

Classificando o desempenho dos candidatos do ENEM 2021 através de dados socioeconômicos

Estatística e probabilidade
Projeto - Classificador Bayesiano



INTEGRANTES DA EQUIPE



KAILANE FELIX



PEDRO CÉSAR



GABRIEL LOPES



LUÍSA MENDES



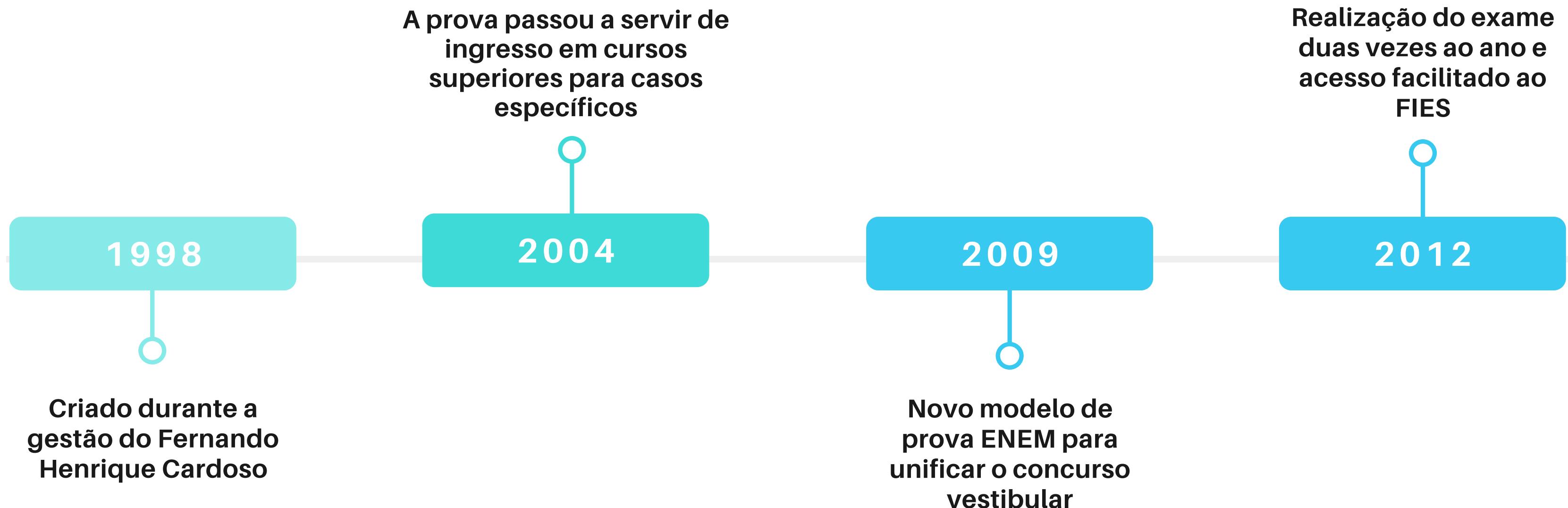
MARIO MOTA

Contexto e motivação

Por que escolhemos esse conjunto de dados?

HISTÓRICO DO EXAME

Como o ENEM surgiu?





CLASSIFICAR DESEMPENHO

Nosso objetivo é classificar o desempenho na prova de matemática dos estudantes que realizaram o ENEM 2021



DISPARIDADE SOCIAL

O Brasil é um dos países mais desiguais do mundo, tanto em termos sociais quanto em termos econômicos



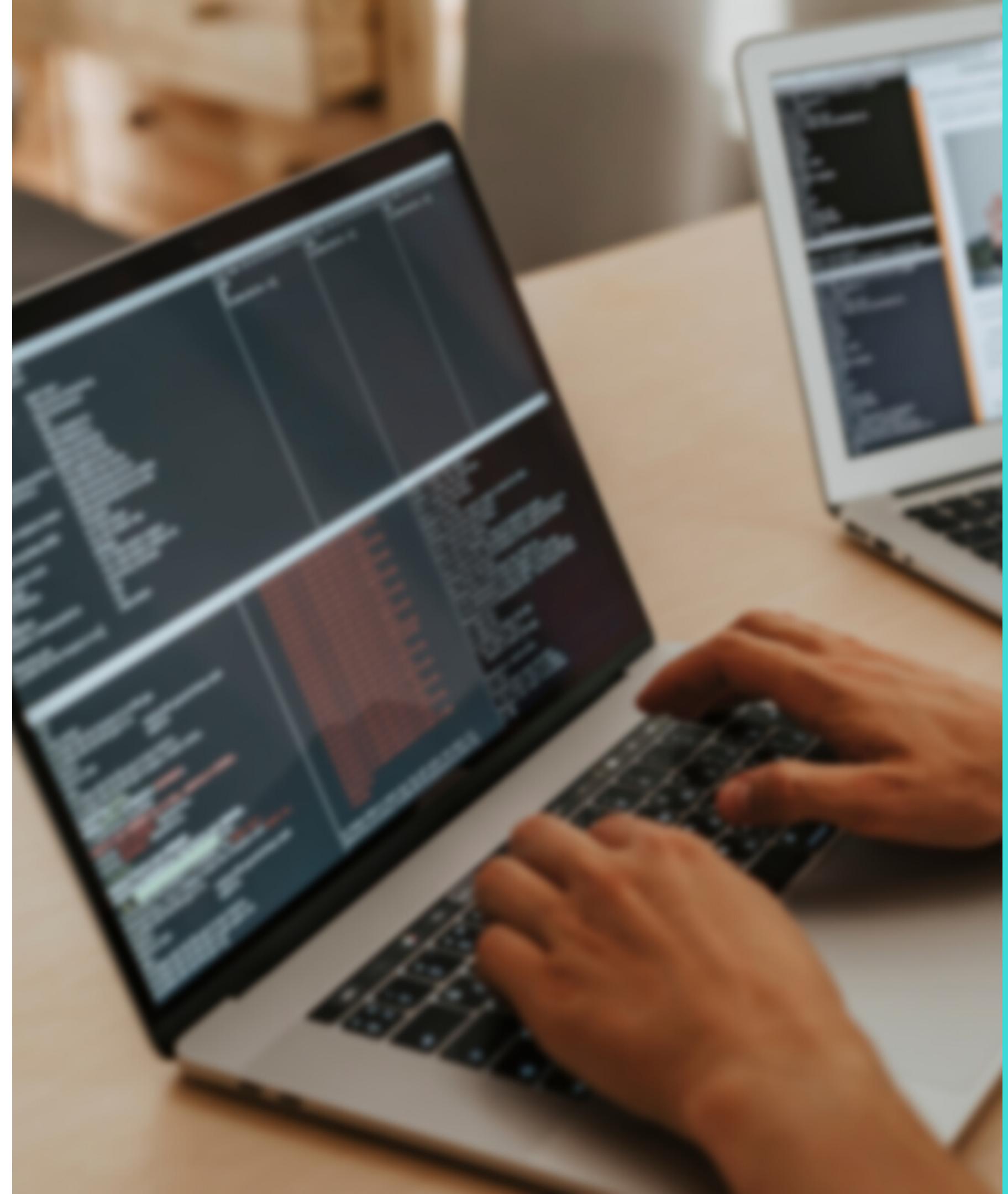
PANDEMIA DO CORONAVÍRUS

A discrepância entre a educação dos alunos mais/menos favorecidos economicamente, foi agravada



UMA ANÁLISE CRÍTICA

Escolhemos analisar esse tema a fim de revelar algo muito além da disparidade de notas no exame



Metodologia

Como iremos desenvolver o projeto?

NOSSOS PASSOS PARA O DESENVOLVIMENTO DO PROJETO

ENTENDER O
PROBLEMA

ADQUIRIR
E TRATAR
OS DADOS

ANÁLISE
EXPLORATÓRIA
DOS DADOS

APLICAR
MODELOS
DE MACHINE
LEARNING

AVALIAR E
APRESENTAR
OS MODELOS

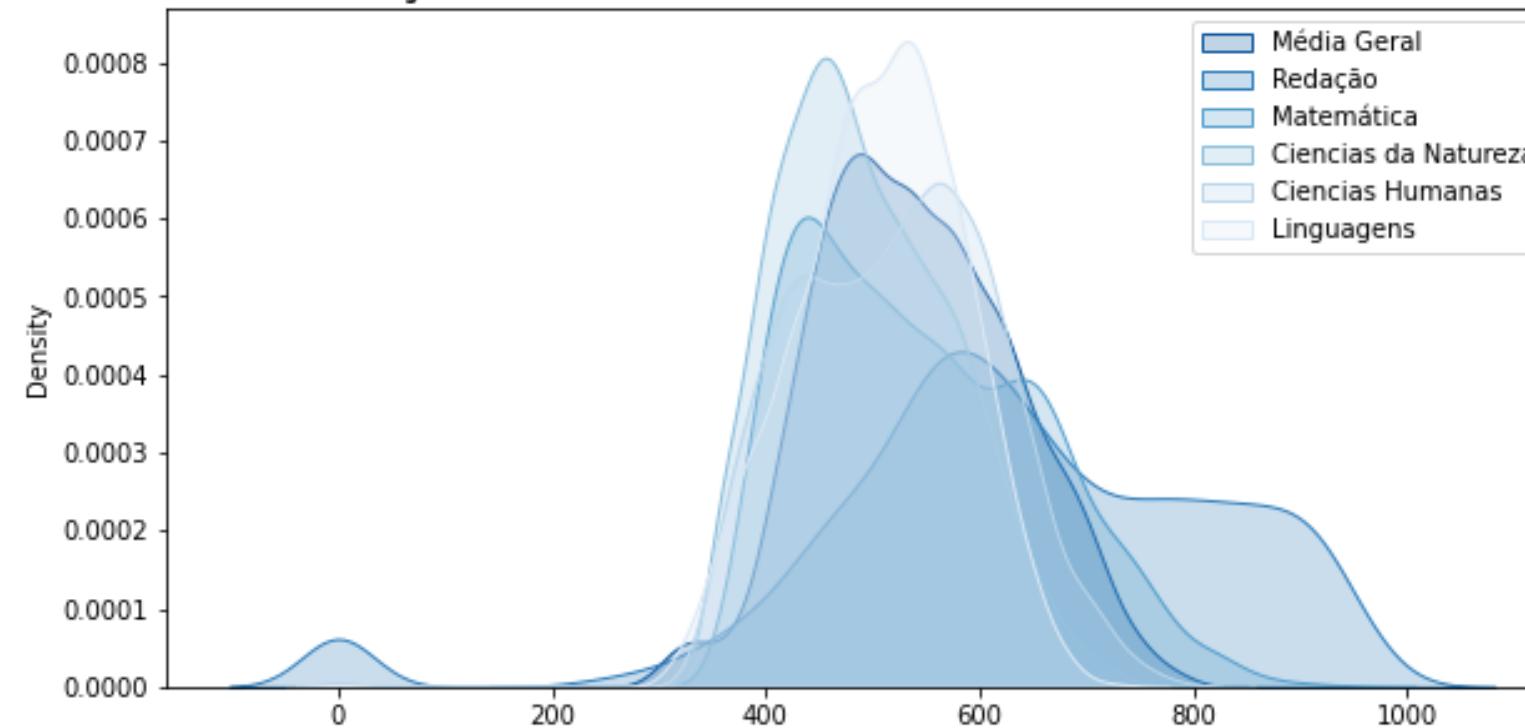
Análise exploratória dos dados

Principais insights das análises

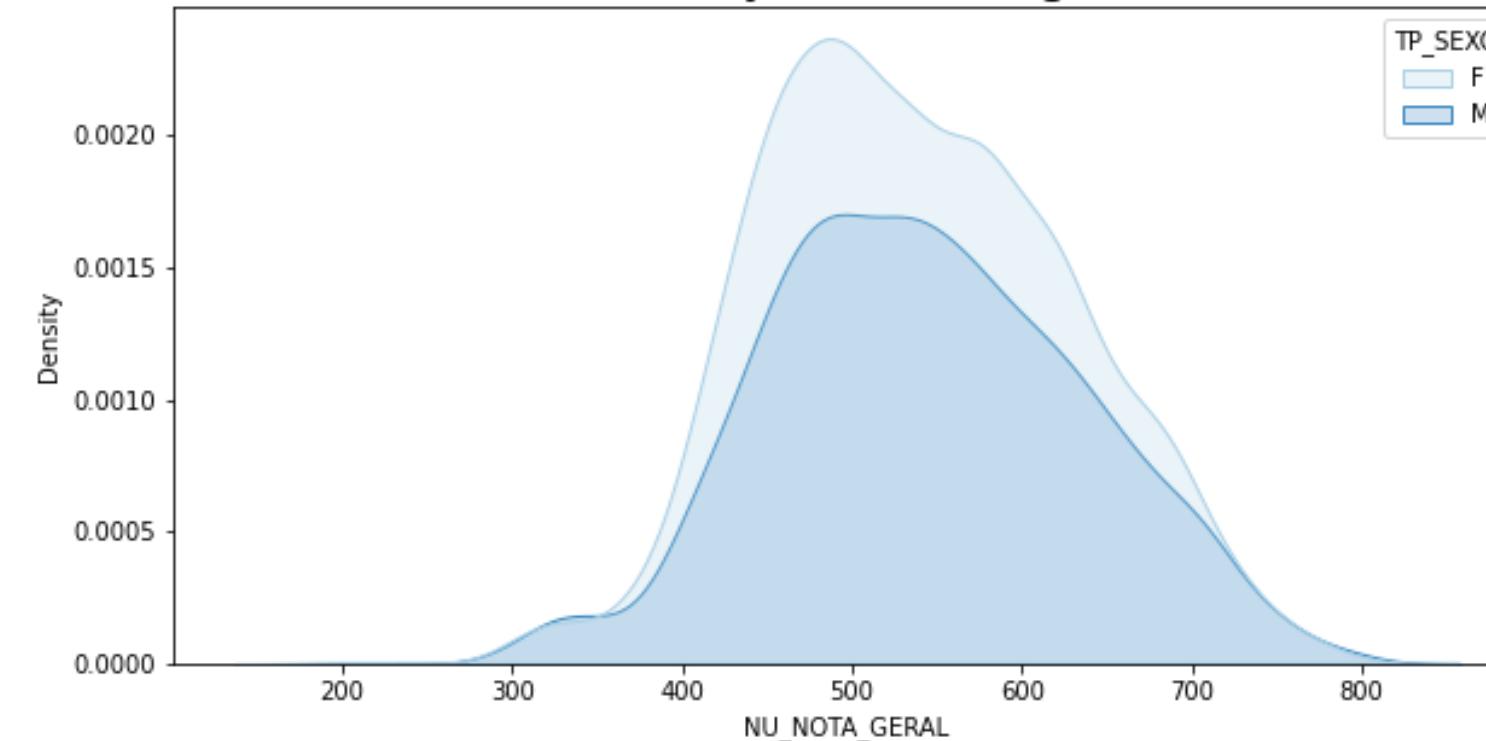
1

QUAL A ARÉA QUE MELHOR REPRESENTA A MÉDIA GERAL?

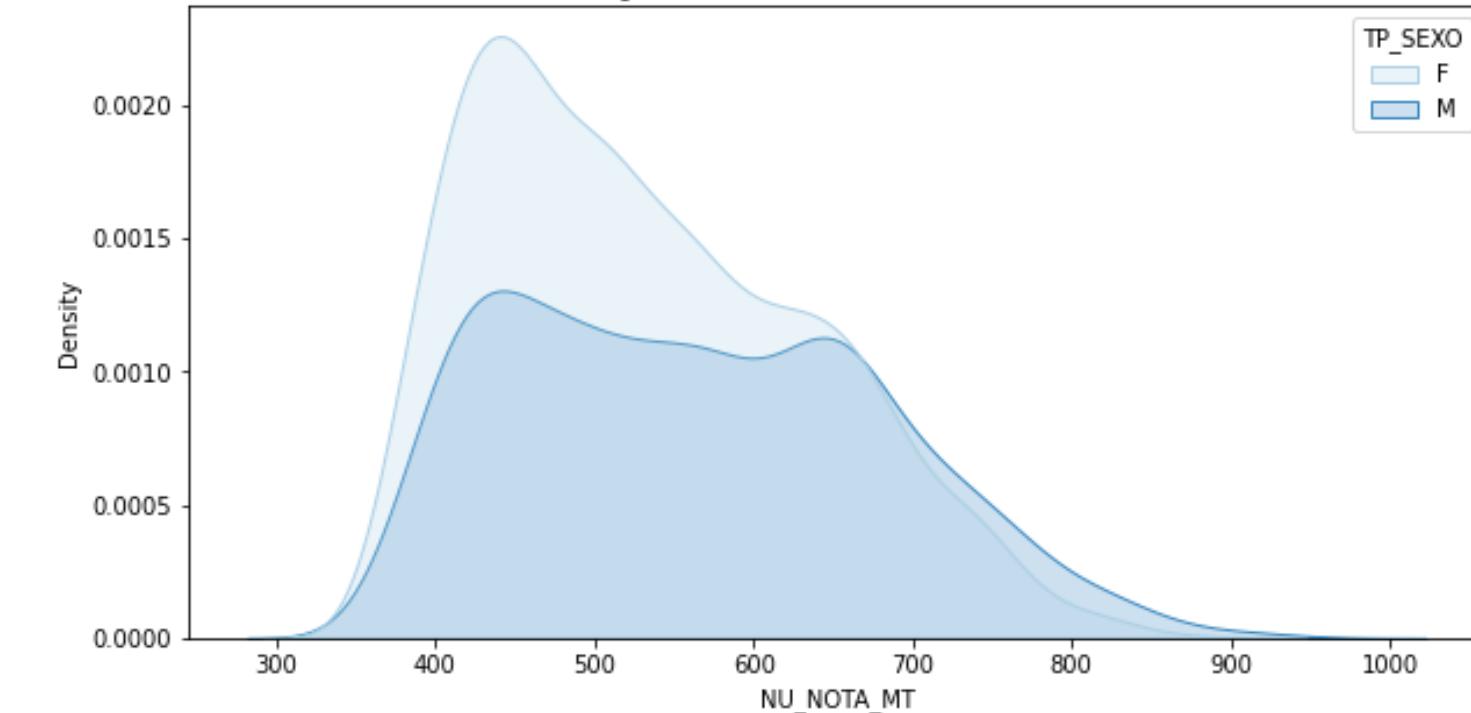
Distribuição das notas de cada área do conhecimento



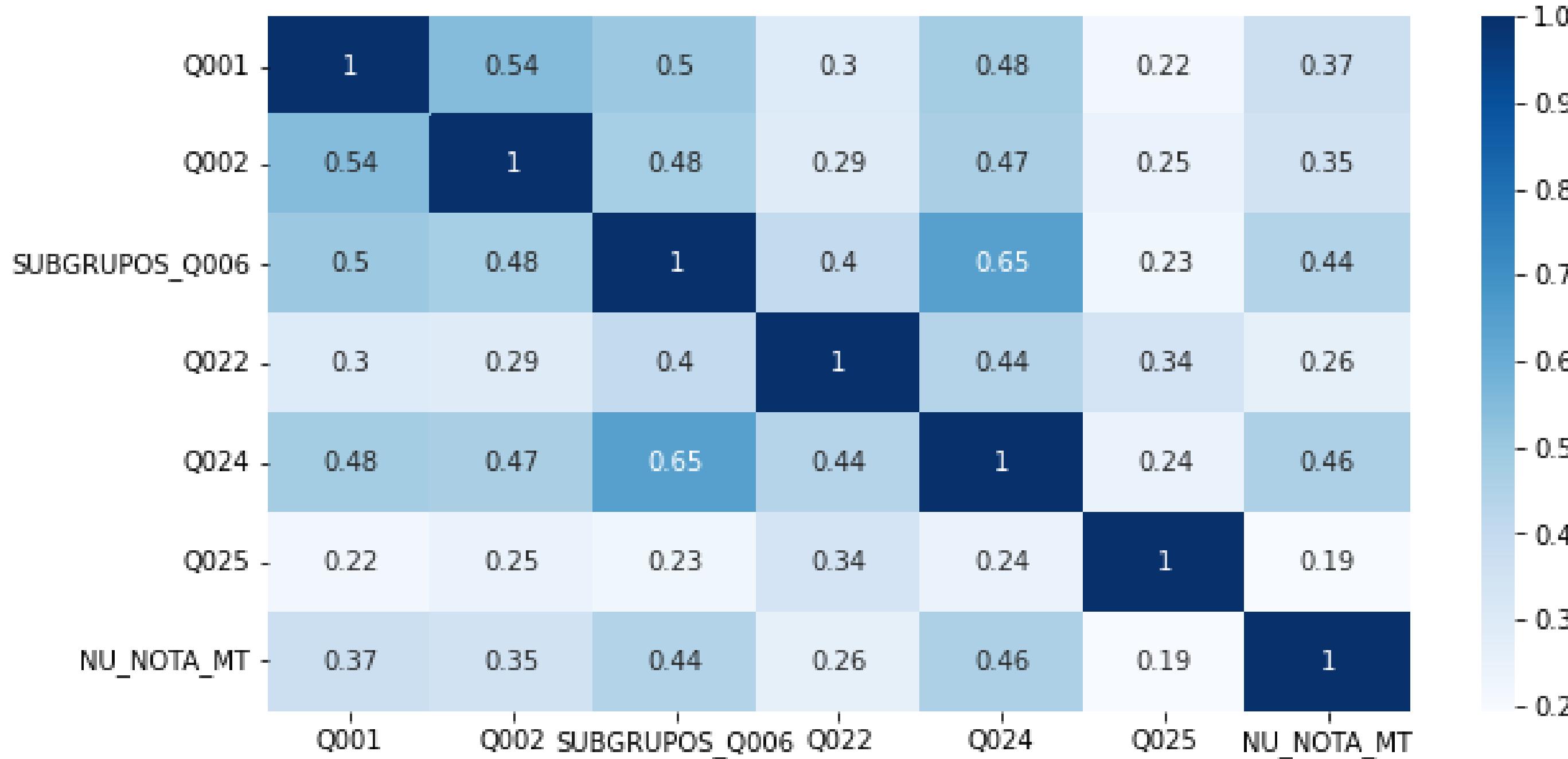
Distribuição da nota geral



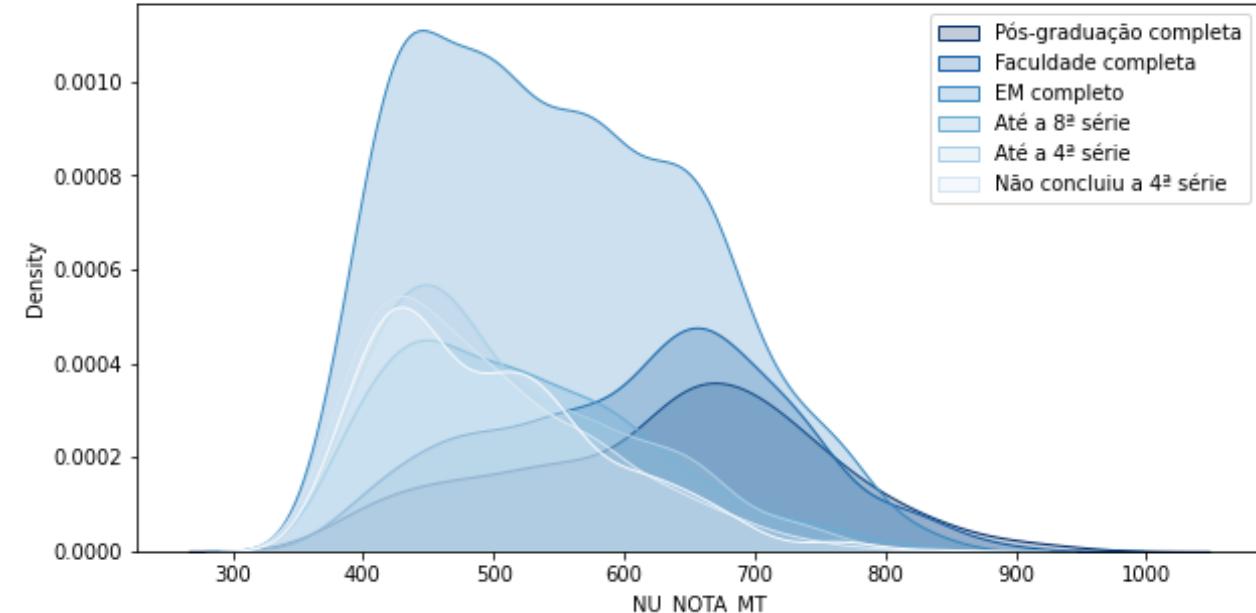
Distribuição da nota de matemática



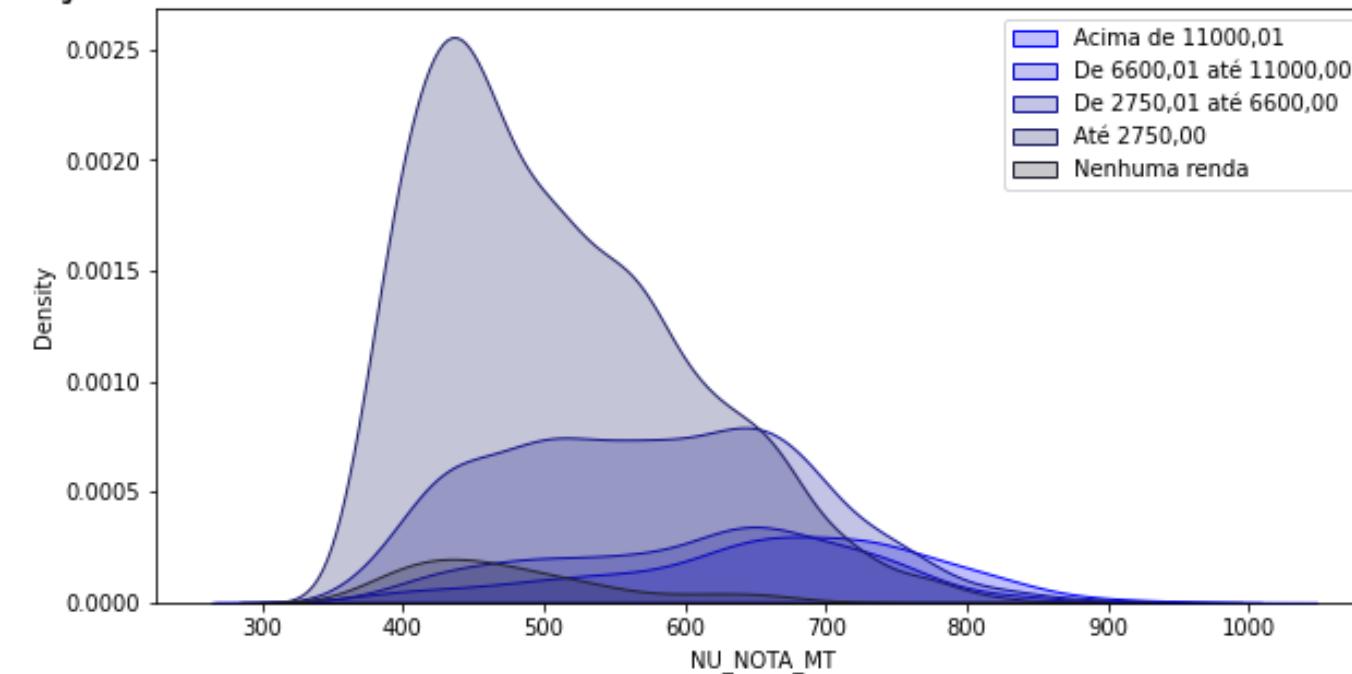
RELAÇÃO ENTRE FATORES SOCIOECONÔMICOS E A NOTA DE MATEMÁTICA



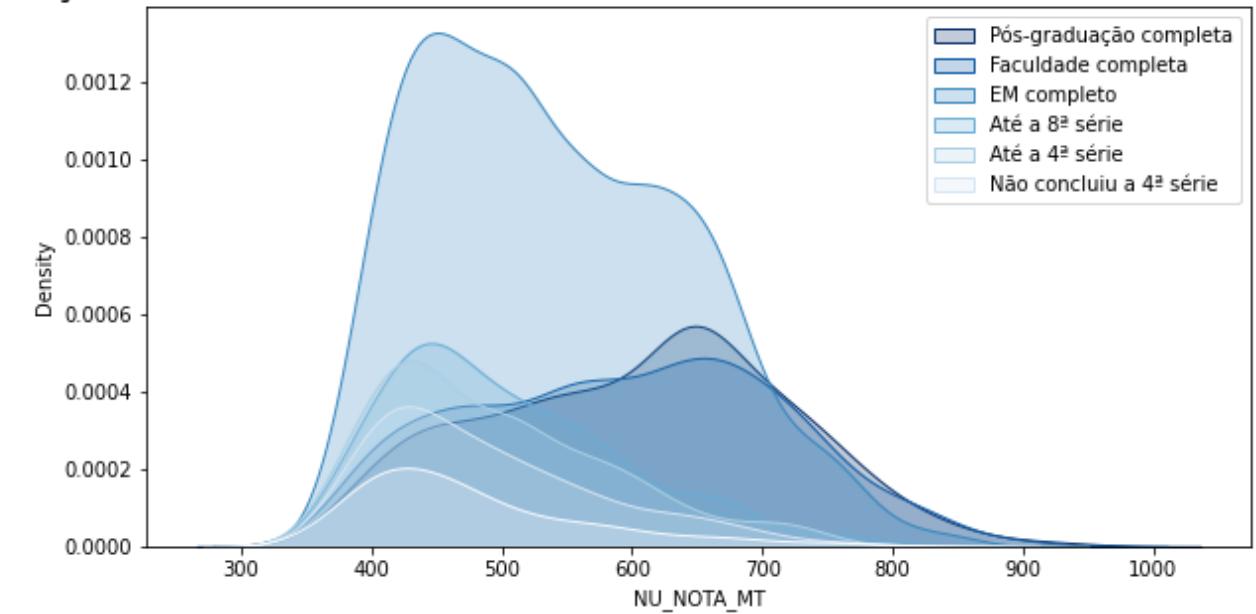
Distribuição das notas de matemática com base no nível de escolaridade do pai



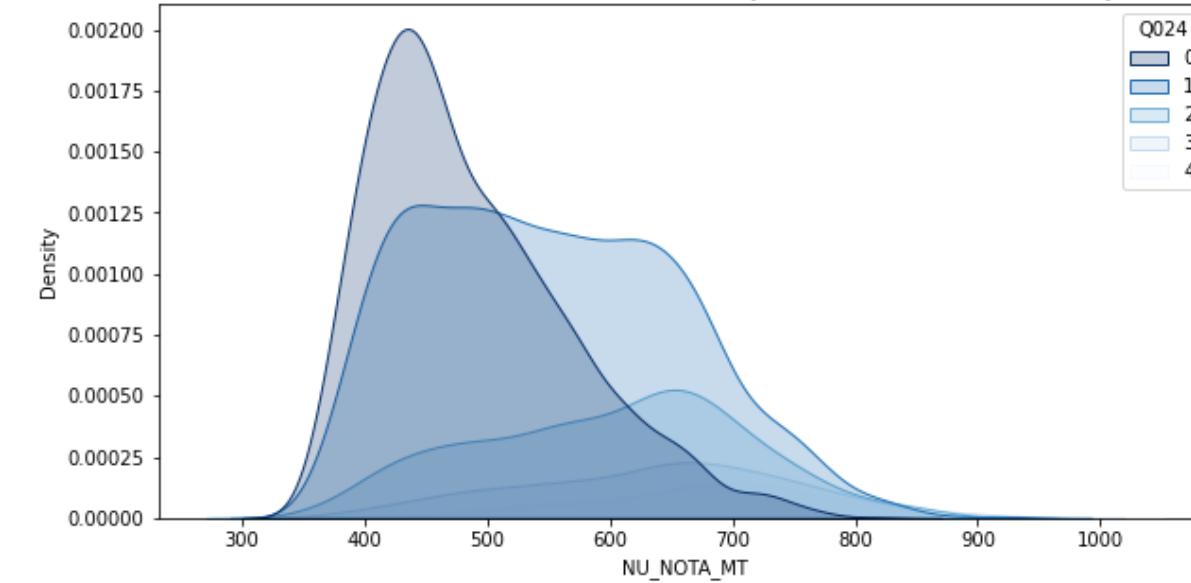
Distribuição das notas de matemática com base na renda mensal total da família



Distribuição das notas de matemática com base no nível de escolaridade da mãe

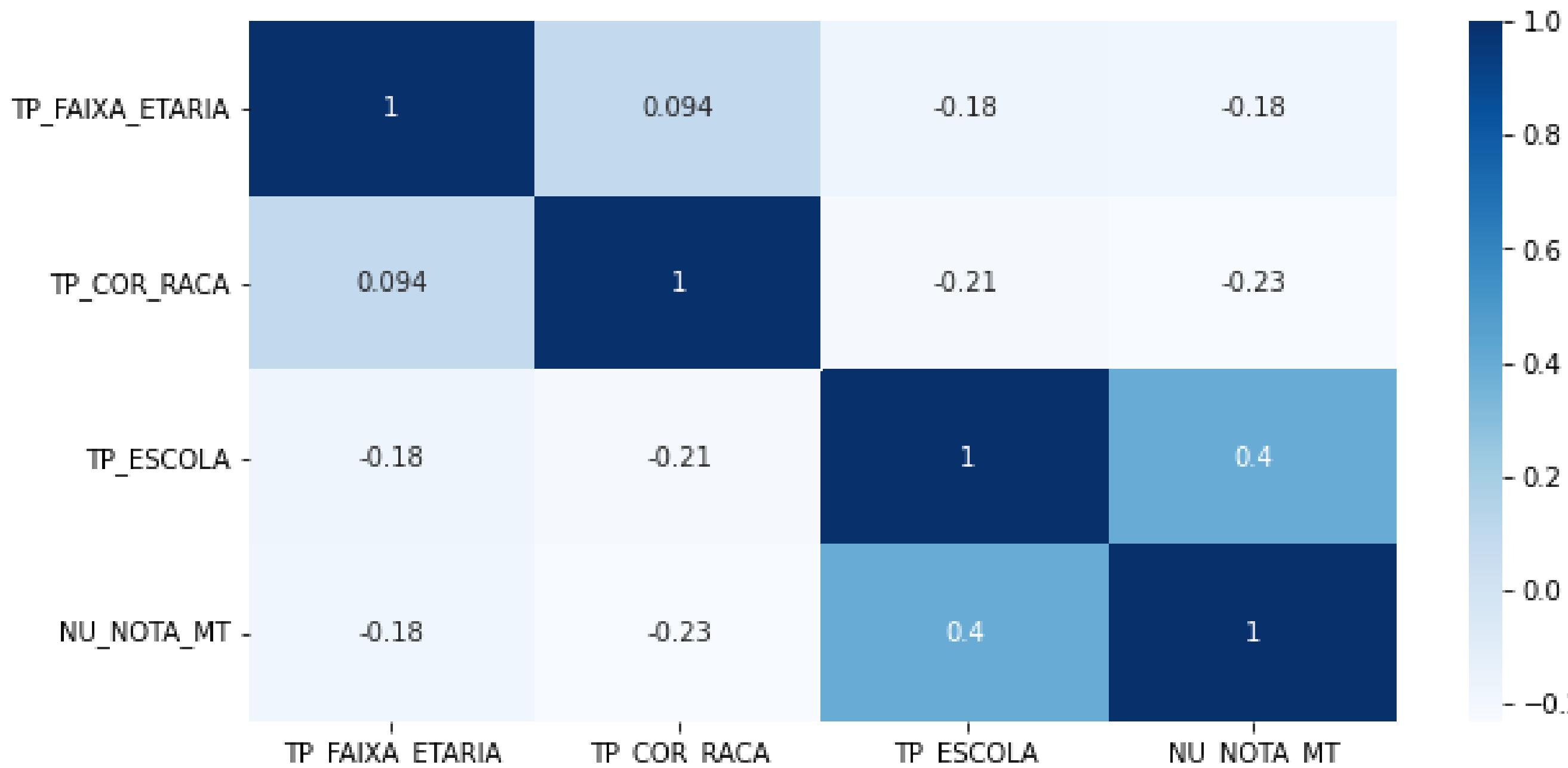


Distribuição das notas de matemática com base na quantidade de computadores que têm na casa



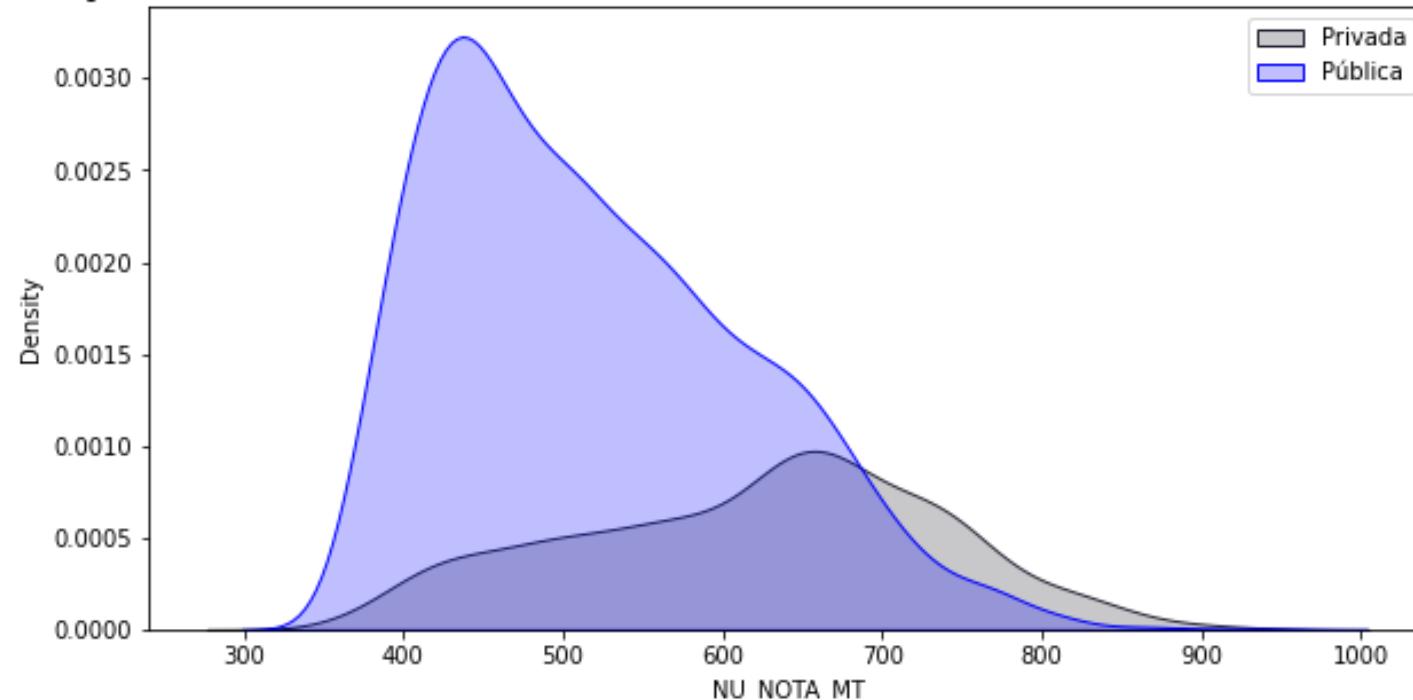
3

RELAÇÃO ENTRE ALGUNS DADOS DO CANDIDATO E A NOTA DE MATEMÁTICA

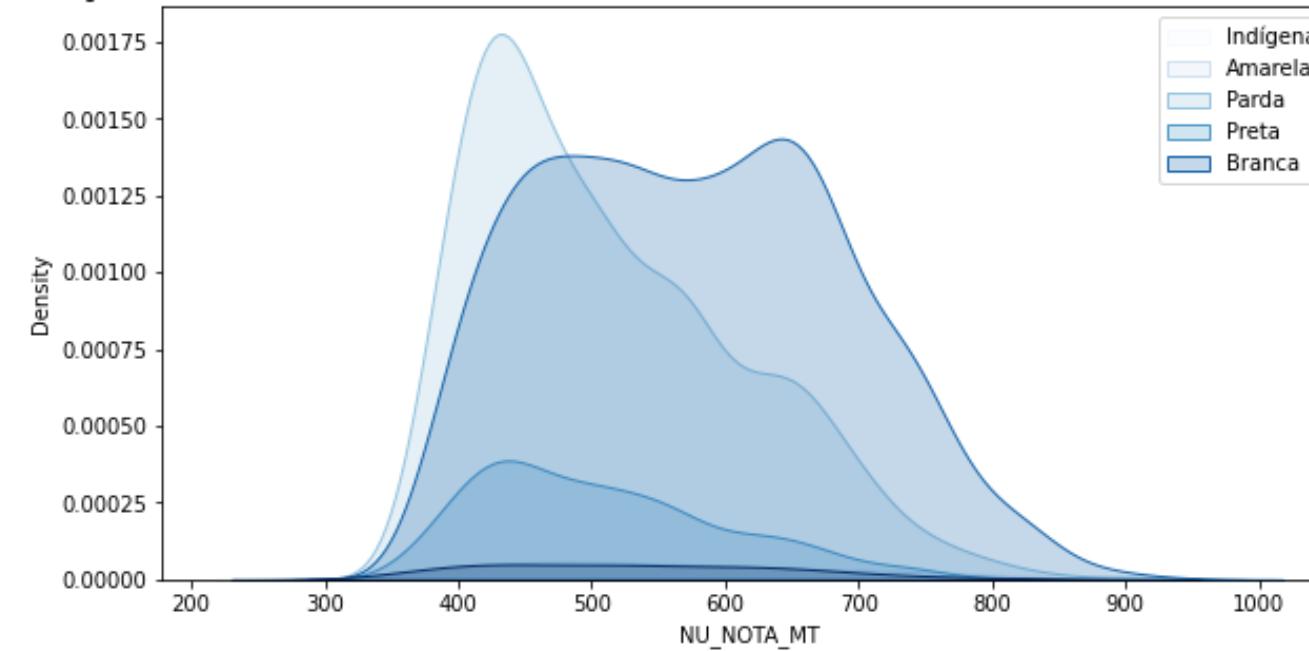


RELAÇÃO ENTRE ALGUNS DADOS DO CANDIDATO E A NOTA DE MATEMÁTICA

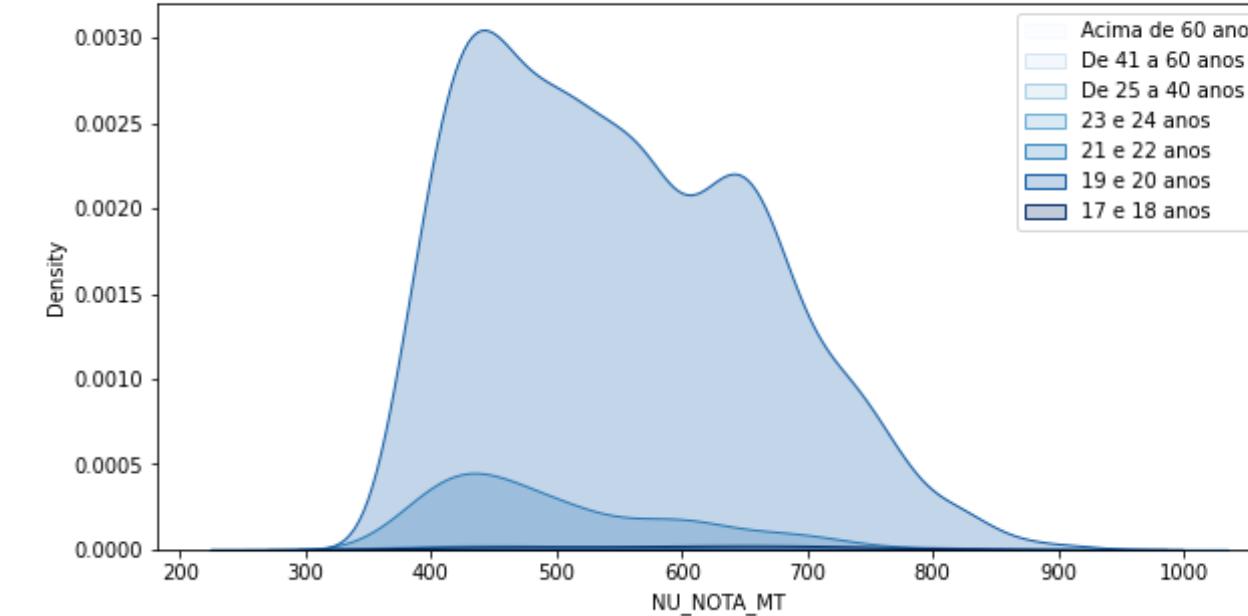
Distribuição das notas de matemática dos candidatos com base na sua escola



Distribuição das notas de matemática dos candidatos com base na sua cor/raça



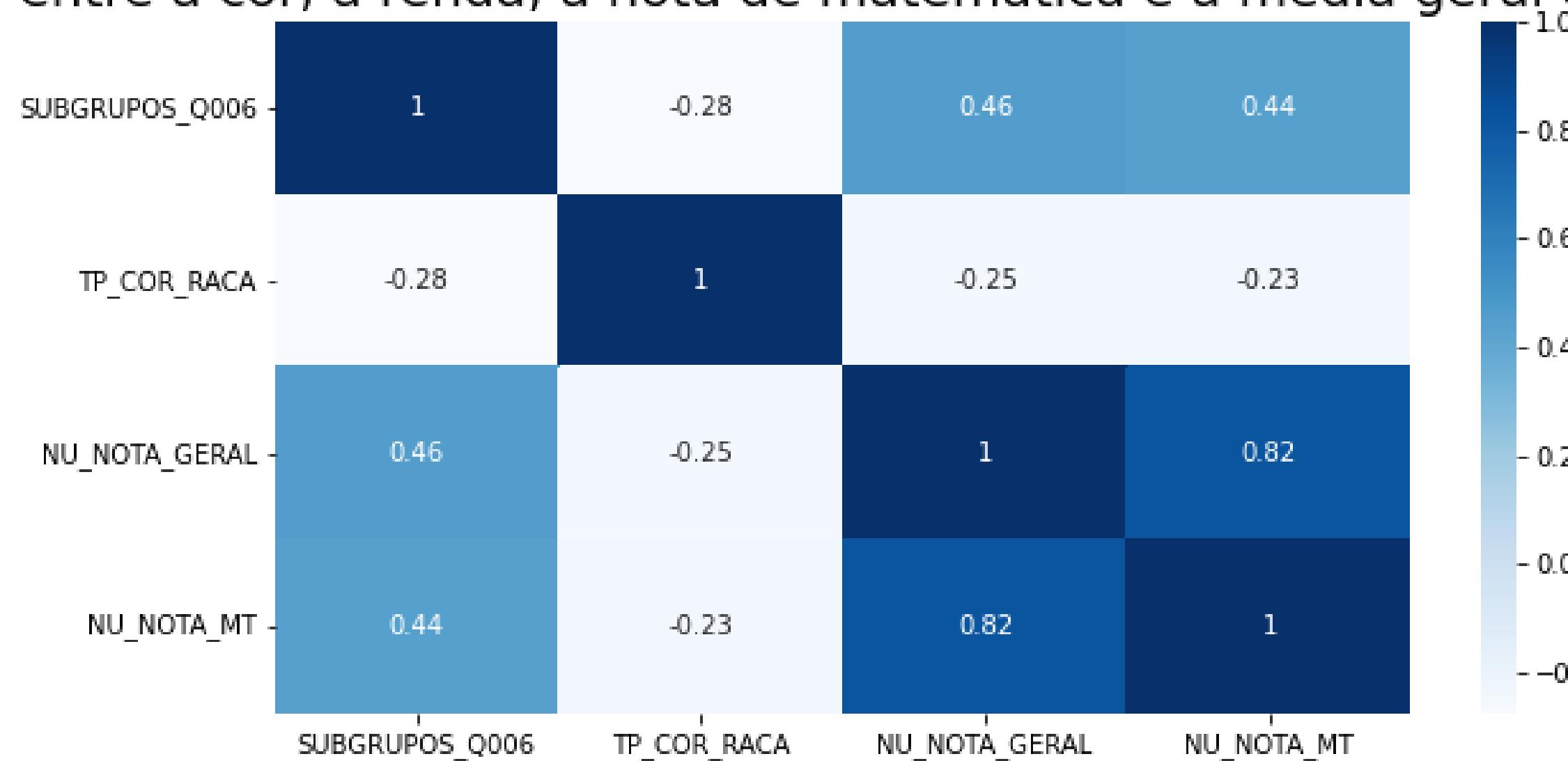
Distribuição das notas de matemática dos candidatos com base na sua faixa etária



4

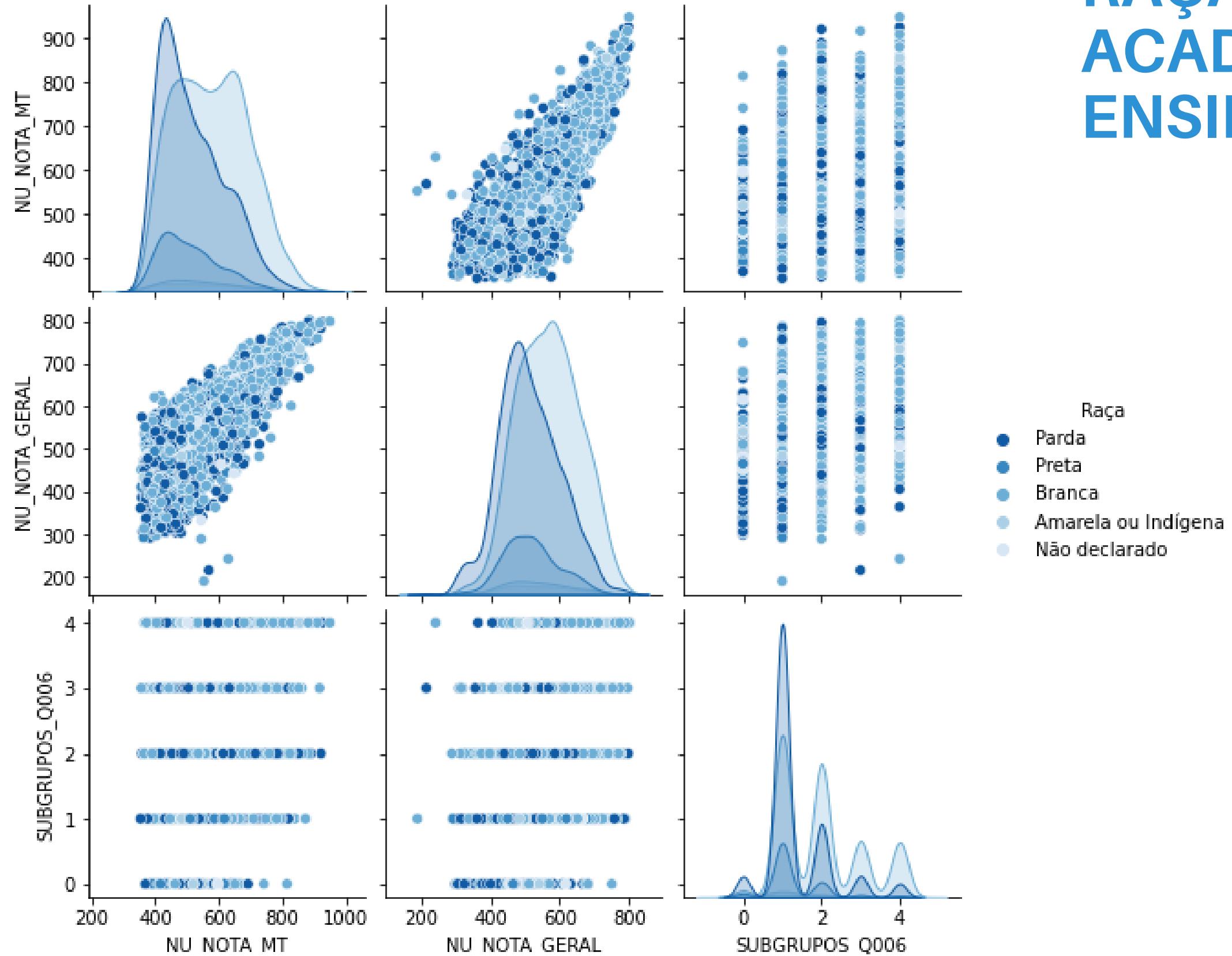
A INTERCESSÃO ENTRE RENDA, RAÇA E DESEMPENHO ACADÊMICO NO ACESSO AO ENSINO SUPERIOR NO BRASIL:

correlação entre a cor, a renda, a nota de matemática e a média geral do participante



4

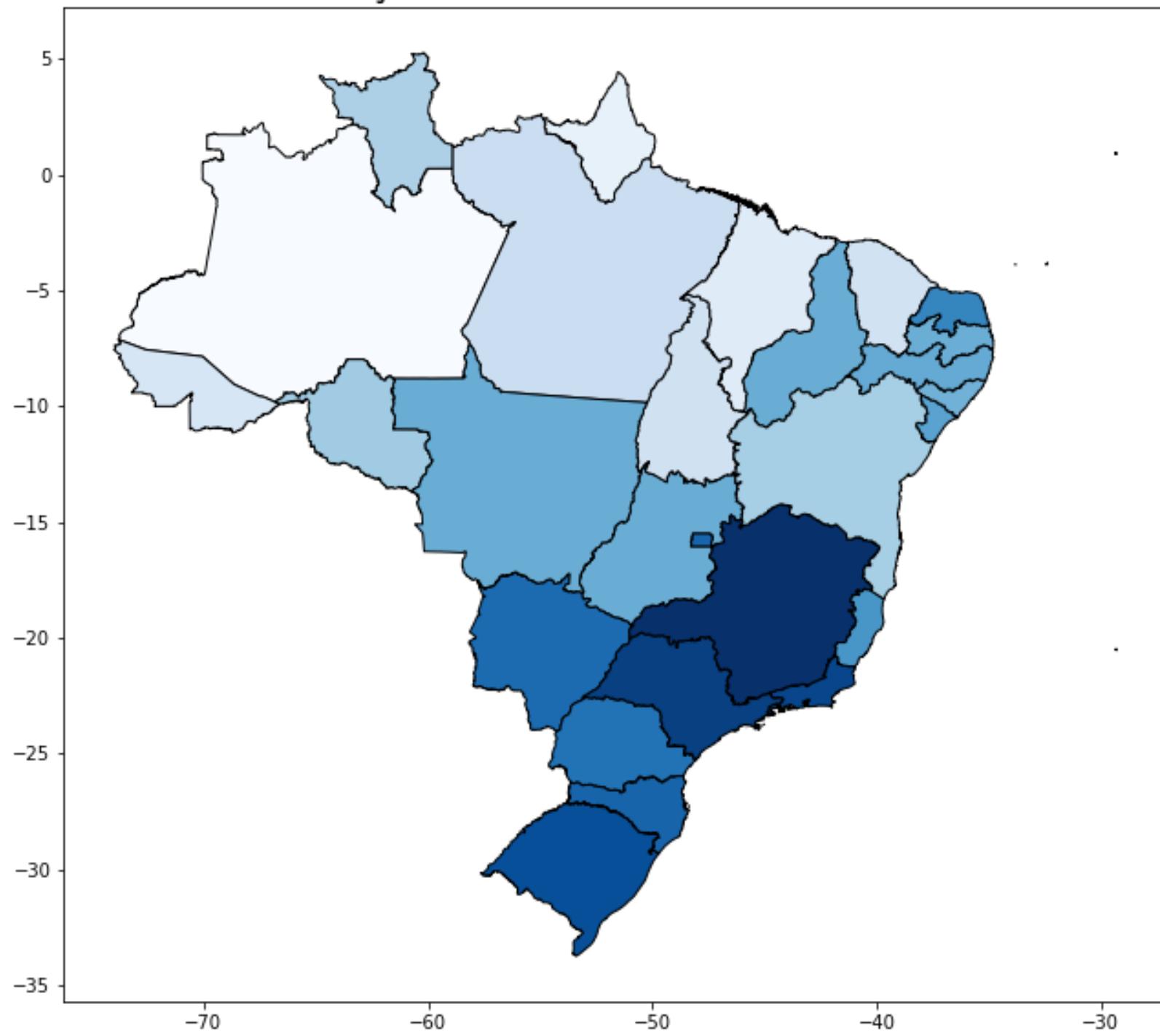
A INTERCESSÃO ENTRE RENDA, RAÇA E DESEMPENHO ACADÊMICO NO ACESSO AO ENSINO SUPERIOR NO BRASIL:



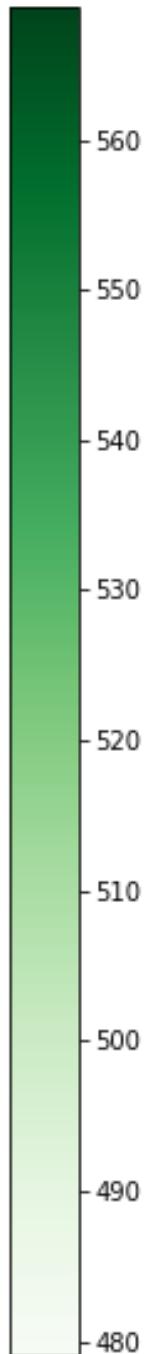
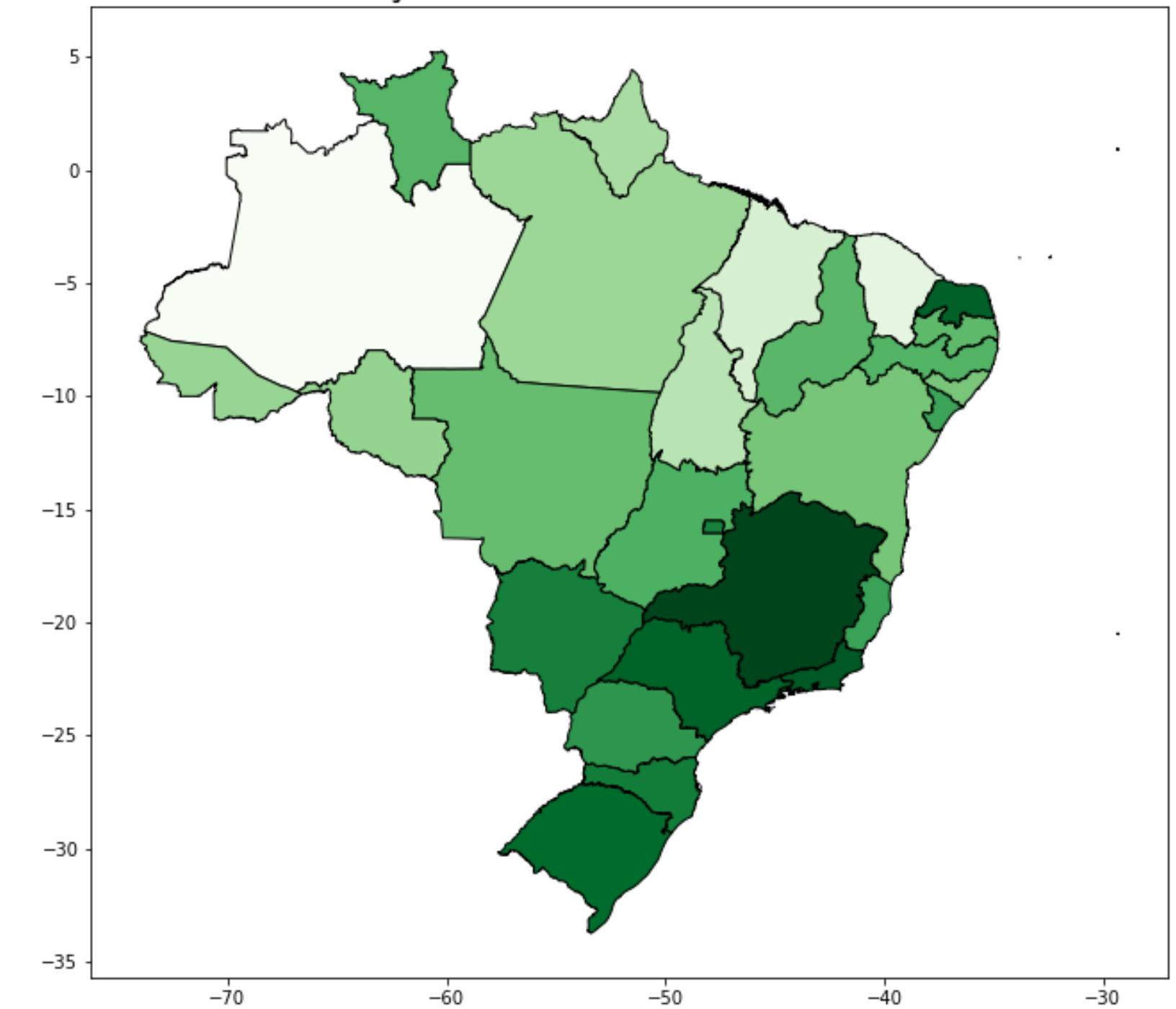
5

DISTRIBUIÇÃO DAS NOTAS PELO BRASIL

distribuição notas de matemática no Brasil



distribuição das médias das notas no Brasil



Modelos estatísticos e de Machine Learning

Como funcionam os algoritmos que utilizamos?

1

NAIVE BAYES



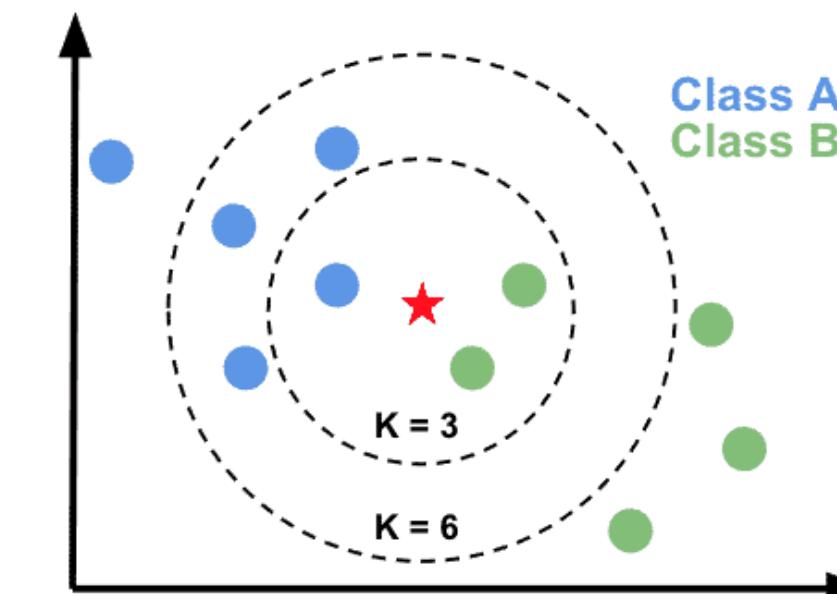
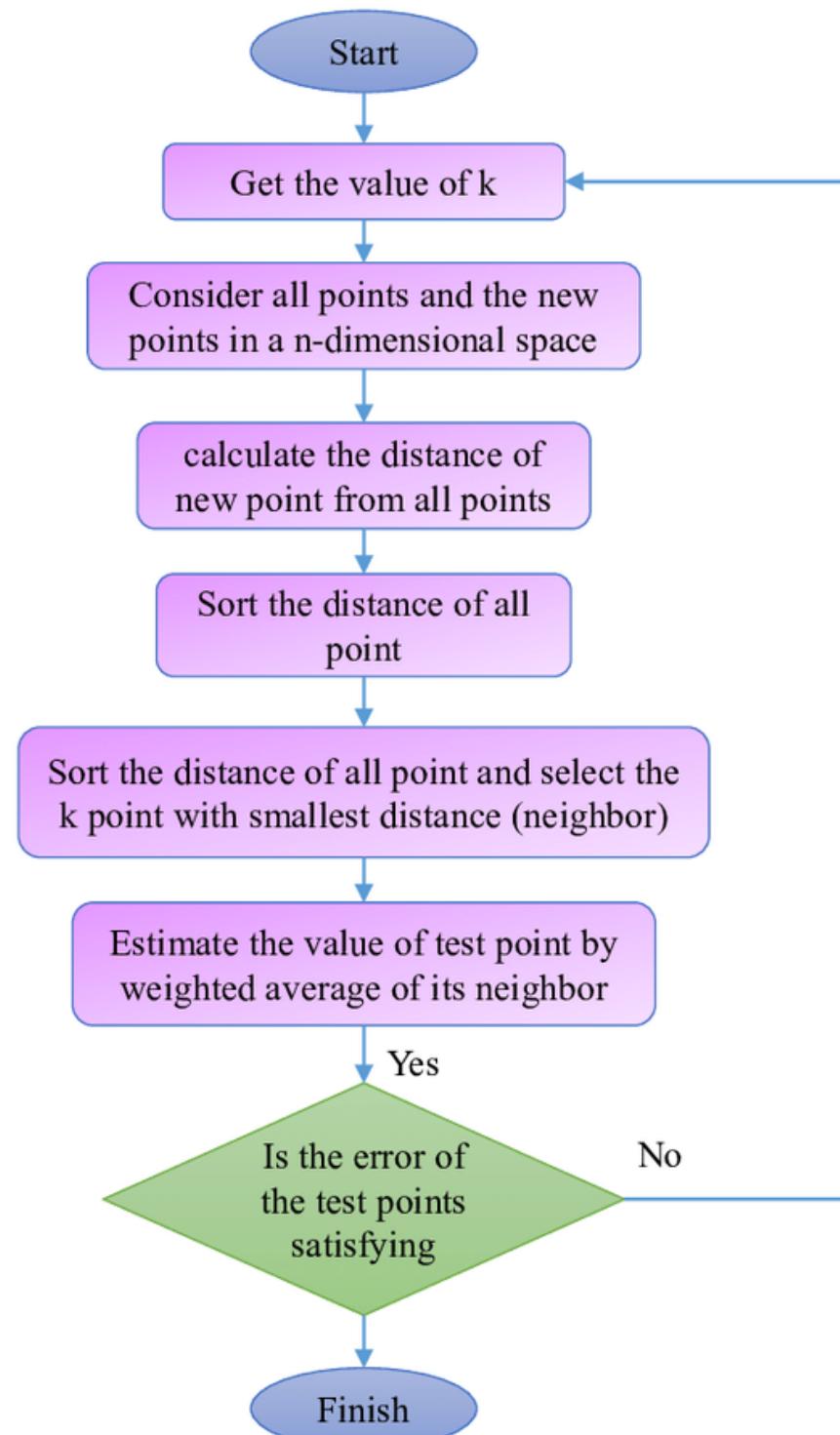
THOMAS BAYES

Responsável pelo teorema que leva o seu nome e utiliza da probabilidade condicional.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

2

K-NEAREST-NEIGHBORS



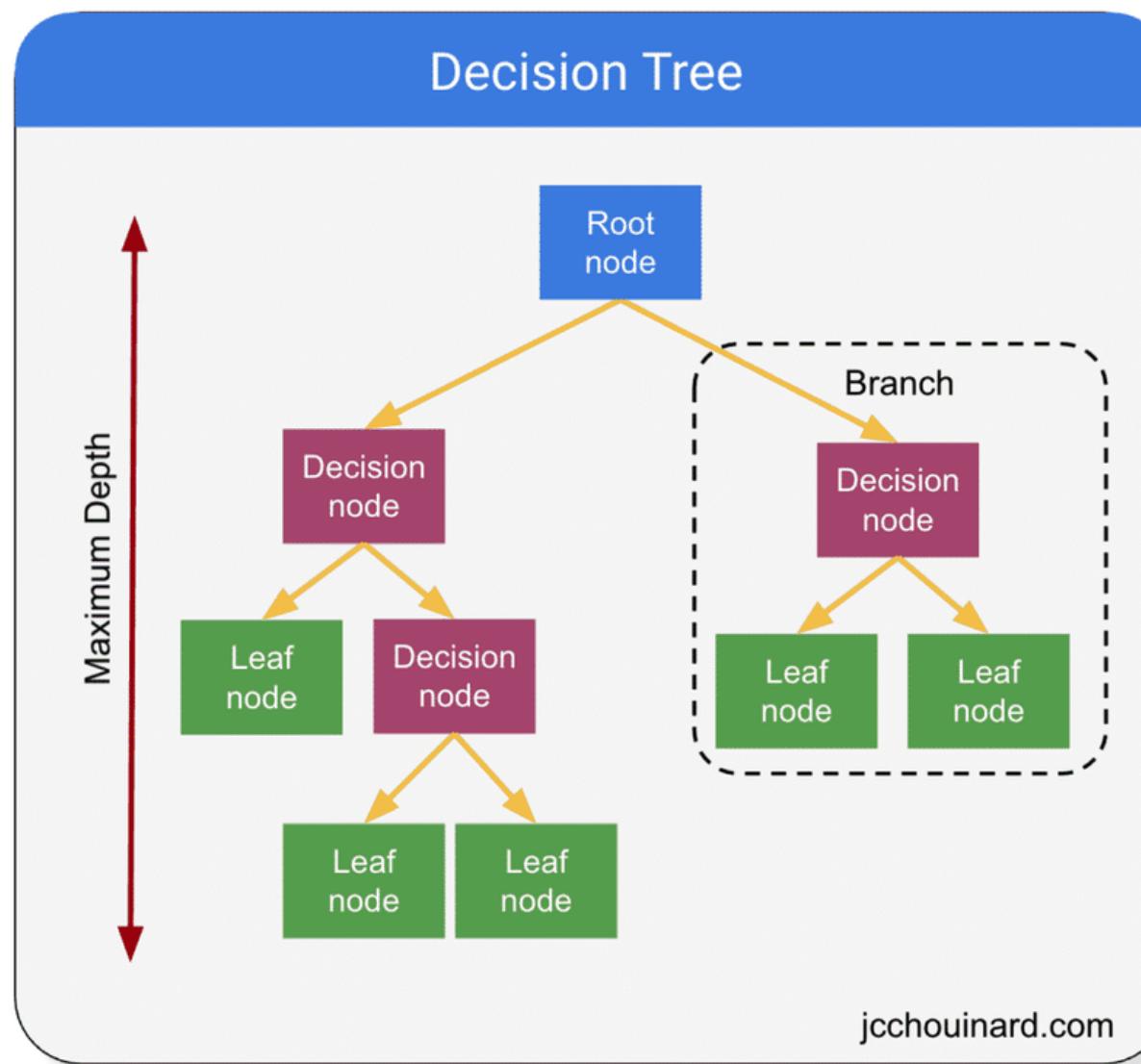
Como é feita a votação para cada classe no gráfico do KNN

COLOCAR ALGO AQUI

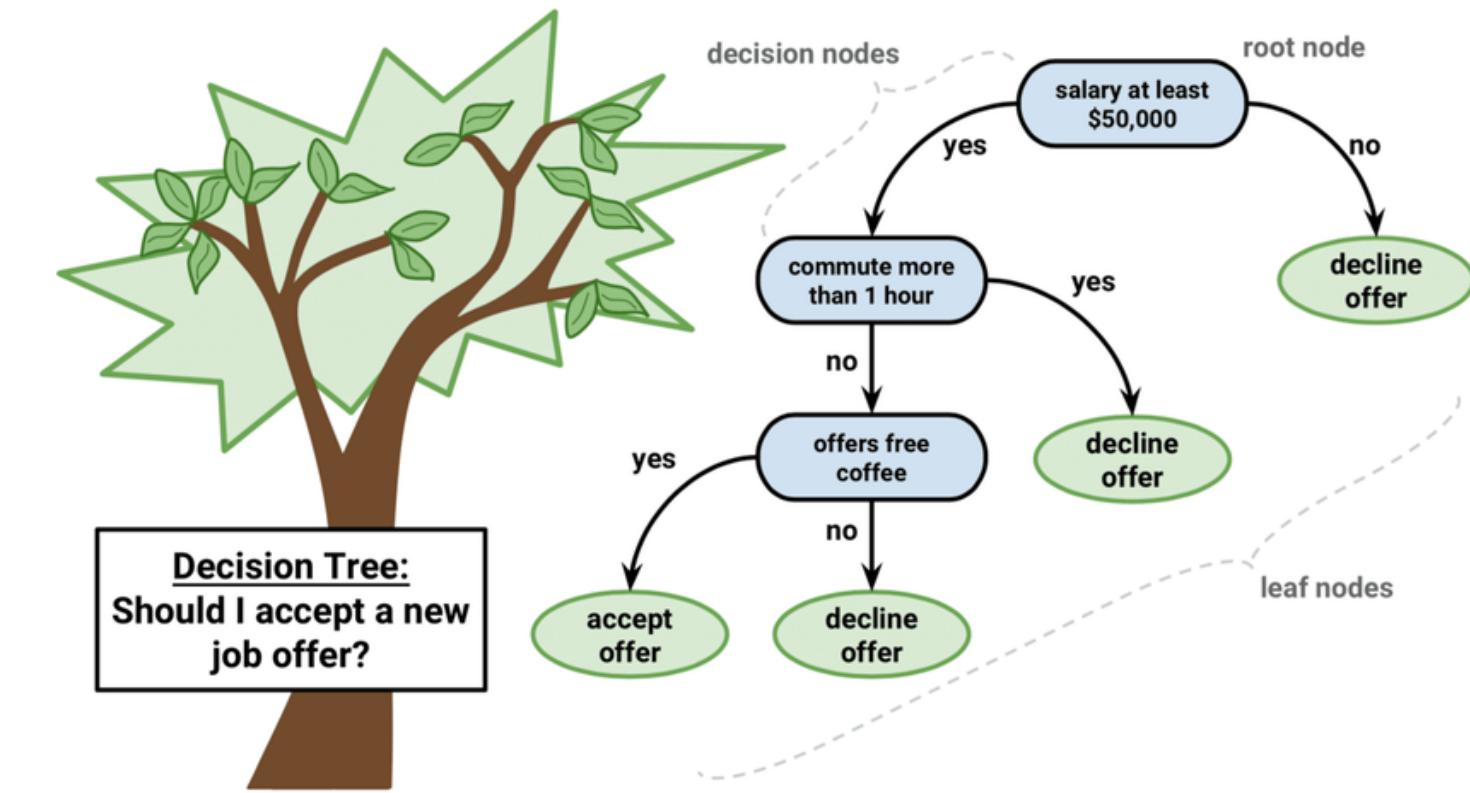
Fluxograma de funcionamento do modelo

3

DECISION TREE



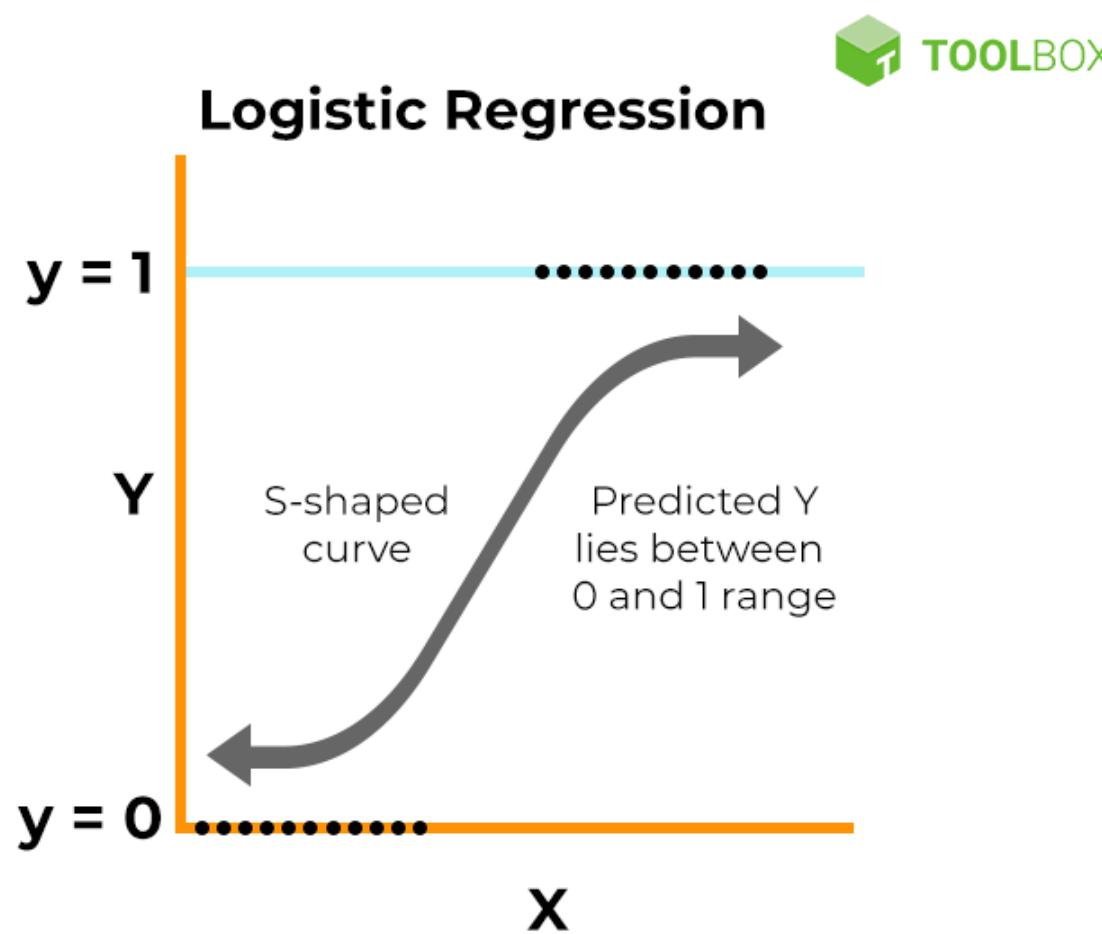
Estrutura de uma árvore de decisões



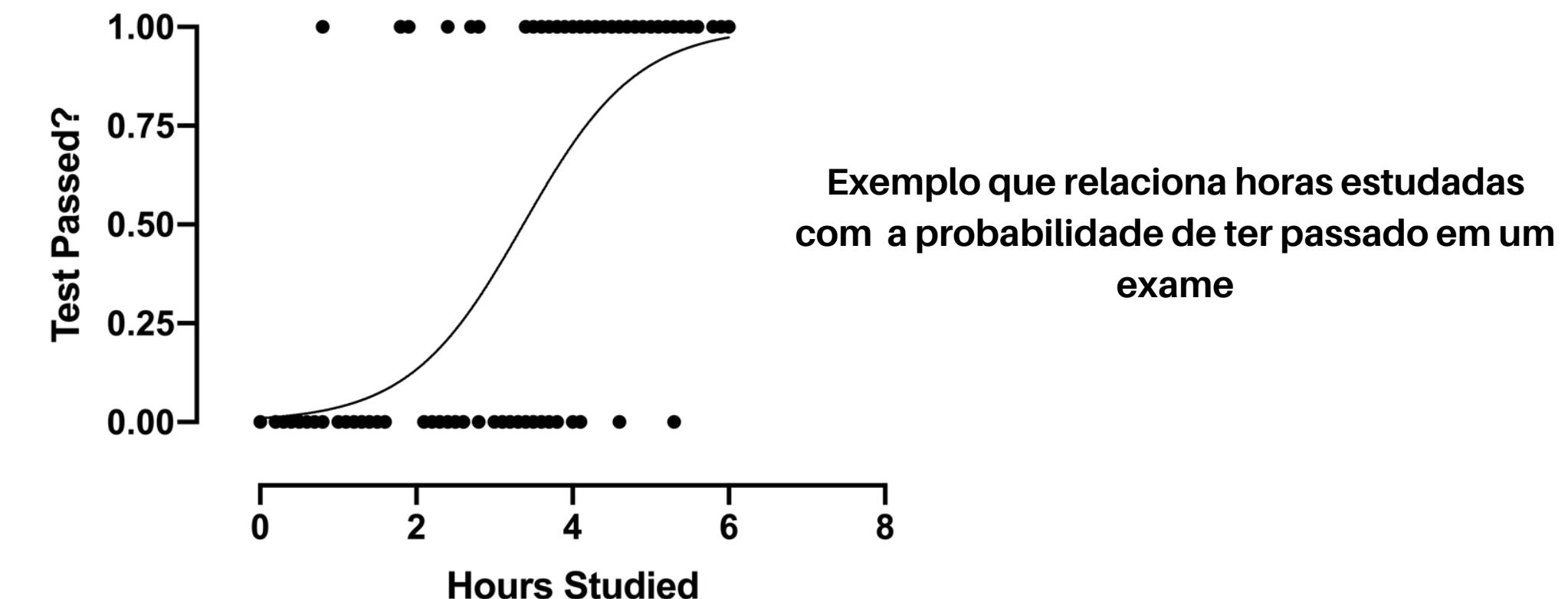
Exemplo de uma decisão realizada pelo algoritmo

4

LOGISTIC REGRESSION

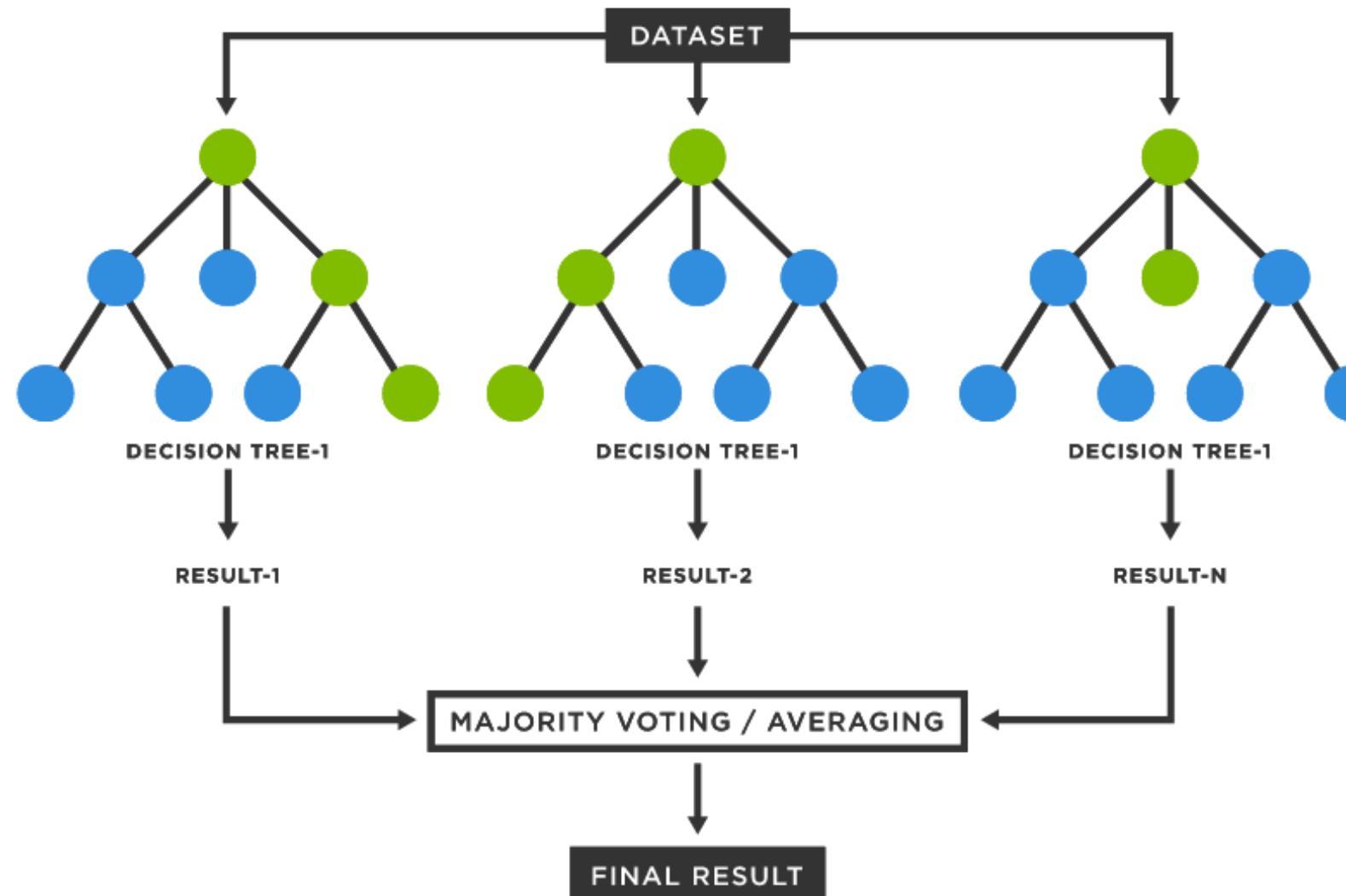


Diferentemente da regressão linear, que a previsão Y pode passar os valores de 0 e 1, a logística se contém entre esses valores



5

RANDOM FOREST



Através do sistema de voto, aquela classificação que obtiver maior número, será a escolhida.

Pré-processamento dos dados

Como preparamos o dataset para aplicar os modelos?



DATA CLEANING

Missing Data

O que são?

Nossa solução

Retirar todas as ocorrências

Valores nulos

NaN

Campos não preenchidos

Na coluna 'Tipo escola', muitos alunos preferiram não responder, optamos por não utilizar essas ocorrências.



FEATURE SELECTION

Originalmente, tínhamos 76 colunas no dataset. Nessa etapa, escolhemos as features que avaliamos serem as mais importantes.

Data columns (total 76 columns):			
#	Column	Non-Null Count	Dtype
0	NU_INSCRICAO	1015521 non-null	int64
1	NU_ANO	1015521 non-null	int64
2	TP_FAIXA_ETARIA	1015521 non-null	int64
3	TP_SEXO	1015521 non-null	object
4	TP_ESTADO_CIVIL	1015521 non-null	int64
5	TP_COR_RACA	1015521 non-null	int64
6	TP_NACIONALIDADE	1015521 non-null	int64
7	TP_ST_CONCLUSAO	1015521 non-null	int64
8	TP_ANO_CONCLUIU	1015521 non-null	int64
9	TP_ESCOLA	1015521 non-null	int64
10	TP_ENSINO	328694 non-null	float64
11	IN_TREINEIRO	1015521 non-null	int64
12	CO_MUNICIPIO_ESC	243782 non-null	float64
13	NO_MUNICIPIO_ESC	243782 non-null	object
14	CO_UF_ESC	243782 non-null	float64
15	SG_UF_ESC	243782 non-null	object
16	TP_DEPENDENCIA_ADMIN_ESC	243782 non-null	float64
17	TP_LOCALIZACAO_ESC	243782 non-null	float64
18	TP_SIT_FUNC_ESC	243782 non-null	float64
19	CO_MUNICIPIO_PROVA	1015521 non-null	int64
20	NO_MUNICIPIO_PROVA	1015521 non-null	object

FEATURES ESCOLHIDAS

TP_ESCOLA	Tipo de escola (pública ou privada)
TP_FAIXA_ETARIA	Grupos de faixas etárias
TP_COR_RACA	Grupos raciais do Brail de acordo com o IBGE
Q001	Escolaridade do pai
Q002	Escolaridade da Mãe
Q005	Quantas pessoas moram na casa do candidato
Q006	Renda mensal da família
Q024	Quantidade de computadores na casa
Q025	Acesso à internet
NU_NOTA_MT	Nota da prova de matemática 



FEATURE ENGINEERING

1

Uma parte das features que escolhemos estavam representadas como categorias alfabéticas, então tivemos que transformar para valores numéricos.



2

Além disso, para features com um grande número de classes, resolvemos agrupá-las.





FEATURE ENGINEERING

Por exemplo, na coluna Renda Mensal (Q006), tínhamos muitas classes. A classe 'C' mapeava aqueles que tinham renda de R\$ 1.100,01 até R\$ 1.650,00.

```
model_df['renda mensal'].value_counts()
```

B	62986
C	37706
D	30796
F	19712
E	16673
G	16660
H	13675
A	13297
I	8687
M	5406
J	5380
Q	4868
K	4218
O	3785
P	3774
L	3576
N	3321

Name: renda mensal, dtype: int64



```
model_df['renda mensal'].value_counts()
```

1	148161
2	58734
3	18580
4	15748
0	13297

Name: renda mensal, dtype: int64



- Grupo 0: Respondeu que a família não possui nenhuma renda
- Grupo 1: Respondeu que a família possui renda entre R\$ 998,01 e R\$ 2.495,00
- Grupo 2: Respondeu que a família possui renda entre R\$ 2.495,00 e R\$ 5.988,00
- Grupo 3: Respondeu que a família possui renda entre R\$ 5.988,00 e R\$ 9.980,00
- Grupo 4: Respondeu que a família possui renda entre R\$ 9.980,00 e R\$ 19.960,00 ou mais

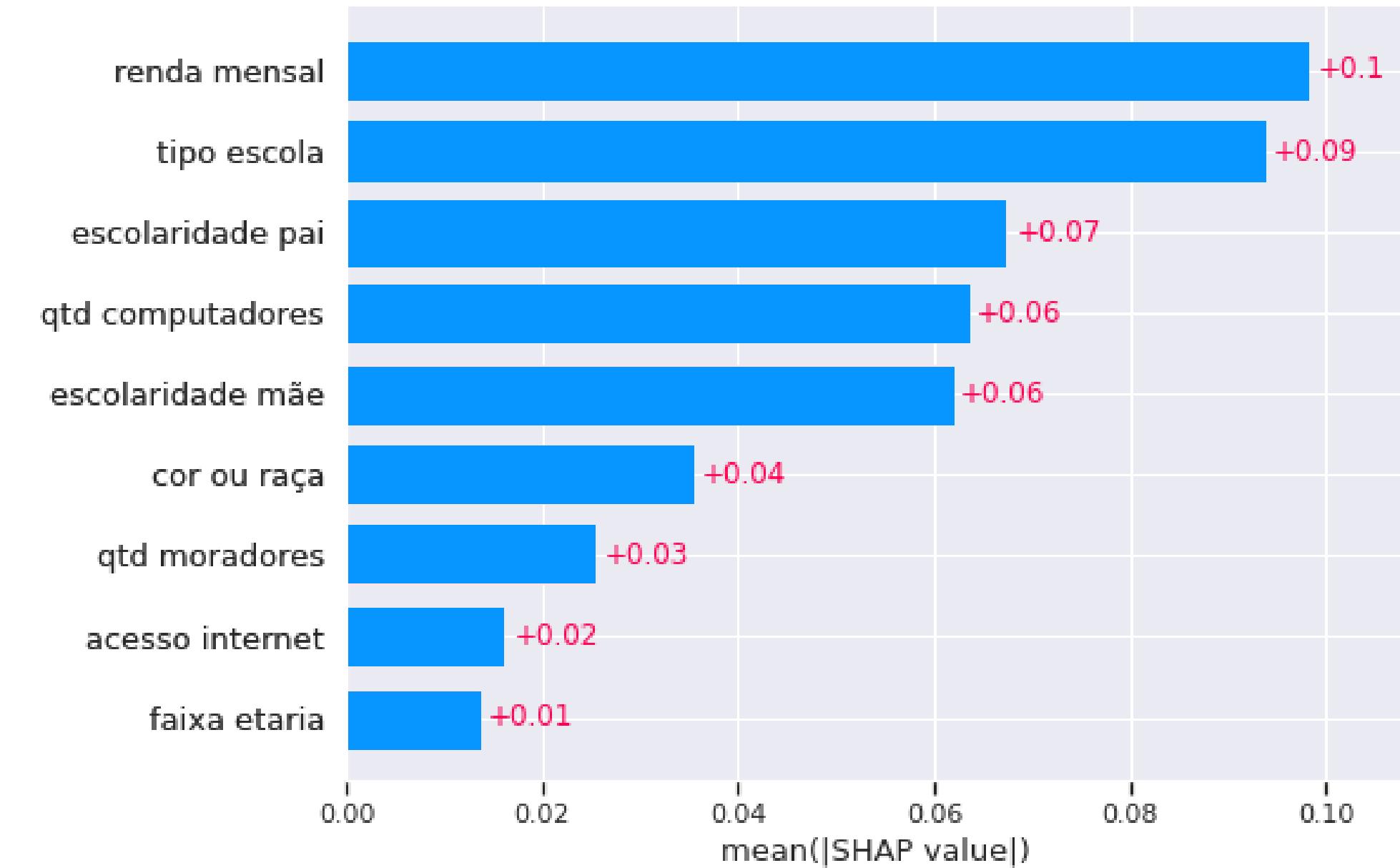
Transformamos todas as classes em 5 grupos, ajustando os intervalos de renda.

Comparação do desempenho dos modelos utilizados

Quais foram as diferenças entre os algoritmos?

ANALISANDO A IMPORTÂNCIA DE CADA FEATURE PARA O NAIVE BAYES

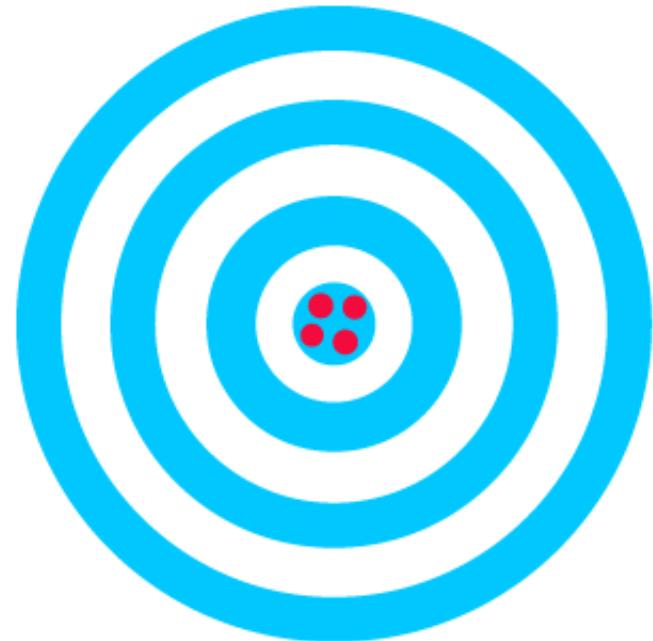
SHAP Values (SHapley Additive exPlanations) é um método baseado na teoria dos jogos usado para aumentar a transparência e a interpretabilidade de modelos de aprendizado de máquina.



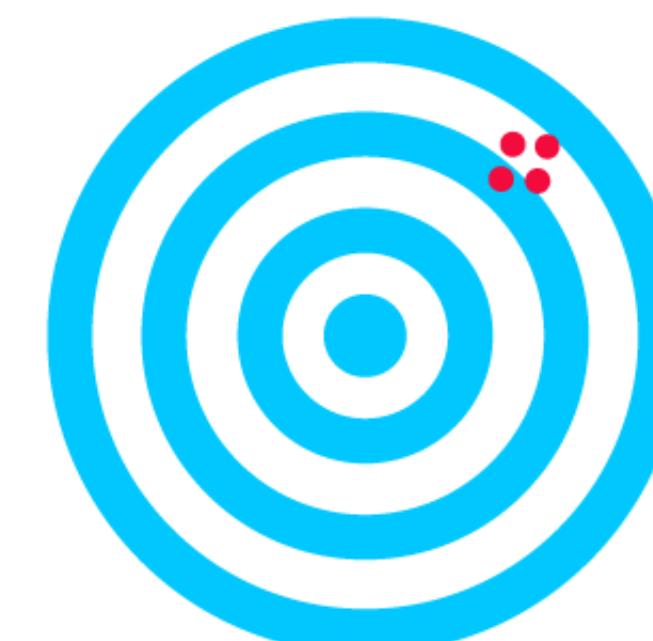
PRINCIPAIS MÉTRICAS PARA AVALIAR UM CLASSIFICADOR

Acurácia vs Precisão

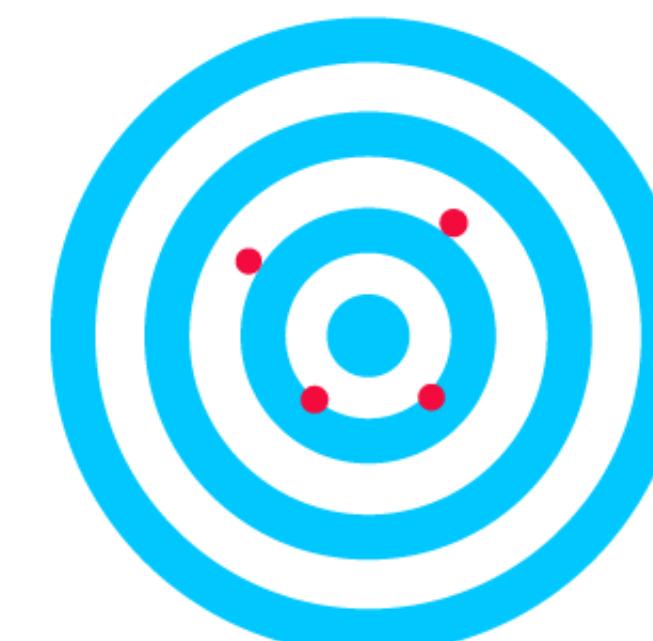
Accurate
Precise



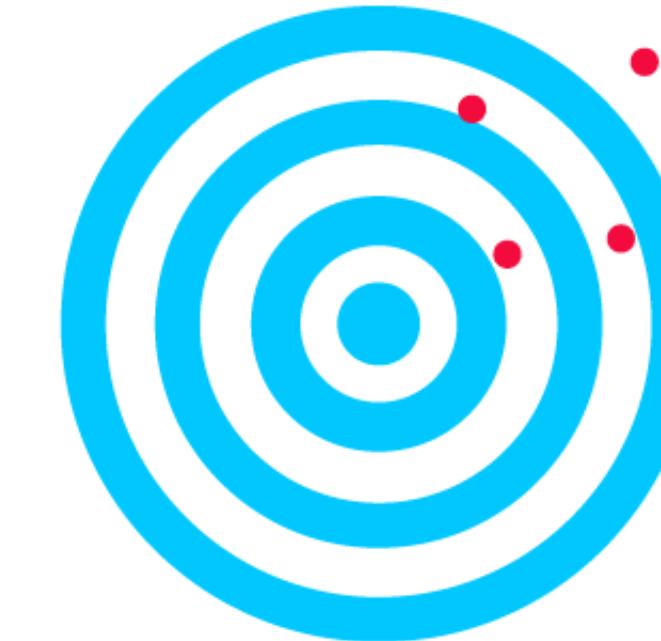
Not Accurate
Precise



Accurate
Not Precise



Not Accurate
Not Precise



PRINCIPAIS MÉTRICAS PARA AVALIAR UM CLASSIFICADOR

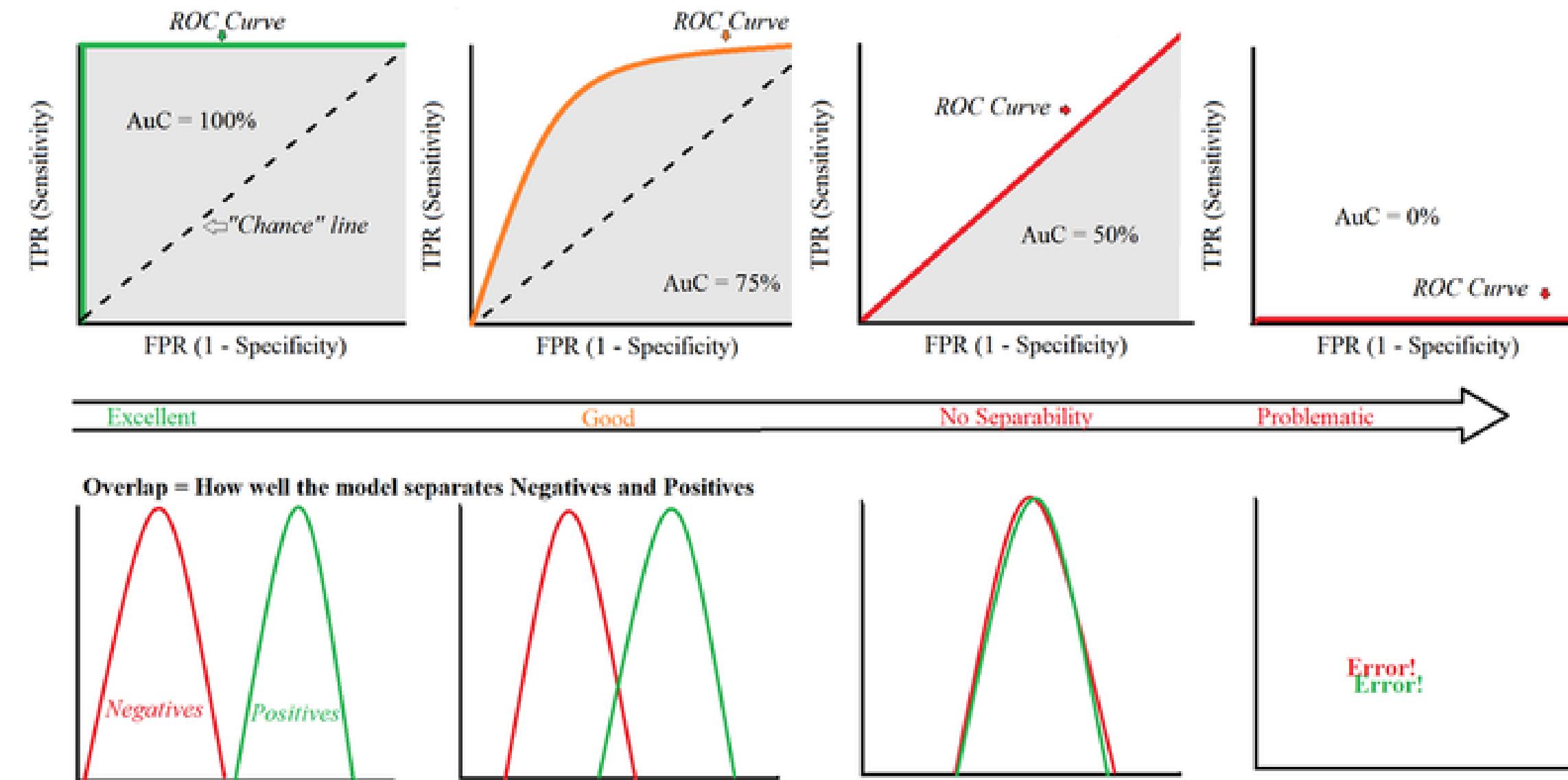
Matriz de confusão

É uma tabela com duas linhas e duas colunas que informa o número de verdadeiros positivos, falsos negativos, falsos positivos e verdadeiros negativos.

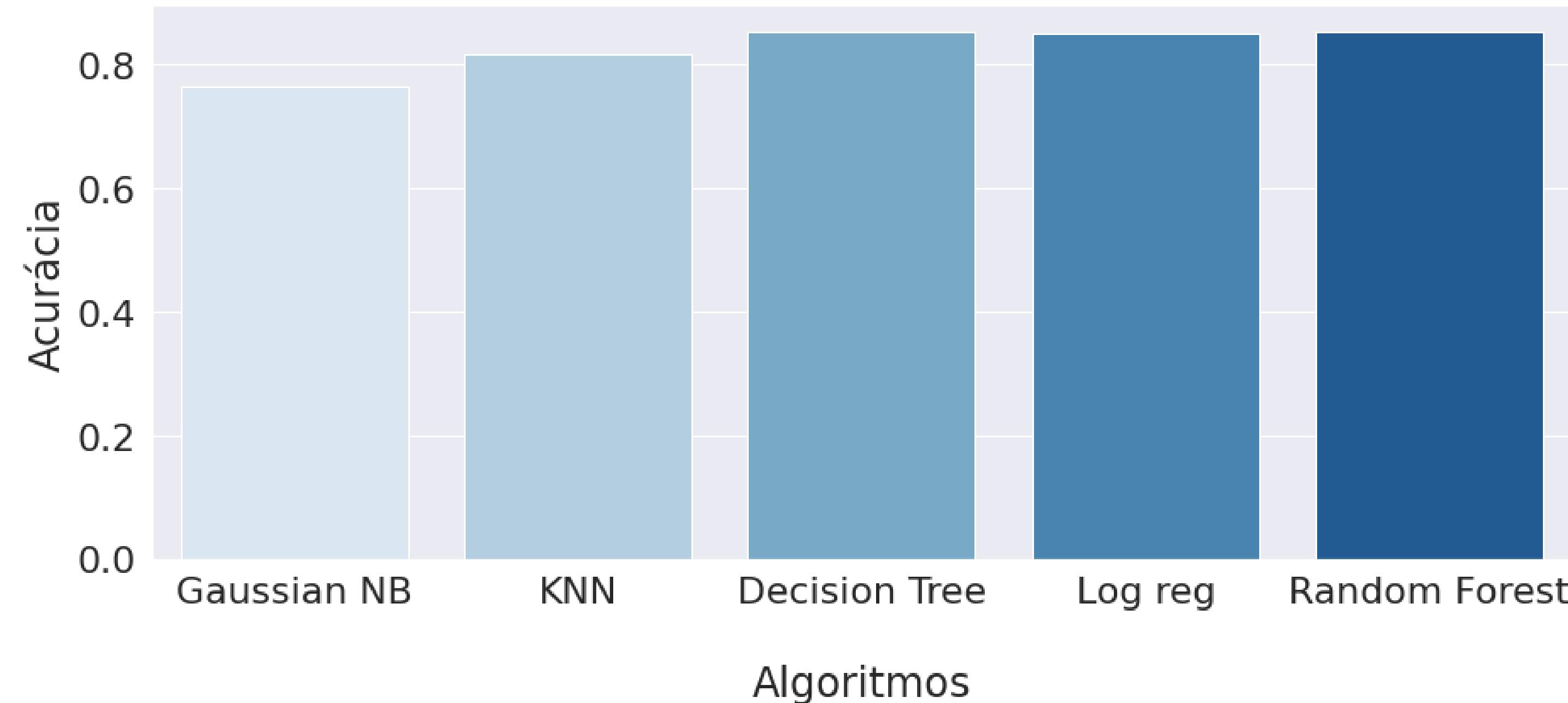
		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

PRINCIPAIS MÉTRICAS PARA AVALIAR UM CLASSIFICADOR

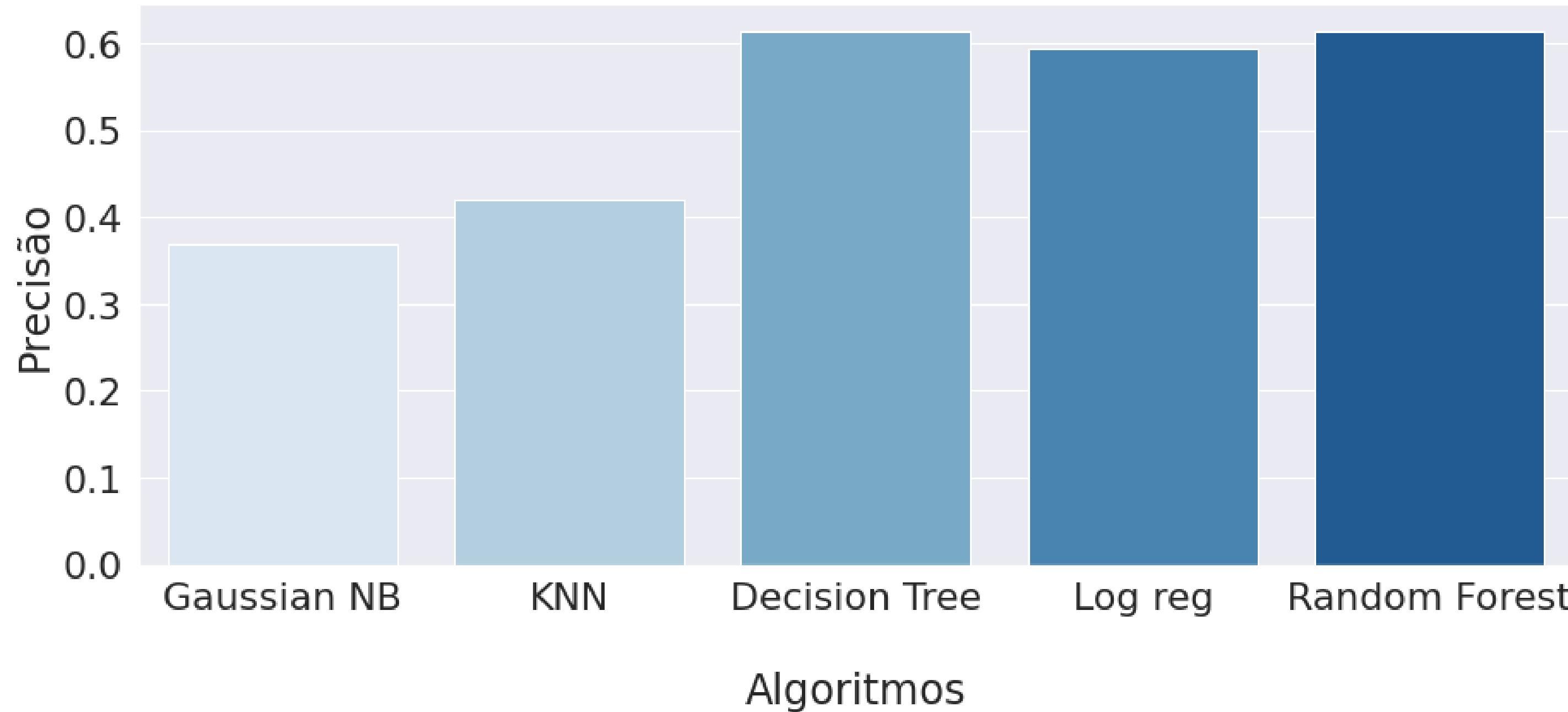
AUC: Area Under the ROC Curve



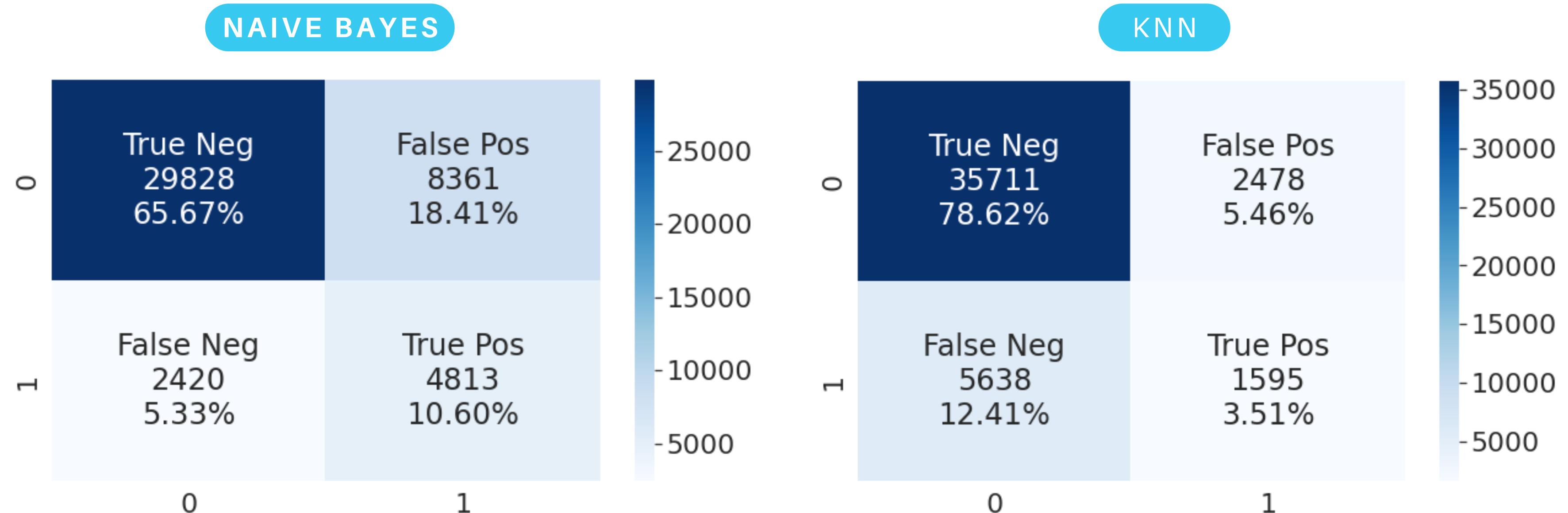
RESULTADOS: ACURÁCIA DE CADA MODELO



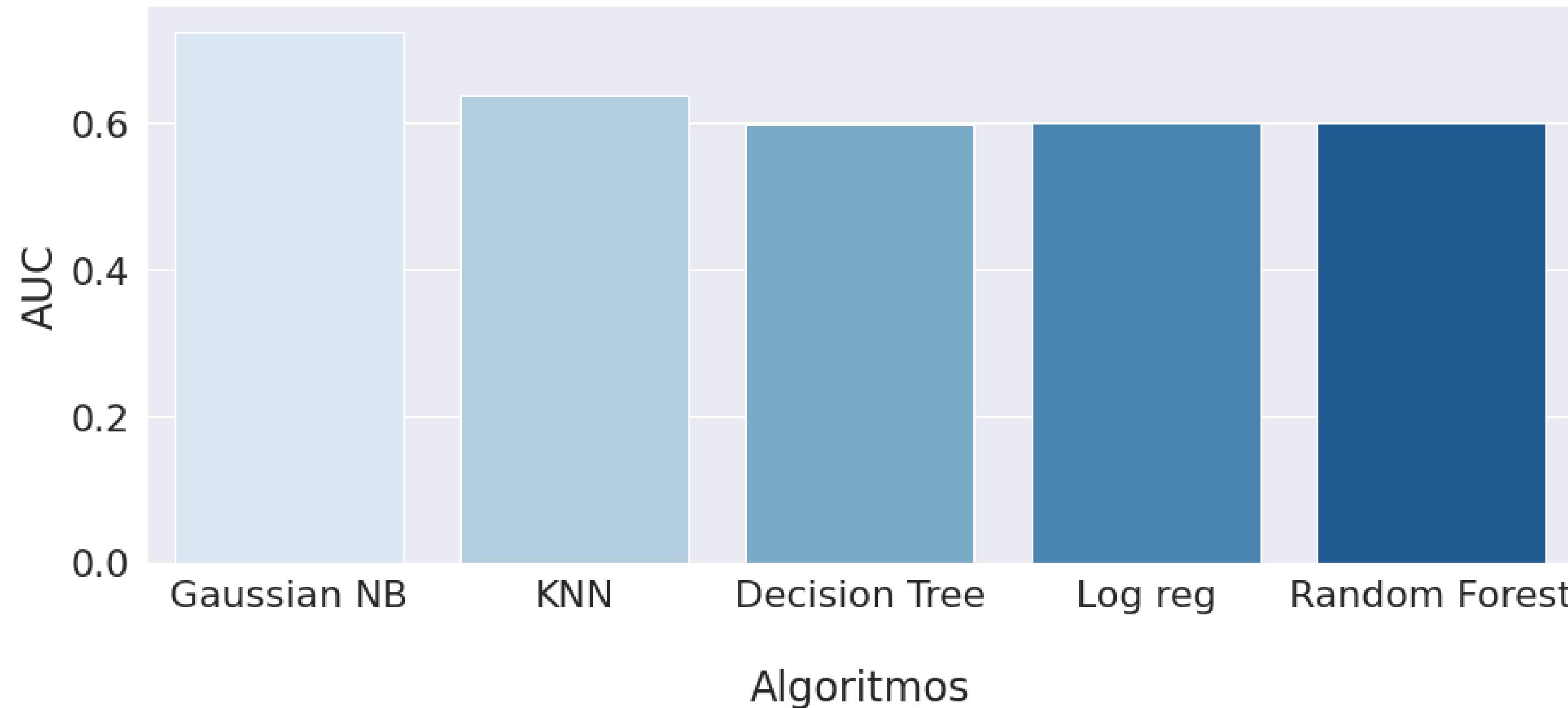
RESULTADOS: PRECISÃO DE CADA MODELO



RESULTADOS: MATRIZ DE CONFUSÃO KNN E NAIVE BAYES



RESULTADOS: AUC DE CADA MODELO

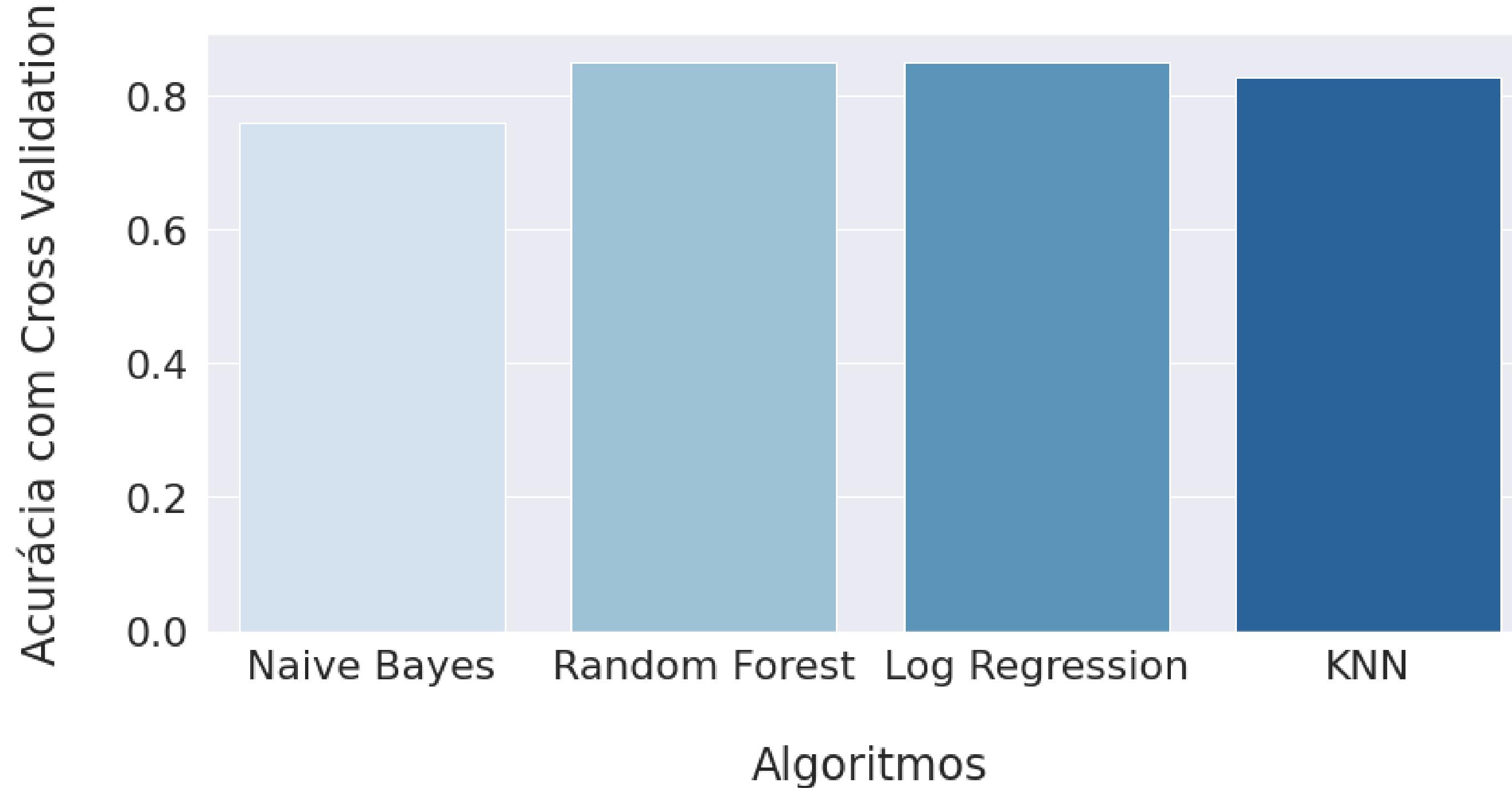


RESULTADOS: ACURÁCIA COM CROSS VALIDATION

Cross Validation é uma técnica muito utilizada para avaliação de desempenho de modelos de Machine Learning. Esse método consiste em particionar os dados em conjuntos, onde um conjunto é utilizado para treino e outro conjunto é utilizado para teste e avaliação do desempenho do modelo.



RESULTADOS: ACURÁCIA COM CROSS VALIDATION



DÚVIDAS E PONTOS DE MELHORIA

Qual será uma possível explicação para o Naive Bayes ter performado melhor na AUC?

Esse problema poderia ser melhor tratado como um problema de regressão?

Tentar enriquecer o dataset com novas variáveis poderia melhorar o desempenho dos modelos?

Utilizar algum tipo de encoding ou scaler poderia melhorar o desempenho dos modelos?

Ter optado por remover todos os valores nulos e não preenchidos pode ter adicionado algum viés aos modelos?

Analizar as acurácia com validação cruzada deve melhorar ou piorar o desempenho dos modelos?

Fim :)

Obrigado!