Predição de Desempenho em Língua Portuguesa de Estudantes do Ensino Fundamental: Um Estudo de Caso em Sergipe

Mellany Linhares Santana, UFS, mellany.linhares@dcomp.ufs.br
ORCID ID: 0009-0009-3223-8148
Renê Pereira de Gusmão, UFAPE, rene.gusmao@ufape.edu.br
ORCID ID: 0000-0002-4806-6506
Cleonides Silva Dias Gusmão, UFPB, cleonides.silva@academico.ufpb.br
ORCID ID: 0000-0002-6094-5642

Resumo: Este estudo teve como objetivo a predição do desempenho de estudantes em língua portuguesa e realizar análises para identificação dos fatores de maior relevância por meio de técnicas de aprendizagem de máquina. Foram analisados dados de estudantes sergipanos do 9º ano do ensino fundamental, obtidos através das bases do Sistema de Avaliação da Educação Básica e do Censo Escolar, ambas do ano de 2019. Foram analisadas informações referentes às condições socioeconômicas dos alunos e situação estrutural das escolas. Observou-se que os modelos construídos utilizando a técnica de floresta aleatória se destacaram, e as variáveis mais relevantes para predição foram as relacionadas ao nível socioeconômico do estudante e à sua estrutura familiar.

Palavras-chave: Predição de Desempenho, Ensino Fundamental, Aprendizagem de Máquina, Censo Escolar, Língua Portuguesa.

Performance Prediction in Portuguese of Elementary School Students: A Case Study in Sergipe

Abstract: This study aimed to predict the performance of students in Portuguese and to carry out analyzes to identify the most relevant factors through machine learning techniques. Data from Sergipe students in the 9th grade of elementary school were analyzed, obtained through the bases of the Basic Education Evaluation System and the School Census, both from the year 2019. Information regarding the socioeconomic conditions of students and the structural situation of schools were analyzed. It was observed that the models built using the random forest technique stood out, and the most relevant variables for prediction were those related to the student's socioeconomic level and family structure.

Keywords: Performance Prediction, Elementary Education, Machine Learning, School Census, Portuguese Language.

1. Introdução

A educação desempenha um papel fundamental no desenvolvimento de um indivíduo, permitindo o seu crescimento e formação como cidadão. Como mostrado em Ayienda, Rimiru e Cheruiyot (2021), os estudantes encontram diferentes fatores que podem afetar seu desempenho acadêmico de maneira negativa, o que resulta em dificuldades e problemas em sua formação acadêmica e pode levar a problemas em suas relações interpessoais, tanto em casa quanto no âmbito escolar. Assim, é necessário analisar e encontrar os fatores que podem gerar tais complicações no desempenho acadêmico dos estudantes, e buscar formas de identificá-los antes que comprometam a vida acadêmica do indivíduo.

Desse modo, ao analisar o contexto do sistema educacional, é possível observar que ele pode ser dividido em educação básica e educação superior. Nesse paradigma, como pode ser visto em Harvey e Kumar (2019), o foco das pesquisas de desempenho

acadêmico é na educação superior. Tendo em vista que a educação básica, mais especificamente o ensino fundamental, é a base do conhecimento do estudante, a falta de trabalhos contemplando as dificuldades relacionadas ao desempenho acadêmico nessa etapa, é algo preocupante e que necessita de mais atenção.

Com o intuito de realizar um estudo sobre a predição de indicadores que possam prejudicar o desempenho acadêmico de estudantes do ensino fundamental, é preciso obter dados sobre esses estudantes, e após obtenção dos dados, utilizar estratégias para obter as informações desejadas. Para tal, como mostrado em Harvey e Kumar (2019), uma solução proeminente é a utilização de técnicas de aprendizagem de máquina supervisionada e mineração de dados, visto que se faz necessário o processo de identificação e classificação de variáveis e padrões de interesse em uma base de dados. Alguns trabalhos que utilizam técnicas semelhantes para investigação de problemas educacionais são (Pradeep; Das e Kizhekkethottam, 2015), (Wandera; Marivate e Sengeh, 2019) e (Qasrawi *et al.*, 2020).

Assim, este trabalho pretende identificar os indicadores e variáveis que interferem no desempenho acadêmico de estudantes do ensino fundamental, utilizando de técnicas de aprendizagem de máquina. Desta forma, será possível realizar a predição de desempenho acadêmico e identificar se a situação socioeconômica do estudante ou a estrutura física das instituições de ensino afetam o desempenho.

2. Trabalhos Relacionados

Kiu (2018) visou identificar e analisar o impacto do perfil socioeconômico, atividades sociais e resultados acadêmicos de estudantes do ensino médio no seu desempenho escolar e realizar a predição do desempenho acadêmica na disciplina de matemática. Foram criados dois modelos, um para classificar apenas em aprovado/reprovado e outro para classificação da nota em diferentes grupos. O algoritmo de ”árvore de decisão”apresentou maior desempenho. Além disso, foi observado que alguns fatores possuem maior relevância para a predição de desempenho, entre eles: reprovações anteriores, suporte educacional extraclasse, nível da saúde.

Ram *et al.* (2021) propõem utilizar algoritmos de aprendizagem de máquina para predizer a nota final de estudantes do ensino médio e comparar a desempenho dos algoritmos utilizados. Dos algoritmos utilizados, o que demonstrou as melhores métricas foi o algoritmo de floresta aleatória. Dos fatores presentes no *dataset*, foi observado que os de maior influência são: as primeiras notas da disciplina e o número de faltas.

Roy e Garg (2017) planejaram identificar os fatores que podem influenciar no desempenho acadêmico de estudantes do ensino médio. Foi observado que os fatores de maior influência para a desempenho do estudante na avaliação final é o seu desempenho das duas primeiras avaliações; porém, ao remover essas duas variáveis, os fatores que prevalecem são: consumo de bebida alcoólica em dias úteis, saúde, relacionamentos românticos e educação dos pais. Ademais, o algoritmo com maior número de instâncias classificadas corretamente foi o classificador J48, com 73.92%.

Kumar *et al.* (2021) implementaram diferentes técnicas de mineração de dados para a predição do desempenho acadêmico de estudantes do ensino médio. Inicialmente, o *dataset* foi testado utilizando todos os fatores presentes e, após isso, utilizando dez fatores de maior influência, identificados através de algoritmos de seleção de fatores. Em ambos os testes o algoritmo de tabela de decisão apresentou maior acurácia, porém após a aplicação da seleção de fatores, todos os algoritmos apresentaram maior acurácia do que na análise anterior.

Stearns *et al.* (2017) visam a utilização de técnicas de aprendizagem de máquina para predizer o desempenho de estudantes na avaliação de matemática do ENEM. Os

algoritmos utilizados foram *gradient boosting* e AdaBoost, sendo que o primeiro foi identificado como o mais eficiente. Uma análise semelhante foi conduzida por (Roslan e Chen, 2022), que propõem a identificação de fatores que afetam o desempenho dos estudantes nas disciplinas de inglês e matemática no MCE. Utilizando dados de desempenhos acadêmicos anteriores dos estudantes e dados demográficos e psicológicos. Analisando cada disciplina separadamente, para a disciplina de matemática os algoritmos árvore de decisão e *Naive Bayes* obtiveram acurácia semelhante, 83.9%, porém o algoritmo *Naive Bayes* superou o primeiro nas outras métricas utilizadas (área abaixo da curva, precisão e f1-score). Para a disciplina de inglês, o algoritmo com maior acurácia foi árvore de decisão, com 87.1%.

3. Metodologia

A metodologia usada neste trabalho baseia-se no modelo de processo CRISP-DM. A metodologia CRISP-DM possui seis fases, a saber: compreensão de domínio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação. As etapas do modelo e como elas foram desenvolvidas serão descritas a seguir.

3.1. Compreensão de Domínio

Esta etapa tem como finalidade definir e entender os objetivos do trabalho. Com esse fim, as seguintes questões norteadoras foram definidas:

- Quais as técnicas de aprendizagem de máquina que apresentam as melhores métricas de avaliação na predição do desempenho acadêmico de estudantes do ensino fundamental?
- Como a situação socioeconômica dos estudantes afeta o desempenho acadêmico?
- Como a situação estrutural das escolas afeta o desempenho acadêmico?

3.2. Compreensão dos Dados

Para esse trabalho serão utilizadas as bases de dados do Sistema de Avaliação da Educação Básica (SAEB) e os microdados do Censo Escolar[*], ambos do ano de 2019. Os microdados do SAEB consistem em um conjunto de dados extraídos de testes e questionários, aplicados pelo Inep a cada dois anos na rede pública e em uma amostra da rede privada de educação (Inep, 2022).

Na base de dados do SAEB há informações sobre o desempenho acadêmico de estudantes do 2º ano, 5º ano e 9º ano do ensino fundamental, e de estudantes do 3º ano do ensino médio, nas disciplinas de língua portuguesa e matemática, além de informações socioeconômicas. Além disso, a base do SAEB possui dados gerais sobre as escolas e informações extraídas de questionários aplicados a professores, diretores e secretários municipais de educação. Já na base de dados do Censo Escolar, são armazenadas informações gerais sobre as instituições de ensino da educação básica.

Para o presente trabalho, serão considerados os dados referentes às turmas do 9º ano do ensino fundamental e informações sobre as escolas advindas do Censo Escolar.

3.3. Preparação dos Dados

Os dados originais foram filtrados para incluir informações referentes ao estado de Sergipe. Após isso, foram formadas três bases de dados: a primeira utilizando apenas os dados referentes aos alunos, a segunda com os dados referentes às escolas e, a terceira com a integração das duas bases anteriores. Os dados dos alunos foram retirados do questionário socioeconômico da base de dados do Saeb, e os dados das escolas foram retirados da base de dados do Censo Escolar.

[*] ⟨https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados⟩

Além disso, não foi possível recuperar o histórico de desempenho escolar dos estudantes para utilização na predição dos dados, já que, para todo ano em que o Saeb é aplicado, a identificação única de cada estudante é alterada, de forma que não é possível realizar a integração das bases de dados de anos diferentes. Portanto foram adicionadas, em cada base, duas novas variáveis, com o desempenho médio dos estudantes do 5º ano do ensino fundamental do ano de 2015 nas disciplinas de língua portuguesa e matemática. É importante ressaltar que esses valores foram calculados por escola, de forma que a média vinculada ao estudante pertence à escola que ele frequenta.

Após isso, foi realizada uma discretização nas notas do Saeb de cada base de dados, criando duas novas variáveis: "Desempenho LP"(representando o desempenho do estudante na disciplina de língua portuguesa) e "Desempenho MT"(representando o desempenho do estudante na disciplina de matemática). No processo de discretização, conforme descrito por (Han; Kamber e Pei, 2012), os valores numéricos das notas do estudante foram substituídos por dois rótulos conceituais, sendo estes: desempenho abaixo da média e desempenho acima da média ou na média.

Na próxima etapa, foram removidas todas as variáveis que possuíam o mesmo valor para todos os estudantes e foram considerados os estudantes que tinham respondido, pelo menos, 70% do questionário socioeconômico do Saeb. Os valores nulos foram substituídos pela moda da variável, em conformidade com o sugerido por (Han; Kamber e Pei, 2012) como técnica de preenchimento de valores faltantes. Após a aplicação dessas etapas, remanesceram 7602 estudantes para a análise.

Além disso, para as bases dos estudantes e para as bases das escolas, foi aplicada a técnica de seleção de fatores para a seleção das variáveis de maior relevância para o desempenho acadêmico das disciplinas analisadas. Como pontuado por Müller e Guido (2016), a utilização das variáveis relevantes torna o modelo mais simples, o que facilita nas análises posteriores, e também contribui na desempenho dos algoritmos de predição.

O método de seleção de fatores aplicado foi o *SelectKBest* pertencente à biblioteca *Scikit-learn*. Para as bases dos estudantes foram consideradas 15 variáveis e para as bases das escolas foram consideradas 25 variáveis. Para a base com ambos os dados, foi realizada uma integração entre as duas bases citadas anteriormente. O número de variáveis de cada tipo de base é definido na 1.

Tabela 1. Quantidade de variáveis consideradas por tipo de base de dados

Base de dados	**Quantidade de variáveis**	**Quantidade de variáveis após a seleção de fatores**
Estudantes	51	15
Escolas	147	25
Integração	196	38

Por fim, foram geradas seis bases de dados: três utilizando todas as variáveis presentes nas bases antes da aplicação da seleção de fatores e três bases utilizando as variáveis retornadas pela seleção de fatores.

3.4. Modelagem

Para o desenvolvimento do trabalho, foi utilizada a linguagem de programação *Python* e as bibliotecas *Pandas* e *Scikit-learn*. Além disso, o ambiente de desenvolvimento escolhido foi a plataforma *Jupyter-notebook*.As técnicas de aprendizagem de máquina supervisionada escolhidas para o desenvolvimento dos experimentos foram: *Naive Bayes*, SVM, Floresta Aleatória, KNN e Árvore de Decisão, por serem as mais utilizadas nos

trabalhos relevantes. Além disso, devido à natureza dos dados, serão utilizados algoritmos de classificação.

É importante ressaltar que os passos descritos à seguir foram repetidos para cada uma das bases formadas na etapa de preparação dos dados. A base de dados original foi dividida aleatoriamente em um conjunto de treinamento (composto por 70% dos dados) e um conjunto de teste (composto por 30% dos dados). Quando necessário, foi aplicada a técnica de *oversample* (duplicar dados da classe minoritária) para combater o desbalanceamento no conjunto de treinamento. Além disso, visando a otimização de cada modelo, foi aplicado o ajuste de hiperparâmetros, que são parâmetros que permitem controlar o processo de treinamento dos modelos (Microsoft, 2023). Os modelos de aprendizagem de máquina foram construídos utilizando o conjunto de treinamento e, por fim, os modelos desenvolvidos foram aplicados no conjunto de teste para avaliação.

3.5. Avaliação

Por último, os resultados encontrados na etapa anterior serão avaliados. Para tal, utilizamos uma matriz de confusão, aplicando acurácia, precisão, *recall* e *f-score* como métricas de avaliação de desempenho, expostas na seção seguinte.

4. Resultados

Neste seção serão apresentados os resultados obtidos após a aplicação das técnicas de aprendizagem de máquina supervisionada para as seis bases geradas. Para facilitar na visualização das tabelas contendo os resultados alcançados, as melhores métricas atingidas serão destacadas em negrito. Os resultados foram listados na Tabela 2, que contém os resultados das bases antes da seleção de fatores, e na Tabela 3, que contém os resultados após a aplicação da seleção de fatores.

Tabela 2. Métricas sem a seleção de fatores

2*Técnica	Estudantes				Escolas				Integração			
	Acurácia	**Precisão**	***Recall***	***F-score***	**Acurácia**	**Precisão**	***Recall***	***F-score***	**Acurácia**	**Precisão**	***Recall***	***F-score***
Floresta aleatória	**82,38%**	**82,55%**	**99,73%**	**90,33%**	54,80%	89,01%	51,62%	65,34%	82,38%	82,72%	99,42%	90,30%
Árvore de decisão	72,95%	83,49%	83,80%	83,65%	50,33%	87,06%	46,79%	60,86%	72,78%	83,67%	83,27%	83,47%
Naive Bayes	73,04%	88,38%	77,54%	82,60%	27,22%	90,55%	13,22%	23,08%	27,58%	90,81%	13,65%	23,73%
KNN	70,36%	84,54%	78,44%	81,38%	54,84%	87,45%	52,89%	65,92%	68,39%	85,25%	74,61%	79,58%
SVM	76,68%	85,13%	86,94%	86,02%	54,45%	89,00%	51,14%	64,96%	75,54%	86,74%	83,06%	84,86%

Tabela 3. Métricas com a seleção de fatores

2*Técnica	Estudantes				Escolas				Integração			
	Acurácia	**Precisão**	***Recall***	***F-score***	**Acurácia**	**Precisão**	***Recall***	***F-score***	**Acurácia**	**Precisão**	***Recall***	***F-score***
Floresta aleatória	76,63%	83,99%	88,58%	86,22%	55,98%	88,59%	53,58%	66,78%	**80,93%**	**83,09%**	**96,55%**	**89,31%**
Árvore de decisão	71,24%	83,47%	81,25%	82,35%	55,50%	88,68%	52,84%	66,22%	74,22%	84,57%	84,12%	84,35%
Naive Bayes	69,97%	89,36%	72,23%	79,88%	19,68%	88,06%	3,13%	6,05%	19,60%	87,69%	3,03%	5,85%
KNN	73,17%	84,63%	82,47%	83,54%	49,19%	85,98%	45,94%	59,88%	72,51%	84,89%	81,15%	82,98%
SVM	68,35%	85,77%	73,92%	79,41%	57,39%	87,93%	56,09%	68,48%	71,33%	86,77%	77,00%	81,60%

É possível observar que a base de dados integrada, após a aplicação da seleção de fatores, obteve os melhores resultados, enquanto as bases de dados contendo apenas dados referentes às escolas não conseguiram atingir resultados satisfatórios na predição de desempenho acadêmico da disciplina.

Além disso, observa-se que a floresta aleatória foi a técnica de aprendizagem de máquina que apresentou as melhores métricas de avaliação, tendo obtido, para as bases de dados integradas, acurácias de 82,38% antes da seleção de fatores e 80,93% após a aplicação da seleção de fatores. Para as bases de dados contendo apenas os dados dos estudantes, a técnica de floresta aleatória apresentou uma acurácia de 82,38% antes da seleção de fatores e 76,63% após a seleção de fatores. Por fim, para as

bases de dados referente aos dados das escolas, a técnica de floresta aleatória retornou acurácias de 54,80% antes da seleção de fatores e 55,98% após a seleção de fatores. As variáveis relevantes referentes aos estudantes utilizadas, junto às suas descrições, podem ser encontradas na Tabela 4.

4.1. Discussão

As variáveis pertencentes à base de dados dos estudantes são as mais relevantes para a predição do desempenho escolar. Dentre elas, destacam-se: o nível socioeconômico da família, a estrutura familiar, as atividades realizadas no tempo livre do estudante e histórico acadêmico. Outrossim, as varíaveis referentes a infraestrutura das escolas, como a localização, tratamento do esgoto e lixo, dependências físicas, presença de equipamentos eletrônicos e acessibilidade, também afetam o desempenho do estudante, embora em menor escala.

Foi observado que, para a disciplina de língua portuguesa, o auxílio nas atividades domésticas tem influência no sucesso acadêmico do estudante. Como evidenciado por (Alberto *et al.*, 2009), a perda da infância tende a ser comum nas classes pobres. As crianças, muitas vezes, necessitam auxiliar o núcleo familiar, sendo uma dessas formas de auxílio o trabalho doméstico, especialmente para membros mais jovens da família. A conciliação das atividades escolares com o trabalho doméstico é de difícil realização, e pode influenciar no desempenho escolar da criança, muitas vezes levando à evasão escolar e repetência.

O mesmo ocorre com trabalhos informais, também uma das variáveis relevantes para a disciplina. As necessidades financeiras influenciam na entrada precoce da criança no mercado de trabalho (Alberto *et al.*, 2009). Como apontado no estudo, essas atividades exigem tempo e esforço demasiado do estudante, o que pode levar à precarização das atividades acadêmicas e, portanto, a um baixo desempenho escolar.

Ademais, é importante mencionar que o tempo destinado pelo estudante ao lazer é relacionado com o seu desempenho. Foi possível perceber que estudantes com baixo desempenho acadêmico destinavam menos de duas horas do seu dia para atividades de lazer. Como descorrido por (Alberto *et al.*, 2009), o pouco tempo destinado para atividades recreativas pode estar ligado à necessidade da criança de destinar todo o seu tempo fora do espaço escolar para atividades domésticas e trabalhos informais, a fim de auxiliar seu núcleo familiar.

Por sua vez, o incentivo familiar foi listado como um fator de influência no desempenho escolar do estudante. Tal influência foi reiterada por (Santos *et al.*, 2019), que afirmam que, embora o aprendizado formal inicie no ambiente escolar, o estímulo da família nas atividades acadêmicas é de extrema importância para o sucesso acadêmico do estudante. A falta do acompanhamento familiar nas atividades escolares pode acarretar em uma dificuldade no aprendizado da criança.

Ainda sobre o núcleo familiar do estudante, a estrutura familiar é um fator que foi identificado como relevante para o desempenho acadêmico; a presença da figura materna foi identificada como variável relevante. Em um estudo elaborado por (Martins e Teixeira, 2021), notou-se que estudantes pertencentes à famílias biparentais apresentam melhor desempenho acadêmico. Isso pode ser explicado pelo fato de que, tendo mais de um responsável presente no núcleo famíliar, a disponibilidade de tempo dos pais para auxiliar nas atividades acadêmicas é maior. Além disso, no mesmo estudo, ao comparar as famílias monoparentais, observou-se que crianças pertencentes à famílias monoparentais maternas possuem melhor desempenho escolar do que crianças pertencentes à famílias apenas com a presença do pai, o que pode significar que a escolaridade da figura materna

Tabela 4. Variáveis relevantes dos estudantes para a disciplina de língua portuguesa

Nome	**Descrição**	**Respostas**
TX_RESP_Q002	Qual é a sua cor ou raça?	Branca, Preta, Parda, Amarela, Indígena, Não quero declarar.
TX_RESP_Q003A	Normalmente, quem mora na sua casa? - Mãe (mães ou madrasta).	Não, Sim.
TX_RESP_Q006C	Com que frequência seus pais ou responsáveis costumam: - Incentivar você a fazer a tarefa de casa.	Nunca ou quase nunca, De vez em quando, Sempre ou quase sempre.
TX_RESP_Q007	Com que frequência sua família paga alguém para auxiliar nos trabalhos domésticos (faxina ou limpeza)?	Nunca ou quase nunca, De vez em quando (uma vez por semana, a cada quinze dias etc.), Sempre ou quase sempre (ex.: três ou mais dias por semana).
TX_RESP_Q008B	Na região que você mora tem: - Água tratada da rua.	Não, Sim.
TX_RESP_Q008C	Na região que você mora tem: - Iluminação na rua.	Não, Sim.
TX_RESP_Q010E	Na sua casa tem: - Garagem.	Não, Sim.
TX_RESP_Q010I	Na sua casa tem: - Freezer (independente ou segunda porta da geladeira).	Não, Sim.
TX_RESP_Q012	Considerando a maior distância percorrida, normalmente de que forma você chega à sua escola?	À pé, De ônibus urbano, De transporte escolar, De barco, De bicicleta, De carro, Outros meios de transporte.
TX_RESP_Q015	Você já foi reprovado?	Não, Sim, uma vez, Sim, duas vezes ou mais.
TX_RESP_Q017A	Fora da escola em dias de aula, quanto tempo você usa para: - Lazer (TV, internet, jogar bola, música etc.).	Não uso meu tempo para isso, Menos de 1 hora, Entre 1 e 2 horas, Mais de 2 horas.
TX_RESP_Q017C	Fora da escola em dias de aula, quanto tempo você usa para: - Fazer trabalhos domésticos (lavar louça, limpar quintal, cuidar dos irmãos).	Não uso meu tempo para isso, Menos de 1 hora, Entre 1 e 2 horas, Mais de 2 horas.
TX_RESP_Q017E	Fora da escola em dias de aula, quanto tempo você usa para: - Trabalhar fora de casa (recebendo ou não um salário).	Não uso meu tempo para isso, Menos de 1 hora, Entre 1 e 2 horas, Mais de 2 horas.
desempenho_lp em 2015	Desempenho dos estudantes do quinto ano em língua portuguesa da escola do estudante em 2015.	Acima da média/na média, Abaixo da média.
desempenho_mt em 2015	Desempenho dos estudantes do quinto ano em matemática da escola do estudante em 2015.	Acima da média/na média, Abaixo da média.

tem maior influência no desempenho acadêmico da criança.

Notou-se, através das variáveis relevantes, que o nível socioeconômico do estudante tem forte influência em seu sucesso acadêmico. Também foi observado que o acesso a água tratada, iluminação na rua onde vive, presença de computador, rede *Wi-Fi* e de eletrodomésticos são relevantes para o desempenho escolar da criança. Mais detalhadamente, foi possível observar que uma parcela significativa dos estudantes que obtiveram desempenho abaixo da média em matemática não tinham acesso à rede *Wi-Fi* no lar.

Outrossim, ao analisar o histórico acadêmico do estudante, é possível notar que a maioria dos estudantes que registraram desempenho escolar abaixo da média reprovaram ao menos uma vez. (Martins e Teixeira, 2021) observaram resultados semelhantes, e destacaram que a repetência afeta negativamente o desempenho acadêmico do estudante.

Como citado anteriormente, não é possível prever o desempenho acadêmico do estudante utilizando apenas as variáveis referentes à escola que ele frequenta. Porém, elas podem ser utilizadas para o complemento das análises realizadas. Foi identificado que uma grande parcela de estudantes em insucesso acadêmico estudavam em colégios em que não há acesso à rede de esgoto pública. Outrossim, a maioria dos estudantes, independente do desempenho escolar, frequentam escolas onde não é feito tratamento de lixo. Outrossim, foi possível notar que mais da metade dos estudantes, independente do desempenho escolar, estudam em colégios onde não há acesso a internet e, como dito anteriormente. Portanto, a falta de acesso ao acervo de uma biblioteca e à internet impacta diretamente nos estudos extraclasse dessas crianças.

Outro fator relevante para o desempenho acadêmico é a estrutura focada para lazer e socialização entre estudantes. Observou-se que áreas destinadas à atividades recreativas, como área verde, piscina e quadra de esportes, por exemplo, tem relevância no desempenho escolar. Resultados similares foram encontrados por (Martins e Teixeira, 2021), que observaram que a presença dessas áreas tem uma correlação positiva com o desempenho acadêmico da criança. Ademais, (Barreto; Rocha e Ferreira, 2017) analisaram a importância do ambiente escolar ser acolhedor para o estudante, com uma estrutura que permita o lazer e convívio amigável, para minar o abandono do ambiente acadêmico, especialmente de estudantes em vulnerabilidade socioeconômica.

Por fim, ao comparar os resultados encontrados com os estudos relevantes, pode-se observar que os resultados encontrados analisando as variáveis referentes às escolas são semelhantes ao estudo realizado por (Wandera; Marivate e Sengeh, 2019), que analisaram a importância da estrutura escolar para o desempenho escolar dos seus estudantes, tendo observado que estudantes de colégios com infraestrutura precária possuem maiores chances de insucesso acadêmico. Ademais, as análises utilizando as variáveis referentes aos estudantes reiteraram os resultados encontrados por (Qasrawi *et al.*, 2020), que identificaram fatores como suporte familiar e status econômico como sendo de grande influência no desempenho dos estudantes.

5. Conclusão

O objetivo do presente trabalho foi a aplicação de técnicas de aprendizagem de máquina supervisionada para a predição do desempenho escolar de estudantes do ensino fundamental, e analisar e identificar possíveis fatores que o afetem. Para isso, foi selecionada a base de dados do Saeb, em conjunto com a base de dados do Censo Escolar, ambas do ano de 2019. Foram utilizados dados de estudantes do 9º ano do ensino fundamental focando na disciplinas de língua portuguesa, filtrando pelo estado de Sergipe. Por fim, para a criação das bases de dados para os experimentos posteriores,

foram aplicadas técnicas de mineração de dados, como limpeza dos dados, criação de novas variáveis, balanceamento das bases de dados e seleção de fatores.

Após a aplicação das técnicas de aprendizagem de máquina selecionadas para língua portuguesa, foi possível observar que a situação socioeconômica do estudante e a sua estrutura familiar possuem grande relevância no desempenho acadêmico da criança. Ademais, percebeu-se que, embora a situação estrutural da escola do estudante não seja fator determinante para o seu desempenho escolar, ela pode ser utilizada como complemento para as análises.

Dos modelos de aprendizagem de máquina construídos, os modelos feitos utilizando a técnica de floresta aleatória apresentaram as melhores métricas de avaliação. Os modelos desenvolvidos utilizando apenas os dados referentes às escolas não apresentaram resultados satisfatórios, enquanto os modelos construídos utilizando apenas os dados dos estudantes retornaram boas métricas de avaliação. Porém, observou-se que a melhor desempenho foi alcançada nas bases de dados que integraram os dois tipos de informação.

Por fim, conclui-se que o trabalho foi bem sucedido, pois as questões norteadoras definidas previamente neste estudo foram respondidas e os objetivos do trabalho foram alcançados. Outrossim, o presente estudo demonstrou a utilidade e importância dos microdados do Inep, ao qual pertencem as duas bases de dados utilizadas no trabalho, para a realização de estudos na área de mineração de dados educacionais. Todavia, é importante salientar que, devido à natureza da base de dados, não foi possível recuperar o desempenho do estudante nos anos anteriores do Saeb e, portanto, não se conseguiu montar o seu histórico acadêmico, o que limitou as análises realizadas.

Para estudos posteriores, recomenda-se a inclusão de outros tipos de dados presentes na base do Saeb, como as informações referentes à professores e outros profissionais da área da educação. Além disso, sugere-se o aumento do escopo da base de dados para mais estados brasileiros e para outros anos do ensino fundamental, para análises mais robustas. Ademais, recomenda-se a utilização de outras técnicas de aprendizagem de máquina para fins de comparação com o presente trabalho.

Referências

Alberto, M. d. F. P. *et al.* Trabalho infantil doméstico: perfil bio-sócio-econômico e configuração da atividade no município de joão pessoa, pb. **Cadernos de Psicologia Social do Trabalho**, v. 12, n. 1, p. 57–73, 2009.

Ayienda, R.; Rimiru, R.; Cheruiyot, W. Predicting students academic performance using a hybrid of machine learning algorithms. In: . [s.n.], 2021. v. 2021-September. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118445388&doi=10.1109%2fAFRICON51333.2021.9571012&partnerID=40&md5=3eb04ad449112c8d7867877030140016⟩.

Barreto, C.; Rocha, S. V.; Ferreira, M. d. F. d. A. A influência da estrutura escolar nos casos de violências e desempenho do aluno na escola. **Colóquio do Museu Pedagógico-ISSN 2175-5493**, v. 12, n. 1, p. 326–331, 2017.

Han, J.; Kamber, M.; Pei, J. **Data mining: concepts and techniques**. [S.l.]: Morgan Kaufmann, 2012.

Harvey, J. L.; Kumar, S. A. P. A practical model for educators to predict student performance in k-12 education using machine learning. In: **2019 IEEE Symposium Series on Computational Intelligence (SSCI)**. [S.l.: s.n.], 2019. p. 3004–3011.

Inep. 2022. Disponível em: ⟨https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb⟩.

Kiu, C.-C. Data mining analysis on student's academic performance through exploration of student's background and social activities. In: **2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA)**. [S.l.: s.n.], 2018. p. 1–5.

Kumar, M.; Nidhi, N.; Majithia, S.; Sharma, N. Predictive model for students' academic performance using classification and feature selection techniques. In: **2021 2nd International Conference on Computational Methods in Science Technology (ICCMST)**. [S.l.: s.n.], 2021. p. 106–111.

Martins, J. C.; Teixeira, E. C. As estruturas familiares afetam o desempenho escolar no brasil? **Revista Econômica do Nordeste**, v. 52, n. 1, p. 65–76, 2021.

Microsoft. 2023. Disponível em: ⟨https://learn.microsoft.com/pt-br/azure/machine-learning/how-to-tune-hyperparameters⟩.

Müller, A. C.; Guido, S. **Introduction to Machine Learning with Python**. [S.l.]: O'Reilly Media, Inc., 2016.

Pradeep, A.; Das, S.; Kizhekkethottam, J. J. Students dropout factor prediction using edm techniques. In: **2015 International Conference on Soft-Computing and Networks Security (ICSNS)**. [S.l.: s.n.], 2015. p. 1–7.

Qasrawi, R.; Abdeen, Z.; Taweel, H.; Younis, M. A.; Al-Halawa, D. A. Data mining techniques in identifying factors associated with schoolchildren science and arts academic achievement. In: **2020 International Conference on Promising Electronic Technologies (ICPET)**. [S.l.: s.n.], 2020. p. 78–83.

Ram, M. S.; Srija, V.; Bhargav, V.; Madhavi, A.; Kumar, G. S. Machine learning based student academic performance prediction. In: **2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)**. [S.l.: s.n.], 2021. p. 683–688.

Roslan, M. H. B.; Chen, C. J. Predicting students' performance in english and mathematics using data mining techniques. **EDUCATION AND INFORMATION TECHNOLOGIES**, 2022. ISSN 1360-2357.

Roy, S.; Garg, A. Predicting academic performance of student using classification techniques. In: **2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)**. [S.l.: s.n.], 2017. p. 568–572.

Santos, I. de A.; Barbosa, V. M. da S.; Silva, S. S.; Ribeiro, G. W. de A. Educação infantil: A influência da família no desempenho escolar dos alunos. 2019.

Stearns, B.; Rangel, F.; Rangel, F.; Faria, F. D.; Oliveira, J. Scholar performance prediction using boosted regression trees techniques. In: . [s.n.], 2017. p. 329–334. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063550732&partnerID=40&md5=84619531eef0d1257efd2c7a42ccdf29⟩.

Wandera, H.; Marivate, V.; Sengeh, M. D. Predicting national school performance for policy making in south africa. In: **2019 6th International Conference on Soft Computing Machine Intelligence (ISCMI)**. [S.l.: s.n.], 2019. p. 23–28.