Abordagem para Identificação dos Fatores Determinantes no Abandono Acadêmico no Ensino Superior Utilizando Algoritmos de IA

Daniel Yuji Yamada, Felipe Mateos Castro de Souza
danielyuji@usp.br
felmateos@usp.br

Abstract

O presente artigo visa relatar uma análise exploratória e o desenvolvimento de um modelo de aprendizado de máquina preditiva, a fim de compreender o problema do êxodo de estudantes no ensino superior. Assim sendo, é possível relacionar os motivos mais prováveis de tal fenômeno para contribuir à discussão sobre problema. Para fins deste estudo optamos por utilizar o conjunto de dados 'Predict students' dropout and academic success" (Realinho and Baptista 2021). Conduzimos os experimentos utilizando uma variedade de técnicas de tratamento e análise de dados e aprendizado de máquina supervisionado. Como produto final obtivemos um classificador baseado em uma Regressão Logistica que alcançou acurácia e medida f1 macro de 82% e 86% respectivamente. Por fim, os resultados mostraram que os motivos principais estão além de fatores internos da universidade, como rendimento acadêmico por exemplo. Mas de fatores e situações socioeconômicas que os estudantes enfrentam durante a graduação. Mais detalhes serão abordados no decorrer do artigo. A implementação prática está disponível em: https://github.com/felmateos/student-dropout-prediction

Introdução

Considera-se o fenômeno da evasão uma das principais preocupações do Ministério da Educação, realizada antes da formação completa do curso e é visto como um alvo a ser combatido ou um índice a ser reduzido (Lima Coimbra and Costa 2021). O número de alunos dentro de faculdades cresceu muito nos últimos dez anos, mas com isso, a taxa de evasão universitária cresceu conjuntamente, e com o abandono crescente de mais alunos, a instituição de ensino superior sofre de vários impactos diretos, podendo apresentar desperdício de recursos e má utilização da infraestrutura disponível.

Alguns dos motivos que podem explicar este fenômeno são a dificuldade de conciliar os estudos com o seu emprego, perca de interesse no curso, uma infraestrutura precarizada por parte da instituição de ensino que não atende aos interesse e expectativas do aluno, entre outros. (Cunha and Morosini 2014)

Além da perda de recursos por parte da instituição de ensino superior, a evasão universitária representa um prejuízo tanto para o aluno - com o investimento financeiro, tempo e esforços frustrados - como também para a sociedade - visto que, há uma perda na oportunidade de produção de conhecimento científico, para além de serviços profissionais mais qualificados no futuro que não irão acontecer.

Nesse trabalho, serão utilizados algoritmos de inteligência artificial para construir modelos que podem predizer se um aluno irá ou não abandonar a graduação. Além disso, faremos uma análise do melhor modelo a fim de encontrar quais foram os atributos mais determianantes para as classificações feitas.

Arcabouço Teórico

Linguagem e ferramentas

O Python é uma das principais linguagens utilizadas para aplicações de modelagens estatísticas no mercado e para o estudo em aprendizado de máquina. Apesar de não ser uma das linguagem mais eficientes, sendo conveniente para aplicações em modelos de larga escala com conjunto de dados maiores do que a utilizada para esse estudo, a linguagem compensa pela disponibilização de várias bibliotecas e outros recursos de aprendizado. Portanto, é conhecida por ter muitas ferramentas disponíveis e outros recursos de estudo que auxiliaram no desenvolvimento deste projeto. Além disso, as principais ferramentas utilizadas que são disponíveis para o Python são Jupyter Notebook e Anaconda.

Conjunto de Dados

O conjunto de dados utilizado foi encontrado na Kaggle e fornece dados sobre alunos de graduação de vários cursos diferentes. Tais como agronomia, design, licenciatura, enfermagem, jornalismo, gestão, serviços sociais e tecnologias. Além disso, a fonte dos dados é de uma instituição de ensino superior, no entanto, a origem dos dados em si foram adquiridos de várias bases de dados distintas. O conjunto de dados é apresentado como própria para modelos de predição, sendo referente ao abandono ou não dos estudantes quanto aos estudos acadêmicos, trazendo como proposta a investigação dos impactos socioeconômicos para o fenômeno.

Os dados consistem em representar cada aluno por registro, e as colunas são suas respectivas informações so-

bre sua trajetória acadêmica, demográfias e aspectos socioeconômicos dos estudantes. Estes resumem 36 variáveis, sendo 16 numéricas e 20 categóricas entre as independentes. A variável dependente, chamada "Target" é dividida entre as categorias "Dropout" (aqueles que saíram da graduação antes da conclusão do curso), "Graduate" (aqueles que se graduaram) e "Enrolled" (aqueles que ainda estão cursando). As variáveis se diferenciam desde notas do primeiro e segundo semestre, formação e grau dos pais, se o estudante é devedor ou não, entre outros.

Bibliotecas

Análise e tratamento de Dados: Para as tarefas descritas nessa sessão, optamos por utilizar as bibliotecas Pandas e Numpy, visto que ela contém estruturas de dados e ferramentas de manipulação e tratamento de dados projetadas para tornar a limpeza e análise de dados rápida e conveniente em Python. (McKinney 2023)

Visualização: A análise visual foi importante para a facilidade de se interpretar os algoritmos e modelos feitos durante o estudo, tendo a necessidade de se utilizar ferramentas que permitem criar mapas e gráficos. Sendo assim, foi utilizado o matplotlib e o seaborn para a visualização de gráficos, matrizes, histogramas, etc.

Pré-processamento e Modelagem: A principal ferramenta para a modelagem foi o SciKitLearn. Com ela tivemos acesso a alguns algoritmos importantes para o estudo e que é amplamente utilizado para estudos e pesquisa envolvendo aprendizado de máquina. Além disso, algumas outras ferramentas utilizadas neste estudo e que são classes de auxílio para ela. (McKinney 2023)

Alguns processos com relação aos dados originais precisaram ser feitos para se adequar melhor ao objetivo do projeto e aos ajustes para os resultados mais proveitosos. Questões com relação a lidar com classificação e classes não balanceadas apropriadamente foi necessário do uso do Imbalanced-Learn.

Algumas estimativas também auxiliaram na compreensão e interpretação de alguns processos durante a modelagem, conduções de testes e disponibilização de meio para um estudo exploratório estatístico. E o StatsModels também foi para cumprir essas demandas.

Interpretabilidade: Optamos por utilizar a biblioteca shap do Python para encontrar quais features tnham mais relevância para o classificador escolhido.

Algoritmos

Foi cogitado para se utilizar três opções de algoritmos de machine learning que se encaixam no aprendizagem do tipo indutivo e supervisionado.

Naive Bayes: O Naive Bayes é um algoritmo classificador e se utiliza de uma técnica de classificação de dados e obtém uma tabela de probabilidades para cada fenômeno ou característica como atributo, se tornando popular na área acadêmica de estatística por ser simples e de trazer bons resultados. Este algoritmo foi baseado no teorema de Bayes, criado por Thomas Bayes. O algoritmo, a partir da análise da tabela de dados, cria classes e oferece uma resposta a partir de critérios estabelecidos. Ela é chamada de "ingênua" também por assumir a priori que todos as variáveis atributo são independentes uma da outra, e todas tem o mesmo peso para os resultados.

Árvore de decisão: A Árvore de Decisão é um outro classificador (também podendo trabalhar com regressão), o qual analisa os dados a fim de criar uma hierarquia de importância entre os nós. Essas decisões são representadas por nós de uma estrutura de dados em árvore, sendo que cada nó é relacionada a um atributo por vez, gerando uma árvore. Seu comportamento é similar a de um fluxograma, onde o nó-raiz da estrutura - por onde o algoritmo começa - inicia a decisão de classificação do dado por meio de uma subclassificação, e depois, ela é passada para outros nós adiante até o nó-folha, por onde a decisão final do modelo é feita. É lá onde a classificação está completa. Esse algoritmo porém, precisa de uma etapa para se decidir quais são as posições de cada nó, e para isso ela é baseada em um cálculo do ganho de informação e entropia de cada atributo e suas categorias.

Regressão Logística: A Regressão Logística é uma técnica que visa relacionar dois fatores de dados por meio de uma técnica de análise de dados. Depois essa relação é utilizada para fazer predições com base em outras relações. Dessa forma é possível de se extrair várias percepções práticas para então realizar uma análise de probabilidade e definir padrões. Esse algoritmo acaba sendo também muito utilizada por ser simples, rápida e flexível na modelagem de um problema.

Outras Técnicas

Análise de Componentes Principais: O PCA é uma das técnicas para modelos estatísticos que visa simplificar a dimensionalidade do problema, criando os componentes, através de procedimentos que se utilizam da transformação ortogonal, e observando a proporção da variância explicada por eles. Sua principal função é encontrar os melhores e mais relevantes parâmetros para o treinamento do modelo.

Os componentes seriam representações de um grupo de variáveis que explicaria uma porção do problema. No entanto neste processo há perda de informação, tornando a análise da variância explicada tão importante. É com essa medida que conseguimos avaliar a variância do modelo original com relação ao feito em componentes. Por isso que a medida que mais componentes são criados, maior será a tendência de se perder menos informação. Ao fim, é preciso buscar um balanceamento entre a simplificação do modelo e a diminuição do tempo de processamento computacional com relação ao quanto de informação é perdida e a alteração da acurácia das predições feitas com o modelo. Sendo assim muito útil para modelos que envolvem um grande número de variáveis com uma extensa quantidade de dados para processar.

Descrição e Modelagem do Problema

Análise Exploratória

Para identificar e priorizar variáveis, realizamos uma análise exploratória detalhada:

Variáveis Categóricas: Avaliamos a associação entre variáveis categóricas e a variável alvo usando o teste Chi-quadrado em tabelas de contingência. O valor V de Cramer foi usado para medir a força da associação. Variáveis com baixa associação foram removidas. Entre as variáveis categóricas removidas estavam Nacionality, Educational special needs, International.

Variáveis Numéricas: Transformamos a variável alvo original (matriculado, graduado, desistente) em uma variável binária indicando "Dropout" (1, positivo) ou "Graduated_or_Enrolled" (0, negativo) dos estudos acadêmicos. Em seguida, aplicamos análise de correlação de Pearson para identificar colinearidade entre variáveis. Utilizamos o Fator de Inflação da Variância (VIF) para eliminar variáveis com valores VIF maior que 5 e alta colinearidade. Variáveis numéricas removidas inclui Curricular units 1st sem (credited), Curricular units 1st sem (enrolled), Curricular units 1st sem (grade), Curricular units 1st sem (evaluations), Curricular units 1st sem (without evaluations), Curricular units 2nd sem (credited), Curricular units 2nd sem (enrolled), Curricular units 2nd sem (evaluations), Curricular units 2nd sem (grade), Curricular units 2nd sem (without evaluations).

Transformações do Conjunto de Dados

Transformações detalhadas foram realizadas para preparar os dados para modelagem:

Remoção de Features Pouco Relevantes: Baseado na análise exploratória, removemos variáveis com baixa relevância, minimizando a complexidade e melhorando a eficiência dos modelos. O dataset original continha 20 variáveis independentes categóricas e 16 variáveis independentes numéricas, para além da variável categórica dependente 'Target', totalizando 35 variáveis. Após a análise inicial das variáveis, foram removidas 3 categóricas e 10 numéricas.

Codificação One-Hot: Algumas das variáveis categóricas foram transformadas em variáveis binárias (pelo one-hot encoding) para que possam ser utilizadas diretamente em algoritmos de aprendizado de máquina, que requerem dados numéricos. Variáveis com três ou mais categorias, foram decompostas em outras variáveis com valores (0) ausente ou (1) presente. Algumas delas já eram binárias, então não precisaram serem decompostas em outras variáveis. Ao final da operação foi totalizado uma dimensionalidade de 220 variáveis para o modelo.

Normalização: As variáveis numéricas foram normalizadas utilizando o MinMaxScaler para garantir que todas as variáveis estejam na mesma escala, o que é crucial para a convergência de muitos algoritmos de aprendizado de máquina e para evitar que variáveis com maior magnitude dominem o modelo.

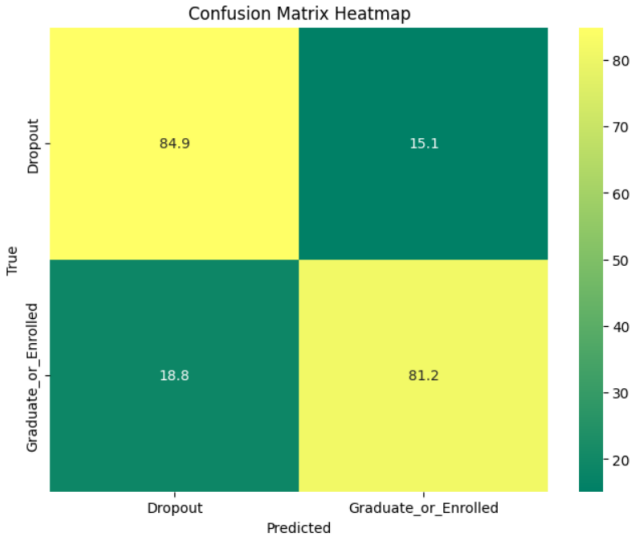

Figure 1: Heatmap para os resultados da Regressão Logística

Rebalanceamento de Classes: Utilizamos uma combinação de TOMEK links e SMOTE para abordar o desequilíbrio de classes. TOMEK links ajuda a limpar as fronteiras entre classes (Jeatrakul, Wong, and Fung 2010), enquanto SMOTE sintetiza novos exemplos da classe minoritária, permitindo que o modelo aprenda melhor com dados balanceados.(Bowyer et al. 2011)

Experimentos

Pipeline de Treinamento e Validação

O pipeline desenvolvido compreende:

Seleção de Features: Utilizamos o SelectKBest para identificar as melhores colunas do dataset, com base na correlação com a variável alvo.

Seleção de Modelos: Realizamos Grid Search com Cross Validation (5 folds) para selecionar o melhor classificador. Utilizamos o F1-score como métrica principal devido à sua capacidade de lidar com o desequilíbrio de classes. O F1-score combina precisão e recall, o que é crucial para capturar a performance do modelo na classe "abandono".

Validação Estratificada: Aplicamos Cross Validation estratificado com 5 folds, focando no F1-score da classe "abandono", assegurando que cada fold represente adequadamente a distribuição das classes.

Teste do Classificador: Avaliamos o classificador no conjunto de teste e geramos uma matriz de confusão para analisar a precisão entre as classes.

Armazenamento do Modelo: O modelo final foi salvo como um arquivo .pkl com o formato a{acurácia}_f{f1_score}.pkl.

Resultados: Os resultados dos experimentos são apresentados na tabela 1, e a matriz de confusão para o melhor modelo (Regressão Logística) é exibida na figura 3.

Table 1: Resultados de Algoritmos com Diferentes Conjuntos de Dados e Hiperparâmetros

Algoritmo	Conjunto de Dados	Hiperparâmetros	Acurácia* (%)	F1* (%)
Naive Bayes	Balanceamento original	`'alpha': 0.1`	74,37	57,74
	Rebalanceado	`'alpha': 0.1`	76,26	76,02
Árvore de Decisão	Balanceamento original	`'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 2`	85,50	75,54
	Rebalanceado	`'max_depth': 3, 'min_samples_leaf': 1, 'min_samples_split': 2`	87,46	87,78
Regressão Logística	Balanceamento original	`'C': 10, 'class_weight': None, 'penalty': 'l1', 'solver': 'liblinear'`	86,12	76,74
	Rebalanceado	`'C': 10, 'class_weight': None, 'penalty': 'l1', 'solver': 'liblinear'`	86,54	86,35

Acurácia* e F1*: Média das medidas de cada fold de uma validação cruzada de 5 folds dentro do conjunto de treinamento.

Justificativa das Medidas de Desempenho

Tendo em vista o objetivo do projeto de promoção e contribuição para a discussão quanto ao problema do abandono dos estudos acadêmicos no ensino superior, há um interesse no trabalho feito pelo grupo de se priorizar os resultados que sejam mais convenientes para a classe de "Dropout". Portanto, a escolha do F1-score como métrica principal se justifica pela natureza desequilibrada do problema, onde a classe "abandono" é menos recorrente nos dados. O F1-score, sendo uma média harmônica entre precisão e recall, oferece uma avaliação balanceada da performance do modelo, garantindo que tanto os falsos positivos quanto os falsos negativos sejam considerados. Isso é especialmente relevante para problemas onde a classe minoritária é de maior interesse, como o abandono acadêmico. A validação cruzada estratificada assegura que a distribuição das classes é preservada em cada fold, proporcionando uma avaliação robusta e realista do desempenho do modelo. Além disso, optamos pela acurácia para termos uma noção geral do desempenho do modelo, assim como a matriz de confusão que nos permite ter uma visão mais holística do desempenho distribuído em cada classe.

Análise de Componentes Principais

Foi feita a aplicação da análise de componentes principais na tentativa de diminuição da dimensionalidade do modelo. No entato, é possível observar que mesmo com a variável de maior interesse (abandono dos estudos) ter tido uma melhora considerável em seus acertos, a outra variável dependente do modelo (não abandono) na modelagem como caiu significativamente, fazendo com que a acurácia geral do modelo decaísse abaixo do satisfatório.

Ainda é possível observar que a Regressão Logística utilizada para este modelo era suficientemente eficiente, já que para ele, haviam 220 variáveis analisadas a princípio, mas que não haviam problemas com tempo de processamento ou escabilidade do volume de dados, então mesmo que com a aplicação do PCA tenha diminuído o tempo de processamento para mais da metade com até 10 componentes o tempo ganho. Além disso, para os valores feitas pelo PCA foi necessária a aplicação de mais de 150 componentes para atingir mais de 90% (como é possível observar na figura 2) de explicação em relação ao modelo original. Com 150 componentes, a diferença na dimensionalidade e a melhora no tempo não foi relevante o suficiente para justificar a utilização da abordagem no estudo.

Interpretabilidade

Para compreender os fatores que mais influenciam o abandono acadêmico, utilizamos a seguinte abordagem de interpretabilidade:

SHAP Values: Utilizamos SHAP (SHapley Additive exPlanations) para calcular a contribuição de cada feature para as previsões do modelo. Os valores SHAP fornecem uma medida unificada da importância das variáveis, mostrando a influência de cada variável em cada previsão individual. A visualização SHAP revelou que variáveis como "Curricular

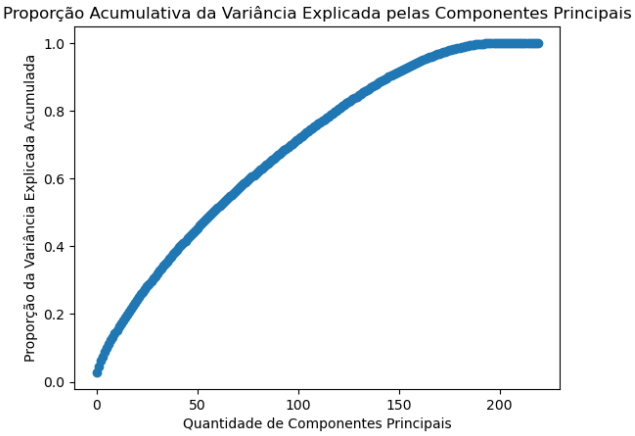

Figure 2: Proporção Acumulativa da Variância Explicada pelas Componentes Principais aplicada ao modelo do estudo.

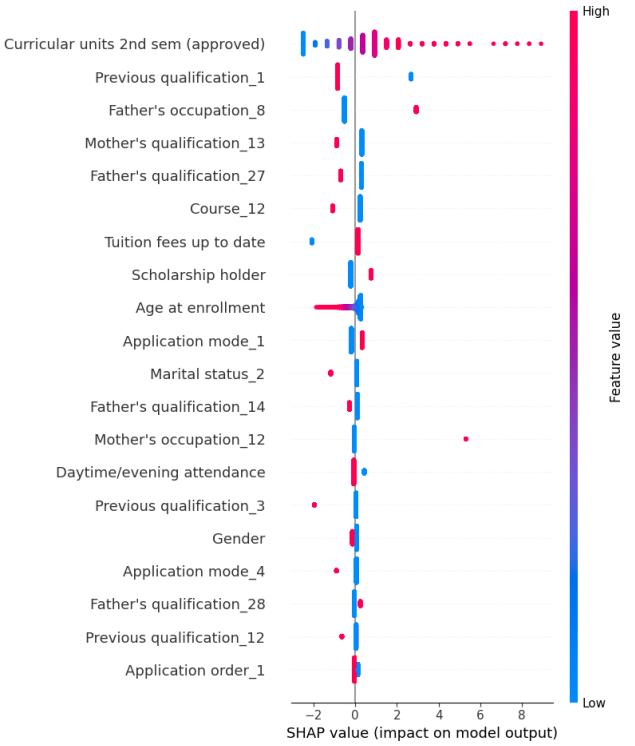

Figure 3: Sumário contendo as features mais relevantes para o modelo escolhido (ordem decrescente)

units 2nd sem (approved)", "Tuition fees up to date", 'Scholarship holder'", "Previous qualification", "Mother's qualification", "Father's qualification", "Mother's occupation", "Father's occupation" também são críticas para a predição do abandono.

Essas análises ajudaram a identificar que tanto o desempenho acadêmico quanto o suporte financeiro e educacional familiar são cruciais para a decisão de um aluno de permanecer ou abandonar o curso.

Conclusão

Este estudo mostra que a Regressão Logística, quando combinada com um pré-processamento adequado e balanceamento de classes, é altamente eficaz na predição do abandono acadêmico. Fatores socioeconômicos e acadêmicos emergiram como determinantes críticos, proporcionando insights valiosos para o desenvolvimento de intervenções que visam aumentar a retenção de estudantes. Para mais detalhes sobre a implementação e os dados utilizados, consulte o repositório no GitHub (https://github.com/felmateos/student-dropout-prediction).

References

[Bowyer et al. 2011] Bowyer, K. W.; Chawla, N. V.; Hall, L. O.; and Kegelmeyer, W. P. 2011. SMOTE: synthetic minority over-sampling technique. *CoRR* abs/1106.1813.

[Cunha and Morosini 2014] Cunha, E. R., and Morosini, M. C. 2014. Evasão na educação superior: uma temática em discussão. *Revista Cocar* 7(14):82–89.

[Jeatrakul, Wong, and Fung 2010] Jeatrakul, P.; Wong, K. W.; and Fung, C. C. 2010. Classification of imbalanced data by combining the complementary neural network and smote algorithm. In Wong, K. W.; Mendis, B. S. U.; and Bouzerdoum, A., eds., *Neural Information Processing. Models and Applications*, 152–159. Berlin, Heidelberg: Springer Berlin Heidelberg.

[Lima Coimbra and Costa 2021] Lima Coimbra, Camila, V. M. M. B. e. S. L., and Costa, N. C. D. 2021. A evasão na educação superior: definições e trajetórias. SciELO Brazil. DOI: https://doi.org/10.1590/S1678-4634202147228764.

[McKinney 2023] McKinney, W., ed. 2023. *Python para análise de dados: Tratamento de dados com Pandas, NumPy e Jupyter*. O'Reilly.

[Realinho and Baptista 2021] Realinho, Valentim, V. M. M. M. J., and Baptista, L. 2021. Predict Students' Dropout and Academic Success. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5MC89.