Uma Análise Experimental Usando Mineração de Dados Educacionais sobre os Dados do ENEM para Identificação de Causas do Desempenho dos Estudantes

Maicon Ribeiro Banni[1], Marcos Vinicius dos P. Oliveira[1], Flavia Cristina Bernardini[1]

[1]Programa de Pós-Graduação do Instituto de Computação – Universidade Federal Fluminense (UFF)

{maiconbanni, marcoso}@id.uff.br, fcbernardini@ic.uff.br

Abstract. *The National High School Examination (ENEM) is one of the ways used to measure the level of knowledge of students at the end of basic education. This paper presents an experimental analysis based on Educational Data Mining, including the use of data visualization techniques and the construction of predictive models to identify the attributes most related to student performance. We used collected data from ENEM 2018 related to all states in Brazil, with a total of more than 5 million records. Our results showed that the socio-economic attributes in fact have a significant relationship with the results of students at ENEM.*

Resumo. *O Exame Nacional do Ensino Médio (ENEM) é uma das maneiras utilizadas para mensurar o nível de conhecimento dos estudantes no fim da educação básica. Este artigo apresenta uma análise experimental baseada em Mineração de Dados Educacionais, incluindo o uso de técnicas de visualização de dados e construção de modelos preditivos para identificar os atributos mais relacionados ao desempenho dos estudantes. Foram usados dados do ENEM de 2018 de todo o Brasil, com um total de mais de 5 milhões de registros. Nossos resultados mostraram que os atributos sócio-econômicos de fato apresentam uma relação significativa com o resultado dos estudantes no ENEM.*

1. Introdução

A fundamental importância no desenvolvimento humano faz do setor educacional um tema de muito interesse para a sociedade em geral. Entretanto, a educação brasileira ainda enfrenta inúmeras dificuldades nos mais diferentes níveis de ensino [Freire 2019]. De acordo com a *Organization for Economic Co-operation and Development* (OECD), o nível da educação brasileira está abaixo da média dos países membros da organização, sendo um fator que pode resultar em inúmeras consequências como desigualdade social e pobreza [OECD 2019].

Uma das medidas adotadas para combater os déficits na educação está na utilização de Mineração de Dados Educacionais (MDE), que usa dados coletados recorrentemente em diferentes ciclos do contexto educacional, seja de forma presencial ou *online* [BAKER et al. 2011]. A MDE é uma área interdisciplinar que utiliza diferentes técnicas e métodos de análise de dados e extração de modelos e padrões para adquirir informações úteis e aplicáveis que possam contribuir para a melhoria dos métodos de ensino-aprendizado e a tomada de decisões dos gestores educacionais. De fato, o objetivo

final da MDE é o aumento da qualidade do ensino e o apoio aos gestores na tomada de decisão [PAIVA et al. 2012] [ALDOWAH et al. 2019].

No contexto brasileiro, o ENEM tem por objetivo avaliar o desempenho do estudante ao fim da escolaridade básica. Além disso, serve como meio de acesso à educação superior na maioria das instituições de ensino do território nacional, bem como, em instituições de Portugal que são conveniadas ao Brasil [INEP, 2020]. Dada essa grande importância do exame, torna-se essencial análises eficientes nos dados obtidos anualmente sobre o conhecimento técnico, informações socioeconômicas e culturais, de modo que tais análises possibilitem mensurar os níveis de conhecimento desses participantes em seus diferentes contextos socioculturais e econômicos. Alguns estudos apontam que (i) os discentes de instituições privadas apresentam melhores rendimentos no exame e escolas com maiores taxas de evasão apresentam resultados piores nas avaliações do ENEM [Sánchez, 2019]; e (ii) há uma tendência de melhor desempenho dos candidatos com maior poder aquisitivo, oriundos de escolas privadas e federais, das raças branca, pardo ou amarelo, evidenciando os reflexos das diferenças socioeconômicas no desempenho dos mesmos [Carmo et al., 2020]. No entanto, é importante evoluirmos na construção de metodologias para análise dos resultados do exame, considerando a grande quantidade de dados disponível para identificar o nível de relação dos atributos sócio-econômicos, dentre outros, com o resultado obtido no ENEM.

Este artigo tem como objetivo apresentar uma análise experimental guiada por uma metodologia baseada em MDE para identificar os principais fatores que influenciam no rendimento dos participantes que realizam o exame. Usamos para isso técnicas de visualização de dados univariadas e bivariadas bem como construímos modelos preditivos, usando aprendizado de máquina, para identificar quais características levam ao rendimento dos candidatos que se submetem ao ENEM. Os modelos preditivos visam encontrar funções cujo domínio são os atributos de entrada ou descritores dos dados, que no nosso caso são os atributos socio-econômicos, dentre outros, e cuja imagem é o atributo classe, que no nosso caso é o desempenho no exame. A vantagem desse tipo de modelo é que é possível fazer uma análise multivariada dessa relação, caso ela exista. Quanto maior a qualidade dos preditores, maior o indicativo de relação entre os atributos de entrada e de saída. A diferença deste trabalho para outros da literatura é que utilizamos todo o conjunto de dados do ENEM 2018, sem nos restringirmos a alguma região do país. Tal estudo é importante para identificar a diferença na performance dos estudantes inclusive em relação aos estados de origem dos estudantes. Com a metodologia, foi possível extrair diversas informações e conhecimentos relevantes baseados nos dados do ENEM, reforçando a questão da desigualdade social em diversos aspectos no Brasil quanto ao acesso ao ensino superior.

Este trabalho está organizado da seguinte forma: a Seção 2 aborda os trabalhos relacionados a este estudo. A Seção 3 apresenta os passos da análise experimental proposta com o detalhamento das etapas e as principais técnicas utilizadas. A Seção 4 explora e discute os resultados obtidos. Por fim, a Seção 5 apresenta as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

Na literatura, são encontrados diversos trabalhos recentes, que apresentam estudos sobre o uso da MDE para análise de dados do ENEM, com o objetivo de extrair padrões desses dados para apoiar gestores da área de educação, tanto na educação básica quanto no

ensino superior. Santos et al. [2019] utilizou o Apache Spark para manipular e analisar dados do exames realizados entre 1998 e 2017 para avaliar as alterações nos perfis dos candidatos que realizaram o exame no período. O estudo se baseia nos dados referentes a faixa etária, sexo, região de nascimento e pontuação em cada área do conhecimento. Os resultados mostraram que o perfil mudou significativamente desde a implantação do ENEM. Franco et al. [2020] buscaram identificar as 20 principais características que influenciam no desempenho dos estudantes nas edições do ENEM de 1998 a 2019. Esse trabalho visava identificar quais atributos faltam para um melhor poder de predição que foram excluídos do formulário do exame com o decorrer dos anos. Os autores aplicaram algoritmos de classificação e seleção de atributos para reduzir a dimensionalidade e evidenciar tais características. Os resultados apontam que alguns atributos considerados importantes nos experimentos foram excluídos do questionário do ENEM dos últimos anos, como por exemplo dados relacionados ao ensino fundamental do estudante. Além disso, os autores verificaram que a importância e correlação dos atributos são alteradas ao longo dos anos. Esse trabalho nos auxiliou a selecionar técnicas de seleção de atributos e algoritmos de aprendizado de máquina para nosso trabalho. No entanto, como nosso objetivo é identificar a relação entre os atributos sócio-econômicos e o rendimento dos estudantes, exploramos mais técnicas de visualização de dados, para além do que foi realizado por Franco et al. [2020]. Já Silva et al. [2020] utilizou técnicas de mineração de dados com os algoritmos de Clusterização e de Regras de Associação, visando identificar o desempenho dos concluintes do ensino médio que realizaram o ENEM de 2019 no Estado de Minas Gerais, a partir das correlações das variáveis relacionadas a aspectos socioeconômicos. O trabalho evidencia a semelhança no desempenho médio de estudantes oriundos de escolas da rede federal em relação aos da rede privada. Carmo et al. [2020] apresentam um estudo sobre o desempenho dos estudantes do Estado do Rio Grande do Sul, baseado em uma análise comparativa no perfil educacional e socioeconômico dos participantes do ENEM de 2019. De acordo com os resultados, há uma tendência de melhor desempenho dos candidatos com maior poder aquisitivo, oriundos de escolas privadas e federais, das raças branca, pardo ou amarelo, evidenciando os reflexos das diferenças socioeconômicas no desempenho dos mesmos. Sánchez [2019] utiliza o algoritmo Apriori para verificar quais fatores estão relacionados ao desempenho dos alunos que realizaram o ENEM, tendo como base a instituição de ensino de origem dos estudantes, visando propor formas de ampliar o bom desempenho desses estudantes. Os resultados apontam evidências que os discentes de instituições privadas apresentam melhores rendimentos no exame. Além disso, o trabalho indica que as escolas com maiores taxas de evasão também apresentam resultados piores nas avaliações do ENEM. Simon e Cazella [2017] e Alves et al. [2018] aplicaram as técnicas de mineração de dados por meio do software WEKA, com o objetivo de encontrar padrões e gerar modelos preditivos a partir dos dados públicos referentes ao ENEM de 2015. Os trabalhos analisam os indicadores de desempenho das notas das escolas do ensino médio. O primeiro estudo se baseia nos dados de ciências da natureza e suas tecnologias e utiliza o algoritmo J48, e o segundo utiliza as notas da prova de Matemática e suas Tecnologias com os algoritmos Naive Bayes e J48. Os resultados mostraram que o desempenho médio dos alunos que compõem os grupos com notas entre 650 e 749,99 são de estudantes de elevado nível socioeconômico.

Os trabalhos descritos apresentam diferentes abordagens para analisar os dados do ENEM. No entanto, os mesmos mostram limitações em relação à abrangência de suas análises, os quais restringem o escopo em uma Região Específica (RE), Área de

Conhecimento Específica (ACE), Conjunto de Atributos Específicos (CAE) e/ou Ano Específico (AE). Ainda, há trabalhos que não usam técnicas para Visualização de Dados (VD) e não objetivam a Identificação de Causas para o desempenho dos estudantes (IC), tendo como base os dados econômicos, culturais e/ou sociais, como mostra a Tabela 1. Nessa tabela, o símbolo ✓ significa que o artigo da referida linha está restrito ao escopo da referida coluna, usa VD ou discute IC. Já o símbolo × significa que o artigo da referida linha não limita o escopo da referida coluna, não usa VD ou não discute IC. Assim, até onde vai nosso conhecimento, não encontramos trabalhos que apresentem a utilização de diferentes técnicas visando o melhor entendimento dos padrões dos estudantes que realizam o ENEM em relação ao seu desempenho considerando todo o território nacional. Uma das possíveis causas desta limitação está relacionada às dificuldades com o processamento do grande volume de dados disponibilizado pelo INEP referente a cada ano do ENEM, onde a manipulação desses dados se torna um gargalo, quando realizada com ferramentas tradicionais ou com recursos computacionais limitados. Nesse sentido, fica nítido que há necessidade de métodos que possibilitem análises sobre a íntegra dos dados disponibilizados. Uma vez que, as desigualdades socioculturais e econômicas no Brasil são expressivas, as quais impactam diretamente o desempenho de cada candidato.

Tabela 1. Comparativos de trabalhos apresentados na literatura

Estudos da literatura	RE	ACE	CAE	AE	VD	IC
[Santos et al., 2019]	×	×	✓	×	×	×
[Franco et al., 2020]	×	×	×	×	×	✓
[Silva et al., 2020]	✓	×	✓	✓	×	✓
[Carmo et al., 2020]	✓	×	✓	✓	×	✓
[Sánchez, 2019]	×	×	✓	✓	×	✓
[Simon e Cazella, 2017]	×	✓	×	✓	×	✓
[Alves et al., 2018]	×	✓	×	✓	×	✓
Este trabalho	×	×	×	✓	✓	✓

3. Metodologia da análise experimental considerando a MDE

Esta seção apresenta as etapas da análise experimental realizada, que combina, simultaneamente, técnicas de mineração e visualização de dados, visando alcançar os objetivos traçados utilizando o *dataset* completo referente ao ENEM (2018) com aproximadamente 3,3 Gigabytes de dados, 5.513.747 registros e 137 atributos. A **primeira etapa** compreende uma das etapas que mais diferencia este artigo dos demais trabalhos da literatura. Essa fase consiste em refinar a base de dados, utilizando técnicas de visualização de dados para viabilizar o emprego de algoritmos de seleção de atributos e modelos de aprendizado de máquina. Essas técnicas possibilitam uma análise global dos dados baseada na correlação e relevância dos atributos de forma prática e sem muito esforço computacional. Esta análise permite reduzir a dimensionalidade dos dados, identificar atributos altamente correlacionados e excluir ou transformá-los, sem perder as características relevantes dos dados originais. Nos trabalhos da literatura por nós analisados, essa etapa foi menos explorada e detalhada. As técnicas de visualização utilizadas foram: *Heatmaps*, gráficos que expressam mapas de calor e transmitem mais facilmente a informação desejada, onde as regiões com maior concentração dos dados possuem as cores mais acentuadas*; BoxPlot*, gráficos que representam a variação de dados observados de uma variável numérica por meio de quartis, sendo possível identificar onde se concentram os valores atípicos ou *outliers; e Violin Plot*, gráficos utilizados para visualizar a distribuição dos dados e sua densidade de probabilidade. Na **segunda etapa**,

utilizamos os resultados obtidos na fase anterior, que consiste em avaliar, compreender, normalizar e reconstruir uma base de dados que atenda às necessidades desejadas, por meio de técnicas de pré-processamento que ajudam a eliminar os registros nulos e fora do escopo da análise, bem como recompor as informações ausentes sempre que possível, sem descaracterizar os dados originais. Na **terceira etapa**, realizamos a aplicação de algoritmos de seleção de atributos usando os seguintes algoritmos: *InfoGainAttributeEval* para avaliar o valor de um atributo medindo o ganho de informações em relação à classe alvo; *GainRatioAttributeEval* para avaliar o valor de um atributo medindo a taxa de ganho em relação à classe alvo*; CorrelationAttributeEval, para* avaliar o valor de um atributo medindo a correlação de Pearson entre ele e a classe. A partir desses algoritmos é possível identificar, de forma quantitativa, o nível de influência de um atributo no desempenho dos participantes do ENEM. No entanto, os atributos com baixa influência, ou seja, com fraca correlação e pouco relevantes, não são automaticamente eliminados, e sim agrupados logicamente, visando manter a originalidade dos dados e, consequentemente, aumentar a precisão do modelo. Porém, caso permaneçam com baixos índices de ganhos e baixa correlação, são eliminados para alcançar as melhorias citadas anteriormente. Outro item que diferencia este estudo dos demais trabalhos da literatura, visto que, nestes a remoção acontece de forma imediata para os registros com baixa correlação. A partir daí, escolhemos, utilizamos e avaliamos algoritmos de aprendizado de máquina para criação dos modelos preditivos. Neste trabalho foram selecionados os algoritmos mais frequentes na literatura apenas para avaliar o desempenho da metodologia, entre eles: *Decision Tree, Random Forest, Logistic Regression, Naive Bayes e K-Nearest Neighbors* (KNN). Realizamos a análise da qualidade dos preditores com as seguintes métricas de avaliação: Acurácia (Eq. 1), Precisão (Eq. 2), *Recall* (Eq. 3) e F1-Score (Eq. 4).

$$Acurácia = \frac{Verdadeiros\ Positivos\ (TP) + Verdadeiros\ Negativo\ (VN)}{Total} \quad (1)$$

$$Precisão = \frac{Verdadeiros\ Positivos\ (TP)}{Verdadeiros\ Positivos\ (TP) + Falsos\ Positivos\ (FP)} \quad (2)$$

$$Recall = \frac{Verdadeiros\ Positivos\ (TP)}{Verdadeiros\ Positivos\ (TP) + Falsos\ Negativos\ (FN)} \quad (3)$$

$$F1 - Score = \frac{2 * Precisão * Recall}{Precisão + Recall} \quad (4)$$

4. Resultados

A partir dos dados usados para nosso estudo, foram realizadas análises estatísticas e visuais para reduzir a dimensionalidade dos dados, identificar tendências e avaliar a densidade dos dados, visando eliminar ou agrupar os atributos com baixa significância, de forma a não causar perda de informações relevantes para este estudo, para facilitar a análise do conjunto de dados[1]. A Figura 2 ilustra algumas das visualizações de dados construídas, e mostramos as que consideramos mais importantes. O gráfico na parte superior à esquerda mostra que conforme varia a raça (0, 1 e 2) e os estados de onde os estudantes vêm, há uma diferença significativa no desempenho do ENEM, indicado pela variação de

[1] Nossos scripts, todas as visualizações construídas e o conjunto de dados utilizado para executar nossa metodologia está disponível em https://github.com/maiconbanni/machine-learning-2020-01

cores (entre amarelo bem claro até vermelho bem escuro). O gráfico de caixas na parte superior à direita mostra que o nível de escolaridade dos pais influencia diretamente na média de desempenho dos estudantes em todos os estados brasileiros. Por fim, o gráfico na parte inferior mostra que a escolha da língua inglesa (idioma 0) possibilita um melhor desempenho também no exame. Esses gráficos mostram o quanto o desempenho dos candidatos varia, dependendo de raça, cor, nível de educação dos pais e localização geográfica, o que enfatiza, assim, a relevância de análises conjuntas quando se busca entender e comparar, de fato, a real situação dos candidatos.

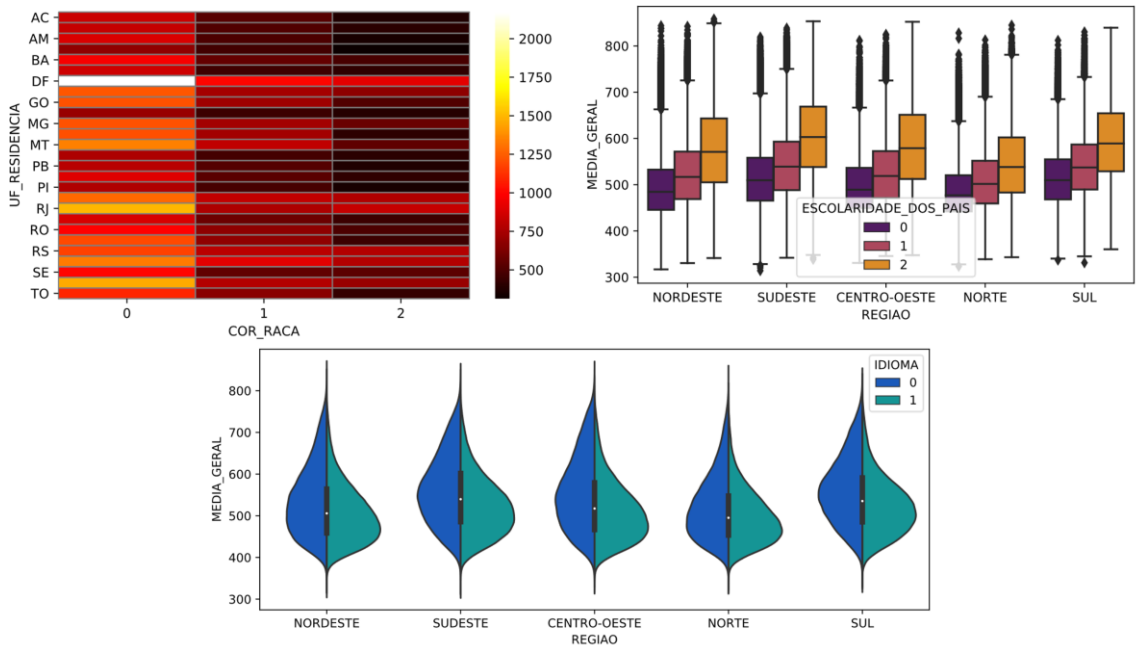

Figura 2. Visualizações de Dados Construídas para a Base de Dados

Após a análise inicial, foram selecionados 18 atributos referentes às informações pessoais, acadêmicas e socioeconômicas dos candidatos. Isso levou a uma redução na dimensão do conjunto de dados (de 137 para 18 atributos), o que facilita a execução dos algoritmos de seleção de atributos InfoGainAttributeEval, GainRatioAttributeEval e CorrelatioAttributeEval, para identificar os atributos mais relevantes para o experimento considerando o atributo classe construído. O atributo classe Média Geral (MG) foi construído com base nos resultados do ENEM nas 5 áreas, que são: NU_NOTA_CN (nota em Ciências da Natureza), NU_NOTA_CH (nota em Ciências Humanas), NU_NOTA_LC (nota em Linguagens e Códigos), NU_NOTA_MT (nota em Matemática), NU_NOTA_REDACAO (nota em Redação). A partir da média dessas 5 notas, dividimos o valor resultante em 3 classes: Insuficiente (<450), Regular (entre 450 e 650) e Excelente (≥650). É importante enfatizar que o número de registros foi mantido (5.513.747). A Figura 3 apresenta os resultados reportados pela ferramenta após a primeira aplicação dos algoritmos. Além disso, construímos dois novos atributos a partir dos apresentados na Figura 3: (i) usando Q0001 e Q0002 (relativo à escolaridade do pai e da mãe do candidato), construímos o atributo ESCOLARIDADE_DOS_PAIS; e (ii) a partir de “Q005” e “Q006”, que revelavam informações financeiras dos familiares, com 20 e 16 categorias respectivamente, construímos o atributo “RENDA_PERCAPITA_FAMILIAR” com apenas 8 categorias, para facilitar o uso do dado para análise.

Fizemos uma combinação dos diversos atributos apresentados na Figura 3 e os 2 atributos construídos para construção de modelos preditivos, usando como atributo alvo o atributo MG e usando algoritmos de aprendizado de máquina implementados na biblioteca *scikit-learn*. Na Figura 4 são ilustrados os cinco atributos construídos e usados da base original que apresentaram os melhores resultados em relação às métricas utilizadas. Selecionamos aleatoriamente 70% do conjunto de dados para treinamento e 30% para teste. Na Figura 6 são ilustrados os resultados obtidos nas quatro métricas de avaliação utilizadas.

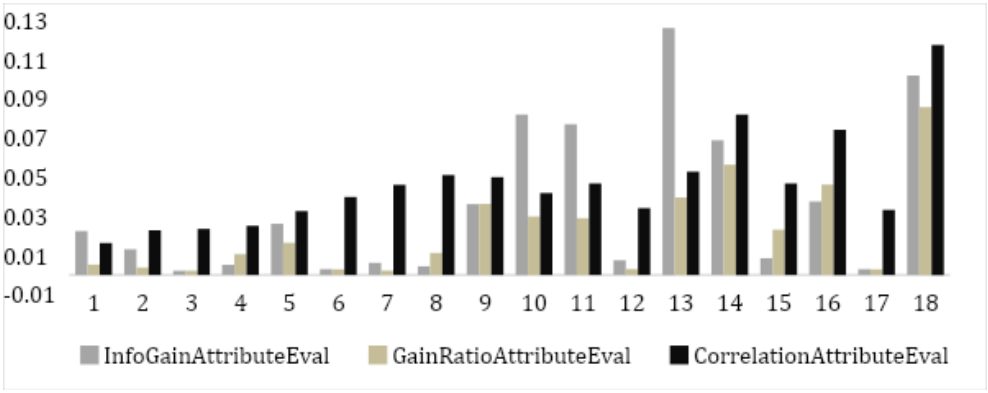

Figura 3. Nível de relevância dos atributos na primeira aplicação dos algoritmos.

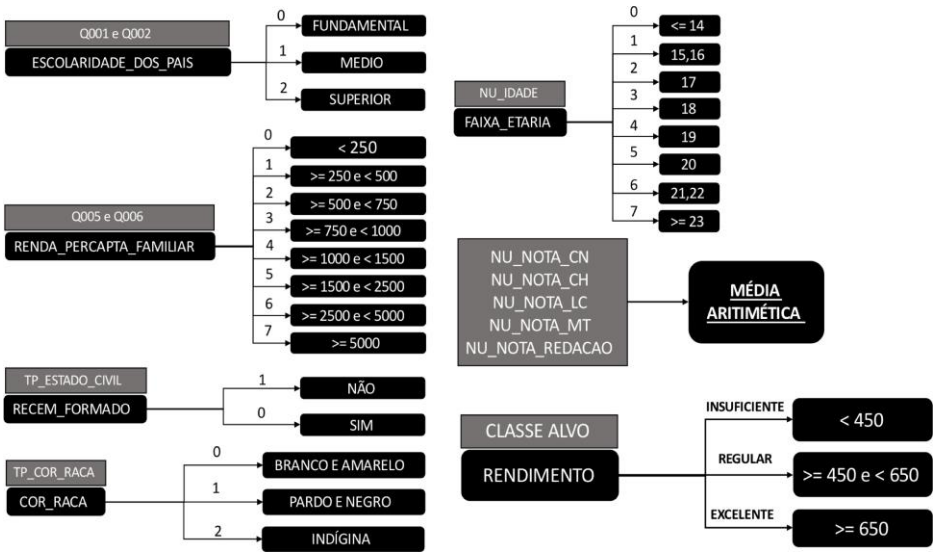

Figura 4 - Normalização e Reutilização de atributos.

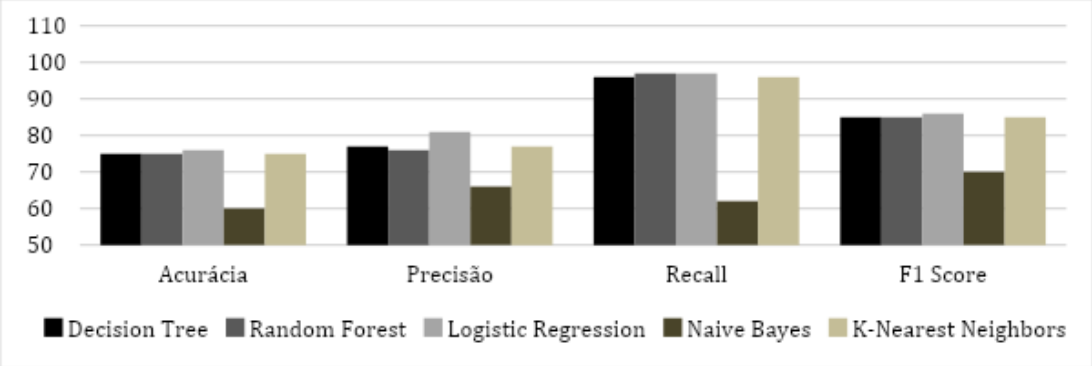

Figura 6 – Métricas de Avaliação do Desempenho dos algoritmos

É importante observar que os algoritmos utilizados na criação dos modelos preditivos apresentaram desempenhos satisfatórios, tendo como base os resultados contidos na literatura, onde todos os algoritmos alcançaram uma precisão média superior a 70%, com destaque para o algoritmo *Logistic Regression*, que alcançou mais de 80% de precisão, superando os resultados dos trabalhos que utilizaram o mesmo algoritmo [Simon e Cazella 2017] [Alves et al. 2018]. Além disso, segundo *F1-Score*, os algoritmos obtiveram médias superiores a 80%, o que indica que mesmo para as 3 classes individualmente a precisão e *recall* são altos para a maioria dos algoritmos. De modo geral, dentre os algoritmos utilizados nos experimentos, a *Logistic Regression* reportou os melhores resultados. Tal algoritmo calcula a regressão logística entre os atributos de entrada e o atributo de saída, indicando alta relação entre esses atributos. Por outro lado, o *Naive Bayes*, que se baseia na independência dos atributos de entrada e em probabilidades de valores individuais entre os atributos de entrada e o atributo classe, obteve os menores índices de desempenho. É importante notar que, portanto, considerando o uso de diferentes algoritmos de aprendizado de máquina, é possível observar uma alta relação entre os atributos Escolaridade dos Pais, Renda Percapta Familiar, se o estudante é Recém Formado, Cor e Raça e Faixa Etária com o resultado obtido no ENEM, reforçando o que havia sido observado anteriormente nas visualizações por atributos.

5. Conclusões

Dada a grande importância do ENEM, este trabalho propõe uma análise experimental usando MDE, baseada em técnicas de visualização de dados, seleção de atributos e construção de modelos preditivos por meio de algoritmos de aprendizado de máquina, no intuito de identificar os fatores mais relevantes para o rendimento dos participantes do exame. Observamos como foi importante a utilização de técnicas de visualização de dados para nos auxiliar a guiar a seleção de atributos. A partir da seleção por meio de visualizações que utilizamos técnicas de seleção de atributos para a construção dos modelos preditivos. As visualizações geradas possibilitaram uma visão global dos dados, o que acelerou a compreensão, ajudou na eliminação de atributos com pouca relevância e, consequentemente, nos auxiliou a reduzir a dimensionalidade dos dados tornando as análises mais simples. É importante observar que o desenvolvimento de novas formas de visualização de dados nos últimos anos foi fundamental para nosso processo de avaliação. Assim, foram notórias as melhorias no desempenho dos algoritmos de seleção e aprendizado de máquina, as quais implicam diretamente na redução dos custos

computacionais. Além disso, foi possível constatar que a utilização de atributos pouco relevantes combinados a atributos mais relevantes pôde resultar em preditores com maior qualidade como, por exemplo, avaliar a renda familiar (atributo com alta relevância), desconsiderando a quantidade de pessoas que compartilham os mesmos recursos (atributo com baixa relevância), pode trazer uma visão equivocada da realidade por trás desta informação. Logo, conclui-se que o ideal deve ser o uso do conhecimento adquirido por meio da etapa de visualização de dados antes da remoção dos atributos de baixa relevância. É importante destacar que nossa proposta não limita o uso específico de técnicas de seleção de atributos ou algoritmos de aprendizado de máquina, mas sim auxilia na compreensão e resolução das dificuldades encontradas durante o experimento. Como resultado do uso de visualizações univariada, bivariada e uso de algoritmos de aprendizado de máquina, foi possível observar uma relação significativa entre atributos sócio-econômicos e desempenho dos estudantes.

Uma das limitações do nosso estudo foi o uso dos dados do ENEM somente do ano de 2018. Isso foi feito para selecionar, dentre os muitos atributos, quais trazem um melhor indicativo da relação entre as características dos estudantes e o resultado no ENEM. É importante observar que a análise experimental é bastante longa, como já apontado por diversos autores da área de mineração de dados em geral. No entanto, nossos resultados mostram um *rationale* interessante para a execução de análise experimental dos dados do ENEM. Além disso, análises comparativas com diferentes discretizações no resultado da classe alvo podem ser realizadas. Nós realizamos diversas discretizações e usamos a que foi apresentada pois foi a que também obteve melhores resultados. Por fim, podem ser realizados cruzamentos dos dados com informações do Programa Universidade para Todos (PROUNI) e do Sistema de Seleção Unificado (SISU), com o objetivo de realizar a predição do resultado e verificar se é possível indicar os possíveis cursos onde o candidato teria condições de disputar uma vaga.

Referências

Aldowah, H., Al-Samarraie, H. and Fauzy, W.M. (2019). Educational Data Mining and Learning Analytics for 21stcentury higher education: A Review and Synthesis Telematics and Informatics.

Alves, R. D., Cechinel, C. e Queiroga, E. (2018). Predição do desempenho de Matemática e Suas Tecnologias do ENEM utilizando técnicas de Mineração De Dados. Em: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. p. 469.

Baker, R.; Isotani, S.; Carvalho, A. (2011). Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação. p.3-13.

Brito, E. C e Damazio, M. R. (2018). Desenvolvimento Econômico No Brasil: Similaridades E Diferenças Entre As Regiões Sul E Nordeste No Período De 2001 A 2015. Revista de Desenvolvimento Econômico – RDE - Ano XX – V. 3 - N. p. 167 – 198.

Carmo. R., V., Heckler. W., F. e Carvalho, J., V. (2020) Uma Análise do Desempenho dos Estudantes do Rio Grande do Sul no ENEM 2019. Revista Novas Tecnologias na Educação. V.18 Nº 2.

Franco, Jacinto José; Miranda, Fernanda Luzia de Almeida; Stiegler, Davi; Dantas, Felipe Rodrigues; Brancher, Jacques Duílio; Nogueira, Tiago do Carmo. (2020). Usando Mineração de Dados para Identificar Fatores mais Importantes do Enem

dos Últimos 22 Anos. Anais do XXXI Simpósio Brasileiro de Informática na Educação. p. 1112-1121.

Freire, P. (2019). Direitos Humanos e Educação Libertadora: gestão democrática da educação pública na cidade de São Paulo. Rio de Janeiro, São Paulo: Paz e Terra.

INEP. (2020). ENEM. Disponível em: http://inep.gov.br/web/guest/enem. Acessado em: 06 de junho.

OECD. (2019), PISA 2018 Results (Volume I): What Students Know and Can Do, PISA, OECD Publishing, Paris. Disponível em: https://doi.org/10.1787/5f07c754-en. Acessado em: 10 de junho.

Paiva, R.; Bittencourt, I. I.; Pacheco, H.; Da Silva, A. P.; Jacques, P.; Isotani, S. (2012) Mineração de dados e a gestão inteligente da aprendizagem: desafios e direcionamentos. Instituto de Computação – Universidade Federal de Alagoas (UFAL), Alagoas – AL.

Sánchez, J. (2019). Desempenho das Instituições de Ensino Brasileiras no ENEM: uma abordagem Usando Mineração de dados. Nuevas Ideas en Informática Educativa, Volumen 15, p. 106 - 113.

Santos, B., Oliveira, C. G., Topin. L. O. H., Mendizabal, O. M. e Barwaldt, R. (2019). Analysis of Candidates Profile for the National Entrance Exams for Admission to Brazilian Universities. Proceedings - Frontiers in Education Conference, FIE.

Silva, V. A. A., Moreira, L. L. O., Gonçalves, L. B., Soares, S. S. R. F., Júnior, R. R. S. (2020). Identificação de Desigualdades Sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados. Anais do XXXI Simpósio Brasileiro de Informática na Educação, p. 72-81.

Simon, A. e Cazella, S. (2017). Mineração de dados educacionais nos resultados do ENEM de 2015. In Anais dos Workshops do Congresso Brasileiro de Informática na Educação, volume 6, p. 754.

Souza, H. V. L., Neiva, D., Cavalcanti, R. P., Rodrigues, R., Gomes, A. S., and Adeodato, P. (2017). Uma análise preditiva de desempenho dos alunos dos cursos no enade ´ com base no perfil socioeconomico e de desempenho no enem. Anais dos Workshops do Congresso Brasileiro de Informática na Educação, volume 6, p. 684.

Stearns, B., Rangel, F., Rangel, F., Firmino, F., and Oliveira, J. (2017). Scholar performance prediction using boosted regression trees techniques. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN).