Previsão da Performance dos Estudantes brasileiros a partir da Média da Nota de Ciências no PISA 2018

Messias Rafael Batista, Universidade de Pernambuco (mrb@ecomp.poli.br)

Dra. Roberta Andrade de Araújo Fagundes, Universidade de Pernambuco (roberta.fagundes@upe.br)

Contexto: Concurso educacionais são avaliações aplicadas em grande escala, seja a um seguimento de alunos de uma nação ou ainda por organismos internacionais em vários países. O objetivo destes concursos é visualizar um panorama da educação dentro de recortes determinados, como ensino superior, ensino médio ou ensino básico, além de fatores socioeconômicos e de infraestrutura da escola, por exemplo. Bases de dados resultante destes concursos podem elucidar sinais de avanço ou pontos de avaliação sobre o recorte aplicado. Motivação: Esta conjectura, permite o desenvolvimento de soluções orientadas a dados, suportando o campo educacional com modelos de *machine learning* que favorecem o processamento de grandes bases de dados. O contexto aplicado, é suportado pelo campo de pesquisa em Mineração de Dados Educacionais (em inglês, *Educational Datamining*), em sua sub-área de *student performance*. Assim, orientando os estudos no sentido de criar métricas ou indicadores da performance dos estudantes e seus fatores de impacto no resultado. Trabalho Relacionados: As técnicas utilizadas nos estudos são orientadas muitas vezes pelos resultados apresentados em relação ao problema estudado e as bases de dados disponíveis. Neste contexto, Cortez e Silva (2008) desenvolvem um trabalho analítico sobre a performance de estudantes a partir de dados acadêmicos e sociais para duas escolas em Portugal. O estudo buscou demonstrar quais as *features* mais influentes sobre o resultado da média obtida no terceiro ano do ensino médio (G3). Evidenciou-se que os resultados das notas anteriores são as *features* que mais impactam no resultado da G3, contrapondo-se as características sociais encontradas na base de dados. Outros estudos norteadores estão relacionados com as técnicas utilizadas, das quais destacam-se o uso de *Deep Learning* (HUSAIN, 2019) ou *Naive Bayes* (JAYAPRAKASH et al 2015), por exemplo. O ponto em comum nos trabalhos que analisam a performance de estudantes é a aplicação da seleção de atributos, no qual os estudos buscam evidenciar as principais *features* para cada cenário de pesquisa. O trabalho de SOKKHEY e OKAZAKI (2020) se trata de um estudo recente que faz uso da seleção de atributos em uma *dataset* com objetivo de melhorar aumentar a performance do modelo. A literatura estudada não apresentou evidências de utilização deste contexto ou base de dados para verificação da performance dos estudantes, assim revelando uma lacuna que resultou nesta pesquisa. Objetivo: Este estudo, portanto, busca compreender quais os principais atributos numéricos que explicam o resultado da nota média na prova de ciências aplicada no Programa Internacional de Avaliação de Estudantes (PISA) 2018 (y), a partir do conjunto de variáveis numéricas (x) presentes no *dataset*. Metodologia: Em busca de alcançar este objetivo, utilizou-se o *dataset* do PISA 2018, que é distribuído pela Organização para Cooperação e Desenvolvimento Econômico (OCDE). A fase de pré-processamento aplicou redução dos dados buscando analisar apenas o cenário brasileiro, e fazendo uso do conjunto de atributos que compunham informações socioeconômicos, de comportamento e de infraestrutura da escola dos estudantes. Em seguida, foi calculada a média da prova de ciências, que foi uma métrica construída como atributo de saída, a partir das notas alcanças na disciplina. Por fim, aplicados dois modelos de *Machine Learninig*, *Multiple Linear Regression* e *Random Forest Regression*, fazendo uso da biblioteca SkLearn. Resultados: Os resultados deste estudo são preliminares de um cenário novo que faz parte de uma pesquisa maior na base de dados do PISA 2018. Assim, demonstrou-se a capacidade explicativa dos dez atributos mais relevantes encontrados pela aplicação do *feature selection*. A aplicação da técnica de seleção *SelectKBest* com *f\_regression*, utilizando como parâmetro k=10, resultaram nos 10 atributos mais explicativos. Que suportaram a aplicação dos modelos de *Machine Learning*, obtendo os resultados da tabela 1 abaixo. O *SelectKBest* foi selecionado por resultar nos K’s parâmetros com pontuação mais alta em relação a variável alvo.

**Tabela 1** – Comparativo de métricas resultantes dos modelos

|  |  |  |  |
| --- | --- | --- | --- |
| Modelo | R2 | RMSE | MAE |
| Multiple Linear Regression | 0.0083 | 83.7515 | 67.6525 |
| Random Forest | 0.0206 | 83.1525 | 67.1253 |

A tabela 1 demonstra as métricas alcançadas pelos modelos aplicados no contexto descrito na metodologia desta pesquisa. As métricas não alcançaram níveis satisfatórios, evidenciando que a utilização do *featured selection* reduzindo os atributos de 1080 para 10 podem não representar aderência a variável alvo (y). **Ameaças e limitações do trabalho:** A construção e aplicação do modelo foram desenhados de acordo com o escopo desta pesquisa não havendo, portanto, evidências de aplicabilidade em outro contexto. **Trabalhos futuros:** A partir dos resultados demonstrados, espera-se que um novo trabalho de ajuste de parâmetros e aplicação de outros algoritmos, possam resultar na continuidade desta pesquisa, alcançando um resultado ótimo na previsão da performance de estudantes.

Palavras-chave: *Mineração de Dados Educacionais; Perfomance de Estudantes; Featured Selection; Regressão.*

Referências

CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações**. São Paulo: Saraiva, 2016.

CORTEZ, Paulo; SILVA, Alice. Using Data Mining to Predict Secondary School Student Performance. *In* A. Brito and J. Teixeira Eds., **Proceedings of 5th Future Business Technology Conference** (FUBUTEC 2008) pp. 5-12, Porto, EUROSIS, ISBN 978-9077381-39-7, Abril, 2008.

HUSSAIN, Sadiq et al. **Prediction Model on Student Performance based on Internal Assessment using Deep Learning**. iJET, v. 14, n. 8, p. 4-22, 2019.

JAYAPRAKASH, Sujith; BALAMURUGAN, E.; CHANDAR, Vibin. **Predicting Students’ Academic Performance Using Naïve Bayes Algorithm**. In: 8th Annual International Applied Research Conference. 2015.

Organization for Economic Co-operation and Development (OCDE). **PISA 2018 Database.** Disponível em <https://www.oecd.org/pisa/data/2018database/#d.en.516012>. Acessado em 05 de maio de 2022.

Scikit-Learn. **Scikit-learn**. Disponível em https://scikit-learn.org/, Acessado em: 07 de julho de 2022.

SUGUNA, R. et al. **Assessment of feature selection for student academic performance through machine learning classification**. Journal of Statistics and Management Systems, v. 22, n. 4, p. 729-739, 2019.

SOKKHEY, Phauk; OKAZAKI, Takeo. **Study on Dominant Factor for Academic Performance Prediction using Feature Selection Methods**. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8, 2020