM são apresentados as provas, gabaritos, informações sobre as questões, notas dos candidatos e também o questionário socioeconômico que foi respondido pelos inscritos em 2016. Essa base de dados é dividida em algumas seções:

- Dados do participante;
- Dados da escola;
- Dados dos pedidos de atendimento especializado;
- Dados dos pedidos de atendimento específico;
- Dados dos pedidos de recursos especializados e específicos para realização das provas;
- Dados dos pedidos de certificação do ensino médio;
- Dados do local de aplicação da prova;
- Dados da prova objetiva;
- Dados da redação;
- Dados do questionário socioeconômico;

Segundo o INEP (2007), para simplificar a utilização dos dados, o arquivo principal, “Microdados _Enem_2016”, contém todos os dados reunidos em um único arquivo. Já o Dicionário de dados mostra a separação de cada seção e contém explicação sobre as variáveis contidas em cada base. Destaca-se ainda que os dados são disponibilizados em formato “.CSV”

e que grande parte das informações como os dados do participante e questionário socioeconômico são coletados durante a realização da inscrição do candidato através das respostas dos mesmos, porém outros dados como as notas das provas objetivas e redação são os resultados das avaliações da prova prestada.

Os dados abertos do ENEM 2016 contam com os dados de todos os alunos que se inscreveram no ENEM, ou seja, a base contém os dados do país inteiro contando com um total de 8627367 instâncias. Restringindo a base de dados, para contar com as informações apenas de alunos de Santa Catarina há uma significativa queda no número de instâncias, passando a contar com 176891 instâncias.

Baseado no contexto da base de dados citada anteriormente e nas informações do INEP acredita-se que a mesma possui potencial para a realização de predições de diversos tipos. Dessa forma este trabalho de conclusão de curso propõem modelos de predição para a prova de redação do ENEM, pois acredita-se que dessa forma os gestores e especialistas da área educacional possam ajudar os estudantes a melhorar seu desempenho no ENEM.

Os quantitativos dos totais de instâncias presentes na base de dados e que foram utilizados neste projeto estão explícitos no Quadro 3, já o Quadro 4 demonstra os totais de instância em relação aos tipos de escolas.

Quadro 3: Quantitativos dos totais de instâncias.

Tipo Registro	Quantidade de registros
Registros no âmbito nacional (total de registros de toda a base de dados)	8627367 instâncias
Sexo masculino no âmbito nacional	3644728 instâncias
Sexo feminino no âmbito nacional	4982639 instâncias
Registros no âmbito estadual (total de registros no estado de Santa Catarina)	176891 instâncias
Sexo masculino no estadual	76082 instâncias
Sexo feminino no âmbito estadual	100809 instâncias

Fonte: Elaborado pelo autor.

Quadro 4: Quantitativos dos totais de instâncias em relação aos tipos de escolas.

Tipo de escola	Quantidade de registros
Pública (âmbito nacional)	1561876 instâncias
Privada (âmbito nacional)	319415 instâncias

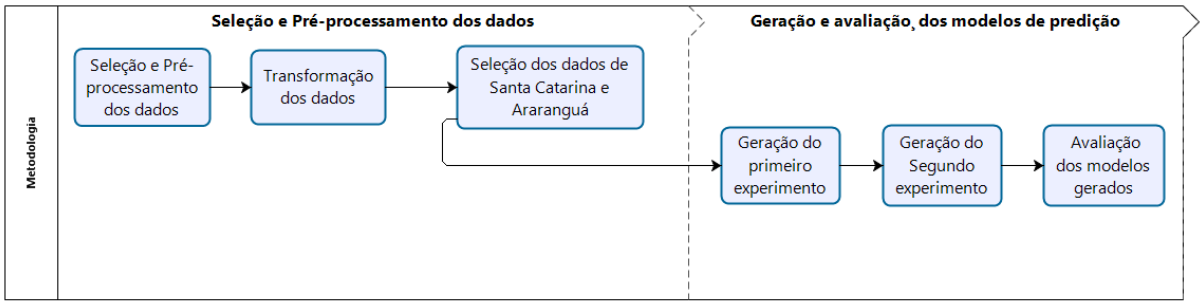
Não foi respondida pelo inscrito (âmbito nacional)	6745091 instâncias
Escola do Exterior (âmbito nacional)	985 instâncias
Pública (âmbito estadual)	111560 instâncias
Privada (âmbito estadual)	53890 instâncias
Não foi respondida pelo inscrito (âmbito estadual)	11409 instâncias
Não foi respondida pelo inscrito (âmbito estadual)	32 instâncias

Fonte: Elaborado pelo autor.

3.2 Metodologia

A metodologia utilizada para a elaboração desse trabalho é composta das seguintes fases: Seleção e Pré-processamento dos dados, geração e avaliação dos modelos de predição. A Figura 3 demonstra de forma mais clara os processos e subprocessos utilizados na metodologia.

Figura 3: Processos utilizados na metodologia.



Fonte: Elaborado pelo autor.

3.3 Seleção e Pré-processamento dos dados

Após o *download* dos microdados do ENEM 2016 no site do INEP, notou-se que diversos erros ocorriam ao tentar abrir o arquivo (que contém todos os dados) com os editores eletrônicos comuns de planilhas e também de texto, Bloco de Notas, Microsoft Excel, Calc, entre outros.

Dessa forma observou-se que os mesmos não eram capazes de abrir corretamente o arquivo, possivelmente devido ao tamanho do arquivo que em seu formato original (.CSV), possui um tamanho de mais de 5 gigabytes e conta com mais de 8 milhões de instâncias. Outra coisa também observada logo no início foi que o software WEKA também possuía uma limitação para carregar os dados provenientes de um ".CSV", limitação essa que impedia a geração de um arquivo ".ARFF" a partir da planilha que apresentava os dados do ENEM 2016, esse problema também está ligado ao tamanho do arquivo disponibilizado pelo INEP.

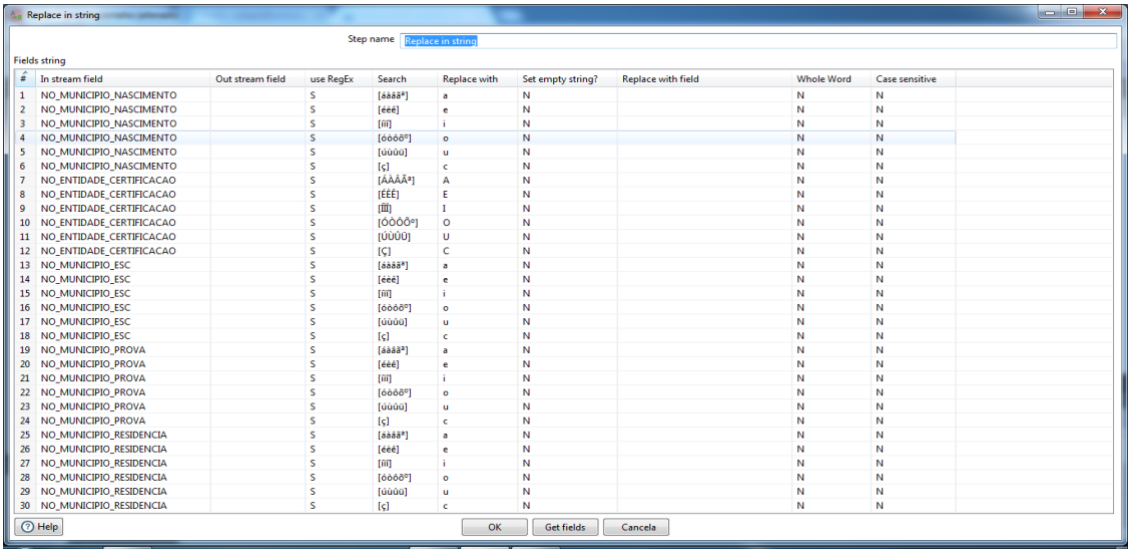
Dessa forma jugou-se necessário a criação de uma base de dados, através de um banco de dados e de uma Linguagem de definição de dados (DDL), para armazenar todas as informações presentes no arquivo original. Devido a isso necessitou-se também fazer o uso de uma ferramenta de Extração Transformação Carregamento (ETL) para preencher o banco de dados, local onde os dados utilizados nesse trabalho ficaram salvos. A ferramenta escolhida para efetuar essa tarefa chama-se Pentaho Data Integration, em sua versão de avaliação.

A primeira etapa realizada no Pentaho, para que a base de dados criada no banco de dados, fosse preenchida se trata da etapa de **CSV file input,** ou seja, nessa etapa deve-se selecionar a base de dados ".CSV", a qual se deseja fazer limpeza dos dados e inserção no banco de dados. A segunda etapa realizada foi a **Replace in string**, que se trata de um recurso do Pentaho que permite fazer a remoção das acentuações, caracterizando-se então como um processo de limpeza de dados. Dessa forma as variáveis a seguir, que continham acentuação em seu conteúdo, passaram por esse processo de remoção dos mesmos:

- NO_MUNICIPIO_NASCIMENTO
- NO_ENTIDADE_CERTIFICACAO
- NO_MUNICIPIO_ESC
- NO_MUNICIPIO_PROVA
- NO_MUNICIPIO_RESIDENCIA

As demais variáveis não passaram pelo processo de remoção dos acentos, pois as mesmas não possuíam acentuação. Como pode-se observar na Figura 4, que demonstra de uma forma mais clara esse processo de configuração para remoção das acentuações, na coluna In stream field é o local onde foi colocado todas as variáveis que iriam passar por esse processo de limpeza, na coluna Search foi posto todos os conjuntos de caracteres acentuados que se desejava buscar para ser substituído. Já na coluna Replace with foi preenchido com o caractere que iria substituir os caracteres buscados através da coluna Search.

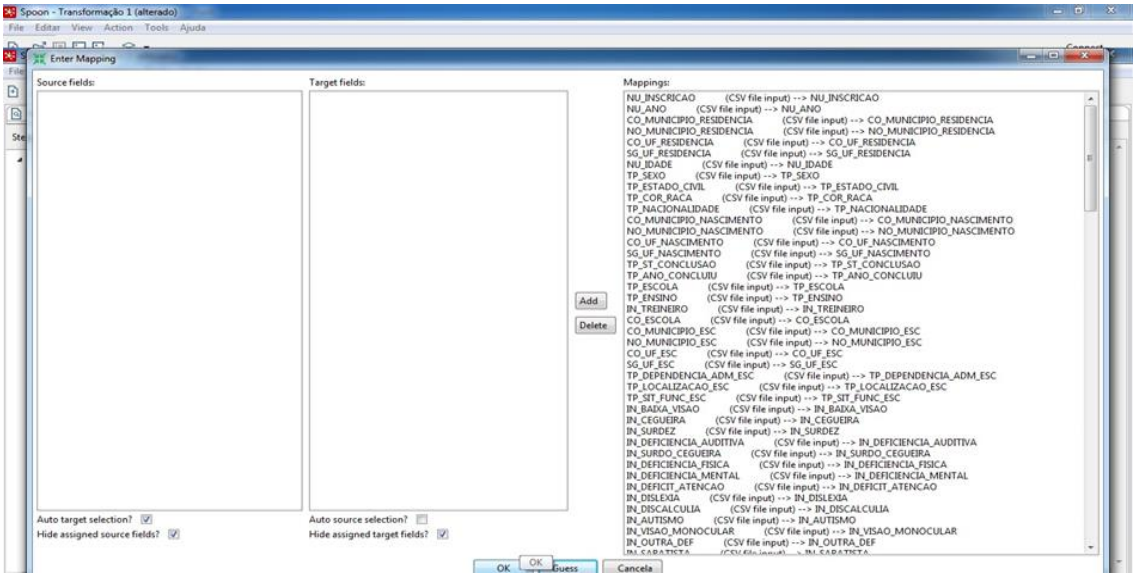
Figura 4: Processo de configuração para remoção das acentuações.



Fonte: Elaborado pelo autor.

A etapa seguinte realizada no Pentaho foi a tarefa de **Table output**, etapa essa a qual deve-se selecionar a base de dados que se deseja preencher com os dados oriundos da etapa de **CSV file input**. Após ter selecionada a base de dados a ser preenchida, foi necessário realizar a etapa de **Table output Mapping**, pois é nessa etapa que é feito o mapeamento, ou seja, deixa-se explicito qual campo/coluna do arquivo selecionado na etapa de **CSV file input,** irá preencher o campo/coluna do base de dados selecionadas na tarefa de **Table output**. A Figura 5 demonstra esse processo de **Table output Mapping**.

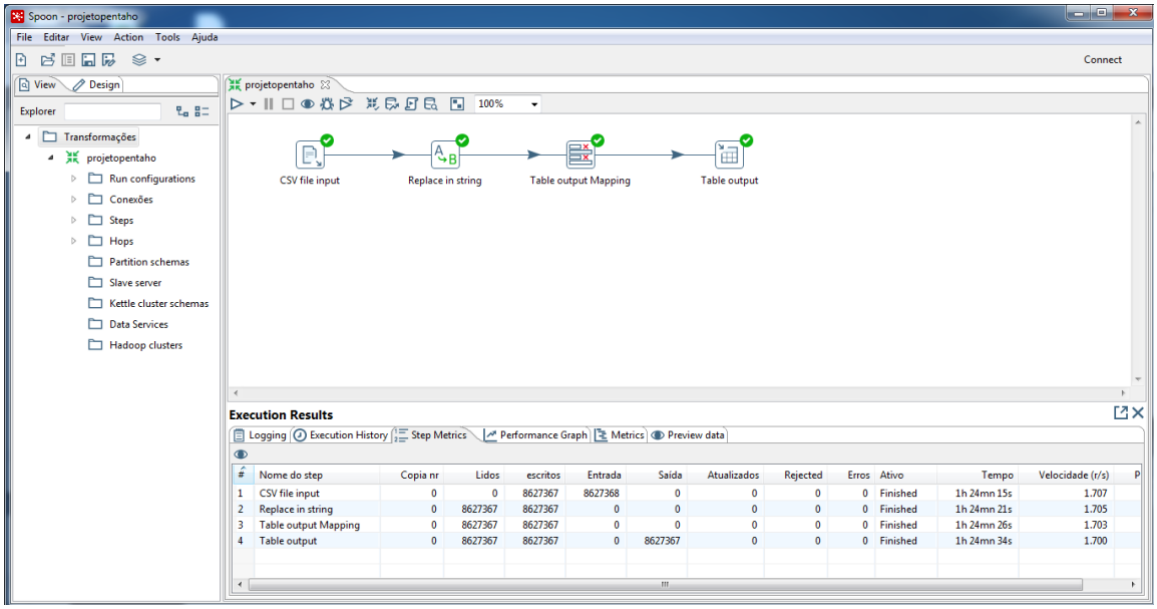
Figura 5: Processo de Table output Mapping



Fonte: Elaborado pelo autor.

A Figura 6, demonstra claramente as etapas realizadas no Pentaho, e o fluxo percorrido pelos dados, para que a base de dados criada no banco de dados fosse preenchida.

Figura 6:Fluxo percorrido pelos dados no Pentaho



Fonte: Elaborado pelo autor.

3.3.1 Transformação dos dados

Com o objetivo de usar os dados, já presentes na base, como input para os algoritmos de mineração, é preciso realizar a etapa de transformação dos dados, para que os mesmos fiquem adequados para esses algoritmos e para possivelmente melhorar o desempenho.

Dessa forma após todas as etapas de pré-processamento dos dados, o próximo passo a ser realizado nesse trabalho foi a transformações dos dados. Uma das principais tarefas realizadas nessa etapa de transformações dos dados foi a categorização, pois a partir da mesma se torna possível categorizar as variáveis fazendo com que os valores da mesma, fiquem divididos em categorias diminuindo então a amplitude desses valores, podendo dessa forma fazer com que os algoritmos de mineração de dados tenham resultados melhores.

A primeira variável a passar por esse processo de categorização foi o atributo a ser predito "NU_NOTA_REDACAO", atributo este que contém as notas tiradas pelos participantes do ENEM na prova de redação. Para a categorização dessa variável foi criado um script em java, este código tem o objetivo de descobrir em quais pontos deveriam serem feitos os cortes

para uma divisão por tercil, que se trata de uma estratégia para categorização onde divide-se um conjunto em 3 partes.

Para a realização dessa divisão, elaborou-se uma função que realizava a importação dos dados presentes no banco de dados, através de uma *Query* SQL (Linguagem de Consulta Estruturada), que selecionava todas as instâncias que não fossem nulas (não possuíssem seu valor como "NULL"), pois as mesmas por si só já formariam um grupo.

A Figura 7, demonstra a função criado para importar as instâncias do banco de dados, deixando clara então, qual foi a *Query* SQL utilizada para a categorização da variável "NU_NOTA_REDACAO". A Figura 8 expõe como foi implementada a função responsável por de fato informar os pontos que deveriam ser realizados cortes para categorização.

Figura 7: Função para importar Instâncias do banco de dados.

```
public void ImportarInstancia() {

    try {
        PreparedStatement stmt;
        String estado = "SC";
        String sql
    = "select NU_INSCRICAO,NU_NOTA_REDACAO from tccenem.enem where SG_UF_RESIDENCIA='SC' and  nu_nota_redacao IS NOT NUll ";

        stmt = con.prepareStatement(sql);
        ResultSet rs = stmt.executeQuery(sql);
        int i = 0;

        while (rs.next()) {
            Enem enem = new Enem((rs.getString("NU_INSCRICAO")), (rs.getString("NU_NOTA_REDACAO")));
            vetor[i] = enem;
            vetorpercentile[i] = Double.parseDouble((rs.getString("NU_NOTA_REDACAO")));
            i++;

        }

        rs.close();
        stmt.close();

        Arrays.sort(vetorpercentile);
        percentile(33);
        percentile(66);

    } catch (SQLException ex) {

        JOptionPane.showMessageDialog(null, "ERRO no importar instancias do bd: " + ex);
    }

}
```

Fonte: Elaborado pelo autor.

Figura 8: Função para descobrir pontos de corte.

```
public void percentile(int descobrir) {

    int k = 0;
    double c;
    int i = descobrir; //posição que se deseja saber
    int n = vetorpercentile.length;

    i = descobrir;
    c = (i * (n + 1));
    c = (c / 100);
    int parteinteira = (int) c;
    double partefracionada = c - parteinteira;
    int arredonda;

    if (parteinteira != c) {
        if (partefracionada >= 0.5) {
            System.out.println(" \n arredonda pra cima ");
            arredonda = parteinteira + 1;
            c = vetorpercentile[parteinteira - 1] + vetorpercentile[arredonda - 1];
            c = c / 2;
            System.out.println(" \n valor percentil " + c);

        } else {
            System.out.println(" \n arredonda pra baixo ");
            arredonda = parteinteira - 1;
            c = vetorpercentile[parteinteira - 1] + vetorpercentile[arredonda - 1];
            c = c / 2;
            System.out.println(" \n valor percentil " + c);
        }
    } else {
        System.out.println(" \n posição inteira: ");
        c = vetorpercentile[parteinteira];
        c = c / 2;
        System.out.println(" \n valor percentil " + c);
    }
}
```

Fonte: Elaborado pelo autor.

Após serem descoberto os valores onde deveriam serem feitas as limitações para cada categoria da classificação, foi criado um novo campo no banco de dados, através de linguagem DDL, chamado de "CATEGORIZACAO_NU_NOTA_REDACAO_NULL". Esse novo campo recebeu os valores de "NU_NOTA_REDACAO" categorizados, através de um *script* SQL executado no próprio banco de dados, como constata a Figura 9.

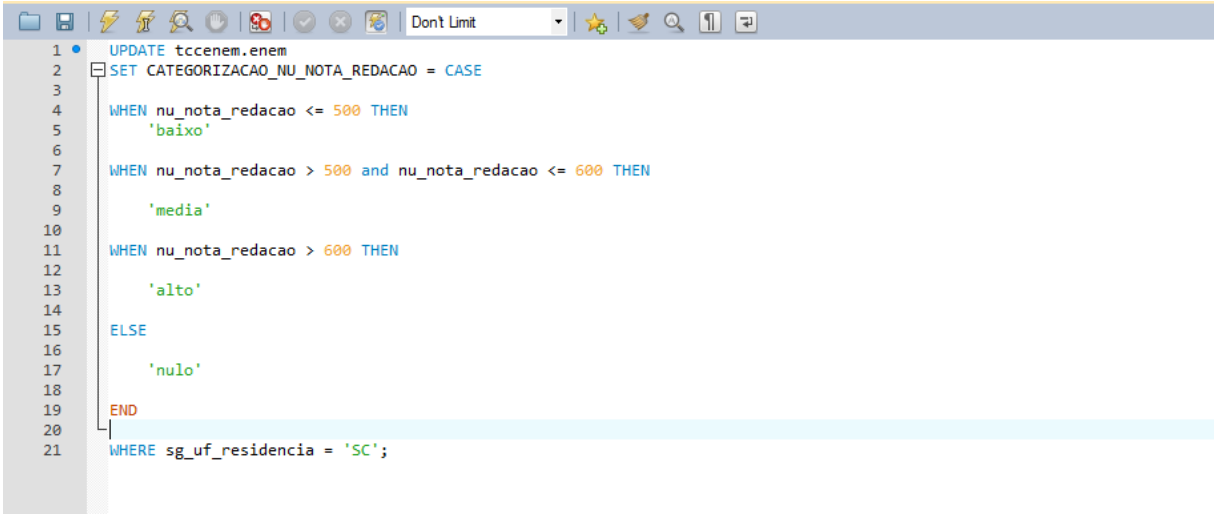
As categorias criadas foram: baixo, media, alto e nulo. Para que as instâncias fossem classificadas como "baixo', deveriam possuir uma nota de até no máximo 500. Já para serem classificadas como "media" a nota deveria ser maior que 500 e menor ou igual a 600. Para receber a classificação como "alto" a nota teria que ser maior que 600. Já a regra para que as instâncias fossem atribuídas como "nulo" a mesma deveria conter valor NULL. O Quadro 5 demonstra com maior clareza as regras para a classificação das notas da redação.

Quadro 5: Regras para a classificação das notas da redação.

Tipo da categoria:	Regra:
baixo	NU_NOTA_REDACAO <=500
media	NU_NOTA_REDACAO>500 e NU_NOTA_REDACAO<=600
alto	NU_NOTA_REDACAO>600
nulo	NU_NOTA_REDACAO=NULL

Fonte: Elaborado pelo autor.

Figura 9: Efetuando a categorização da variável "NU_NOTA_REDACAO".



Fonte: Elaborado pelo autor.

Em relação as variáveis socioeconômicas, dos 50 atributos, um total de 40 foram categorizadas, cada uma delas obedecendo um conjunto de regras específicos, essas informações estão presentes no Apêndice A. As demais variáveis socioeconômicas não possuíam a necessidade de serem categorizadas.

Dos 40 atributos categorizados, da seção de dados do questionário socioeconômico, destaca-se a categorização da variável Q005, que possui os valores do número de moradores da residência do candidato, a mesma foi categorizada pois possuía uma amplitude de valores muito grande, como pode-se observar através do **Anexo A**. O outro atributo enfatizado é a variável Q006, a mesma armazena a renda mensal da família do candidato, e também foi categorizada, pelo mesmo motivo do atributo Q005, ou seja, uma grande amplitude como demonstrado no **Anexo A**. Já o Quadro 6 e o Quadro 7 demonstra como ficaram classificadas e quais as regras utilizadas para a classificação das variáveis Q005 e Q006 respectivamente.

Quadro 6: Regras para a classificação da variável Q005.

Tipo da categoria:	Regra:
QtdMorador1	Q005=1
QtdMorador2	Q005>=2 e Q005 <=4
QtdMorador3	Q005>=5 e Q005 <=7
QtdMorador4	Q005>=8

Fonte: Elaborado pelo autor.

Quadro 7: Regras para a classificação da variável Q006.

Tipo da categoria:	Regra:
MuitoBaixo	Q006=A OR B
Baixo	Q006= C OR D OR E
Medio	Q006= F OR G OR H OR I
Alto	Q006 # dos valores presentes nas regras anteriores

Fonte: Elaborado pelo autor.

Da seção dos dados do participante as únicas variáveis a ser categorizadas foi o atributo “NU_IDADE” e o atributo “TP_ANO_CONCLUIU”. O campo “NU_IDADE”

armazena a idade do participante, já em relação a variável “TP_ANO_CONCLUIU”, é o local onde está registrado o ano de conclusão do ensino médio do participante. A seção de dados da prova objetiva, contou com as seguintes variáveis categorizadas: “NU_NOTA_CN”, “NU_NOTA_CH”, “NU_NOTA_LC”, “NU_NOTA_MT”.

O procedimento utilizado para criar os campos que armazenam a classificação de cada instância das variáveis da seção de: dados do participante e dados da prova objetiva, foi o mesmo utilizado pela classificação da variável “NU_NOTA_REDACAO” presente na seção de dados da redação. Assim como nas variáveis oriundas da seção de dados do questionário socioeconômico, cada atributo obedeceu um conjunto de regras específicos, que estão informados no Apêndice A**.**

Para categorizar a variável idade por exemplo, foi utilizada a regra de que se o valor presente no campo “NU_IDADE” fosse menor ou igual a 18, a variável “CATEGORIZACAO_IDADE”, seria classificada como “idade1”. Já se o valor fosse maior que 18 e menor ou igual a 25 a mesma receberia a classificação de “idade2”. Para ser classificada como “idade3” deveria possuir um valor maior que 25 e menor ou igual a 30. A idade também poderia ser classificada como “idade4” e para isso, deveria ter um valor maior que 30. O Quadro 8 demonstra com maior clareza as regras para a classificação da idade.

Quadro 8: Regras para a classificação da idade.

Tipo da categoria:	Regra:
Idade1	NU_IDADE<=18
Idade2	NU_IDADE>18 e NU_IDADE<=25
Idade3	NU_IDADE>25 e NU_IDADE<=30
Idade4	NU_IDADE>30

Fonte: Elaborado pelo autor.

3.3.2 Seleção dos dados de Santa Catarina e Araranguá

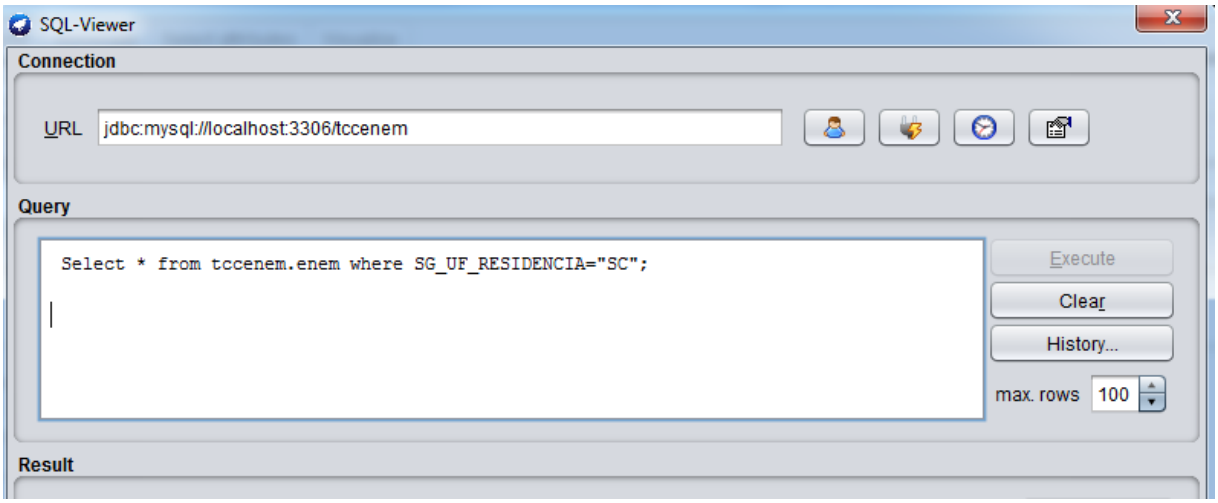
Nesta etapa deste trabalho de conclusão de curso, foi realizado a seleção ou subdivisão da base de dados. No software WEKA por meio da opção “Open DB” foi realizado a conexão com a base de dados, criada em etapas anteriores. Após concluída a conexão, a base foi de fato dividida em duas partes:

1. Base referente a todo o estado de Santa Catarina;

2. Base referente a cidade de Araranguá - SC.

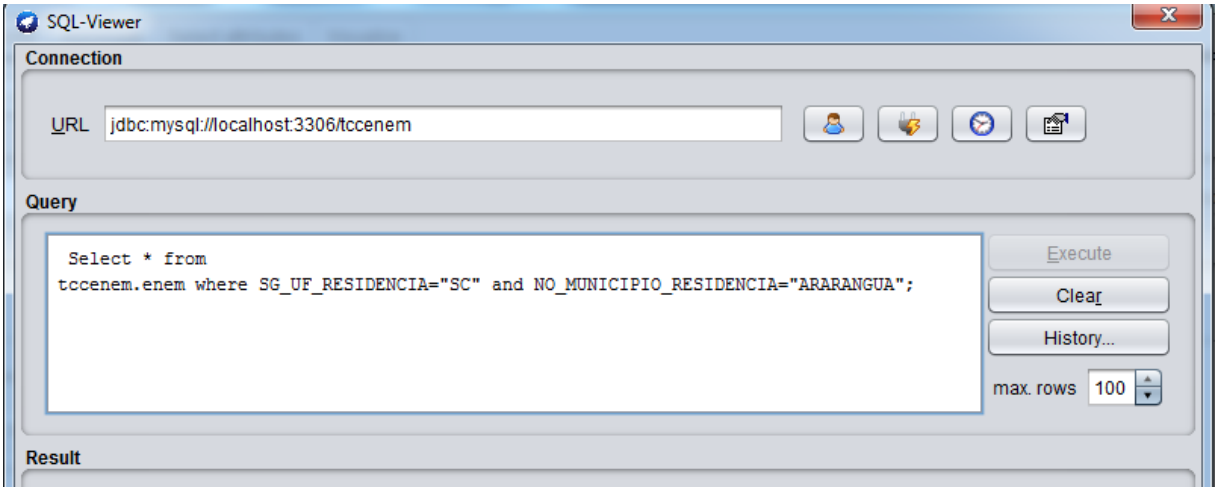
Dessa forma, dois arquivos ".ARFF' foram gerados por meio de linguagem SQL que permitiu essas duas seleções distintas, como demonstra a Figura 10 e a Figura 11.

Figura 10: Seleção dos dados do estado de Santa Catarina.



Fonte: Elaborado pelo autor.

Figura 11: Seleção dos dados da cidade de Araranguá.



Fonte: Elaborado pelo autor.

3.4 Geração e avaliação¸ dos modelos de predição

Esse processo corresponde a etapa de mineração de dados, que é o eixo desse trabalho, a tarefa de busca de informação aplicada nesse trabalho, foi feita por meio das técnicas de classificação.

Os experimentos realizados nesse trabalho de conclusão curso foram realizados, focados no objetivo inicial. Dessa forma, as variáveis escolhidas foram testadas para observar-se se os resultados seriam positivos ou negativos em relação a sua eficiência.

Os experimentos foram feitos utilizando o software WEKA e em seguida será esclarecido como os mesmos foram realizados e as particularidades em relação aos dados usados em cada uma delas. Foi escolhido o software WEKA pois o mesmo possui uma interface descomplicada e de fácil uso, além de contar com diversos algoritmos preditivos como o J48 e o Naive Bayes.

Ao longo da elaboração desse trabalho de conclusão de curso, inúmeros experimentos foram feitos precedentemente a delimitação total da metodologia usada no mesmo. Dessa forma duas experimentações ganharam destaque em relação as demais e por esse motivo serão rapidamente descritas nesta seção do trabalho, sendo elas:

1. Utilização dos dados abertos do ENEM para a predição do desempenho da redação do ENEM, no estado de Santa Catarina
2. Utilização dos dados abertos do ENEM para predição do desempenho da redação do ENEM, na cidade de Araranguá

Para a realização desses experimentos, foi feita a escolha dos algoritmos a ser utilizados, sendo eles: J48 e Naive Bayes. A próxima etapa feita, foi um estudo mais profundo, através do dicionário de dados, sobre cada atributo presente na base, para que dessa forma fosse possível escolher as variáveis que possivelmente dariam o maior desempenho.

O **primeiro experimento**, tratava-se de gerar um modelo preditivo da redação do ENEM por meio do conjunto de dados ligados apenas ao estado de Santa Catarina, como mostra a figura 10. Baseado nisso, através do software WEKA, diversas variáveis, foram escolhidas para compor o input dos algoritmos utilizados, onde a classe/variável a ser predita escolhida sempre foi a "CATEGORIZACAO_NU_NOTA_REDACAO_NULL". Conforme o conjunto de variáveis de input mudavam, ocorriam oscilações na acurácia do experimento. O conjunto de atributos que apresentou o melhor desempenho em relação ao número de acertos, possuía as seguintes variáveis:

- TP_SEXO
- TP_ESCOLA
- TP_ENSINO
- IN_TREINEIRO

- TP_DEPENDENCIA_ADM_ESC
- TP_LOCALIZACAO_ESC
- Q045
- Q047
- CATEGORIZACAO_IDADE
- CATEGORIZACAO_Q005
- CATEGORIZACAO_Q007
- CATEGORIZACAO_Q027
- CATEGORIZACAO_Q001
- CATEGORIZACAO_Q002
- CATEGORIZACAO_Q006
- CATEGORIZACAO_TP_ANO_CONCLUIU
- CATEGORIZACAO_Q040

As configurações utilizadas no algoritmo J48 e Naive Bayes, nesse experimento, foram a padrão dos mesmos. Em relação a separação dos dados para treino e teste, foram usados 70% dos dados para treino e 30% para teste.

O **segundo experimento** tinha como objetivo gerar um modelo preditivo da redação do ENEM a partir da seleção dos dados da cidade de Araranguá como demonstra a figura 11. Como feito no primeiro experimento, diversos conjuntos de atributos foram testados para ser usado como input dos algoritmos. As configurações utilizadas nos algoritmos J48 e Naive Bayes, usados nesse experimento, foram as padrões de cada um deles. Já em relação a porcentagem final utilizada para treinamento e teste, foi de 75% e 25% respectivamente. O conjunto de atributos final definido para input foi:

- TP_SEXO
- Q047
- CATEGORIZACAO_Q005
- CATEGORIZACAO_NU_NOTA_REDACAO_NULL

A variável/classe a ser predita foi a "CATEGORIZACAO_NU_NOTA_REDACAO_NULL". Durante as experiencias, conforme

mudava os atributos de input e os valores definidos para treinamento e teste, a acurácia sofria modificações tanto positivas como negativas.

4 RESULTADOS E DISCUSSÕES

Após a realização dos experimentos, iniciou-se a análise e discussão dos resultados obtidos, onde primeiramente buscou-se compreender e apresentar os resultados de cada experimento individualmente, através das relutâncias obtidas por meio dos algoritmos J48 e Naive Bayes, fazendo um comparativo entre os dois algoritmos. Em um segundo momento comparou-se os resultados do primeiro experimento com os do segundo experimento. Já em um terceiro momento a comparação deu-se por meio dos resultados das pesquisas dos trabalhos relacionados. Esse capítulo apresenta ainda árvores de decisão geradas em ambos os experimentos.

4.1 Resultados do primeiro experimento

No primeiro experimento, que como dito anteriormente nesse trabalho de conclusão de curso, buscou-se gerar um modelo preditivo da redação do ENEM por meio do conjunto de dados ligados apenas ao estado de Santa Catarina, conseguiu-se atingir uma classificação correta de 57.1523% das instâncias analisadas, através do algoritmo J48 como mostra a Figura 12. Já o algoritmo Naive Bayes, como demonstra a Figura 14, acertou 56.3872% dos dados testados.

O Quadro 9 mostra a acurácia detalhada do primeiro experimento, por meio do algoritmo J48. Observa-se então no Quadro 9 que o TP-Rate (True-Positive Rate) demonstra que a situação mais fácil de ser predita são as das redações classificadas como “nulo”, e a mais difícil de predizer são as classificadas como “alto”. Por meio do algoritmo Naive Bayes, o cenário mais fácil de prever, como mostra o Quadro 10, também foi em relação as instâncias classificadas como “nulo”. Já a situação mais complicada de predizer, é aquela em que a redação é classificada como “media”.

Por meio da Matriz de Confusão presente na Figura 13 (gerada pelo J48) e também na Figura 15 (gerada pelo Naive Bayes) é possível entender de forma mais clara, quais instâncias os modelos gerados mais acertam e quais eles mais erram.

Figura 12: Resumo dos resultados obtidos por meio do algoritmo J48 relacionados aos dados de Santa Catarina..

```
Correctly Classified Instances       30329               57.1523 %
Incorrectly Classified Instances     22738               42.8477 %
Kappa statistic                          0.4216
Mean absolute error                      0.2489
Root mean squared error                  0.3578
Relative absolute error                 66.8077 %
Root relative squared error             82.9023 %
Total Number of Instances            53067
```

Fonte: Elaborado pelo autor.

Quadro 9: Acurácia detalhada do primeiro experimento, utilizando o algoritmo J48 e apenas os dados de Santa Catarina.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,394	0,089	0,511	0,394	0,445	0,338	0,787	0,438	alto
	0,482	0,182	0,478	0,482	0,480	0,299	0,762	0,468	baixo
	0,412	0,220	0,401	0,412	0,406	0,190	0,707	0,389	media
	0,916	0,083	0,816	0,916	0,863	0,806	0,979	0,958	nulo
Weighted Avg.	0,572	0,146	0,562	0,572	0,564	0,424	0,815	0,583	

Fonte: Elaborado pelo autor.

Figura 13: Matriz de Confusão gerada por meido do algoritmo J48 relacionados aos dados de Santa Catarina.

```
    a     b     c     d   <-- classified as
 3999  1968  3549   638 |     a = alto
 1150  6560  4504  1408 |     b = baixo
 2493  4632  5756  1106 |     c = media
  185   562   543 14014 |     d = nulo
```

Fonte: Elaborado pelo autor.

Figura 14: Resumo dos resultados obtidos por meio do algoritmo Naive Baye e relacionados aos dados de Santa Catarina.

```
Correctly Classified Instances       29923               56.3872 %
Incorrectly Classified Instances     23144               43.6128 %
Kappa statistic                          0.4131
Mean absolute error                      0.2346
Root mean squared error                  0.3716
Relative absolute error                 62.9593 %
Root relative squared error             86.1214 %
Total Number of Instances            53067
```

Fonte: Elaborado pelo autor.

Quadro 10: Acurácia detalhada do primeiro experimento, utilizando o algoritmo Naive Bayes e apenas os dados de Santa Catarina.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,450	0,113	0,485	0,450	0,467	0,347	0,789	0,458	alto
	0,527	0,214	0,460	0,527	0,491	0,300	0,765	0,487	baixo
	0,306	0,162	0,404	0,306	0,348	0,159	0,705	0,395	media
	0,908	0,094	0,797	0,908	0,849	0,785	0,980	0,958	nulo
Weighted Avg.	0,564	0,146	0,547	0,564	0,552	0,412	0,816	0,593	

Fonte: Elaborado pelo autor.

Figura 15: Matriz de Confusão gerada por meio do algoritmo Naive Baye e relacionada aos dados de Santa Catarina.

```
    a     b     c     d   <-- classified as
 4569  2346  2564   675 |     a = alto
 1505  7178  3325  1614 |     b = baixo
 3061  5398  4281  1247 |     c = media
  285   688   436 13895 |     d = nulo
```

Fonte: Elaborado pelo autor.

4.2 Resultados do segundo experimento

No segundo experimento, em que foi gerado um modelo preditivo da redação do ENEM, através de instancias pertencentes apenas ao município de Araranguá - SC , obteve-se uma acurácia de 61.1227%, por meio do algoritmo J48 como mostra a Figura 16, e uma acurácia de 61.7464% através do Naive Bayes, como mostra a Figura 18.

Por meio do Quadro 11, baseado nos resultados obtidos com o J48, nota-se que o TP-Rate (True-Positive Rate) mostra que o cenário mais simples para predição das redações são daquelas classificadas como "baixo", enquanto o mais complexo fica por conta daquelas classificadas como "media". Já quando o algoritmo utilizado foi o Naive Bayes, a situação mais fácil de predizer também é aquela onde a redação possui o valor de "baixo", e a mais complexa é a que a redação e classificada como "media" (assim como no uso do J48). Essa situação está melhor detalhada no Quadro 12.

Através da Matriz de Confusão presente na Figura 17 (gerada pelo J48) e também na Figura 19 (gerada pelo Naive Bayes) é possível compreender de uma melhor forma, quais instâncias foram mais acertadas quando o modelo gerado confrontou as mesmas com o conjunto teste.

Figura 16: Resumo dos resultados obtidos por meio do algoritmo J48 e relacionados aos dados de Araranguá.

Correctly Classified Instances	294	61.1227 %
Incorrectly Classified Instances	187	38.8773 %
Kappa statistic	0.4719	
Mean absolute error	0.238	
Root mean squared error	0.3476	
Relative absolute error	64.0843 %	
Root relative squared error	80.7195 %	
Total Number of Instances	481	

Fonte: Elaborado pelo autor.

Quadro 11: Acurácia detalhada do primeiro experimento, utilizando o algoritmo J48 e apenas os dados de Araranguá.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,491	0,249	0,363	0,491	0,417	0,219	0,708	0,343	alto
	0,908	0,043	0,908	0,908	0,908	0,865	0,984	0,961	baixo
	0,417	0,050	0,636	0,417	0,504	0,437	0,782	0,432	media
	0,496	0,174	0,531	0,496	0,513	0,329	0,788	0,508	nulo
Weighted Avg.	0,611	0,128	0,631	0,611	0,615	0,493	0,831	0,601	

Fonte: Elaborado pelo autor.

Figura 17: Matriz de Confusão gerada por meio do algoritmo J48 e relacionada aos dados de Araranguá.

```
  a   b   c   d   <-- classified as
 53   3  18  34 |   a = media
  8 138   0   6 |   b = nulo
 26   3  35  20 |   c = alto
 59   8   2  68 |   d = baixo
```

Fonte: Elaborado pelo autor.

Figura 18: Resumo dos resultados obtidos por meio do algoritmo Naive Baye relacionado aos dados de Araranguá.

```
Correctly Classified Instances         297               61.7464 %
Incorrectly Classified Instances       184               38.2536 %
Kappa statistic                          0.4817
Mean absolute error                      0.2357
Root mean squared error                  0.3408
Relative absolute error                 63.4622 %
Root relative squared error             79.1397 %
Total Number of Instances              481
```

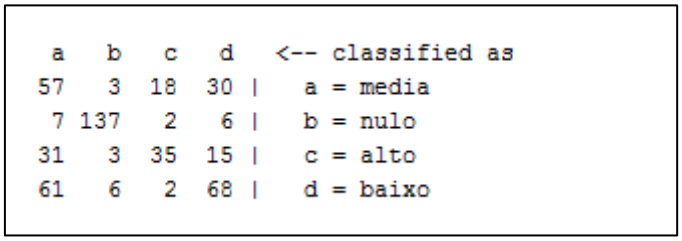
Fonte: Elaborado pelo autor.

Quadro 12: Acurácia detalhada do primeiro experimento, utilizando o algoritmo Naive Bayes e apenas os dados de Araranguá.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,528	0,265	0,365	0,528	0,432	0,234	0,714	0,361	alto
	0,901	0,036	0,919	0,901	0,910	0,870	0,987	0,973	baixo
	0,417	0,055	0,614	0,417	0,496	0,424	0,818	0,476	media
	0,496	0,148	0,571	0,496	0,531	0,364	0,817	0,562	nulo
Weighted Avg.	0,617	0,123	0,643	0,617	0,623	0,505	0,848	0,632	

Fonte: Elaborado pelo autor.

Figura 19: Matriz de Confusão gerada por meio do algoritmo Naive Baye e relacionado aos dados de Araranguá.



Fonte: Elaborado pelo autor.

4.3 Comparativo dos resultados do primeiro experimento com os do segundo experimento

O melhor resultado obtido no primeiro experimento foi oriundo do uso do algoritmo J48, que teve uma acurácia de 57.1523%. Já no segundo experimento o resultado mais satisfatório foi proveniente da utilização do algoritmo Naive Bayes, que conseguiu obter uma acurácia de 61.7464%.

O primeiro experimento usou 17 variáveis para input, enquanto o segundo utilizou apenas 4 atributos para essa mesma função. Com base nesse fato, e nos resultados obtidos em cada um dos experimentos, conclui-se que um grande número de variáveis de entrada não garantem uma boa acurácia. Provavelmente a qualidade dos dados utilizados para o input e o

relacionamento entre os mesmo tem um maior impacto positivo no resultado do processo de mineração de dados, do que o volume de variáveis utilizadas no input.

Apesar de o segundo experimento ter obtido uma acurácia maior que o primeiro, a diferença não é tão grande, ou seja, os resultados obtidos foram similares nos dois experimentos.

4.4 Comparação com os Trabalhos Relacionados

Como dito anteriormente, Simon e Cazella (2017) realizaram um estudo que tinha como principal objetivo: gerar um modelo preditivo do indicador de desempenho médio na área de ciências da natureza e suas tecnologias dos alunos de escolas do ensino médio através dos dados abertos do ENEM 2015. No trabalho de Simon e Cazella (2017), os mesmos utilizaram o algoritmo j48 para predizer a variável "Média Escola", o modelo dos mesmos conseguiu acertar em 77,02% das instâncias testadas.

Comparado aos modelos gerados por esse trabalho de conclusão de curso, o modelo gerado por Simon e Cazella (2017) é significativamente superior, pois a pesquisa dos mesmos tem 77,02% de acertos, enquanto o melhor modelo desse trabalho de conclusão de curso apresenta 61.7464% de acertos, realizados no segundo experimento usando o algoritmo Naive Bayes.O melhor resultado usando o algoritmo J48 também foi no segundo experimento, alcançando a acurácia de 61.1227%. Destaca-se que apesar do trabalho de Simon e Cazella (2017) utilizar dados do ENEM para predição, o objetivo do que se deseja predizer é muito diferente do desse trabalho, com base nisso a superioridade do modelo de Simon e Cazella (2017), comparadas com essa pesquisa, é justificada.

Uma das poucas pesquisa que encontrou-se com o objetivo de trabalhar com a predição ligada a redação do ENEM, assim como nesse trabalho de conclusão de curso, é a pesquisa feita por Gomes (2015). Na pesquisa realizada por Gomes (2015), o mesmo realizou uma análise de redações com notas zero, porem o atributo classificador utilizado foi "STATUS_REDACAO", desviando-se então do objetivo desse trabalho de conclusão de curso. Como Gomes (2015) não informou os valores de acurácia e nem da Matriz de Confusão dessa predição, tornou-se inviável a comparação de resultados.

4.5 Árvores de Decisão

Esse tópico apresenta as arvores geradas por meio dos dois experimentos, já relatados anteriormente. O motivo pelo qual essas arvores foram geradas, é permitir que gestores e especialista em educação, tenham mais uma ferramenta que os ajude a identificar atributos que causam um maior impacto no desempenho dos alunos na prova de redação. Desta forma os especialistas poderiam dar mais atenção a essas variáveis de grande impacto, podendo então melhorar o desempenho dos alunos na redação do ENEM caso esse atributo possa ser controlado e modificado pelos gestores da educação.

4.5.1 Árvore de Decisão gerada no primeiro experimento

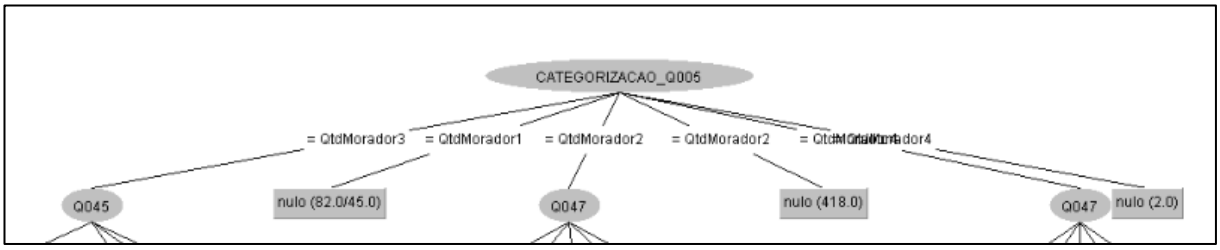
Quando aplicou-se o algoritmo J48 no primeiro experimento, descrito anteriormente, uma arvore de decisão foi gerada, a mesma se encontra no Apêndice B. Através de uma breve análise do autor deste trabalho, notou-se que a mesma possui como variável principal a variável "CATEGORIZACAO_Q005". Esse atributo aparece mais 5 vezes ao longo da árvore de decisão, o que leva a crer que nessa predição a variável "CATEGORIZACAO_Q005" é a mais importante. Outras variáveis consideradas importante são: Q045, Q047 e CATEGORIZACAO_Q040.

4.5.2 Árvore de Decisão gerada no segundo experimento

No momento em que se fez a predição do segundo experimento através do algoritmo J48, a árvore de decisão apresentada no Apêndice B foi gerada. Assim como feito com a arvore gerada pelo primeiro experimento, o autor desse TCC, realizou uma pequena análise da arvore. Assim como na arvore do primeiro experimento, nessa arvore o atributo principal também é a "CATEGORIZACAO_Q005", outras variáveis consideradas bem importantes para essa árvore são as seguintes: Q045, Q047 e TP_SEXO.

A Figura 20 demonstra uma visualização parcial da árvore de decisão obtida no processo de mineração de dados em relação aos dados da cidade de Araranguá.

Figura 20: Visualização parcial da árvore de decisão da cidade de Araranguá.



Fonte: Elaborado pelo autor.

5 CONCLUSÃO

Podemos notar que apesar do crescimento na realização de pesquisas na área de mineração de dados educacionais, ainda é uma área que necessita ser mais explorada e estudada. Uma prova disso é que não se encontrou trabalhos com uma problemática bem próxima da desse trabalho, apesar de como relatado anteriormente, as redações do ENEM serem decisivas para o desempenho final do candidato e serem uma das maiores preocupações de alunos e professores.

Através dos modelos de predição gerados nesse trabalho conclui-se que é possível prever com um mínimo de exatidão o desempenho dos alunos na redação do ENEM. Porém destaca-se que é necessário trabalhos futuros voltados ao mesmo assunto e objetivo desse trabalho de conclusão de curso, para que se possa evoluir significativamente os modelos de predição.

As variáveis que mais causaram impacto foram praticamente as mesmas para os dois experimentos realizados, o que leva a crer que os gestores e especialistas da área da educação devem analisar as mesmas. Ainda em relação as variáveis destaca-se que as mesmas precisam ser coletadas de uma forma mais eficiente, tendo em vista que muitas vezes os candidatos deixaram em branco as perguntas do questionário socioeconômico.

Para trabalhos futuros, julga-se importante a integração da base de dados utilizadas nesse trabalho com outras bases, para que as mesmas possam agregar em relação a novos atributos, onde os mesmos podem ajudar a aumentar o desempenho dos modelos preditivos.

O objetivo principal desta pesquisa foi realizar um estudo e aplicar as técnicas e métodos de mineração de dados, para gerar modelos para predição do desempenho da redação do ENEM, por meio dos microdados do ENEM 2016 e algoritmos de classificação como o J48 e o Naive Bayes, e também gerar árvores de decisão. Baseado nisso, esse trabalho buscou auxiliar na melhora do desempenho dos alunos na prova de redação, por meio de modelos para predição do desempenho da redação do ENEM.

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: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2014. p. 891.

ANDRIOLA, Wagner Bandeira. Doze motivos favoráveis à adoção do Exame Nacional do Ensino Médio (ENEM) pelas Instituições Federais de Ensino Superior (IFES). **Ensaio: avaliação e políticas públicas em educação**, v. 19, n. 70, 2011.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineraçao de dados educacionais: Oportunidades para o brasil. **Brazilian Journal of Computers in Education**, v. 19, n. 02, p. 03, 2011.

BAKER, Ryan SJD; YACEF, Kalina. The state of educational data mining in 2009: A review and future visions. **JEDM| Journal of Educational Data Mining**, v. 1, n. 1, p. 3-17, 2009.

BERNERS-LEE, Tim. **Linked Data.** 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 27 maio 2018.

BERNERS-LEE, Tim. **Putting government data online**. 2009. Disponível em: < https://www.w3.org/DesignIssues/GovData.html>. Acesso em: 29 maio 2018.

BERRY, Michael J. A.; LINOFF, Gordon S. Data mining techniques. USA: Wiley Publishing Inc, 2004. 2ª edição.

BRASIL. MEC. . **Qual o critério de desempate?** 2018. Disponível em: <Ministério da Educação. Qual o critério de desempate? 2018. Disponível em: . Acesso em: 12 jun. 2018.>. Acesso em: 12 jun. 2018.

CHIMIESKI, Bruno Fernandes; FAGUNDES, Rubem Dutra Ribeiro. Association and classification data mining algorithms comparison over medical datasets. **Journal of health informatics**, v. 5, n. 2, 2013.

DETONI, Douglas; CECHINEL, Cristian; MATSUMURA ARAÚJO, Ricardo. Modelagem e Predição de Reprovação de Acadêmicos de Cursos de Educação a Distância a partir da Contagem de Interações. **Revista Brasileira de Informática na Educação**, v. 23, n. 3, 2015.

DINIZ, Vagner. **Como conseguir dados governamentais abertos**. 2010.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **Aaai Press**, p.37-54, 1996.

GARCIA, Patricio et al. Automatic detection of team roles in computer supported collaborative work. **IEEE Latin America Transactions**, v. 11, n. 4, p. 1066-1074, 2013.

GLOBO. Cai número de alunos com nota mil na redação do Enem e sobe total de zero. 2016. Disponível em: <https://g1.globo.com/educacao/noticia/cai-numero-de-alunos-com-nota-mil-na-redacao-do-enem-e-sobe-total-de-zero.ghtml>. Acesso em: 25 jun. 2018.

GOMES, Tancicleide Carina Simões. **Descoberta de Conhecimento Utilizando Mineração de Dados Educacionais Abertos.** 2015. 67 f. TCC (Graduação) - Curso de Sistemas de

Informação, Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco, Recife, 2015.

HALL, Mark et al. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, v. 11, n. 1, p. 10-18, 2009.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining:** Concepts and Techniques. 3. ed. Amsterdam: Elsevier, 2011. 744 p.

INEP. **Dúvidas Frequentes.** 2018. Disponível em: <https://enem.inep.gov.br/#/faq?_k=glm11t>. Acesso em: 30 maio 2018.

INEP. **Conheça o Enem**. Disponível em: <https://enem.inep.gov.br/#/antes?_k=l13hcw>. Acesso em: 12 maio 2018.

INEP. **Inep divulga os microdados do Enem 2016.** 2017. Disponível em: <http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/inep-divulga-os-microdados-do-enem-2016/21206>. Acesso em: 30 maio 2018.

KAMPFF, Adriana Justin Cerveira. **Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente**. 2009.

LIBRELOTTO, Solange Rubert; MOZZAQUATRO, Patricia Mariotto. Análise dos algoritmos de mineração J48 e Apriori aplicados na detecção de indicadores da qualidade de vida e saúde. **Revista Interdisciplinar de Ensino, Pesquisa e Extensão**, v. 1, n. 1, 2014.

LUAN, Jing. **Data Mining Applications in Higher Education**. 2007.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

OLIVEIRA, Thiago L. de; JESUS, Thaynara A. M. de; BRAZ, Fernando José. DESENVOLVIMENTO DE AMBIENTE PARA A GESTÃO DO CONHECIMENTO RELACIONADO AOS DADOS PRODUZIDOS PELO SISTEMA DE GERENCIAMENTO DE TRANSITO DA CIDADE DE JOINVILLE/SC - PARTE I. **Mostra Científica e Tecnológica (mct)**, Araquari, p.1-4, 2015.

OLIVEIRA, Vinícius. **Pentaho - Visão Geral.** 2018. Disponível em: <https://www.binapratica.com.br/visao-pentaho>. Acesso em: 09 jun. 2018.

OPEN KNOWLEDGE BRASIL. **Índice de Dados Abertos da Open Knowledge indica pouco progresso por parte dos governos em abrir dados chave.** 2014. Disponível em: <https://br.okfn.org/2014/12/09/indice-de-dados-abertos-da-open-knowledge-indica-pouco-progresso-por-parte-dos-governos-em-abrir-dados-chave/>. Acesso em: 27 maio 2018.

PICHILIANI, MAURO. Data mining na prática: Árvores de Decisão. **Disponıvel em: http://imasters. com. br/artigo/5130/sql-server/data-mining-na-pratica-arvores-de-decisao**, 2008.

QUEIROGA, Emanuel Marques. **Geração de modelos de predição para estudantes em risco de evasão em cursos técnicos a distância utilizando técnicas de mineração de dados**. 2017. Dissertação de Mestrado. Universidade Federal de Pelotas.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, [S.l.], v.1, n.1, p.81–

106, 1986.

RIBEIRO, Claudio Jose Silva; ALMEIDA, Reinaldo Figueiredo de. Dados Abertos Governamentais (Open Government Data): instrumento para exercício de cidadania pela sociedade. **XII Enancib-Políticas de Informação para a Sociedade-Anais. Brasília: Thesaurus**, p. 2568-2580, 2011.

RODRIGUES, Rodrigo Lins et al. A literatura brasileira sobre mineração de dados educacionais. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2014. p. 621.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational Data Mining: A Review of the State of the Art. **Ieee Transactions On Systems, Man, And Cybernetics, Part C (applications And Reviews)**, [s.l.], v. 40, n. 6, p.601-618, nov. 2010. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tsmcc.2010.2053532.

SIMON, Augusto; CAZELLA, Sílvio. Mineração de Dados Educacionais nos Resultados do ENEM de 2015. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2017. p. 754.

TRAVITZKI, Rodrigo. **ENEM: limites e possibilidades do Exame Nacional do Ensino Médio enquanto indicador de qualidade escolar**. 2013. Tese de Doutorado. Universidade de São Paulo.

UNIVERSIDADE METODISTA DE SÃO PAULO (São Paulo). **Professores apontam dificuldades dos alunos na hora da redação.** 2011. Disponível em: <http://www.metodista.br/rronline/noticias/cidades/2011/10-1/temporario/redacao-e-a-maior-dificuldade-dos-vestibulandos>. Acesso em: 12 jun. 2018.

WEKA. **Weka 3:** Data Mining Software in Java. 2018. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 09 jun. 2018.

WITTEN, I. H.; FRANK, E. **Data mining: practical machine learning tools and techniques**. 2 ed. San Francisco: Morgan Kaufmann Publishers, 2005.

ZAKI, Mohammed J.; MEIRA JUNIOR, Wagner. **Data Mining and Analysis:** Fundamental Concepts and Algorithms. New York: Cambridge University Press, 2014.

APÊNDICE A – Informações das variáveis categorizadas

As informações e regras de como as variáveis foram categorizadas, estão disponíveis no link: https://onedrive.live.com/?id=5632EEA93CF4F3A0%211434&cid=5632EEA93CF4F3A0

Optou-se por disponibilizar os mesmo dessa maneira devido a grande quantidade de informações.

APÊNDICE B – Árvores de Decisão

As árvores de decisão, estão disponíveis no link: https://onedrive.live.com/?id=5632EEA93CF4F3A0%211434&cid=5632EEA93CF4F3A0

Optou-se por disponibilizar as mesmas dessa maneira devido ao grande volume de informações

ANEXO A – Recorte do Dicionário de Dados do ENEM 2016

DICIONÁRIO DE VARIÁVEIS - ENEM 2016					
NOME DA VARIÁVEL	Descrição	Variáveis Categóricas		Tamanho	Tipo
		Categoria	Descrição		
Q005	Incluindo você, quantas pessoas moram atualmente em sua residência?	1	1, pois moro sozinho(a).	2	Numérica
		2	2		
		3	3		
		4	4		
		5	5		
		6	6		
		7	7		
		8	8		
		9	9		
		10	10		
		11	11		
		12	12		
		13	13		
		14	14		
		15	15		
		16	16		
		17	17		
		18	18		
		19	19		
		20	20		
Q006	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)	A	Nenhuma renda.	1	Alfanumérica
		B	Até R$ 880,00.		
		C	De R$ 880,01 até R$ 1.320,00.		
		D	De R$ 1.320,01 até R$ 1.760,00.		
		E	De R$ 1.760,01 até R$ 2.200,00.		
		F	De R$ 2.200,01 até R$ 2.640,00.		
		G	De R$ 2.640,01 até R$ 3.520,00.		
		H	De R$ 3.520,01 até R$ 4.400,00.		
		I	De R$ 4.400,01 até R$ 5.280,00.		
		J	De R$ 5.280,01 até R$ 6.160,00.		
		K	De R$ 6.160,01 até R$ 7.040,00.		
		L	De R$ 7.040,01 até R$ 7.920,00.		
		M	De R$ 7.920,01 até R$ 8.800,00.		
		N	De R$ 8.800,01 até R$ 10.560,00.		
		O	De R$ 10.560,01 até R$ 13.200,00.		
		P	De R$ 13.200,01 até R$ 17.600,00.		
		Q	Mais de R$ 17.600,00.		