

Projeto de Estatística

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



Nossa Equipe:



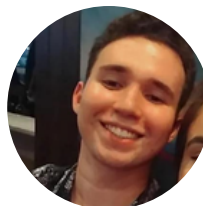
Ênio Henrique Nunes Ribeiro



Filipe Maciel Leicht



Matheus Augusto Monte Silva



Thiago José Grangeiro Costa



Victória Xavier Queiroz

Fases da Pesquisa

1

Entendendo o Problema

2

Processamento da Base de Dados

3

Análise Estatística dos Dados

4

Construção e Treinamento do Modelo

5

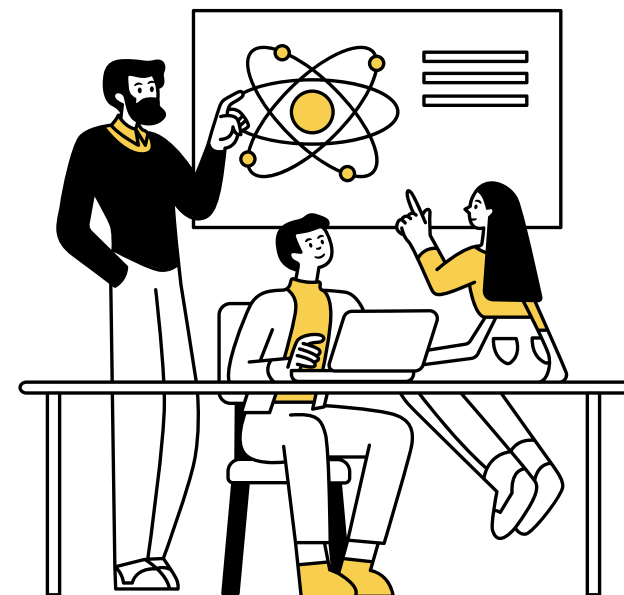
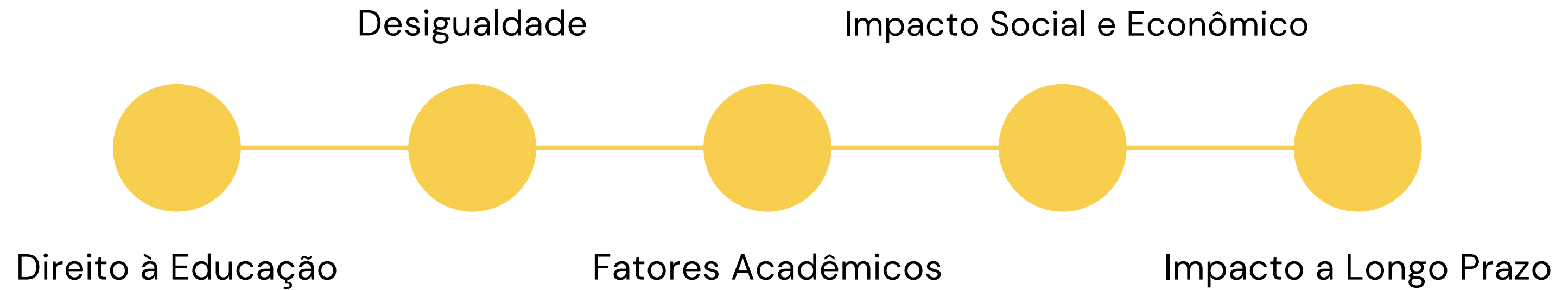
Conclusões

1

Entendendo o Problema



Contexto



2

Processamento da Base de Dados



Fases de Manipulação da Base de Dados

1

Obtenção da Base de Dados

2

Filtro de Instâncias

3

Engenharia de atributos

4

Seleção de Atributos

Obtenção da Base de Dados



Predict students' dropout and academic success

Donated on 12/12/2021

A dataset created from a higher education institution (acquired from several disjoint databases) related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and...

▼

Dataset Characteristics	Subject Area	Associated Tasks
Tabular	Other	Classification
Attribute Type	# Instances	# Attributes
-	4424	36

Creators

Valentim Realinho
vrealinho@
Instituto Politécnico de Portalegre

Mónica Vieira Martins
mvmartins@ipportalegre.pt
Instituto Politécnico de Portalegre

Jorge Machado
jmachado@ipportalegre.pt
Instituto Politécnico de Portalegre

Luís Baptista
lmbt@ipportalegre.pt
Instituto Politécnico de Portalegre

Predict students' dropout and academic success



Atributos da Base de Dados

Classes e seus Atributos

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

Pré-Processamento da Base de Dados

Filtro de Instâncias

Na base de dados há 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 e, por isso, foi deixada de fora.

ANTES → DEPOIS

Target	Target
Dropout	Dropout
Graduate	Graduate
Enrolled	Dropout
Graduate	Graduate
Dropout	Graduate

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



Pré-Processamento da Base de Dados

Engenharia de Atributos

1

Simplificação do Nome de Alguns Atributos



Remover caracteres especiais para otimizar o nome das variáveis em geral

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

Pré-Processamento da Base de Dados

Engenharia de Atributos

2

Agrupamento de Dados



Tornar os dados mais familiares para o entendimento do Naive Bayes

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

Pré-Processamento da Base de Dados

Engenharia de Atributos

3

Categorizar Variáveis



Metrificar dados. Neste caso, notas, em 5 categorias.

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

Pré-Processamento da Base de Dados

Engenharia de Atributos

4

Unificar Atributos



Criar um parâmetro único para classificar a performance do aluno em um semestre, a partir da fusão das colunas referentes a desempenho acadêmico

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

Pré-Processamento da Base de Dados

Seleção de Atributos

1	Modo de Entrada	5	Notas (Qualificação Anterior)	9	Gênero
2	Curso	6	Necessidades Educacionais Especiais	10	Idade no Ano de Entrada
3	Turno	7	Devedor	11	Performance no 1º Semestre
4	Qualificações Anteriores	8	Mensalidade em Dia	12	Performance no 2º Semestre

Pré-Processamento da Base de Dados

Seleção de Atributos



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


3

Análise Estatística dos Dados



Estudo Exploratório

[View in Power BI](#)



3630

Matriculados



1421

Abandonos

39,15%



2209

Graduados

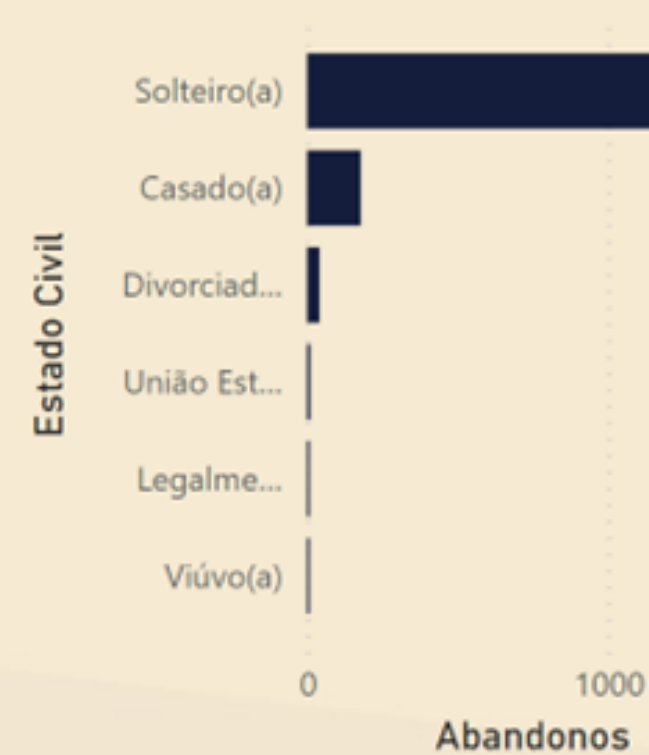
60,85%

Abandonos por Sexo

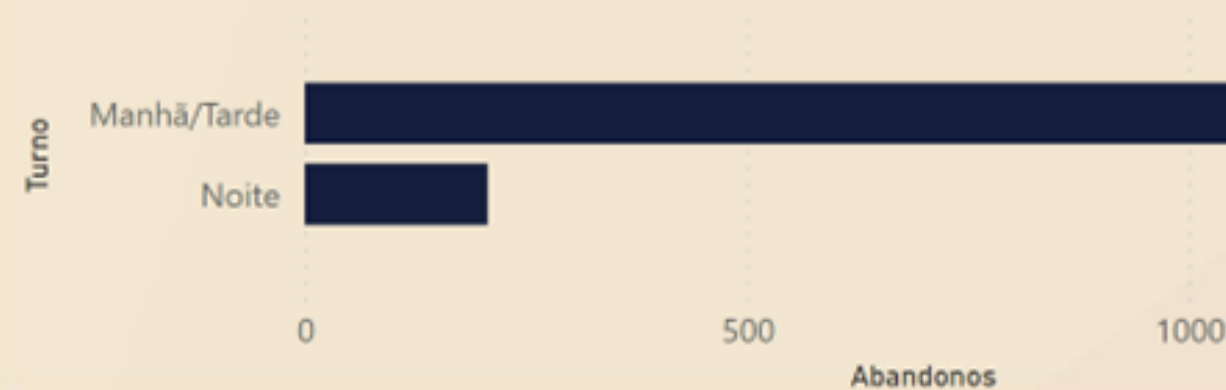


Sexo ● Feminino ● Masculino

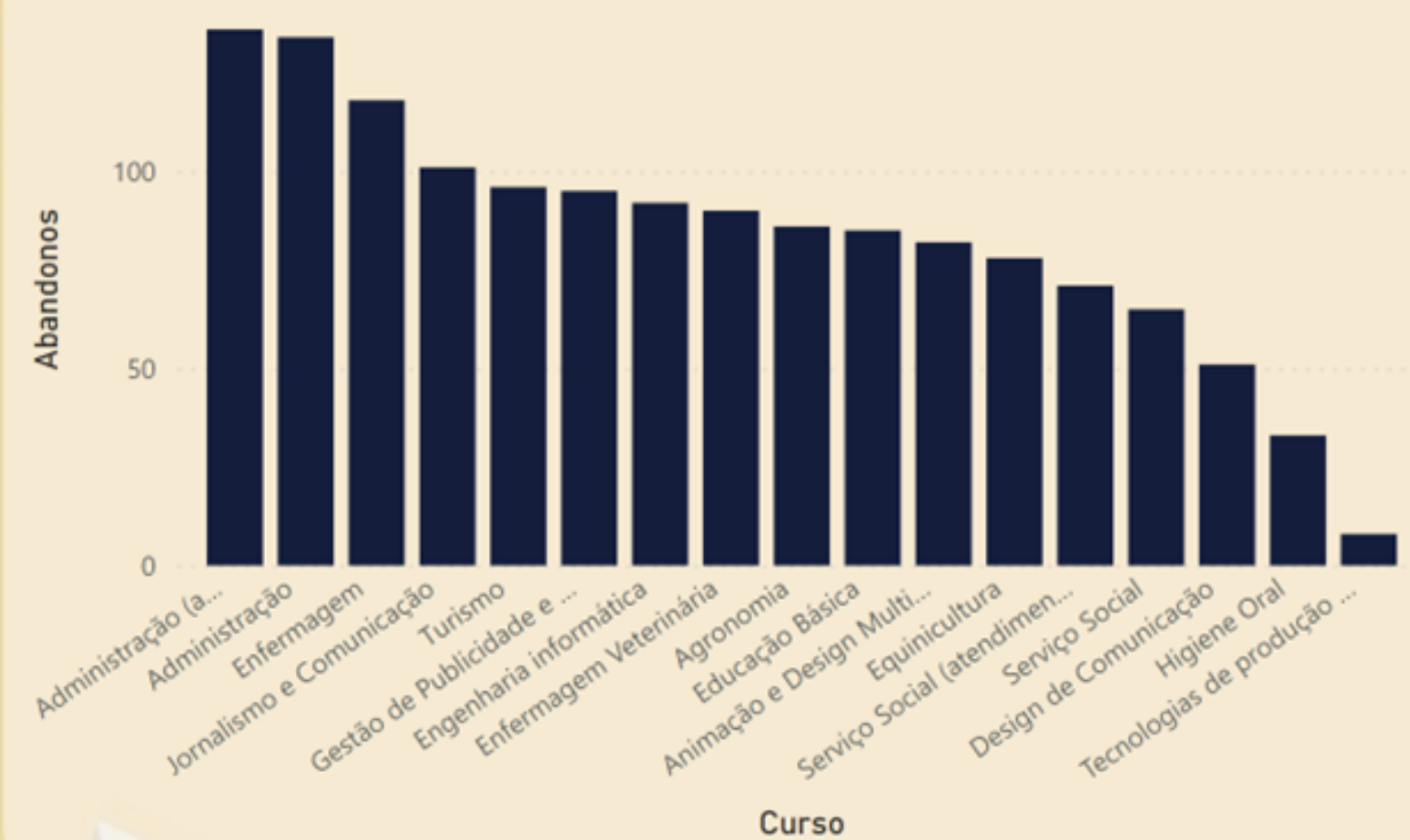
Abandonos por Estado Civil



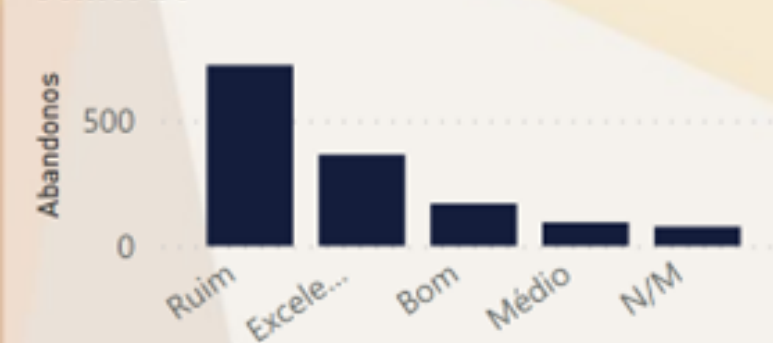
Abandonos por Turno



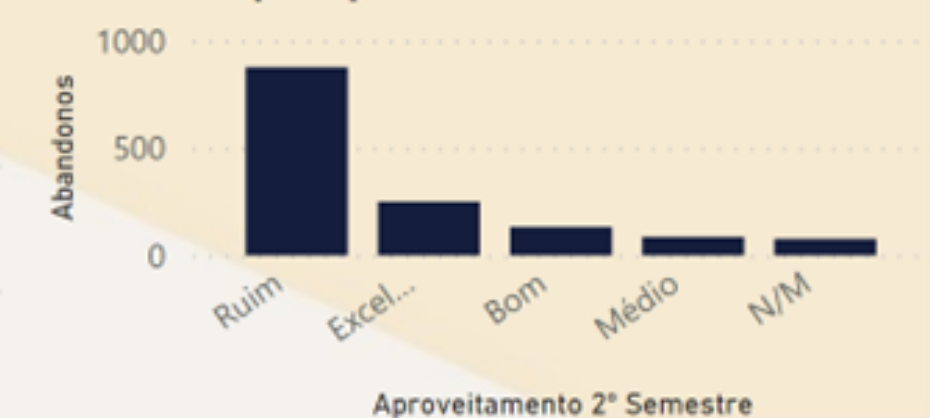
Abandonos por Curso



Abandonos por Aproveitamento 1º Semestre



Abandonos por Aproveitamento 2º Semestre



Estudo Exploratório

[View in Power BI](#)



3630

Matriculados



1421

Abandonos

39,15%

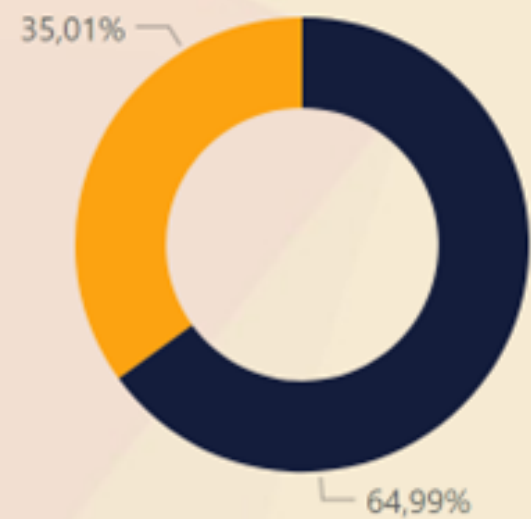


2209

Graduados

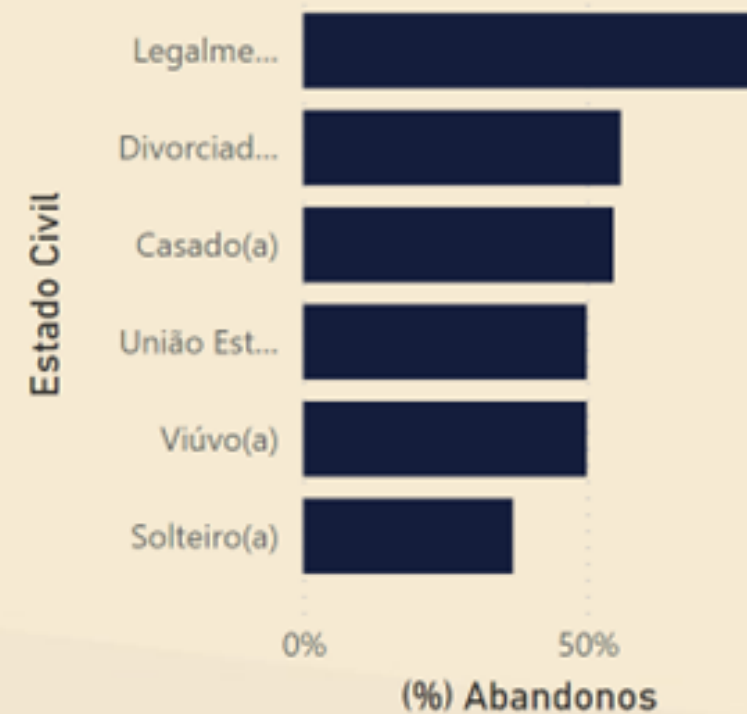
60,85%

(%) Abandonos por Sexo

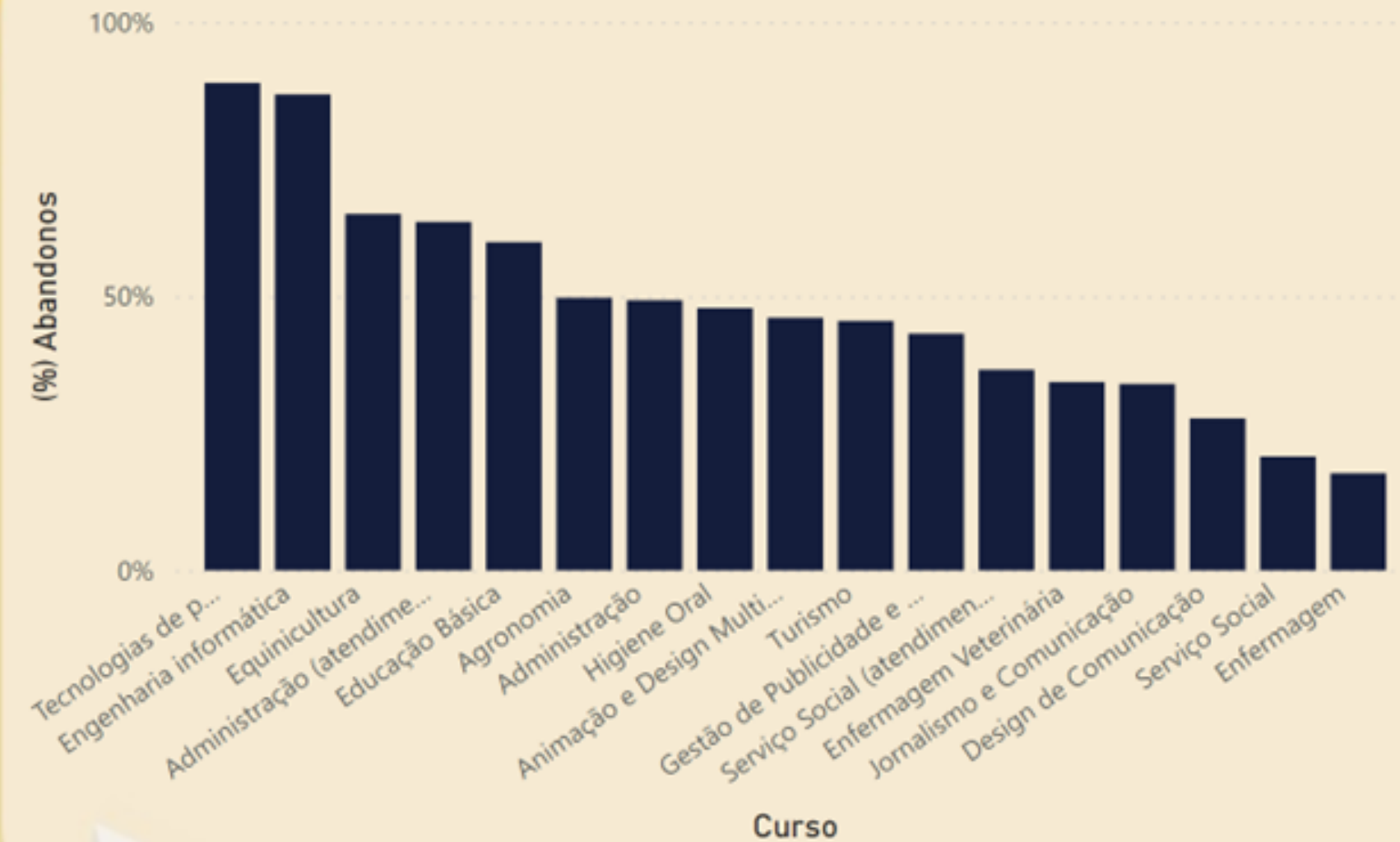


Sexo ● Masculino ● Feminino

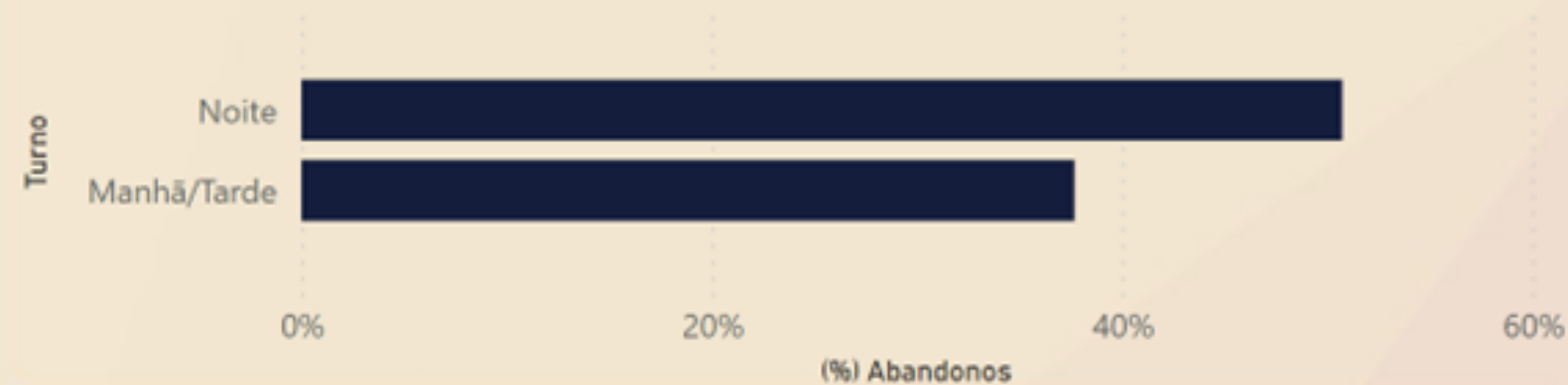
(%) Abandonos por Estado Civil



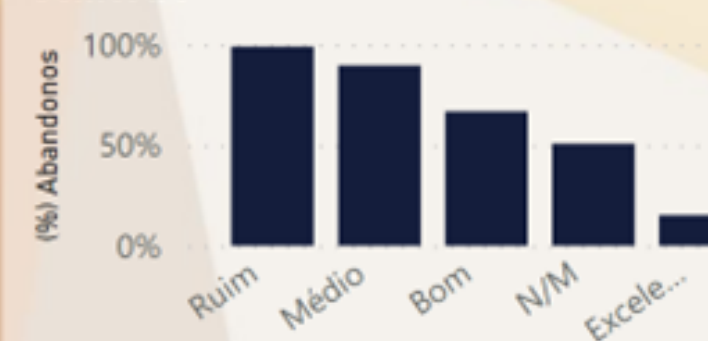
(%) Abandonos por Curso



(%) Abandonos por Turno



(%) Abandonos por Aproveitamento 1º Semestre



(%) Abandonos por Aproveitamento 2º Semestre



4

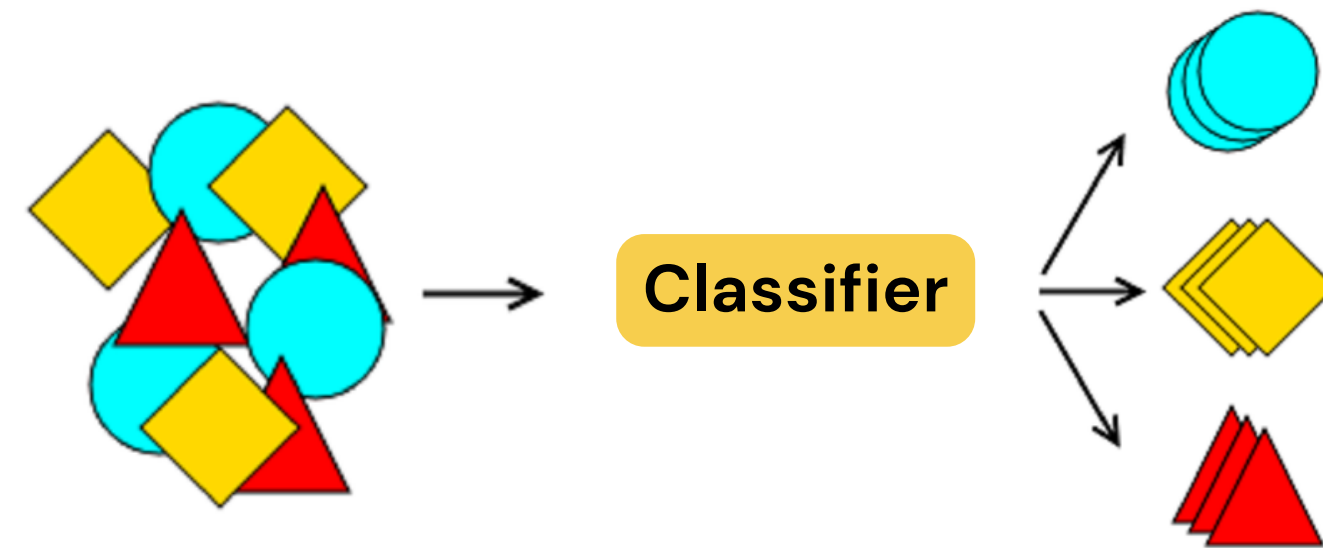
Construção e Treinamento do Modelo



Naive Bayes

Algoritmo de classificação probabilístico baseado no **Teorema de Bayes**

- 1 Gaussian Naive Bayes
- 2 Multinomial Naive Bayes
- 3 Complement Naive Bayes
- 4 Bernoulli Naive Bayes
- 5 Categorical Naive Bayes



Categorical Naive Bayes

É uma variação do algoritmo **Naive Bayes** projetada especificamente para lidar com dados categóricos

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



```
1 import numpy as np
2 import pandas as pd
3 from sklearn.naive_bayes import CategoricalNB
```

Implementação

Nossa implementação do algoritmo aplicado à base de dados estudada:



[Google Colaboratory](#)



5

Conclusões

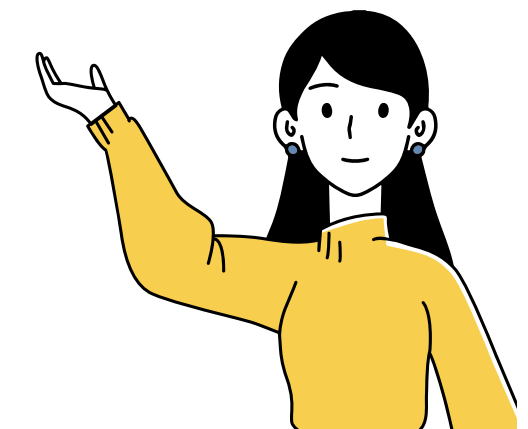


Nossas Conclusões

Notamos a necessidade de alterar a forma como os dados eram apresentados para **maximizar a precisão** do modelo utilizado

Percebemos que analisar os dados relativos à vida de um estudante para prever se ele vai ou não abandonar o curso nos leva a **conclusões imprecisas**

O dado mais impactante no nosso modelo é o **desempenho do semestre do estudante**, o resto não apresentou impacto significativo



Obrigado pela Atenção!

