Predição do desempenho de Matemática e Suas Tecnologias do ENEM utilizando técnicas de Mineração De Dados

Rafael Damiani Alves[1], Cristian Cechinel[2], Emanuel Marques Queiroga[3]

[1]Universidade Federal de Santa Catarina

[2]Universidade Federal de Santa Catarina

[3]Universidade Federal de Pelotas

{rafapk@hotmail.com, contato@cristiancechinel.pro.br, emanuelmqueiroga@gmail.com

Abstract. *The objective of this research is to find patterns and generate a predictive model of the performance indicator of the marks of the Mathematics test and its Technologies of the secondary schools, through the open data referring to the National High School Examination (ENEM) of 2015. The objective in question is based on data from the Program for International Student Assessment (PISA) of the year 2015, which demonstrate a worrying scenario of low performance of Brazilian elementary school students. The various techniques of data mining have been used to discover patterns that allow improvement in many areas. In this context, experiments were performed through educational data mining (EDM), where the data were categorized, to obtain a better result in the application of the algorithms. The predicted class was the school average that was categorized as: low, medium, high. The final models were trained and tested using the Naive Bayes and J48 algorithms. These algorithms were used through the WEKA software package.*

Resumo. *O objetivo desta pesquisa propõe-se a encontrar padrões e gerar um modelo preditivo do indicador de desempenho das notas da prova de Matemática e suas Tecnologias das escolas do ensino médio, por meio dos dados abertos referentes ao Exame Nacional do Ensino Médio (ENEM) de 2015. O objetivo em questão fundamenta-se nos dados do Programme for International Student Assessment (PISA) do ano de 2015, os quais demonstram um cenário preocupando de baixo desempenho dos alunos brasileiros de ensino básico. As diversas técnicas de mineração de dados vêm sendo utilizadas para realizar descoberta de padrões que permitem a melhoria em muitas áreas. Neste contexto, foram realizados experimentos por meio da mineração de dados educacionais (MDE), onde os dados foram categorizados, para obter um melhor resultado na aplicação dos algoritmos. A classe predita foi a média da escola que foi categorizada como: Baixa, Media, Alta. Os modelos finais foram treinados e testados por meio dos algoritmos: Naive Bayes e J48. Esses algoritmos foram utilizados por meio do pacote de software WEKA.*

DOI: 10.5753/cbie.wcbie.2018.469

1. Introdução

Conforme os dados de 2015 do Programme for International Student Assessment (PISA), criado e dirigido pela Organização para Cooperação e Desenvolvimento Econômico (OCDE), o Brasil se encontra em um cenário alarmante de baixo desempenho dos alunos do ensino básico, ou seja, o sistema educacional brasileiro nos últimos anos não possui motivos para comemorações e tão pouco razões para conforto.

Segundo OECD (2016), o desempenho dos alunos no Brasil está abaixo da média dos alunos em países da OCDE. Ainda conforme OECD (2016), em ciências os estudantes brasileiros atingiram 401 pontos enquanto à média da OCDE é de 493 pontos, em leitura (407 pontos, em relação à média de 493 points) e em matemática (377 pontos, comparados à média de 490 pontos).

Destaca-se ainda que os alunos do Brasil na área de matemática, apresentaram uma queda de 11 pontos se comparados a média de 2012 à média de 2015. A Figura 1 e a Figura 2 demonstra com maior clareza a situação da média em matemática do Brasil, comparada com a média de alguns países membros da OCDE.

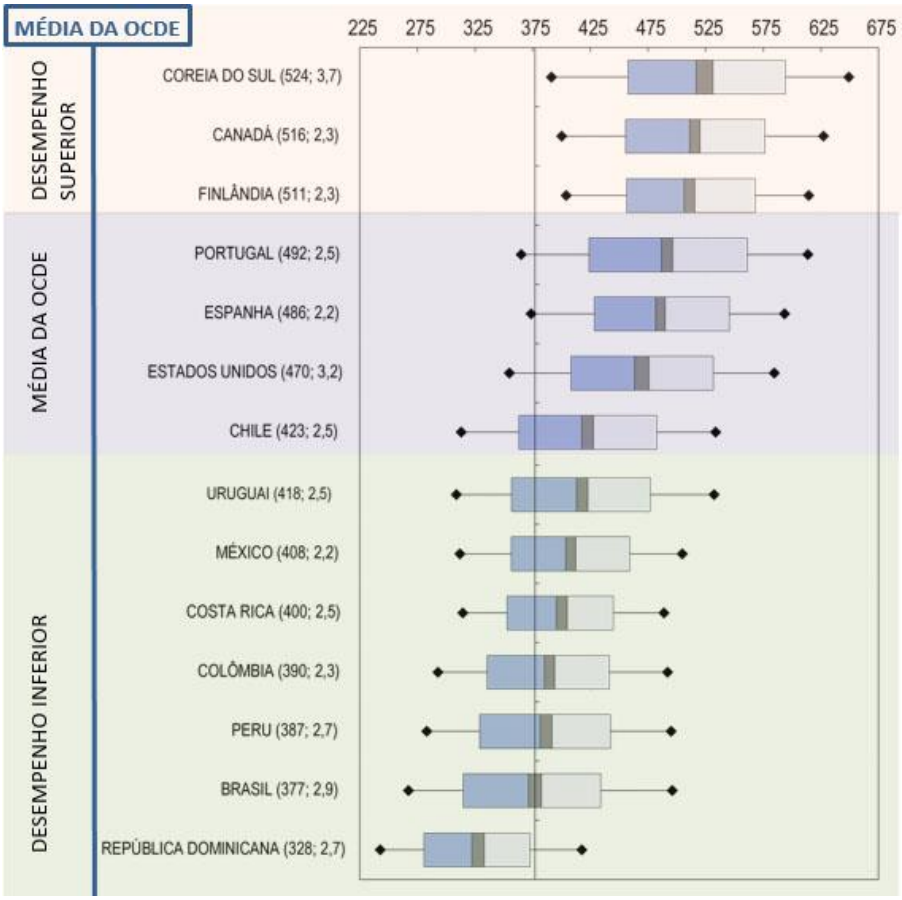

Figura 1. Desempenho dos brasileiros em matemática.

Fonte: OCDE, Inep.

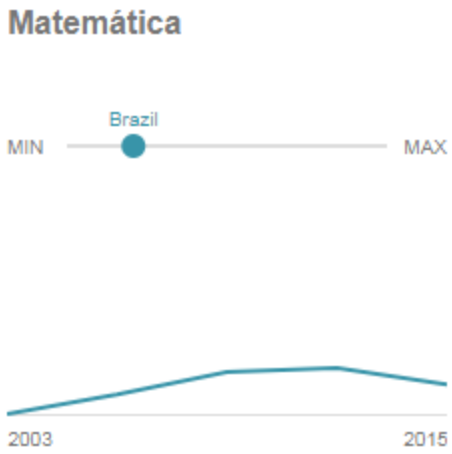

Figura 2. Desempenho médio do Brasil em relação à média da OCDE.

Fonte: OCDE.

O avanço das tecnologias da informação e comunicação tem proporcionado o armazenamento de bases de dados cada vez maiores. Essas bases de dados podem ser dos mais variados tipos inclusive bases de dados educacionais. Conforme Fayyad et al. (1996), as diversas áreas existentes vêm gerando um grande aumento nos seus bancos de dados.

Segundo Fayyad et al. (1996), devido ao grande aumento de dados nas bases de dados, se torna difícil analisar os mesmos, necessitando do uso de novas ferramentas e técnicas para análise automática e inteligente de bancos de dados. Com base nesse Cenário, conforme Luan (2007), a mineração de dados pode ser usada para descoberta de tendências e padrões ocultos, gerando resultados os quais os analistas e estudiosos da área podem observar os mesmos e tomar decisões mais explícitas e inteligentes, podendo focar a atenção nos pontos ou casos mais críticos e específicos.

De acordo com Romero e Ventura (2010), uma das subáreas da mineração de dados, é a mineração de dados educacionais (MDE) a qual se trata de um segmento de pesquisa interdisciplinar que possui métodos e técnicas para estudar dados originados a partir do cenário educacional.

Esta pesquisa possui o objetivo de encontrar padrões e gerar um modelo preditivo do indicador de desempenho das notas da prova de Matemática e suas Tecnologias das escolas do ensino médio, por meio das técnicas de mineração de dados, a partir dos dados públicos referentes ao Exame Nacional do Ensino Médio (ENEM) de 2015 (esses dados são disponibilizados pelo INEP e estão agrupados por escolas).

2. Mineração de dados

Segundo Zaki e Meira Junior (2014) a mineração de dados (Data Mining) é o processo de descoberta de padrões importantes e novos (desconhecidos), assim como modelos descritivos, compreensíveis e preditivos por meio de grandes bases de dados.

Segundo Han, Kamber e Pei (2011) muitos enxergam a mineração de dados como uma etapa fundamental no processo de descoberta do conhecimento (do inglês Knowledge Discovery in Databases - KDD). Para Gomes (2015) o KDD contém diversas fases que possivelmente são o caminho que os dados percorrem até virarem conhecimento. A Figura 3 demonstra com maior clareza quais são essas etapas.

Figura 3. O ciclo do processo de KDD.

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

De acordo com Fayyad et al. (1996), o primeiro passo para a descoberta de conhecimento é a compreensão do problema e estabelecimento do objetivo que almejasse alcançar. Já no segundo instante a etapa a ser realizada é a seleção. A seleção é o processo onde deve-se selecionar o conjunto de dados a ser utilizado nas etapas seguintes de KDD.

Segundo Fayyad et al. (1996), a terceira etapa chama-se pré-processamento de dados, etapa está que é responsável pela remoção de ruídos, formatação dos dados, ou seja, uma limpeza de modo geral. A quarta etapa denomina-se transformação, a mesma possui o propósito de redução da dimensionalidade dos dados ou até mesmo complementar os dados.

Para Fayyad et al. (1996), A quinta etapa é a mineração de dados, está etapa busca encontrar padrões por meio de algoritmos específicos de aprendizagem, como por exemplo algoritmos para efetuar: associação, classificação, clusterização e etc. A sexta e última fase trata-se da documentação do conhecimento e interpretação dos padrões pelos especialistas.

2.1. Mineração de dados Educacionais

De acordo com Baker, Carvalho e Isotani (2011), a mineração de dados educacionais é estabelecida como a área de pesquisa que tem como meta o desenvolvimento de métodos para estudar e entender conjuntos de dados coletados em ambientes educacionais.

Para Baker e Yacef (2009), uma das mais importantes áreas de atividade da MDE tem sido a aperfeiçoamento dos modelos de estudantes, esses modelos demonstram diversas informações sobre as características de um aluno ou estado, como o conhecimento atual do aluno, motivação, metacognição e atitudes.

Para que as metas da EDM possam ser concretizadas é preciso utilizar métodos e técnicas, a seguir, encontra-se as principais técnicas e métodos utilizados na MDE que conforme Baker, Carvalho e Isotani (2011) são: predição, agrupamento, mineração de relações, destilação de dados para facilitar decisões humanas, descobertas com modelos.

3. Trabalhos Relacionados

Esta seção do trabalho apresentara projetos relacionados ao ENEM e a MDE, os quais possuem propósitos e objetivos parecidos com o dessa pesquisa.

De todos os trabalhos encontrados o que apresentou mais similaridade com esta pesquisa é o trabalho de Simon e Cazella (2017). O mesmo buscou gerar um modelo preditivo do indicador de desempenho médio em ciências da natureza e suas tecnologias dos alunos de escolas do ensino médio, por meio dos dados ao Exame Nacional do Ensino Médio (ENEM) de 2015 os quais apresentam valores médios de aproveitamento dos estudantes agrupados por escolas, a base de dados utilizada nesse estudo contém 15599 instancias.

Ainda sobre o trabalho de Simon e Cazella (2017), o software usado para o processo de MDE foi o Waikato Environment for Knowledge Analysis (WEKA versão 3.8.1), a técnica de mineração de dados escolhida foi a árvore de decisão e o algoritmo utilizado foi o j48. Além disso, Simon e Cazella (2017) realizaram a categorização de algumas variáveis inclusive da variável a ser predita "Média Escola", por meio dessas técnicas e métodos a pesquisa de Simon e Cazella (2017) alcançou 77,02% de acurácia.

Outro trabalho que possui semelhança com esta pesquisa é o trabalho de Alves (2018), tendo em vista que o mesmo buscou gerar modelos para predição do desempenho da redação do ENEM, por meio dos microdados do ENEM 2016 e algoritmos de classificação.

A base de dados utilizada por Alves (2018) em sua pesquisa, continha 8627367 de instâncias. Neste cenário, foram realizados testes onde os dados foram categorizados, para alcançar um melhor resultado no uso 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. O uso desses algoritmos foi realizado por meio do software WEKA. O modelo com a maior eficácia conseguiu prever 61.7464% das amostras presentes na base de dados do ENEM 2016.

Este trabalho, diferente das pesquisas citadas no texto acima, foca apenas nos dados referentes a Matemática e suas Tecnologias buscando encontrar padrões e variáveis influenciadoras. Pois este setor da educação demonstrar estar muito fragilizado no Brasil como demonstra os dados da OCDE ja citadas nesse artigo.

4. Metodologia

Esta seção relata o cenário em que os dados foram utilizados, assim como os processos usados na metodologia, do desenvolvimento desse trabalho, que foram: Seleção e Pré-processamento dos dados, Transformação dos dados, Seleção dos dados no WEKA, Geração e avaliação dos modelos de predição. Além disso essa seção expõe a delimitação dos experimentos realizados. A Figura 4 apresenta de forma mais objetiva os processos e subprocessos usados na metodologia.

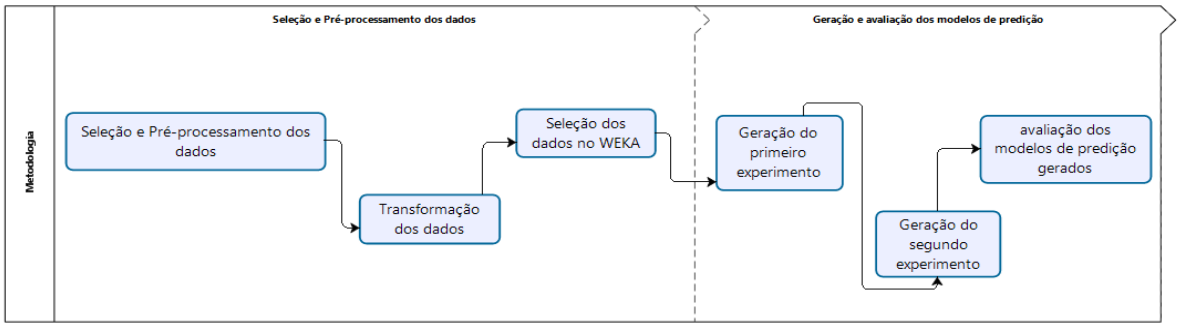

Figura 4. Processos utilizados na metodologia.

Fonte: Elaborado pelo autor.

4.1 Contexto

Para a realização deste trabalho foram usados e analisados os dados abertos do Exame Nacional do Ensino Médio 2015 (ENEM) os quais apresentam valores médios de aproveitamento dos estudantes agrupados por escolas. Esses dados foram obtidos na página web do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). De acordo com o INEP (2015), os dados do ENEM 2015 por Escola apresentam as médias e os percentuais de alunos em cada um dos quatro níveis de proficiência e da redação dos estudantes que participaram do Enem, por escola, para cada uma das áreas do conhecimento consideradas.

Ainda segundo INEP (2015), esses dados foram divulgados para as escolas que cumpriram, os critérios a seguir: possuir pelo menos 10 (dez) alunos concluintes do ensino médio regular seriado participantes do Enem 2015 e possuir pelo menos 50% de alunos participantes do Enem 2015, de acordo com os dados do Censo Escolar 2015. Os dados utilizados nessa pesquisa, são disponibilizados no formato ".xlsx", essa base contém um total de 15598 instâncias.

4.2 Seleção e Pré-processamento dos dados

Para a produção desse trabalho foram utilizados e analisados os dados abertos do ENEM 2015 por escola, logo após o download desta base, observou-se que devido ao formato do arquivo, o WEKA não tinha capacidade de utilizar os mesmos.

Devido a esse cenário adotou-se a estratégia de criar uma base de dados, por meio de um banco de dados e de uma Linguagem de definição de dados (DDL), para armazenar todas as instâncias (existentes no arquivo original) que pertenciam a prova de Matemática e suas Tecnologias.

Assim sendo, necessitou-se também utilizar uma ferramenta de Extração Transformação Carregamento (ETL) para preencher o banco de dados. A ferramenta escolhida para essa tarefa chama-se Pentaho Data Integration ®, em sua versão de avaliação.

Outro processo feito por meio do Pentaho ® foi a remoção das acentuações das Strings, contidas no Arquivo ".xlsx" que iriam compor a base de dados criada, etapa esse que se caracteriza como limpeza de dados.

4.2.1 Transformação dos dados

Uma das mais fundamentais etapas feitas na fase de transformações dos dados foi a categorização, pois por meio desta é possível categorizar os atributos fazendo com que os valores dos mesmos, fiquem segmentados em categorias, condensando então a amplitude desses valores. Dessa forma os algoritmos de mineração de dados podem alcançar resultados mais efetivos.

A primeira variável a passar pela etapa de categorização foi o atributo a ser predito "MEDIA_ESCOLA", atributo esse que armazena as notas recebidas pelos participantes do ENEM na prova de Matemática e suas Tecnologias.

Os valores dessa categorização foram baseados em um script em Java que tem objetivo de descobrir em quais pontos deveriam serem feitos os cortes para uma divisão por tercil. O tercil trata-se de uma estratégia de "balanceamento" para categorização onde divide-se um conjunto em 3 partes. Dessa forma a classe predita ficou "balanceada" ou seja com as quantidades de ocorrências praticamente iguais para cada caso.

Após serem conhecidos os valores onde seriam realizadas as divisões para cada categoria da classificação, foi criado uma nova variável no banco de dados, por meio de linguagem DDL, a qual foi denominada de "CATEGORIZACAO_MEDIA_ESCOLA".

Essa nova variável herdou os valores de "MEDIA_ESCOLA" categorizados, por meio de um script SQL executado no banco de dados. As categorias criadas foram: baixo, media, alto. Para que as instâncias fossem classificadas como "baixo', teriam que possuir uma nota de até no máximo 451.Para serem classificadas como "media" a nota deveria ser maior que 451 e menor ou igual a 502. Já para ser classificada como "alto" a nota deveria ser maior que 502.

Diversos atributos passaram pelo processo de categorizadas, cada um deles obedecendo critérios definidos especialmente para aquela variável com o propósito de que pudessem alcançar um desempenho maior na execução dos algoritmos.

4.2.2 Seleção dos dados no WEKA

Nesta etapa foi feito a seleção da base de dados no WEKA. No WEKA na opção "Open DB" foi realizado a conexão com a base de dados (criada nos processos anteriores). Após efetuada a conexão, foi criado por meio de linguagem SQL, um arquivo ".ARFF", que foi salvo e sendo assim então a base de dados utilizada nas etapas seguintes.

4.3 Geração e avaliação¸ dos modelos de predição

Essa etapa nada mais é do que o processo de mineração de dados, o qual foi efetuado usando a ferramenta WEKA. Nessa etapa dois experimentos foram feitos. Seguindo uma tendência verificada nos estudos levantados, optou-se por utilizar o algoritmo de classificação J48 e devido aos estudos comparativos terem demonstrados resultados com técnicas de redes Bayesianas o algoritmo Naive Bayes também foi utilizado. Nos dois experimentos a classe/variável a ser predita foi a "CATEGORIZACAO_MEDIA_ESCOLA".

O primeiro experimento buscou gerar um modelo preditivo do indicador de desempenho das notas da prova de Matemática e suas Tecnologias do ENEM, por meio do algoritmo J48. Já o segundo experimento possuía o mesmo objetivo, porém através do uso do algoritmo Naive Bayes.

As configurações usadas nos algoritmos J48 e Naive Bayes, foram os padrões dos mesmos. Em relação a separação dos dados para treino e teste, foram utilizados 70% dos dados para treino e 30% para teste tanto no primeiro experimento, quanto no segundo.

Cerca de 15 variáveis, foram escolhidas para compor o input dos algoritmos usados. Durante os testes de acordo com o conjunto de variáveis de input utilizados, ocorriam oscilações na acurácia do experimento.

5. Resultados e Discussão

Após realizar os dois experimentos, começou-se a análise e discussão dos resultados obtidos. Primeiramente tentou-se interpretar e exibir os resultados de cada um dos experimentos separadamente, e em seguida realizando uma comparação entre os dois experimentos.

5.1 Resultados do primeiro experimento

No **primeiro experimento**, atingiu-se uma acurácia de 71.9384 % das instâncias analisadas, por meio do algoritmo J48. O TP-Rate (True-Positive Rate), mostrou que a situação mais simples de ser predita são as notas classificadas como “Alta”, e as de maior complexidade de prever são as classificadas como “Media”.

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,856	0,109	0,806	0,856	0,830	0,737	0,928	0,851	alta
	0,580	0,196	0,584	0,580	0,582	0,384	0,755	0,553	media
	0,712	0,114	0,756	0,712	0,733	0,607	0,885	0,757	baixa
Weighted Avg.	0,719	0,139	0,718	0,719	0,718	0,581	0,858	0,724	

Figura 5. Acurácia detalhada do primeiro experimento.

Fonte: Elaborado pelo autor.

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
 1388  208   25 |    a = alta
  301  873  332 |    b = media
   33  414 1105 |    c = baixa
```

Figura 6. Matriz de Confusão gerada no primeiro experimento.

Fonte: Elaborado pelo autor.

5.2 Resultados do segundo experimento

No segundo experimento, o acerto alcançado foi de 68.0701 % das instâncias testadas, por meio do algoritmo Naive Bayes, com base nos resultados obtidos com o mesmo, observou-se que o TP-Rate (True-Positive Rate) demonstra que o cenário mais simples para prever as notas de Matemática e suas Tecnologias, são as classificadas como "Alta", enquanto o caso mais complexo fica por conta daquelas classificadas como "Media".

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,853    0,161    0,738      0,853    0,791      0,672    0,920     0,852     alta
                 0,442    0,171    0,550      0,442    0,490      0,289    0,743     0,527     media
                 0,733    0,147    0,712      0,733    0,723      0,582    0,898     0,820     baixa
Weighted Avg.    0,681    0,160    0,669      0,681    0,671      0,519    0,856     0,736
```

Figura 7. Acurácia detalhada do segundo experimento.

Fonte: Elaborado pelo autor.

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
 1382  201   38 |    a = alta
  419  665  422 |    b = media
   72  342 1138 |    c = baixa
```

Figura 8. Matriz de Confusão gerada no segundo experimento.

Fonte: Elaborado pelo autor.

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

O melhor resultado foi proveniente do primeiro experimento, ou seja, do uso do algoritmo J48, que obteve uma acurácia de 71.9384 %, enquanto o segundo experimento que utilizou o algoritmo Naive Bayes alcançou apenas 68.0701 % de acertos. Apesar de não haver uma diferença tão grande na acurácia dos dois experimentos, o uso do J48 se mostra mais efetivo que o uso do Naive Bayes pois por meio da Figura 5 e da Figura 7, é possível notar que o TP-Rate do algoritmo J48 apresentou melhores resultados para praticamente todas as categorias.

5.4 Árvore de Decisão

O primeiro experimento utilizou o algoritmo J48, o mesmo tem a capacidade gerar arvores de decisões. Com a arvore de decisão gerada no primeiro experimento notou-se que a variável mais importante para essa árvore foi o atributo “DEPENDENCIA_ADMINISTRATIVA”. Outras variáveis consideradas importante para o experimento são: INDICADOR_DE_NIVEL_SOCIOECONOMICO e CATEGORIZACAO_TAXA_DE_PARTICIPACAO. A Figura 9 apresenta o conteúdo armazenado por cada uma das variáveis citadas.

Nome da Variável:	Conteúdo armazenado:
DEPENDENCIA_ADMINISTRATIVA	Tipo da escola: -Privada; -Municipal ; -Estadual –Federal.
INDICADOR_DE_NIVEL_SOCIOECONOMICO	Nível socioeconômico da escola: -Muito Baixo; -Baixo; -Médio Baixo; –Médio; -Médio Alto; –Alto; -Muito Alto.
CATEGORIZACAO_TAXA_DE_PARTICIPACAO	Taxa de participação: Foram criadas 4 categorias a partir de certos intervalos numéricos.

Figura 9. Conteúdo armazenado pelas variáveis.

Fonte: Elaborado pelo autor.

5.5 Resultado para educação

A literatura evidencia que modelos preditivos podem ajudar nas estratégias educacionais. Dessa forma destaca-se que por meio da análise do modelo preditivo gerado nesse estudo, os especialistas da área de educação e do ensino de Matemática e suas Tecnologias podem traçar estratégias para a identificação de pontos críticos e para a melhoria do ensino dessa área da educação. As variáveis DEPENDENCIA_ADMINISTRATIVA, INDICADOR_DE_NIVEL_SOCIOECONOMICO e CATEGORIZACAO_TAXA_DE_PARTICIPACAO foram as variáveis mais significativas para os algoritmos, dessa forma os especialistas podem traçar estratégias como um olhar mais detalhado para cada situação das mesmas.

6. Considerações Finais

A meta principal deste trabalho foi efetuar um estudo e aplicar as técnicas e métodos de mineração de dados, para gerar modelos para predição do desempenho da redação do ENEM, a partir dos dados do ENEM 2015 e algoritmos de classificação como o J48 e o Naive Bayes, e também gerar uma arvore de decisão.

Por meio dos modelos de predição gerados nesse trabalho destaca-se que é possível predizer com um mínimo de exatidão o desempenho da prova de Matemática e suas Tecnologias. Contudo ressalta-se que é preciso pesquisas futuras voltadas ao mesmo assunto e objetivo desse trabalho, para que se melhorar os modelos de predição.

As variáveis que mais causaram impacto foram DEPENDENCIA_ADMINISTRATIVA, INDICADOR_DE_NIVEL_SOCIOECONOMICO e CATEGORIZACAO_TAXA_DE_PARTICIPACAO, devido a isso conclui-se que os especialistas da área da educação devem prestar atenção e analisar essas variáveis.

Destaca-se que o primeiro experimento obteve o melhor resultado dessa pesquisa atingindo uma acurácia de 71.9384%. Com base nesse cenário, esse trabalho procurou colaborar na melhora do desempenho dos alunos na prova de Matemática e suas Tecnologias, por meio de modelos para predição do desempenho Matemática e suas Tecnologias gerados nessa pesquisa.

Referências

ALVES, Rafael Damiani. PREDIÇÃO DO DESEMPENHO DA REDAÇÃO DO ENEM UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS. 2018. 67 f. TCC (Graduação) - Curso de Tecnologias da Informação e Comunicação, Universidade Federal de Santa Catarina, Araranguá, 2018.

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.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data

Mining to Knowledge Discovery in Databases. Aaai Press, p.37-54, 1996.

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.

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

INEP. Enem Por Escola. 2015. Disponível em: <http://portal.inep.gov.br/web/guest/enem-por-escola>. Acesso em: 20 set. 2018.

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

OECD. Resumo de resultados nacionais do PISA 2015. 2016. Disponível em: <https://www.oecd.org/pisa/PISA-2015-Brazil-PRT.pdf>. Acesso em: 18 set. 2018.

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.

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