Predição do Desempenho Escolar Baseado em Machine Learning

José H. C. A. Junior[1], Warley M. Silva[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco (IFPE)
54.080-000 – Jaboatão dos Guararapes – PE – Brasil

jhcaj@discente.ifpe.edu.br, wms11@discente.ifpe.edu.br

Abstract. *This study applied Machine Learning models to predict student academic performance, achieving significant improvements in certain models. Nevertheless, the absence of detailed information regarding previous datasets raises questions about direct comparisons. The findings underscore the relevance of AI in education and its potential contributions to students' academic success.*

Resumo. *Este estudo aplicou modelos de Machine Learning para prever o desempenho escolar dos alunos, com melhorias notáveis em alguns modelos. No entanto, a falta de informações detalhadas sobre bases de dados anteriores levanta dúvidas sobre comparações diretas. Os resultados ressaltam a relevância da IA na educação e suas potenciais contribuições para o sucesso acadêmico dos alunos.*

1. Introdução

A predição do desempenho escolar dos alunos é uma questão de grande relevância dentro do contexto da educação. A compreensão antecipada dos fatores que influenciam o sucesso dos alunos não apenas auxilia na identificação de estudantes em risco, como também proporciona a oportunidade de intervenções educacionais. Uma abordagem inovadora para verificar essa questão é a aplicação de algoritmos de *Machine Learning*, que têm a capacidade de analisar grandes volumes de dados e extrair padrões significativos.

A aplicação de técnicas de *Machine Learning* na predição do desempenho escolar dos alunos desempenha um papel central na intersecção entre a educação e a ciência da computação. Esse campo envolve a análise de dados educacionais, conhecido como *Education Data Mining (EDM). O EDM* oferece um valor fundamental às instituições de ensino e às entidades que estão envolvidas em diferentes processos de aprendizagem, [MANJARRES et al 2018].

O objetivo deste estudo é aplicar técnicas de *Machine Learning* para desenvolver modelos de previsão do desempenho escolar dos alunos em disciplinas de leitura, escrita e matemática. Buscaremos identificar os fatores que podem influenciar no sucesso escolar e avaliar a eficácia desses modelos na antecipação de dificuldades de aprendizagem, fornecendo dados valiosos para instituições de ensino.

2. Trabalhos Relacionados

Na literatura, existem alguns estudos que investigam a aplicação de Machine Learning na previsão do desempenho escolar dos alunos. Nesse contexto, destacam-se pesquisas

que abordam a influência de variáveis como raça, gênero, acesso a programas de alimentação escolar e nível de escolaridade dos pais.

Naicker et al. [Naicker et al. 2020] conduziram uma investigação abrangente sobre os fatores que influenciam o desempenho escolar dos alunos. Eles destacaram que questões relacionadas à raça, gênero e o programa de almoço escolar podem ser fatores significativos nesse contexto. O estudo constatou que a educação dos pais não demonstrou ter uma influência direta no desempenho acadêmico dos alunos. Esses resultados levantam questões importantes sobre as disparidades educacionais e apontam para a necessidade de intervenções específicas.

Já Chen, [Chen et al. s.d], realizaram um estudo que explorou a relação entre o gênero dos alunos e seu desempenho escolar. Suas descobertas sugerem que meninas tendem a demonstrar um nível de autodisciplina mais elevado em comparação com os meninos, o que pode influenciar positivamente seu sucesso acadêmico. Esse estudo também destaca a importância de uma possível monitorização mais atenta dos meninos na escola. Além disso, observou-se que o nível de escolaridade dos pais desempenha um papel significativo nas notas dos alunos, com notas correlacionadas com o grau de escolaridade dos pais.

3. Metodologia

Para realizar esse projeto, iremos usar o *Colab* do Google, que é um ambiente online onde serão feitas todas as análises e treinamentos da nossa ferramenta. A base de dados está disponível gratuitamente no site da "*Kaggle*" (https://www.kaggle.com/spscientist/students-performance-in-exams/activity). Não foi necessária nenhuma limpeza de dados.

A base de dados é dividida em oito colunas na seguinte ordem: gênero, raça/etnia, nível de escolaridade dos pais, acesso ao almoço, preparação de testes, pontuação em matemática, leitura e escrita. Para atingir nosso objetivo, foi preciso criar uma nova coluna que recebeu o nome de 'situacao', onde, dependendo da média das notas obtidas em leitura, escrita e matemática, os alunos serão divididos em 'reprovado[0]' e 'aprovado[1]'. Em seguida, foram criadas duas variáveis, x e y. A variável x recebeu tudo o que está contido no DataFrame, exceto as colunas 'mediasNotas' e 'situacao', pois a primeira foi criada para receber as médias das notas, e 'situacao' é nossa coluna alvo, que será atribuída à variável y. Usamos as funções *OneHotEncoder, OrdinalEncoder* e *make_column_transformer* para converter as informações das colunas *'gender', 'race/ethnicity', 'parental level of education', 'lunch' e 'test preparation course'*. O modelo foi treinado, colocando 20% como teste e os 80% como treinamento, utilizando alguns métodos de aprendizado de máquina.

3.1. *Support Vector Classifier*

O *SVC*, também conhecido como *Support Vector Machine (SVM)*, é utilizado para resolver problemas de classificação, que é o nosso caso. Assim como os métodos mencionados anteriormente, o *SVC* lida com problemas de decisão binária, como verdadeiro ou falso, sim ou não, aprovado ou reprovado, entre outros.

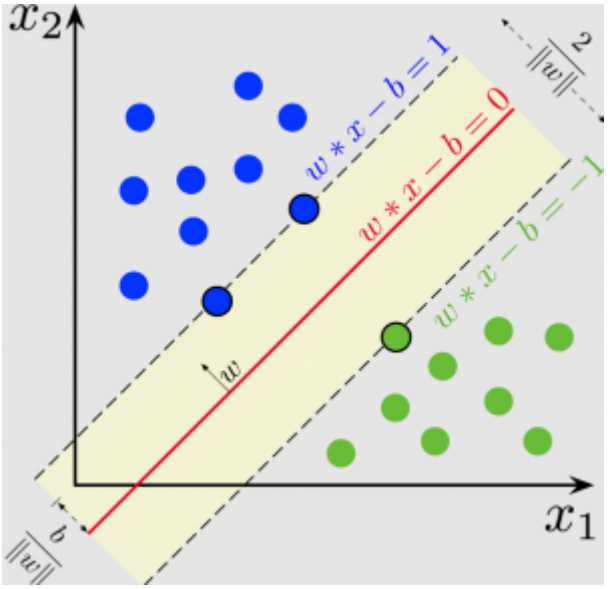

Figura 1. ***Support Vector Classifier.*** **Fonte:** Ajitesh Kumar

3.2. *CatBoost Classifier*

O *CatBoost* é um algoritmo de aprendizado de máquina projetado para resolver problemas de classificação, especialmente adequado para lidar com variáveis categóricas, o que se alinha com o nosso caso de determinar se um aluno será aprovado ou reprovado. O nome '*CatBoost*' é uma combinação das palavras 'categoria' (*Cat)* e 'aprimoramento' (*Boosting*). Esse é uma implementação de *Gradient Boosting*, que essencialmente consiste na combinação de várias árvores de decisão mais fracas para criar um modelo mais robusto.

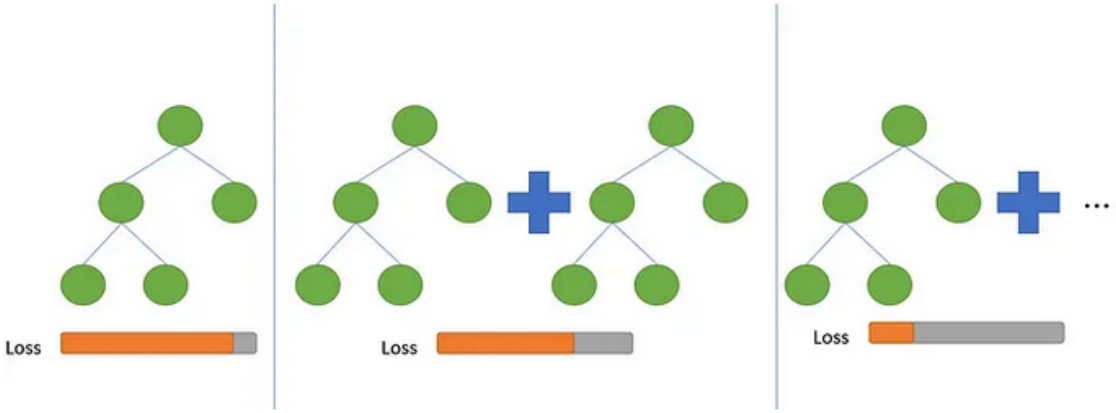

Figura 2. ***Gradient Boosting.*** **Fonte:** Medium

3.3. *Logistic Regression*

A *Logistic Regression* é um algoritmo de classificação binária simples, porém altamente eficaz, amplamente utilizado em problemas do mundo real devido à sua interpretabilidade e capacidade de fornecer probabilidades. É por essas razões que

escolhemos usá-lo em nosso teste.

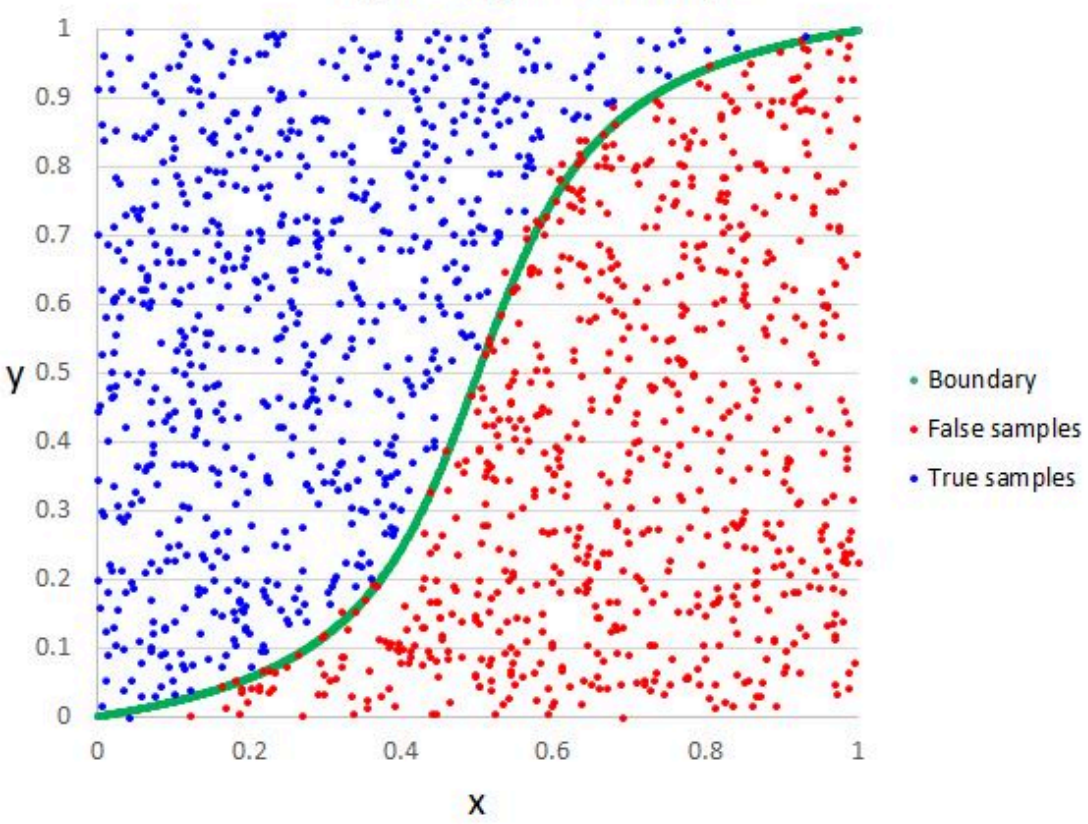

Figura 3. *Linear Regression* Fonte: AI Wiki

3.4. *Decision Tree Classifier*

Também optamos por utilizar a Árvore de Decisão em todos os três níveis. Este modelo de aprendizado de máquina é amplamente conhecido por sua interpretabilidade, flexibilidade e eficiência ao lidar com uma variedade de problemas. Uma de suas vantagens é a capacidade de controlar a profundidade da árvore, o que é importante para evitar o *overfitting*, algo que consideramos em nosso projeto.

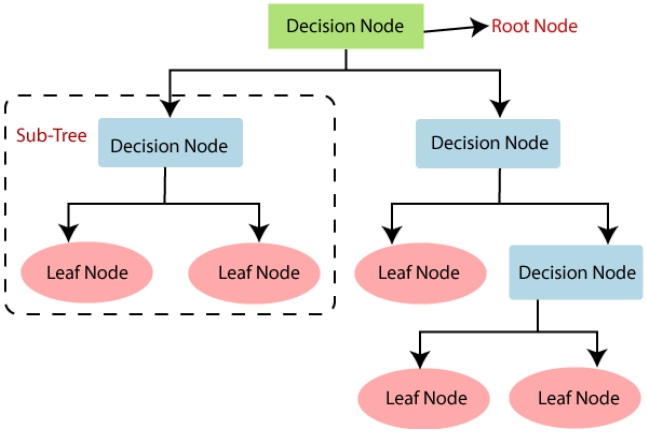

Figura 4. *Decision Tree Classifier.* Fonte: Java T Point

3.5. *Gaussian Naive Bayes*

O *GNB,* um algoritmo de classificação, baseia-se no Teorema de Bayes e faz a suposição de independência entre as características, embora essa suposição nem sempre seja verdadeira. Ele é conhecido por sua simplicidade e eficiência, especialmente quando a base de dados é pequena, o que também se aplica ao nosso caso.

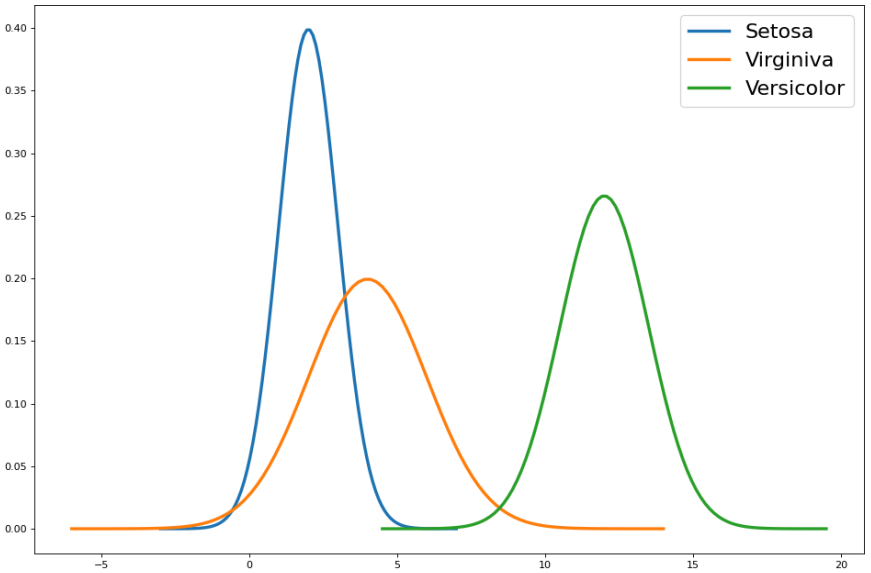

Figura 4. *Gaussian Naive Bayes.* Fonte: Eloquent Arduino

4. Resultados

Usamos um total de cinco tipos de modelos de aprendizagem de máquina neste projeto os resultados obtidos estão na tabela abaixo.

	Resultados							
	Regressão logística	Árvore de decisão fina	Árvore de Decisão Média	Árvore de Decisão Grosseira	SVC linear	SVM poly	Gaussian Naive Bayes	CatBoost
Artigos Pesquisados	89.4	83.7	82.4	75.2	90.1	85.7	68.4	94,45
Nosso Artigo	72,5	71	74,5	71	75,5	73,5	71,5	74

Tabela 1. Foram considerados dados dos artigos pesquisados para serem comparados com o nosso artigo

Como é possível observar, obtivemos ganhos significativos em apenas um dos modelos de aprendizagem de máquina, o '*Gaussian Naive Bayes*', no qual conseguimos aumentar a precisão em 3.1% em comparação com os resultados mencionados no artigo [Naicker et al. 2020]. Nosso melhor desempenho foi com o SVC linear, onde alcançamos uma acurácia de 75.5%, no entanto, em comparação com a acurácia observada no mesmo artigo, perdemos por cerca de 14.6%. Nenhum dos artigos pesquisados revelou os métodos utilizados para alcançar os resultados apresentados.

Ao analisar os artigos e ler as discussões no "*Kaggle*", notamos que a base de dados usada provavelmente não era composta de dados reais, e as pessoas podem ter adicionado informações extras dependendo da situação. Por essa razão, levantamos a hipótese de que nossos resultados não se aproximaram dos dos artigos devido à falta de acesso aos parâmetros provavelmente presentes na base usada. Isso fica evidente, por exemplo, no artigo [Fan et al. 2023], que apresenta um dos melhores resultados em nossa pesquisa, alcançando uma acurácia de 94.45%. Neste artigo, várias colunas adicionais foram incluídas, como razão, educação e ocupação dos pais, além de informações sobre a conexão à internet, entre outros. Por esse motivo, consideramos que qualquer comparação direta com bases de dados tão diferentes seria inviável.

5. Conclusões

Em resumo, este artigo explorou uma variedade de modelos de aprendizado de máquina aplicados a uma base de dados para prever o sucesso ou insucesso dos alunos em uma escola. Essas previsões têm o potencial de ser uma ferramenta valiosa nas mãos dos profissionais da educação. Ao reduzir as taxas de reprovação e melhorar a experiência educacional, essas análises podem tornar o conteúdo mais atrativo e eficaz para os alunos. Como resultado, os esforços para utilizar a inteligência artificial e o aprendizado de máquina no campo da educação podem desempenhar um papel fundamental na promoção do sucesso acadêmico e no aprimoramento do ensino.

Referências

[Chen et al. s.d] CHEN, B. et al. Analytics Groupwork: Analysing Students' Academic Performance on Free Lunch Programme. Disponível em: <https://rstudio-pubs-static.s3.amazonaws.com/697418_eeee5b9bb7bc45b2a19269b730825e17.html>.

[Fan et al. 2023] FAN, Z.; GOU, J.; WANG, C. Predicting secondary school student performance using a double particle swarm optimization-based categorical boosting model. Engineering Applications of Artificial Intelligence, v. 124, p. 106649–106649, 1 set. 2023.

[MANJARRES et al. 2018] MANJARRES, A. V.; SANDOVAL, L. G. M.; SUÁREZ, M. S. Data mining techniques applied in educational environments: Literature Review. **Digital Education Review**, n. 33, p. 235–266, 2018.

[Naicker et al. 2020] NAICKER, N.; ADELIYI, T.; WING, J. Linear Support Vector

Machines for Prediction of Student Performance in School-Based Education. Mathematical Problems in Engineering, v. 2020, p. 1–7, 1 out. 2020.