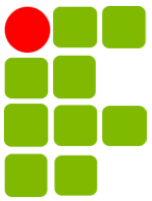


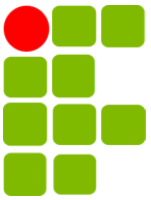
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Campus Campina Grande
Coordenação do Curso Superior de Engenharia de Computação

Análise e predição de evasão dos alunos usando Aprendizado de Máquina

Alexandre dos Santos Oliveira

Orientador: Prof. Danyllo Wagner Albuquerque, Msc.

Campina Grande, Fevereiro de 2024



Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Campus Campina Grande
Coordenação do Curso Superior de Engenharia de Computação

Análise e predição de evasão dos alunos usando Aprendizado de Máquina

Alexandre dos Santos Oliveira

Trabalho de Conclusão de Curso apresentado ao Curso Engenharia de Computação, do Instituto Federal da Paraíba – Campus Campina Grande, em cumprimento às exigências parciais para a obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Prof. Danyllo Wagner Albuquerque, Msc.

Campina Grande, Fevereiro de 2024

O48a

Oliveira, Alexandre dos Santos

Análise e predição de evasão dos alunos usando Aprendizado de Máquina / Alexandre dos Santos Oliveira. - Campina Grande, 2024.

19f. : il.

Trabalho de Conclusão de Curso (Curso Superior em Engenharia de Computação.) - Instituto Federal da Paraíba, 2023.

Orientador: Prof. Msc. Danyllo Wagner Albuquerque,

1. Evasão escolar 2. Aprendizado de máquina 3. Algoritmo de previsão I. Albuquerque, Danyllo Wagner II. Título.

CDU 004.8

Análise e predição de evasão dos alunos usando Aprendizado de Máquina

Trabalho de Conclusão de Curso apresentado ao Curso Engenharia de Computação, do Instituto Federal da Paraíba – Campus Campina Grande, em cumprimento às exigências parciais para a obtenção do título de Bacharel em Engenharia de Computação.

Aprovado em 06/12/2023

Prof. Danyllo Wagner Albuquerque, Msc.
Orientador

Emanuel Dantas Filho, Sc. M
Membro da Banca

Paulo Ribeiro Lins Júnior, D.Sc.
Membro da Banca

Campina Grande, Paraíba, Brasil
Fevereiro/2024

"Confie ao Senhor tudo que você faz, e seus planos serão bem sucedidos."
– Provérbios 16:3

Agradecimentos

Primeiramente, gostaria de expressar minha gratidão a Deus, cuja orientação e providência estiveram sempre presentes ao longo desta jornada acadêmica. Sua graça e amor foram a luz que iluminou meu caminho, dando-me força nos momentos mais desafiadores e motivando-me a alcançar meus objetivos.

À minha família, em especial aos meus pais, Francisco Lopes de Oliveira e Ana Lucia dos Santos Oliveira, meu eterno agradecimento pelo apoio incondicional, amor e sacrifícios que fizeram para que eu pudesse chegar até este momento. Seu apoio constante e incentivo foram fundamentais para minha trajetória acadêmica. Dedico a vocês este trabalho como uma singela forma de reconhecimento de todo o amor e dedicação que sempre demonstraram por mim.

Aos meus professores do Instituto Federal da Paraíba (IFPB), gostaria de expressar minha profunda gratidão. Suas aulas, orientações e conselhos foram cruciais para o meu desenvolvimento acadêmico e pessoal. Agradeço especialmente ao meu orientador, Prof. Danyllo Wagner Albuquerque, pela sua orientação, dedicação, paciência e apoio ao longo deste trabalho.

Agradeço também aos demais professores que contribuíram para a minha formação, tanto dentro como fora da sala de aula. Seu conhecimento e experiência foram inspiradores e enriquecedores para o meu aprendizado.

A todos que de alguma forma contribuíram para esta conquista, meu sincero agradecimento. Que este trabalho possa servir como uma pequena forma de retribuição por todo o apoio e incentivo que recebi ao longo desta jornada.

Sumário

Lista de Figuras

Lista de Tabelas

Análise e predição de evasão dos alunos usando Aprendizado de Máquina

Alexandre dos Santos Oliveira

[1]Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Campus Campina Grande – PB – Brasil

[2]Coordenação do Curso Superior de Bacharelado em Engenharia de Computação

{alexandre.oliveira}@academico.ifpb.edu.br

Abstract. *In this article, a case study was presented on the use of machine learning algorithms to predict the academic situation of students on IT courses at IFPB Campina Grande. The data was obtained from the SUAP platform and pre-processed to remove null values and convert categorical variables into numerical ones. Three different algorithms were used: naive bayses, the SVM algorithm, Decision Tree Classifier, XGBoost and MLPClassifier. The results showed that XGBoost had the best accuracy, precision, recall and F1-score for data from IT courses, while Decision Tree Classifier obtained slightly worse results and naive bayses had the worst performance. The study presented in this article may be useful for the institution to predict the academic situation of its students to provide additional support to students at risk of dropping out.*

Resumo. *Neste artigo, foi apresentado um estudo de caso sobre a utilização de algoritmos de aprendizado de máquina para prever o nível de evasão dos cursos de Telemática e Engenharia de Computação do IFPB Campina Grande. Os dados foram obtidos a partir da plataforma SUAP e pré-processados para remover valores nulos e converter variáveis categóricas em numéricas. Foram utilizados cinco algoritmos diferentes: naive bayses, o algoritmo SVM, Decision Tree Classifier, XGBoost e MLPClassifier. Os resultados mostraram que o XGBoost teve a melhor acurácia, precision, recall e F1-score para os dados dos cursos de TI, enquanto o Decision Tree Classifier obteve resultados um pouco piores e o naive bayses teve a pior performance. O estudo apresentado neste artigo pode ser útil para a instituição prever a situação acadêmica de seus alunos para fornecer suporte adicional aos alunos em risco de evasão.*

1. Introdução

Segundo a comissão especial de estudos sobre evasão (1996, p. 56) " 1) Evasão de curso seria aquela que ocorre quando o estudante se desliga do curso superior em situações diversas, tais como: abandono (deixa de matricular-se), desistência (oficial), transferência ou reopção (mudança de curso), exclusão por norma institucional; 2) evasão da instituição seria quando o estudante se desliga da instituição na qual está matriculado; e 3) evasão do sistema aconteceria quando o estudante abandona de forma definitiva ou temporária o ensino superior. A evasão estudantil em cursos de graduação é um problema comum entre as universidades e instituições de ensino superior no Brasil. De acordo com Arantes e Pinho (2020), a evasão de alunos em cursos de graduação é um problema frequente nas universidades brasileiras. E de fato, a evasão escolar é uma questão que preocupa não só o Brasil, mas também outros países, e tem

sido alvo de diversos estudos e pesquisas ao longo dos anos.” Em se tratando de cursos superiores no Brasil, a taxa de evasão tem aumentado de maneira expressiva. Segundo dados do Sindicato das Entidades Mantenedoras de Estabelecimentos de Ensino Superior no Estado de São Paulo (SEMESP), “em 2021, a taxa de evasão chegou aos 36,6% nas modalidades de ensino a distância (EaD) e presencial. O percentual equivale a 3,42 milhões de alunos”. E em cursos de graduação na área de TI mais da metade dos alunos matriculados não chegam à conclusão do curso. Segundo uma pesquisa elaborada pela Associação Brasileira das Empresas de Tecnologia da Informação e Comunicação (Brasscom), “cerca de 69% dos universitários nas áreas de TI não se formam” (BRASSCOM, 2021). Um dado preocupante já que a demanda por profissionais na área está em alta.

Geralmente, a identificação da intenção do estudante em abandonar o curso ocorre tardiamente, quando já não há mais possibilidades de recuperá-lo, o que prejudica não apenas os discentes, mas também as próprias instituições, que têm uma queda na regularidade e qualidade de ensino. Essa situação é ainda mais preocupante quando se considera que a demanda por profissionais na área de TI está em alta. Em instituições da rede pública e principalmente nas redes privadas têm consequências nos seus orçamentos, já que o número de alunos matriculados contribui significativamente para compor a matriz orçamentária de uma instituição.

1.1 Objetivos

O objetivo deste estudo é propor um modelo preditivo usando algoritmos de aprendizado de máquina para prever quais variáveis têm maior peso na evasão do aluno do curso.

Os objetivos específicos consistem em: Analisar os algoritmos de Machine Learning: Naive Bayes, SVM (Support Vector Machine), Decision Tree (DecisionTreeClassifier), Artificial Neural Network – MultiLayer Percetron (MLPClassifier) e XGBoost (XGBClassifier). Identificar quais variáveis têm o maior peso na evasão do discente. Aplicar o modelo de previsão de evasão escolar desenvolvido para analisar os dados dos alunos disponibilizados.

1.2 Relevância

A análise da evasão do curso é importante para compreender o problema e com isso tomar medidas para evitar essas evasões. E uma maneira eficaz é prever quais discentes podem evadir do curso e tomar essas medidas antecipadas. Este estudo também irá beneficiar diretamente o IFPB, utilizando métodos eficazes de Machine Learning poderá identificar quais alunos estão em risco de evasão e ajudar os gestores a adotar as estratégias mais adequadas no sentido de mitigar a desistência de alunos.

1.3 Contribuições

Sabendo do poder computacional, a variedade de técnicas e algoritmos de

Machine Learning com a grande quantidade de informação disponibilizada digitalmente atualmente, é interessante que esses métodos possam ser utilizados para o auxílio para diminuir o número de evasões de alunos durante o curso, favorecendo os gestores a tomarem decisões eficazes e de forma antecipada.

As contribuições deste estudo incluem o desenvolvimento de modelos preditivos para a previsão de evasão de alunos em cursos universitários, além do fornecimento de insights sobre as variáveis mais relevantes para a evasão.

2. Referencial Teórico

2.1 Machine Learning

Machine learning (aprendizado de máquina) investiga como computadores podem aprender (ou melhorar o seu desempenho) com base em dados analisados. A principal área de pesquisa é desenvolver algoritmos para aprenderem a reconhecer padrões complexos e tomar decisões inteligentes baseadas nos dados. Por exemplo, é um problema de aprendizado de máquina programar um computador para que ele possa reconhecer automaticamente códigos postais manuscritos no correio após a aprendizagem de um conjunto de exemplos [Han et al., 2011].

As técnicas de machine learning têm sido aplicadas aos mais variados problemas do mundo real [Faceli et al. 2011, Witten et al. 2016]. Mas para ter resultados satisfatórios, é necessário conhecer bem a base de dados disponível, para depois aplicar de forma eficaz e corretamente as técnicas de machine learning.

2.2 Técnicas de Classificação

Classificação, em machine learning, é a tarefa de organizar objetos em uma entre diversas categorias pré-definidas. Essas técnicas são aplicadas em problemas onde as instâncias de uma base de dados, representados por um conjunto de atributos, precisam ser enquadrados em um conjunto predefinido de possíveis rótulos (classes). Por exemplo, um programa de e-mail pode tentar classificar um e-mail como "legítimo" ou como "spam", usando a classificação baseada em emails anteriormente recebidos e rotulados [Han et al., 2011]

Duas etapas principais são realizadas no processo de classificação: (a) a aprendizagem, na qual dados de treinamento são analisados por um algoritmo classificador, em que são atribuídos os rótulos de classe e o modelo classificador é representado sob a forma de regras de classificação; e (b) a classificação, na qual os dados de teste são usados para estimar a acurácia das regras de classificação. Se a acurácia for considerada aceitável, as regras podem ser aplicadas para a classificação de novos dados [Han et al., 2011]. As principais classes de algoritmos são agrupados em cinco categorias principais: árvores de decisão, classificadores baseados em regras, bayesianos, classificadores de vizinho mais próximo, redes neurais artificiais e Support Vector Machine (SVM) [Witten et al., 2011].

2.3 Métricas

São usados os rótulos (P - Positivo, N-Negativo) para diferenciar a classe real e a classe prevista. Dado um classificador é uma instância a classificar, há quatro resultados possíveis: Se a instância é positiva e é classificada corretamente como positiva, ela é contada como um verdadeiro positivo (VP); Se a instância é positiva e é classificada incorretamente como negativa, é contada como um falso negativo (FN); Se a instância é negativa e é classificada corretamente como negativa, é contada como um verdadeiro negativo (VN); e Se a instância é negativa e é classificada incorretamente como positiva, é contada como um falso positivo (FP).

Após a classificação de todas as instâncias da base de testes, pode-se, então, obter as seguintes métricas que avaliam o poder preditivo do classificador [Fawcett, 2006] [Tan et al., 2009] [Silva et al., 2016]:

Acurácia (Accuracy) - Representa a taxa de acerto de todo o classificador e é obtida pela razão entre a soma dos acertos das duas classes e o número total de instâncias classificadas. É obtida pela expressão: Acurácia = (VP + VN) / (VP + VN + FP + FN). Obviamente, o melhor caso esperado para uma matriz de confusão é o preenchimento apenas da diagonal principal, o que resultaria em uma acurácia de 100%.

Precisão (Precision) - Representa a preditividade positiva, que é o percentual de acertos de verdadeiros positivos dentre todos os exemplos classificados como positivos. Sua expressão é: Precisão = VP / (VP + FP). Quanto maior a precisão, menor o erro de falsos positivos cometidos pelo classificador.

Sensibilidade (Recall) - Indica a taxa de verdadeiros positivos, ou seja, o percentual de VP previstos corretamente. É obtida por Recall = VP / (VP + FN). Um alto recall indica que o classificador produziu poucos exemplos positivos classificados como falso negativos.

3. Trabalhos Relacionados

Alguns trabalhos relataram técnicas de machine learning para estabelecer perfis e realizar tarefas de predição da evasão de alunos do ensino superior. O trabalho de [Lanes and Alcantara 2018], os autores utilizaram dados acadêmicos referentes ao primeiro ano dos cursos de graduação para identificar alunos em risco de evasão. Os dados de 12 cursos de graduação foram extraídos do sistema acadêmico da Universidade Federal do Rio Grande (FURG) no período de 2012 até 2017. O algoritmo J48 (Árvore de Decisão) foi utilizado como gerador de regras para o sistema de classificação. O mesmo método e objetivos também podem ser vistos no trabalho de [Paz and Cazella 2017].

Brito et al. (2015) propõem modelos baseados nos algoritmos NaiveBayes, J48 e AdaBoostM1, para a identificação de estudantes com risco de evasão a partir da observação das notas obtidas pelos estudantes na prova de ingresso na instituição e nas disciplinas do primeiro semestre do curso de graduação.

Em um outro trabalho [Digiampietri et al. 2016], os autores também utilizaram somente os dados referentes ao desempenho nas disciplinas do primeiro ano de curso. Foram analisados os dados extraídos dos históricos de mais de 1000 alunos do Bacharelado de Sistemas de Informação da EACH-USP entre os períodos 2005 e 2015. Para este trabalho, foi utilizado o classificador Rotation Forest, que por padrão aplica seleção de atributos a partir do Principal Component Analysis (PCA).

A pesquisa de Rigo et al. (2014) utilizou dados de disciplinas em três semestres de três cursos de graduação com o intuito de obter inferências em relação à evasão no curso. Os resultados da identificação de perfis de alunos com risco de evasão tiveram taxas de acerto na ordem de 87%, utilizando o algoritmo RNA Multilayer Perceptron.

Como a maior parte dos estudos é restrita a uma base de dados específica de um curso ou universidade, espera-se que o presente trabalho contribua para um enriquecimento dos estudos sobre tema e traga uma perspectiva geral sobre os principais fatores relacionados à evasão em cursos superiores do Instituto Federal da Paraíba do campus Campina Grande.

4. Metodologia

Para esse trabalho foram utilizados os dados dos cursos de Engenharia de Computação e do curso de Telemática que foram disponibilizados pela a instituição para fins de pesquisa. Assim, a instituição encaminhou dois arquivos no formato ".csv" com os dados necessários para a pesquisa.

A análise e interpretação dos dados se deu por meio de algoritmos de Aprendizado de Máquina, esses algoritmos se utilizam de técnicas estatísticas para realizar as previsões de riscos se algo vier acontecer. Um resumo do processo utilizado neste trabalho pode ser visto na Figura 1. Todas as implementações foram desenvolvidas com o uso da linguagem Python e das ferramentas disponíveis no pacote scikit-learn(Pedregosa, et al., 2011).

E para criar modelos de classificação foram escolhidos cinco algoritmos de Aprendizado de Máquina desde o mais básico e mais simples de implementar ao mais avançado que utiliza técnicas de redes neurais com a intenção de comparar seus resultados e verificar quais variáveis tem maior peso. Os algoritmos são o naive bayses, o algoritmo SVM, as árvores de decisão, do inglês Decision Tree (DT), o XGBoost (eXtreme Gradient Boosting) e o algoritmo de Multilayer Perceptron da biblioteca Sklearn para classificação é o MLPClassifier. E esses modelos serão avaliados e comparados com base em sua taxa de acerto e nos seus resultados a fim de obter o melhor desempenho para predição de estudantes com risco de evasão. As implementações serão desenvolvidas com o uso da plataforma Google Colab.

Inicialmente, foi realizado um pré-processamento dos dados, de forma a organizar a estrutura das informações para a aplicação das técnicas escolhidas. Foram removidos dados faltantes e desconsideradas as características irrelevantes para o estudo em questão, como dados de informação pessoal, dados que identificassem o aluno e dados repetidos. Os dados crus disponibilizados, somando os dois arquivos, dos cursos de Engenharia de Computação e Telemática contêm 2053 linhas com 46 colunas. Ao fim da limpeza o arquivo ficou com o total de 1373 linhas e 14 colunas. Após uma etapa

de balanceamento, a base foi separada em duas partes: dados para treinamento e dados para testes. A base de treinamento, que representa os dados efetivamente usados no processo de aprendizagem, foi composta por 70% dos dados, selecionados aleatoriamente. A base de testes foi composta pelos 30% restantes. Ela foi usada como dados novos, nunca vistos pelo sistema.

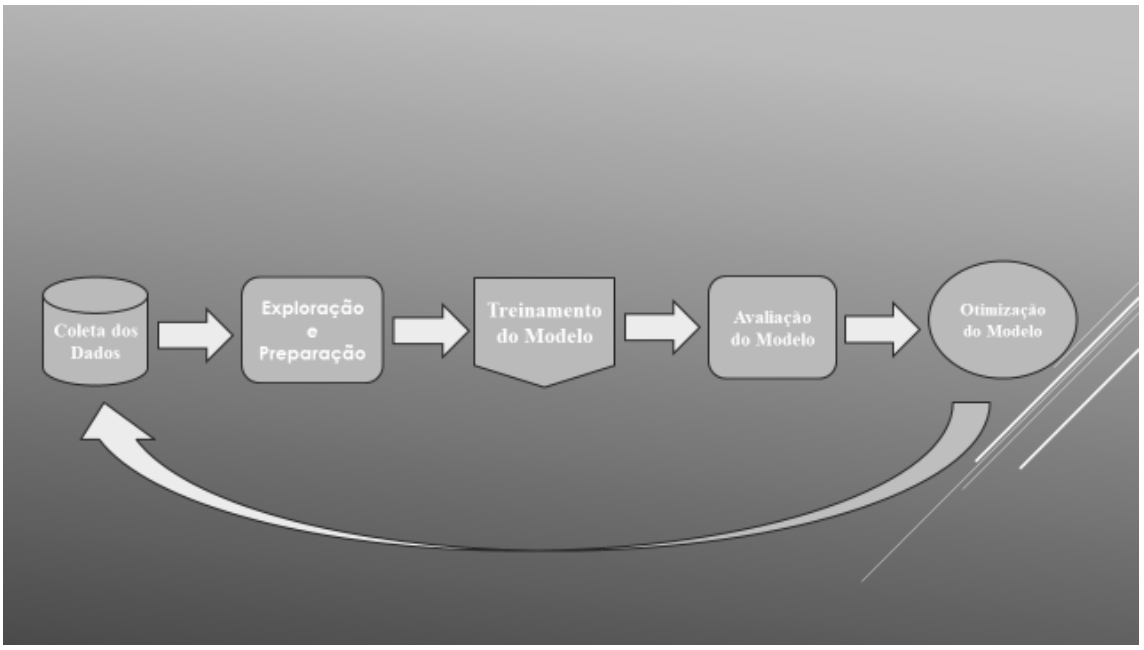


Figura 1 - Processo de funcionamento.

4.1 Técnicas de Machine Learning

Para criar o modelo preditivo de evasão, utilizamos técnicas de aprendizado de máquina (Machine Learning). Como este trabalho tem como objetivo a identificação das variáveis mais determinantes na classificação de um estudante como evasão ou sucesso, a utilização de diferentes técnicas foi necessária para que o melhor resultado possível para a classificação fosse alcançado. E também, este presente trabalho não procura comparar as técnicas em si, mas encontrar o melhor resultado para o problema abordado. Estas técnicas foram escolhidas por sua ampla utilização em problemas de classificação. Para cada uma das técnicas, diferentes configurações de parâmetros foram testados a fim de melhorar os resultados e reduzir a ocorrência de overfitting.

Utilizamos esses algoritmos em conjunto com técnicas de validação cruzada para garantir que o modelo não seja enviesado pelos dados de treinamento. Além disso, utilizamos métricas como a acurácia, a precisão, o recall e a pontuação F1 para avaliar o desempenho do modelo em dados de teste. Com a aplicação dessas técnicas, obtivemos um modelo preditivo para prever a evasão de alunos em curso.

4.2 Variáveis

Para a realização do estudo, foram selecionadas as variáveis da tabela 1 referente ao curso de Engenharia de Computação e de Telemática, e essas variáveis foram utilizadas como entrada dos modelos de Aprendizado de Máquina. Foram definidas as variáveis de entrada do modelo dada por x = (situacao_ult_periodo, ano_ingresso, ivs_valido, coef_progresso, cota_sistec, cre, cidade, faixa_renda, forma_ingresso, periodo_atual, periodo_ingresso, sexo, tipo_escola_ant, zona_residencial, ultimo_periodo_letivo, idade_disc, ingresso_disc) e variável de saída do modelo dada por y = (situacao).

Tabela 1. Variáveis utilizadas

Variável	***Descrição***
situacao	Situação atual do aluno no curso
situacao_ult_periodo	Situação do aluno no último período cursado
ano_ingresso	Ano de ingresso do aluno
ivs_valido	Índice de vulnerabilidade social válido
coef_progresso	Coeficiente de progressão do aluno
cota_sistec	Tipo de cota do aluno (Sistema de Cotas)
cre	Coeficiente de rendimento escolar do aluno
cidade	Cidade de residência do aluno
faixa_renda	Faixa de renda do aluno
forma_ingresso	Forma de ingresso do aluno no curso
periodo_atual	Período atual do aluno no curso
periodo_ingresso	Período de ingresso do aluno no curso
sexo	Gênero do aluno
tipo_escola_ant	Tipo de escola de ensino médio que o aluno frequentou
zona_residencial	Zona de residência do aluno
ultimo_periodo_letivo	Último período letivo do aluno
idade_disc	Idade do aluno quando cursou a disciplina
ingresso_disc	Período de ingresso do aluno na disciplina

Fonte: Elaborado pelo autor deste artigo (2023)

5. Análise e Resultados Obtidos

Para avaliar o desempenho dos modelos de aprendizado de máquina, foram utilizadas as métricas de precisão, recall e F1-score. A precisão mede a proporção de verdadeiros positivos em relação ao número total de exemplos classificados como positivos pelo modelo. O recall mede a proporção de verdadeiros positivos em relação ao número total de exemplos positivos na base de dados. O F1-score é uma medida combinada de precisão e recall. A tabela 2 mostra os resultados obtidos com os algoritmos utilizados. A tabela 2 a seguir resume as métricas de desempenho de diferentes modelos de aprendizado de máquina utilizados para prever a evasão nos cursos de Engenharia de Computação e Telemática:

Modelo	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)
Naive Bayes	**95**	**87**	**93**	**90**
SVM	**91**	**95**	**65**	**77**
Decision Tree	**98**	**97**	**95**	**96**
XGBoost	**98**	**97**	**94**	**96**
MultiLayer Perceptron	**96**	**91**	**90**	**91**

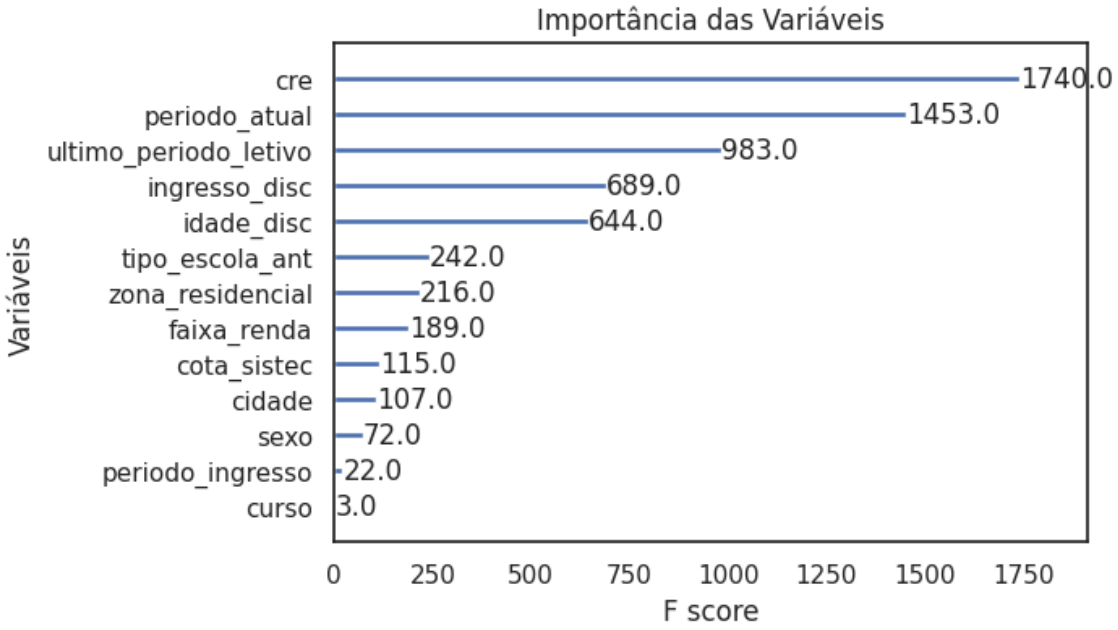
Tabela 2 Fonte: Elaborado pelo autor deste artigo(2023)

Naive Bayes: Este modelo apresentou uma alta acurácia (95%) e um bom equilíbrio entre precisão (87%) e recall (93%), resultando em um F1-Score de 90%. Isso indica uma boa capacidade geral de previsão de evasão. SVM (Support Vector Machine): Embora este modelo tenha uma precisão elevada (95%), o seu recall relativamente baixo (65%) sugere que ele pode não estar identificando todos os casos de evasão eficientemente, o que é refletido no F1-Score mais baixo (77%). Decision Tree: Este modelo se destaca com a maior acurácia (98%) e um excelente equilíbrio entre precisão (97%) e recall (95%), levando a um F1-Score alto (96%). Isso sugere uma

capacidade superior na identificação correta de casos de evasão. XGBoost: Com resultados semelhantes ao do modelo Decision Tree, o XGBoost também mostra um desempenho excepcional em todas as métricas, com destaque para a sua alta acurácia e F1-Score (ambos 96%). MultiLayer Perceptron (MLP): Este modelo, baseado em redes neurais, também demonstrou um desempenho robusto com uma acurácia de 96%, equilíbrio entre precisão (91%) e recall (90%), resultando em um F1-Score de 91%. Em resumo, os modelos Decision Tree e XGBoost mostraram-se os mais eficazes, com altas pontuações em todas as métricas. O SVM, apesar de sua alta precisão, teve um desempenho inferior em termos de recall. Naive Bayes e MLP apresentaram um equilíbrio consistente entre todas as métricas.

Para avaliar a importância das variáveis, o algoritmo XGBoost oferece algumas funções que fornecem o gráfico e uma lista com as variáveis a partir do nível de relevância no modelo. Nesse caso, foi utilizada a função plot_importance, a qual mostra a importância das variáveis a partir do peso. Nos dados do curso de Engenharia de Computação e Telemática analisados (figura 3) a variável que apresentou maior relevância foi cre e a menor relevância foi curso. Neste caso, é possível perceber a importância de cada variável utilizada.

Figura 2. Importância das variáveis Usando XGBoost



Fonte: Elaborado pelo autor deste artigo(2023)

6. Conclusões

Com base nos resultados obtidos, pode-se afirmar que os modelos de Aprendizado de Máquina apresentam uma boa capacidade de predição da situação acadêmica dos estudantes dos cursos de Engenharia de Computação e em Telemática. O

modelo baseado em XGBoost obteve as melhores métricas de desempenho, com destaque para a acurácia e a precisão. Com a métrica de avaliação dos modelos, foram utilizados Accuracy, Precision, Recall, F1-Score e Matriz de confusão. Os dois modelos (XGBoost e MLP) se mostraram melhores que os demais modelos, porém o modelo XGBoost se mostrou melhor. No dataframe as métricas do algoritmo XGBoost foram de (Acurácia de 98% , Precisão de 97%, Recall de 94% e de F1-Score de 96%). De forma geral, o algoritmo XGBoost apresentou-se melhor que os demias algoritmos analisados neste trabalho. Foi analisada a importância das variáveis, em que, para o dataframe a variável que apresentou maior relevância foi o cre, seguida de periodo_atual e ultimo_periodo_letivo. Já as três piores variáveis foram curso, periodo_ingresso e sexo.

O pré-processamento dos dados e a seleção das variáveis mais relevantes foram essenciais para a obtenção de um modelo com boa capacidade de generalização. Ainda assim, é possível que a inclusão de outras variáveis possa melhorar ainda mais o desempenho dos modelos.

Uma das principais limitações do modelo foi conseguir utilizar dados apenas do sistema acadêmico onde não tinha dados mais específicos relacionados às disciplinas em que o discente pagou ou não. Outro limitante foi a quantidade de dados ausentes encontrados nas bases de dados. Assim, se a instituição conseguir melhorar a manipulação de seus dados, é possível melhorar a qualidade e a precisão das previsões de evasão escolar. Dessa forma o trabalho contribui para que a Instituto Federal da Paraíba Campus Campina Grande possa formar uma lista de características importantes para avaliação, fazendo com que os estudantes permaneçam no curso de graduação até sua conclusão e possibilitando que a instituição foque em melhorar essas variáveis em seus sistemas acadêmicos.

6.1 Trabalhos futuros

Uma possível sugestão para trabalhos futuros seria a realização de uma análise mais aprofundada das variáveis selecionadas, com o objetivo de identificar quais delas apresentam maior importância na predição da situação acadêmica dos estudantes. Essa análise poderia ser realizada por meio de técnicas de feature selection.

Outra sugestão seria a realização de testes com outros algoritmos de machine learning e a realização de uma comparação mais detalhada do desempenho dos diferentes modelos. Além disso, seria interessante também explorar outras técnicas de pré-processamento de dados, como a normalização ou padronização de variáveis.

Por fim, seria interessante também realizar uma análise mais aprofundada dos resultados obtidos para entender melhor os padrões e as características dos estudantes que apresentam maiores chances de evasão ou reprovação. Essa análise poderia ajudar a identificar possíveis fatores de risco que poderiam ser alvo de intervenções preventivas por parte das instituições de ensino.

7. Referências Bibliográficas

BRASSCOM. Evasão escolar: 69% dos alunos dos cursos de tecnologia não se formam. Disponível em: https://brasscom.org.br/evasao-escolar-69-dos-alunos-dos-cursos-de-tecnologia-nao-se-formam/. Acesso em: 10 abr. 2023.

Lanes, M. and Alcântara, C. (2018). Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), volume 29, page 1921.

Digiampietri, L., Nakano, F., and Lauretto, M. (2016). Mineração de dados para identificação de alunos com alto risco de evasão: Um estudo de caso. Grad+ Revista

de Graduação da USP, 1:17–23.

SESU/MEC; ANDIFES; ABRUEM. Diplomação, retenção e evasão nos cursos de graduação em IES públicas: Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras. Brasília, DF: [s. n.], 1996. Disponível em: http://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf. Acesso em: 04 jan. 2021.

Paz, F. and Cazella, S. (2017). Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. In Anais dos Workshops do Congresso Brasileiro de Informática na Educação, volume 6, page 624.

RIGO, S. J. et al. Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. Revista Brasileira de Informática na Educação, v. 22, n.01, p. 132, 2014.

Brito, D., Pascoal, T., Araújo, J., Lemos, M. e Rêgo, T. (2015) “Identificação de

estudantes do primeiro semestre com risco de evasão através de técnicas de Data

Mining.” In TISE 2015 - XX Congresso Internacional de Informática Educativa.

Santiago.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

Zhang, Y., Yang, Q., & Xu, W. (2018). Deep neural networks for multi-target regression. Neurocomputing, 275, 1237-1244.

Zhou, J., Wang, F., Hu, J., & Liu, X. (2020). A hybrid forecasting model using machine learning for estimating residential electricity consumption. Applied Energy, 279, 115871.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Campina Grande
R. Tranqüilino Coelho Lemos, 671, Dinamérica, CEP 58432-300, Campina Grande (PB)
CNPJ: 10.783.898/0003-37 - Telefone: (83) 2102.6200

Documento Digitalizado Ostensivo (Público)

TCC corrigido

Assunto:	TCC corrigido
Assinado por:	Henrique Nascimento
Tipo do Documento:	Comprovante
Situação:	Finalizado
Nível de Acesso:	Ostensivo (Público)
Tipo do Conferência:	Documento Original

Documento assinado eletronicamente por:

- **Henrique do Nascimento Cunha**, **COORDENADOR(A) DE CURSO - FUC1 - CCEC-CG**, em 24/02/2024 09:17:21.

Este documento foi armazenado no SUAP em 24/02/2024. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1092404
Código de Autenticação: 8ad7796c0d