Mineração de Dados Educacionais usando base de dados do Enem

João Vítor de Castro Martins Ferreira Nogueira (joao.nogueira@estudante.ufjf.br)
Lorenza Leão Oliveira Moreno (lorenza@ice.ufjf.br)
Stênio Sã Rosário Furtado Soares (ssoares@ice.ufjf.br)
Luciana Brugiolo Gonçalves (lbrugiolo@ice.ufjf.br)
Departamento de Ciência da Computação - Universidade Federal de Juiz de Fora

Resumo

Os microdados do Enem [1] (Exame Nacional do Ensino Médio) são dados abertos disponibilizados pelo Inep (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) em que a fonte dos dados são resultados das informações dos estudantes que se inscreveram para o exame. Nesses dados, estão contidas informações dos participantes relacionadas a prova e sobre aspectos socioeconômicos presente no questionário obrigatório para os inscritos da prova. Com isso, muitas fontes ricas de informações estão disponíveis nessas bases de dados, considerando o papel central que o exame tem para a entrada de estudantes no ensino superior no país.

Métodos

A partir dessa base de dados, foi realizado no trabalho a estratégia de KDD (*Knowledge Discovery in Database)* para guiar a metodologia da pesquisa, em que se segue o processo de seleção, pré-tratamento e tratamento dos dados, seguidos da aplicação dos algoritmos de mineração de dados e por fim da análise dos resultados. Um dos objetivos centrais da pesquisa foi analisar os impactos da pandemia de Covid-19 no exame. Isso foi possível através do acesso a base de dados de antes do início da pandemia e dos 2 anos iniciais de pandemia (2020-2021).

Foi feito uma análise inicial dos conjuntos dos dados, analisando os aspectos quantitativos de variáveis sociais relevantes do ponto de vista de desempenho [2] e sobre aspectos das abstenções na prova. Foi utilizado para os algoritmos de mineração de dados as bases de 2018 até 2021, usando os registros (estudantes) de Minas Gerais que são concluintes do ensino médio e que estiveram presentes nos exames, e que pertencem a colégios de escola estadual, tendo em vista que foi a variável que mais teve mudança de valores entre as consideradas, e que a presença de alunos de colégio estadual teve uma queda significativa quando comparado os valores pré-pademia com os valores durante a pandemia.

Na parte da mineração de dados, foi utilizado o algoritmo de clusterização de dados categóricos *K-modes*, utilizando o *método do cotovelo* para a escolha do número de clusters, e o algoritmo de classificação *Árvore de Decisão,* que utiliza os resultados do algoritmo de clusterização para fazer a classificação. Para o algoritmo de classificação foi usado os dados de 2019 para criar o modelo de classificação e os dados dos outros anos foram usados para aplicar o modelo gerado (**Figura 2**).

Tanto para a manipulação inicial dos dados, quanto para tratamento e implementação dos algoritmos de mineração de dados foi usado a linguagem *python* com as bibliotecas *pandas* (manipulação da base de dados), *sklearn* (algoritmos de mineração de dados), *seaborn* (visualização de dados), *k-modes* (do algoritmo *kmodes*)

Resultados

Ao utilizar o *método do cotovelo* para identificar o melhor número de clusters, o método retornou 5 clusters (C1, C2, C3, C4, C5), em que após análises estatísticas desses clusters com os atributos da base, percebe-se a concentração de valores associados à baixo desempenho [2] no cluster C4. Após a aplicação do algoritmo *árvore de decisão* para a base de 2019, e aplicar o modelo gerado para os outros anos, tem-se o resultado na **Figura 1**, em que se tem um decréscimo relativo maior para o cluster 4 em relação aos outros clusters. Considerando que nos anos 2018 e 2019 tinha-se que o cluster 4 possuía 23.5% e 24.4% do total dos dados, enquanto esse percentual cai para os anos de 2020 e 2021, já nos anos de pandemia, para 18.6% e 15% respectivamente

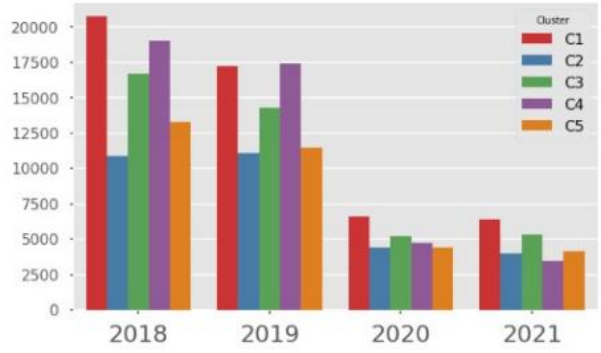

Figura 1: Resultado gráfico da classificação aplicada para os anos de 2018-2021

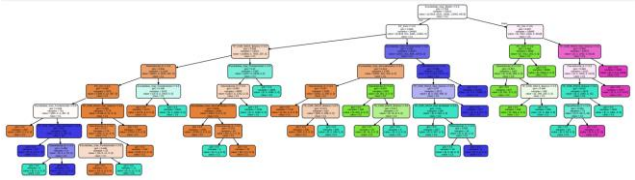

Figura 2: Árvore de decisão gerada a partir da base de 2019

Referências bibliográficas

[1] **Inep – Microdados do Enem.** URL: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

[2] da Silva, V. A. A., Moreno, L. L. O., Gonçalves, L. B., Soares, S. S. R. F., & Júnior, R. R. S. (2020, November). **Identificação de Desigualdades Sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados**. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação* (pp. 72-81). SBC.