Uso do Teorema de Bayes para Classificação de Alunos com Alto Risco de Evasão Escolar

1st Ênio Henrique Nunes Ribeiro
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
ehnr@cin.ufpe.br

2nd Filipe Maciel Leicht
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
fml2@cin.ufpe.br

3rd Matheus Augusto Monte Silva
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
mams4@cin.ufpe.br

4th Thiago José Grangeiro Costa
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
tjgc@cin.ufpe.br

5th Victória Xavier Queiroz
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
vxq@cin.ufpe.br

***Abstract*—Este relatório tem como propósito documentar o projeto feito para a cadeira Estatística e Probabilidade, no qual o Teorema de Bayes foi usado para analisar e estudar dados estatísticos de uma base de dados que diz respeito a alunos com alto risco de evasão escolar.**

***Index Terms*—Alunos, Evasão, Prever, Bayes, Classificador**

I. Introdução

A evasão escolar é uma preocupação persistente que afeta sistemas educacionais em todo o mundo. Ela se refere ao fenômeno em que os alunos abandonam a escola antes de completar seu ciclo de estudos, seja no ensino fundamental, médio ou superior.

É notório que além de impactar o futuro educacional e profissional dos indivíduos, a evasão escolar também tem implicações sociais e econômicas, contribuindo para a desigualdade e limitando o potencial de desenvolvimento de uma sociedade. Portanto, compreender as causas e implementar estratégias eficazes para prevenir a evasão escolar é fundamental para garantir que todos os alunos tenham a oportunidade de uma educação de qualidade e o apoio necessário para alcançar seu pleno potencial.

II. Objetivos

Este relatório tem como objetivo investigar a evasão escolar sob uma perspectiva estatística, buscando compreender sua dinâmica por meio de análises quantitativas e identificando tendências significativas. Examinaremos não apenas a prevalência da evasão escolar, mas também as variáveis e fatores estatísticos que a influenciam.

Ao combinar dados empíricos com análises estatísticas, pretendemos revelar *insights* valiosos sobre as causas e os impactos da evasão escolar. Através da criação de um modelo feito com o Classificador Ingênuo de Bayes, será discutido como as informações estatísticas podem vir a servir como base para prever a probabilidade de um aluno de deixar sua Instituição Escolar.

III. Justificativas

Durante o percurso da disciplina, discutiu-se diversos modelos classificadores os quais podem desempenhar um papel crucial na previsão da evasão escolar de estudantes. A evasão escolar é um problema educacional de alcance global que afeta inúmeros alunos a cada ano. Fatores como ambiente familiar, fatores socioeconômicos, desempenho acadêmico e desmotivação estão frequentemente associados a esse fenômeno.

Portanto, é de incomensurável importância realizar o desenvolvimento de um modelo capaz de antecipar e alertar sobre o risco de evasão escolar, utilizando os conhecimentos adquiridos na cadeira, incluindo o uso do classificador ingênuo de Bayes e técnicas de aprendizado de máquina. Ao aplicar esses métodos, poderão ser criadas estratégias de prevenção mais eficazes e políticas educacionais direcionadas, assegurando que todos os alunos tenham a oportunidade de concluir sua educação de forma bem-sucedida.

IV. Metodologia

A. *Obtenção da Base de Dados*

A equipe escolheu a base de dados no *site* recomendado pelo professor da disciplina, *UC Irvine Machine Learning Repository*. É um *dataset* de uma pesquisa feita com os alunos do Instituo Politécnico de Portalegre, em Portugal. Nele, não há campos em branco e cada instância representa um estudante.

As fontes de dados utilizadas consistem em dados internos e externos da instituição e incluem dados de: Sistema de Gestão Acadêmica (AMS) da instituição, Sistema de Apoio à Atividade de Ensino da instituição (desenvolvido internamente e chamado de PAE), Dados anuais da Direção-Geral do Ensino Superior (DGES) sobre admissão através do Concurso Nacional de Acesso ao Ensino Superior (CNAES) e do Banco de Dados Contemporâneo de Portugal (PORDATA) contendo dados macroeconômicos.

Os dados se referem a registros de alunos matriculados nos anos acadêmicos de 2008/2009 (após a implementação do Processo de Bolonha no ensino superior na Europa) até 2018/2019. Isso inclui dados de 17 cursos de graduação de diferentes áreas do conhecimento, como agronomia, design, educação, enfermagem, jornalismo, gestão, serviço social e tecnologias.

B. Atributos da Base de Dados

O conjunto de dados inclui informações demográficas, dados socioeconômicos, dados macroeconômicos e dados acadêmicos referentes ao aluno.

A base de dados utilizada consiste em 4424 registros com 37 atributos, não contendo valores ausentes. A Tabela I descreve resumidamente os atributos utilizados no conjunto de dados, agrupados por classe: demográficos, socioeconômicos, macroeconômicos e dados acadêmicos.

TABLE I: Tabela de Classes e Atributos

Classe	Atributos
Dados Demográficos	Status civil Nacionalidade Gênero *Displaced* Intercambista Idade no ano de entrada
Dados Socioeconômicos	Ocupação e Qualificação dos pais Necessidades Educacionais Especiais Devedor Mensalidade em dia Bolsista
Dados Macroeconômicos	Produto Interno Bruto Taxa de Inflação Taxa de Desemprego
Dados Acadêmicos	Curso Turno Modo de Entrada Ordem de Escolha Qualificações Anteriores Nota da Qualificação Anterior Dados Curriculares do 1º Semestre Dados Curriculares do 2º Semestre

C. Pré-processamento da Base de Dados

*1) **Filtro de Instâncias**:* Esta etapa consiste em remover instâncias que não são consideradas relevantes ao estudo. Na base de dados pode ser observado um atributo chamado ***Target***, referente ao status do estudante em relação à Universidade, o qual poderia ter 3 classificações: Graduado, Desistente e Matriculado. Como o estudo busca saber se o aluno concluiu o curso de fato ou não, a classificação Matriculado não é interessante para nós no momento de fornecer os dados para o Classificador e, por isso, foi deixada de fora nessa fase da pesquisa *(Ver Apêndice)*.

*2) **Engenharia de Atributos**:* Esta etapa consiste em transformar os atributos para melhorar a qualidade dos dados e torná-los mais adequados para a modelagem de aprendizado de máquina. Foram realizadas 3 tipos de transformações nos dados: a simplificação no nome de alguns atributos (para trabalhar melhor com o *Dataset*), o agrupamento de dados (tornar os dados mais familiares para o modelo) e a unificação de atributos (criação de um parâmetro único a partir de dois ou mais atributos do banco de dados) *(ver Apêndice)*.

*3) **Seleção de Atributos**:* Esta etapa consiste em selecionar os atributos considerados úteis e que serão utilizados para treinar o algoritmo de aprendizado de máquina. Na Tabela II constam os 12 atributos que foram selecionados para realização do estudo (*ver Apêndice*), são eles:

TABLE II: Atributos selecionados

Atributos	
Curso	Devedor
Turno	Mensalidade em dia
Gênero	Qualificações anteriores
Necessidades Educacionais Especiais	Notas das qualificações anteriores
Idade no ano de entrada	Performance no 1º semestre
Modo de entrada	Performance no 2º semestre

D. Análise dos Dados

Para a análise dos dados, além do uso de notebooks Python para modelagem, foram empregadas ferramentas de *Business Intelligence* para visualização. Nesse contexto, os dados já modelados e transformados foram posteriormente exportados para o Microsoft Power BI, ferramenta de visualização, análise e compartilhamento de dados. No Power BI, foram elaborados Gráficos e Relatórios destinados a apresentar as estatísticas relevantes dos dados.

E. Algoritmo de Aprendizado de Máquina

Para o *Dataset* foi escolhido o modelo *Categorical Naive Bayes*, fornecido pela biblioteca "*Scikit-Learn*", o

qual funciona melhor em análises nas quais os parâmetros utilizados estão divididos em categorias. A estrutura dos dados do *Dataset* favorece essa categorização, uma vez que é possível transformar muitos dos parâmetros contínuos em discretos, além de simplificar outros parâmetros mais complexos em classes, para melhor utilização do modelo.

1) ***Naive Bayes****:* É um algoritmo de classificação de dados utilizado para "*Machine Learning*". Como um classificador, ele analisa os dados fornecidos e oferece uma resposta com base em critérios estabelecidos pelo próprio algoritmo, fazendo a suposição que todos os parâmetros passados são independentes entre si.

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Fig. 1: Teorema de Bayes

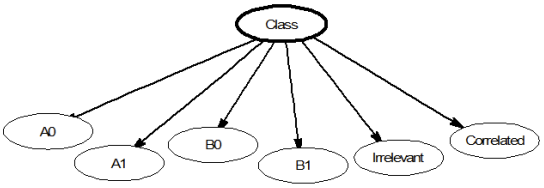

Fig. 2: Representação gráfica de um classificador Naive Bayes.

2) ***Categorical Naive Bayes****:* Esse modelo do *Naive Bayes* implementa o algoritmo do *Categorical Naive Bayes* para dados categoricamente distribuidos. Ele assume que cada item descrito pelo index tem a sua própria distribuição, que a matriz está codificada e que todas as categorias estão representadas por números positivos. As mudanças realizadas no pré-processamento da base de dados tiveram como objetivo molda-los para melhor uso nesse modelo.

$$P(x_i = t \mid y = c\,;\, \alpha) = \frac{N_{tic} + \alpha}{N_c + \alpha n_i}$$

Fig. 3: Algoritmo "Categorical Naive Bayes"

V. Análise Exploratória dos Dados

A análise exploratória de dados (AED) desempenha um papel crucial na preparação e implementação de algoritmos preditivos, permitindo a identificação de padrões, a detecção de *outliers*, a seleção de recursos relevantes e a validação de suposições sobre os dados. Além disso, a AED fornece *insights* valiosos sobre as relações entre as variáveis, auxiliando na adaptação dos algoritmos ao contexto específico da pesquisa e na melhoria da interpretação dos resultados, tornando-se, assim, uma etapa fundamental para o sucesso das análises preditivas.

Dessa maneira, ao analisar os Dados do *Dataset* escolhido, foi realizada dois tipos de análise, a primeira considerando as quantidades absolutas dos estudantes desistentes e a segunda considerando a % de desisitencias por cada classe de dados. dessa maneira, foi possível inferir conclusões e insights determinantes para a implementação dos algoritmos preditivos.

A. *Análise Demográfica*

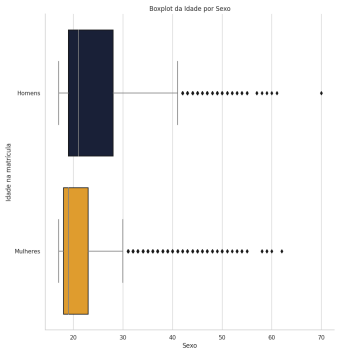

Fig. 4: Gráfico 1

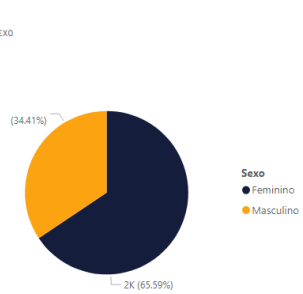

Fig. 5: Gráfico 2

Na primeira instância, procedemos à análise dos dados demográficos gerais presentes no conjunto de dados. Através da análise realizada nos Gráficos 1 e 2, podemos obter algumas conclusões preliminares significativas. O Gráfico 1 revela que a maioria dos registros está concentrada na faixa etária entre 17 e 30 anos, destacando-se a presença considerável de valores discrepantes (*outliers*). Por outro lado, o Gráfico 2 evidencia que a maioria dos estudantes é do sexo feminino, um fato que possivelmente terá implicações de interesse ao longo da análise.

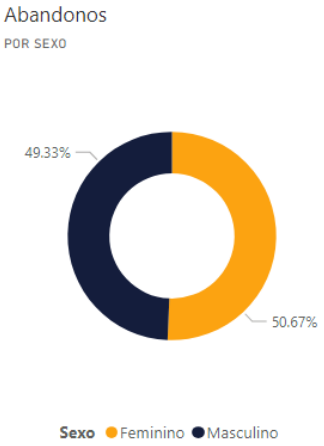

Fig. 6: Gráfico 3

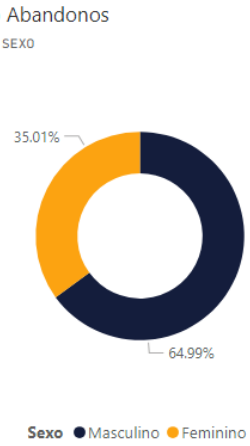

Fig. 7: Gráfico 4

Ao comparar os dois gráficos, é possível notar uma diferença significativa no percentual de abandono em cada caso. No Gráfico 3, estamos analisando os abandonos absolutos por sexo, e podemos observar que a quantidade de abandonos escolares é praticamente igual entre os sexos.

No entanto, ao analisarmos o Gráfico 4, podemos notar que o percentual de abandono é substancialmente maior entre os homens. Isso se deve ao fato de a quantidade de mulheres ser maior em relação ao homens no total geral, como descrito no Gráfico 1. Dessa maneira, podemos supor que a probabilidade de um homem abandonar é maior do que a de uma mulher, o que torna esse dado relevante para o algoritmo.

B. Análise Acadêmica

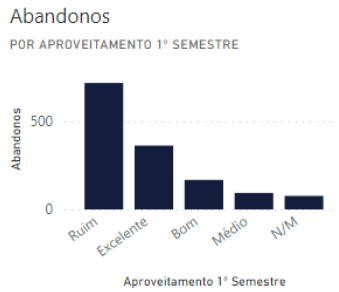

Fig. 8: Gráfico 5

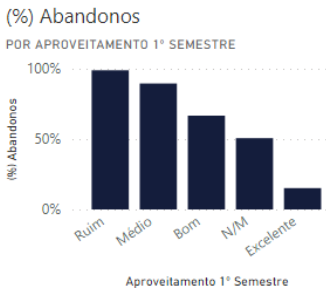

Fig. 9: Gráfico 6

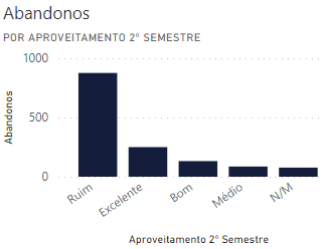

Fig. 10: Gráfico 7

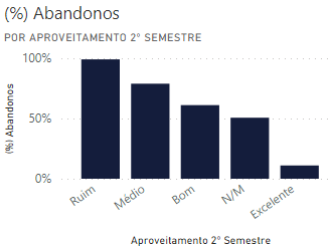

Fig. 11: Gráfico 8

Ao analisar o desempenho nos 1º e 2º semestres, destaca-se um achado intrigante. Nos Gráficos 5 e 7, observa-se que, em termos quantitativos, os alunos que alcançaram um desempenho excelente, de acordo com o tratamento dos dados, apresentam a segunda maior quantidade de casos de abandono, sendo superados apenas pelos alunos com desempenho ruim, conforme previsto. No entanto, ao examinar a taxa percentual de evasão em cada categoria, constata-se que os alunos com desempenho excelente exibem a menor taxa de evasão, um fenômeno que se repete no 2º semestre. Além disso, é notável que a quantidade de abandonos no 2º semestre supera a do 1º, o que sugere uma previsão mais precisa à medida que mais dados relacionados ao desempenho escolar são incorporados.

VI. Análise de Resultados

Apenas aplicar o modelo sem o pré-processamento apresentou resultados pouco satisfatórios, com um modelo de precisão de aproximadamente 87,13%. O tratamento dos dados levou a um aumento de 2,25% de precisão, com a maior parte desse ganho devido à unificação dos atributos referentes a quantidade de cadeiras que o aluno se matriculou e a quantidade de cadeiras nas quais foi aprovado nos primeiros semestres. Certos atributos mostraram reduzir a precisão do modelo, como *GDP*, qualificação dos pais e *Inflation Rate*. Remover esses atributos levou a mais um aumento na acurácia, elevando-a a 90,16%. Esses resultados sugerem que esses atributos não são tão relevantes para predizer se um estudante irá se graduar ou não, principalmente quando comparados com o desempenho do aluno nos primeiros semestres.

VII. Conclusões Finais

A predição do abandono escolar é importante para que instituições de ensino possam auxiliar alunos que tem alto risco de evasão antes que eles deixem seus cursos. Foi utilizada uma base de dados portuguesa, cujos atributos eram propícios para a utilização do *Categorical Naive Bayes* utilizado como modelo para o *Machine Learning*. Após o tratamento dos dados ficou aparente que o desempenho do aluno tinha muito mais eficácia para prever o seu futuro acadêmico do que qualquer atributo referente a qualidade de vida ou educação prévia. Embora seja possível que as condições de vida de uma pessoa afetem negativamente suas possibilidades de ingressar em uma instituição de ensino superior, aquelas que conseguem ingressar devem ter suas chances de graduação avaliadas não por fatores externos, mas por seu desempenho atual na instituição. Um estudo mais aprofundado pode ser feito com o auxílio de especialistas em educação e comportamento humano para filtrar melhor os parâmetros que podem ser utilizado e que categorias abrangem melhor os diferentes dados dos atributos, para criar um modelo mais eficaz e preciso.

References

[1] V. Realinho, M. Vieira Martins, J. Machado, e L. Baptista, “Predict students’ dropout and academic success”, UC Irvine Machine Learning Repository [https://doi.org/10.24432/C5MC89], 2021.
[2] Scikit-learn Machine Learning in Python - Naive Bayes [https://scikit-learn.org/stable/modules/naive_bayes.html]
[3] StackExchange - Understanding Naive Bayes [https://stats.stackexchange.com/questions/21822/understanding-naive-bayes]
[4] MDPI - Predicting Student Dropout and Academic Success [https://www.mdpi.com/2306-5729/7/11/146]

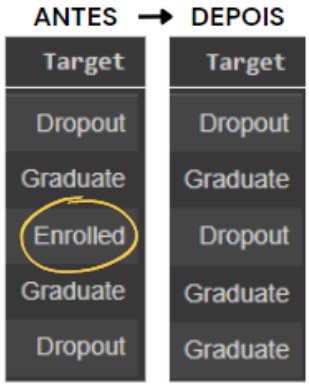

Fig. 12: Comparativo entre as instâncias no atributo *Target* antes e depois da aplicação de um filtro de instâncias.

```python
#Aplicando um filtro para retirar a coluna "Target" que é o nosso objetivo
filter1 = dados[(dados['Target'] != 'Enrolled')]
```

Fig. 13: Implementação do filtro de instâncias.

```python
# Renomeando atributos
dados.rename(columns = {'Application mode':'ApplicationMode'}, inplace = True)
dados.rename(columns = {'Previous qualification':'PreviousQualification'}, inplace = True)
dados.rename(columns = {'Previous qualification (grade)':'PreviousGrade'}, inplace = True)
dados.rename(columns = {'Curricular units 1st sem (enrolled)':'Enrolled1st'}, inplace = True)
dados.rename(columns = {'Curricular units 1st sem (approved)':'Approved1st'}, inplace = True)
dados.rename(columns = {'Curricular units 2nd sem (enrolled)':'Enrolled2nd'}, inplace = True)
dados.rename(columns = {'Curricular units 2nd sem (approved)':'Approved2nd'}, inplace = True)
dados.rename(columns = {'Approved1st':'1st semester performance'}, inplace = True)
dados.rename(columns = {'Approved2nd':'2nd semester performance'}, inplace = True)
```

Fig. 14: Alguns atributos foram renomeados.

```python
# Agrupando dados para melhorar o CategoricalNB
for a in range(4424):
  #Modo de entrada
  if dados.ApplicationMode[a] in [1, 2, 10, 17, 18]: dados.ApplicationMode[a] = 0 #Entrada normal
  elif dados.ApplicationMode[a] in [5, 15, 16, 26]: dados.ApplicationMode[a] = 1 #Entrada especial
  elif dados.ApplicationMode[a] in [27, 42, 43, 51, 57]: dados.ApplicationMode[a] = 2 #Mudança de curso/Transferência
  else: dados.ApplicationMode[a] = 3 #Já tem diploma
  #Previous education
  if dados.PreviousQualification[a] in [2, 3, 4, 5, 6, 39, 40, 42, 43]: dados.PreviousQualification[a] = 0 #Ensino superior
  elif dados.PreviousQualification[a] in [9, 10, 12, 14, 15]: dados.PreviousQualification[a] = 2 #Ensino médio incompleto
  elif dados.PreviousQualification[a] != 1: dados.PreviousQualification[a] = 3 #Fundamental incompleto
```

Fig. 15: Alguns atributos foram agrupados.

```python
# Tornando as variáveis "contínuas" em categorias para facilitar o CatergoricalNB
for a in range(4424):
  if dados.PreviousGrade[a] < 111: #Muito ruim
    dados.PreviousGrade[a] = 0
  elif dados.PreviousGrade[a] < 131: #Ruim
    dados.PreviousGrade[a] = 1
  elif dados.PreviousGrade[a] < 151: #Médio
    dados.PreviousGrade[a] = 2
  elif dados.PreviousGrade[a] < 171: #Bom
    dados.PreviousGrade[a] = 3
  else: dados.PreviousGrade[a] = 4 #Muito bom
```

Fig. 16: Alguns atributos foram categorizados.

```python
# Converter as colunas de matrícula/aprovação dos semestres em uma única coluna de proporção entre elas
for a in range(4424):
  if dados.Enrolled1st[a] == 0: dados.Approved1st[a] = 4 #Não se matriculou
  else:
    ratio1st = dados.Approved1st[a] / dados.Enrolled1st[a]
    if (ratio1st > 0.9) & (ratio1st <= 1): dados.Approved1st[a] = 0 #Excelente desempenho
    elif (ratio1st > 0.7) & (ratio1st <= 0.9): dados.Approved1st[a] = 1 #Bom desempenho
    elif (ratio1st > 0.5) & (ratio1st <= 0.7): dados.Approved1st[a] = 2
    else: dados.Approved1st[a] = 3 #Mal desempenho
  if dados.Enrolled2nd[a] == 0: dados.Approved2nd[a] = 4 #Não se matriculou
  else:
    ratio2nd = dados.Approved2nd[a] / dados.Enrolled2nd[a]
    if (ratio2nd > 0.9) & (ratio2nd <= 1): dados.Approved2nd[a] = 0 #Excelente desempenho
    elif (ratio2nd > 0.7) & (ratio2nd <= 0.9): dados.Approved2nd[a] = 1 #Bom desempenho
    elif (ratio2nd > 0.5) & (ratio2nd <= 0.7): dados.Approved2nd[a] = 2 #Médio desempenho
    else: dados.Approved2nd[a] = 3 #Mal desempenho
```

Fig. 17: Alguns atributos foram unificados.

```python
#Filtrando os atributos que não serão utilizados
filter.drop(['Nacionality', 'Marital status', 'Application order', 'Admission grade',
             'Target', "Mother's qualification", "Father's qualification",
             "Mother's occupation", "Father's occupation", 'GDP', 'Displaced',
             'Scholarship holder', 'International', 'Curricular units 1st sem (credited)',
             'Curricular units 1st sem (evaluations)', 'Curricular units 1st sem (grade)',
             'Curricular units 1st sem (without evaluations)', 'Curricular units 2nd sem (credited)',
             'Curricular units 2nd sem (evaluations)', 'Curricular units 2nd sem (without evaluations)',
             'Inflation rate', 'Unemployment rate'], inplace=True, axis=1)
```

Fig. 18: Atributos retirados do modelo para testes.

```python
import numpy as np
import pandas as pd
from sklearn.naive_bayes import CategoricalNB
```

Fig. 19: Bibliotecas importadas utilizadas no projeto.