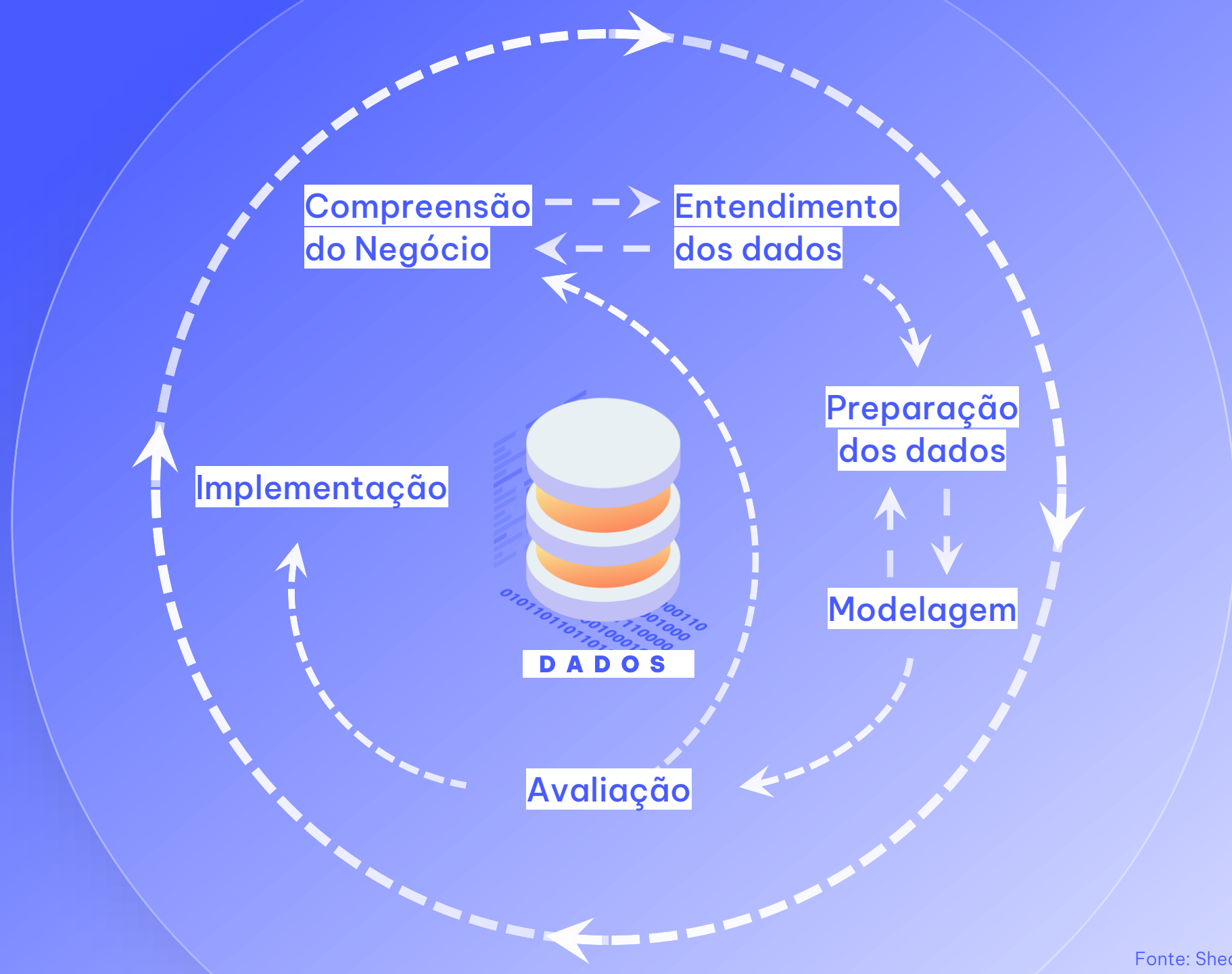




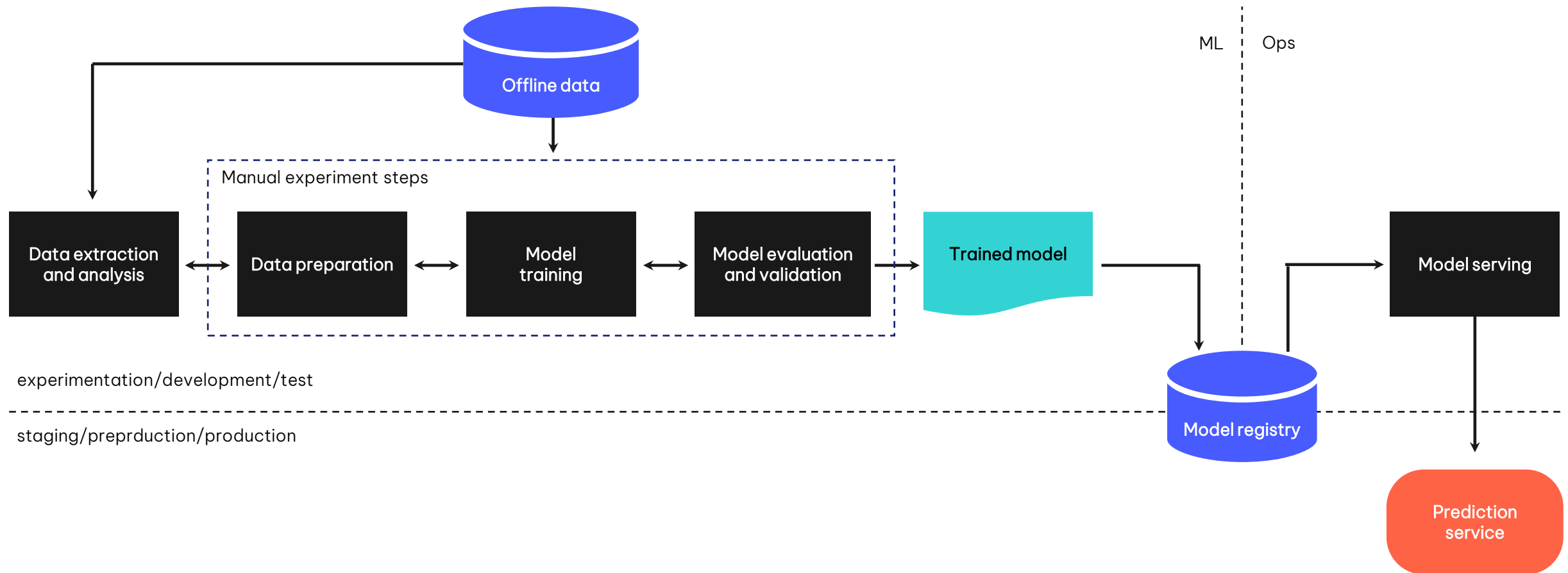
Projeto Final Tutorial MLE

BRENO GARCIA FERRAZ

METODOLOGIA



Pipeline de MLOps de nível 0: processo manual



Fonte: Documentação Google arquitetura IA e ML - MLOps

Compreensão do **Negócio**

Problema

Prever as notas da prova de ciências humanas do ENEM de 2023 a partir dos dados socioeconômicos

Entender como os dados socioeconômicos afetam a nota

Entendimento dos dados

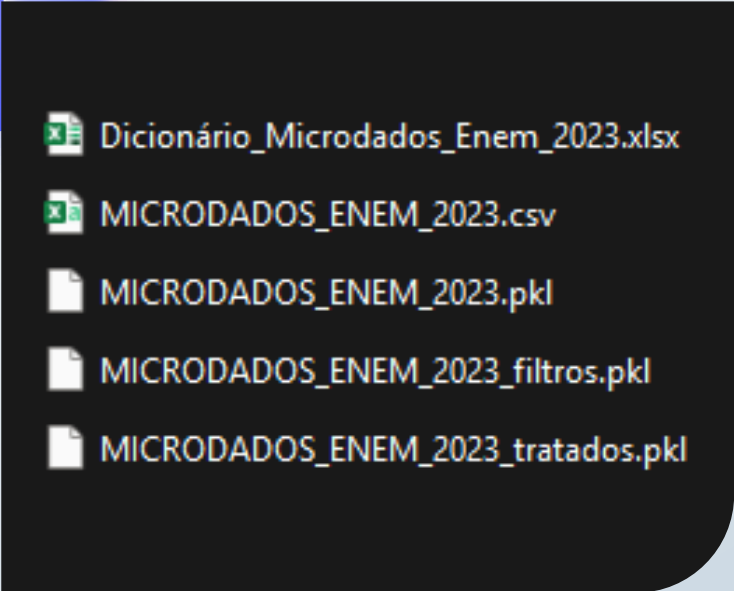
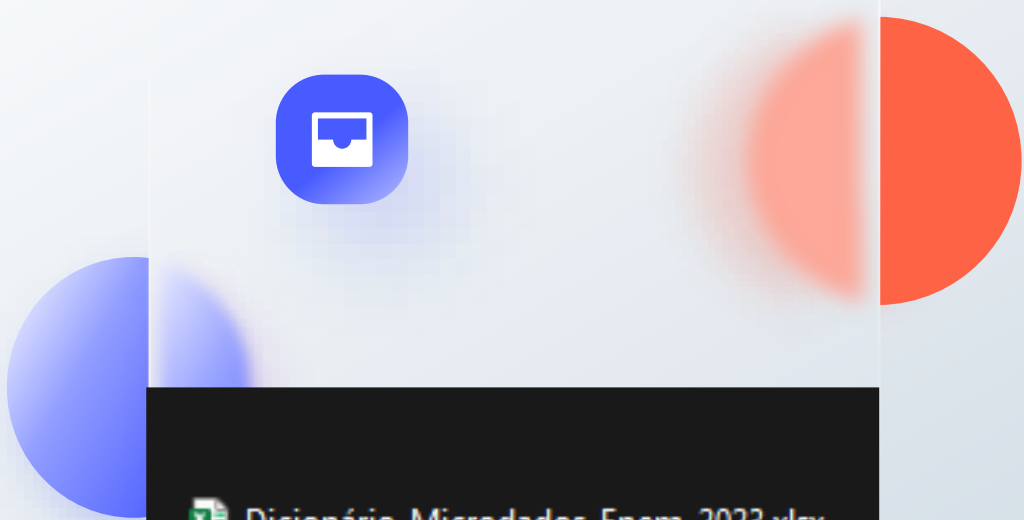

(data extraction and analysis)

- ✓ Site INEP – ENEM 2023
- ✓ (<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>)
- ✓ Dicionário de dados

DICIONÁRIO DE VARIÁVEIS - ENEM 2023					
NOME DA VARIÁVEL	Descrição	Variáveis Categóricas		Tamanho	Tipo
		Categoria	Descrição		
DADOS DO PARTICIPANTE					
NU_INSCRICAO	Número de inscrição ¹			12	Numérica
NU_ANO	Ano do Enem			4	Numérica
TP_FAIXA_ETARIA	Faixa etária ²	1	Menor de 17 anos	2	Numérica
		2	17 anos		
		3	18 anos		
		4	19 anos		
		5	20 anos		
		6	21 anos		
		7	22 anos		
		8	23 anos		
		9	24 anos		
		10	25 anos		
		11	Entre 26 e 30 anos	2	Numérica
		12	Entre 31 e 35 anos		
		13	Entre 36 e 40 anos		

Entendimento do dados

- ✓ Leitura da base de dados (.csv)
- ✓ Verificação de tipos, 'describe' e 'info'
- ✓ Salvar como '.pkl'



Dicionário_Microdados_Enem_2023.xlsx
MICRODADOS_ENEM_2023.csv
MICRODADOS_ENEM_2023.pkl
MICRODADOS_ENEM_2023_filtros.pkl
MICRODADOS_ENEM_2023_tratados.pkl

Entendimento dos dados

- ✓ Explicação das colunas
- ✓ Dados socioeconômicos do formulário do ENEM
- ✓ 76 colunas originais, foram selecionadas 38
- ✓ Nota prevista CH

PRESENCA_ASPIRADOR
PRESENCA_DVD
PRESENCA_TV_ASSINATURA
PRESENCA_TEL_FIXO,
PRESENCA_INTERNET
COR_RACA
CO_MUNICIPIO_ESC
CO_UF_ESC
DEPENDENCIA_ADM_ESC
ENSINO
ESCOLA
ESTADO_CIVIL
FAIXA_ETARIA
LINGUA
LOCALIZACAO_ESC
NACIONALIDADE
OCUPACAO_PAI
OCUPACAO_MAE
SEXO
SIT_FUNC_ESC
GRAU_ESTUDO_PAI
GRAU_ESTUDO_MAE
QTD_RESIDENTES
RENDA_MENSAL_FAMILIA
FREQ_EMPREGADO
QTD_BANHEIRO
QTD_QUARTO
QTD_CARRO
QTD_MOTO
QTD_GELADEIRA
QTD_FREEZER
QTD_MAQ_LAVAR_ROUPA
QTD_MAQ_SECAR
QTD_MICROONDAS
QTD_MAQ_LAVAR_LOUCA
QTD_TELEVISOR
QTD_CELULAR
QTD_COMPUTADOR

Entendimento dos dados

— Exploração dos dados por tipo (entrada e saída):



Numéricos

- Describe: Assimetria, Outliers e Normalidade
- Correlações
- Relação entre alvo e variáveis



Binários

- Proporção verdadeiros/falsos
- Relação entre alvo e variáveis



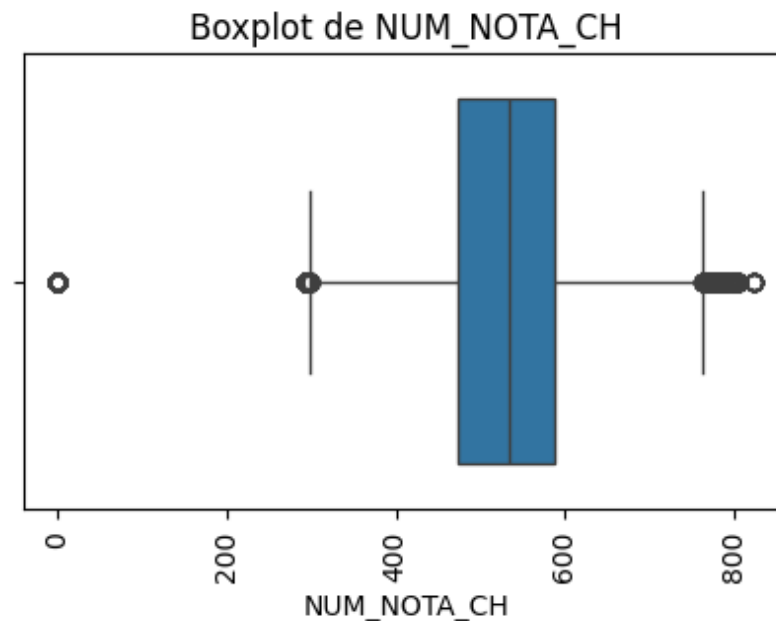
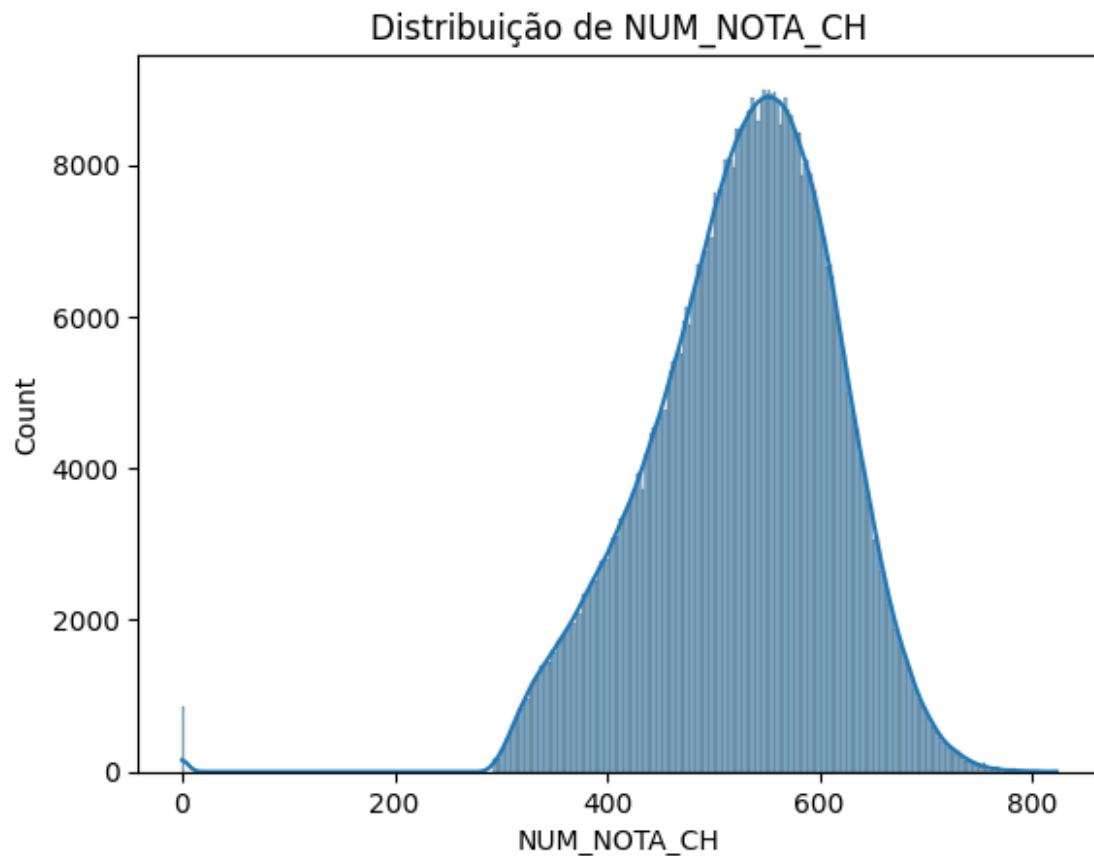
Catégoricos

- Classes desbalanceadas e padrões claros
- Relação entre alvo e variáveis
- Crosstab

NUMÉRICOS

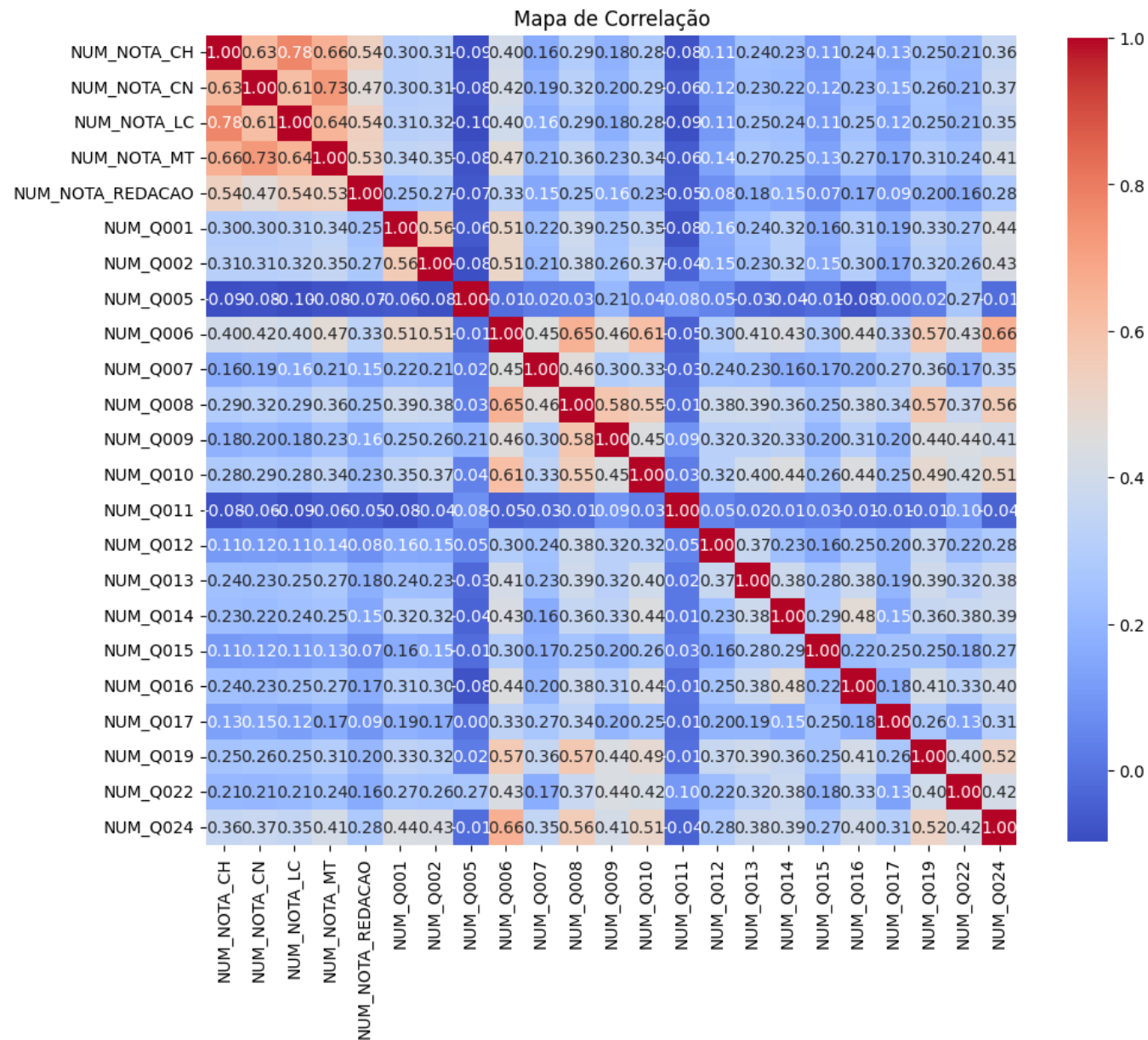
Describe: Assimetria, Outliers e Normalidade

Nota disciplina Ciências Humanas



NUMÉRICOS

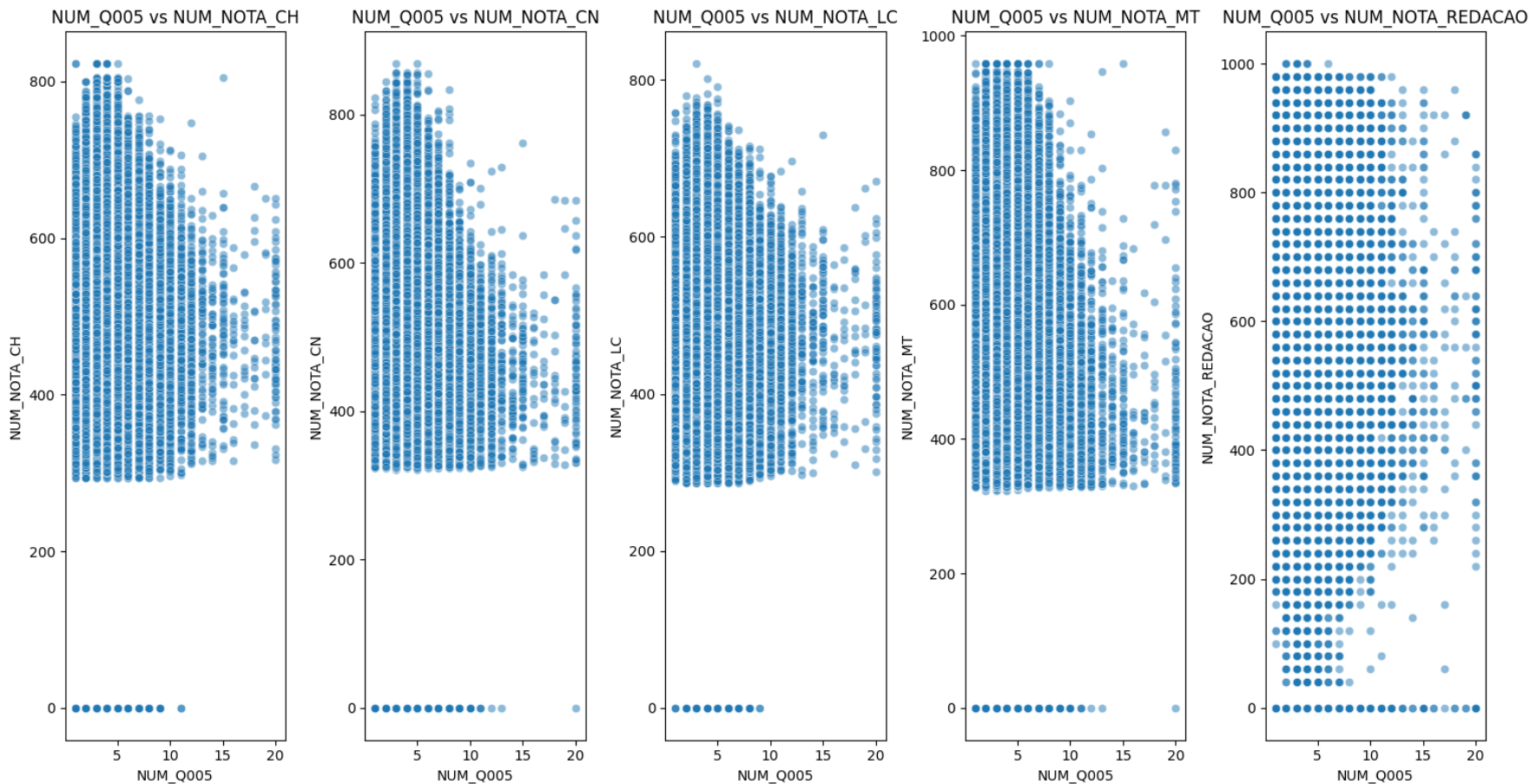
Correlações



NUMÉRICOS

Relação entre alvo e variáveis

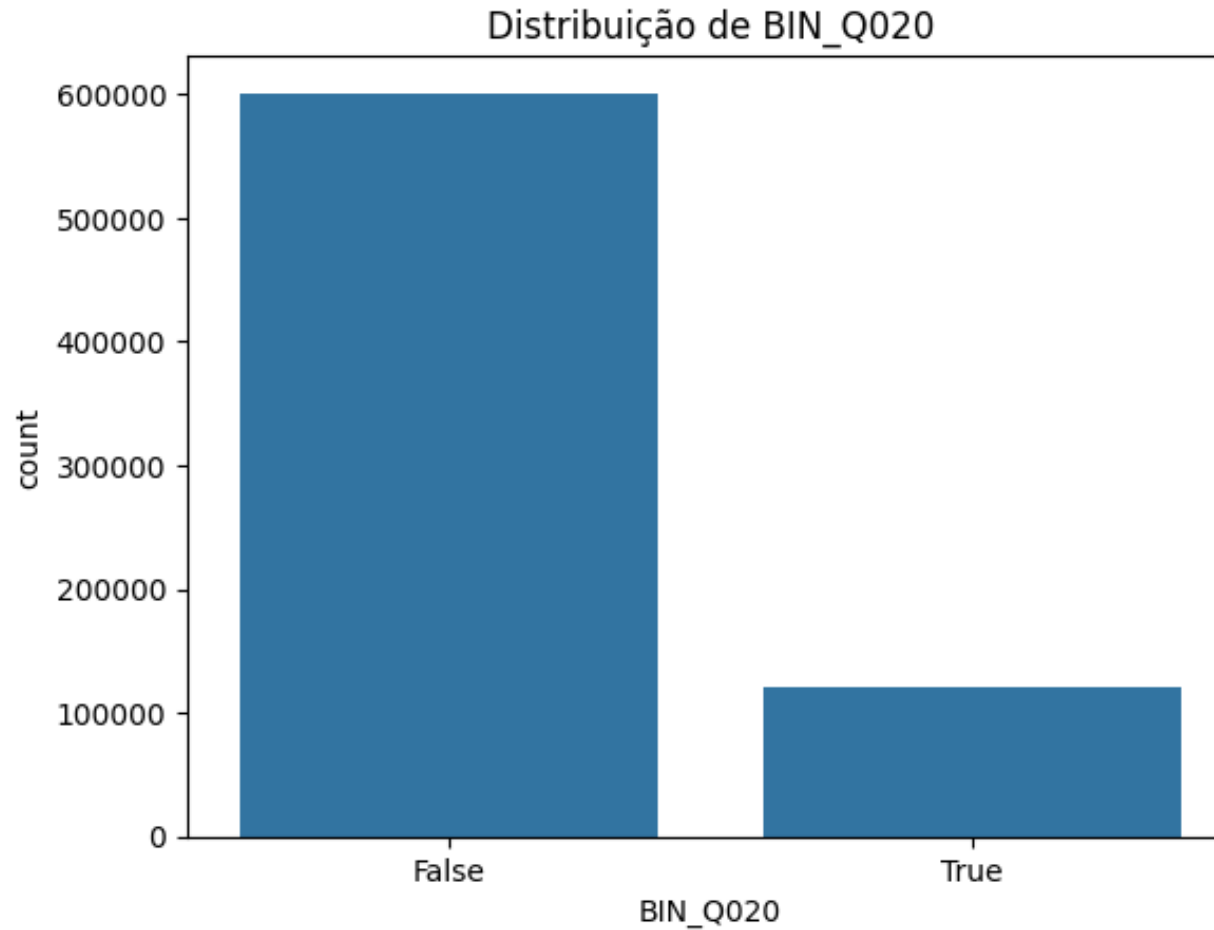
Relação entre Notas ENEM e Q005
(número de residentes)



BINÁRIOS

Proporção verdadeiros/falsos

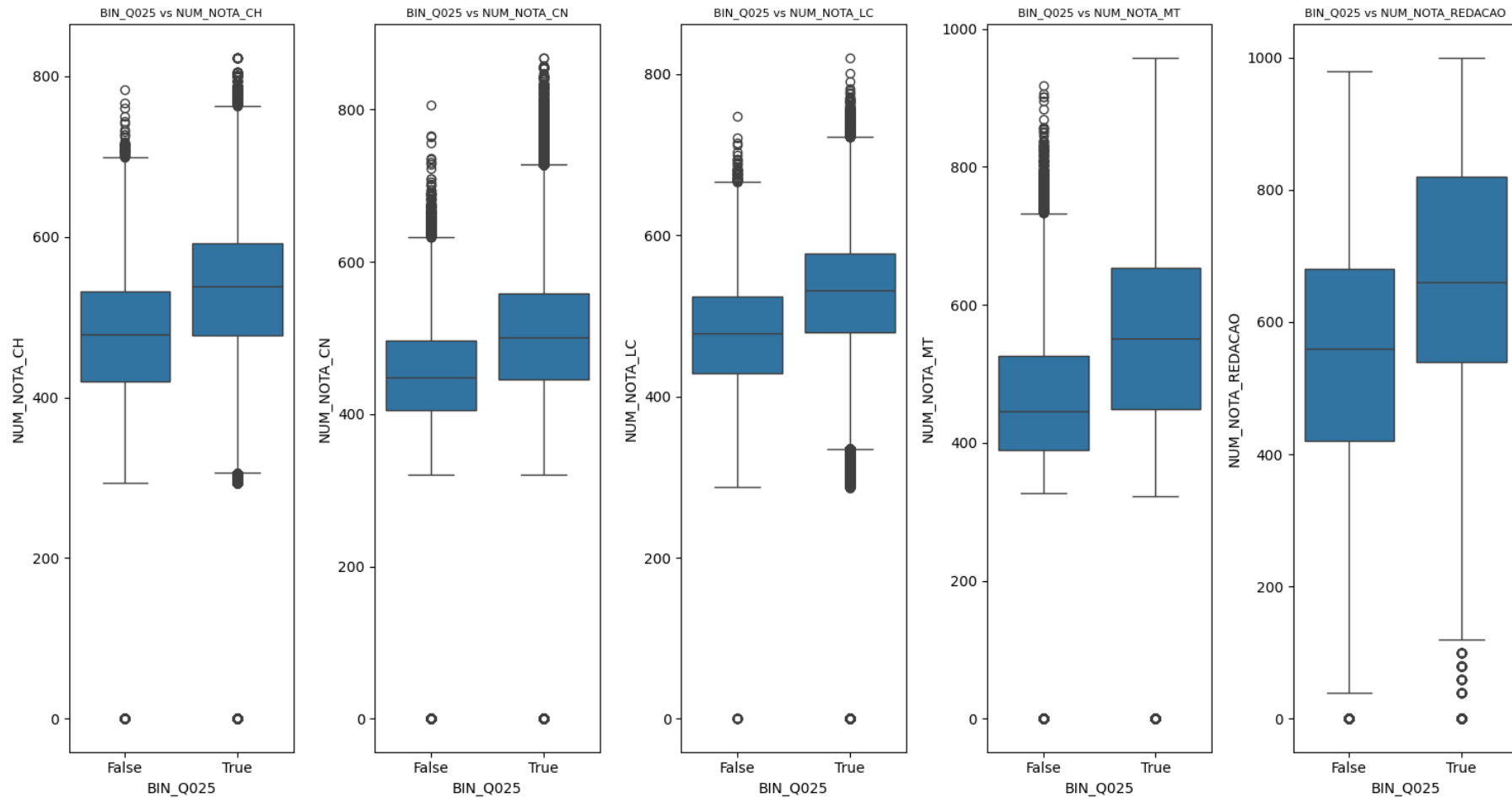
Contagem de verdadeiro de falsos Q020
(Possui aparelho de DVD)



BINÁRIOS

Relação entre alvo e variáveis

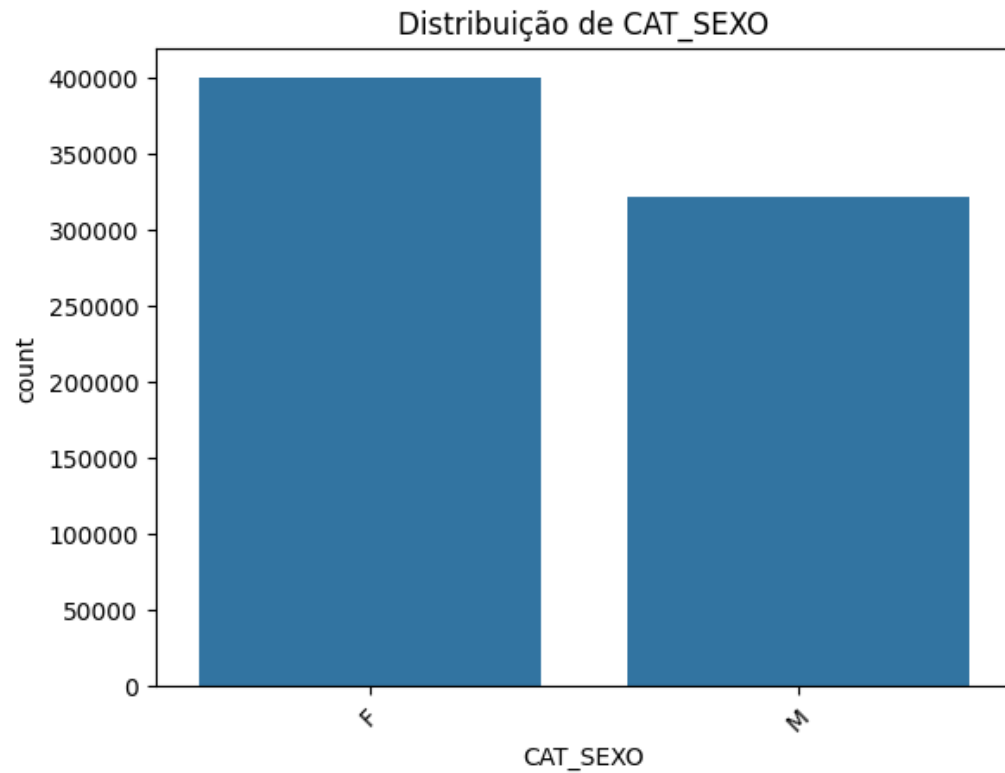
Relação entre Notas ENEM e Q025
(número de residentes)



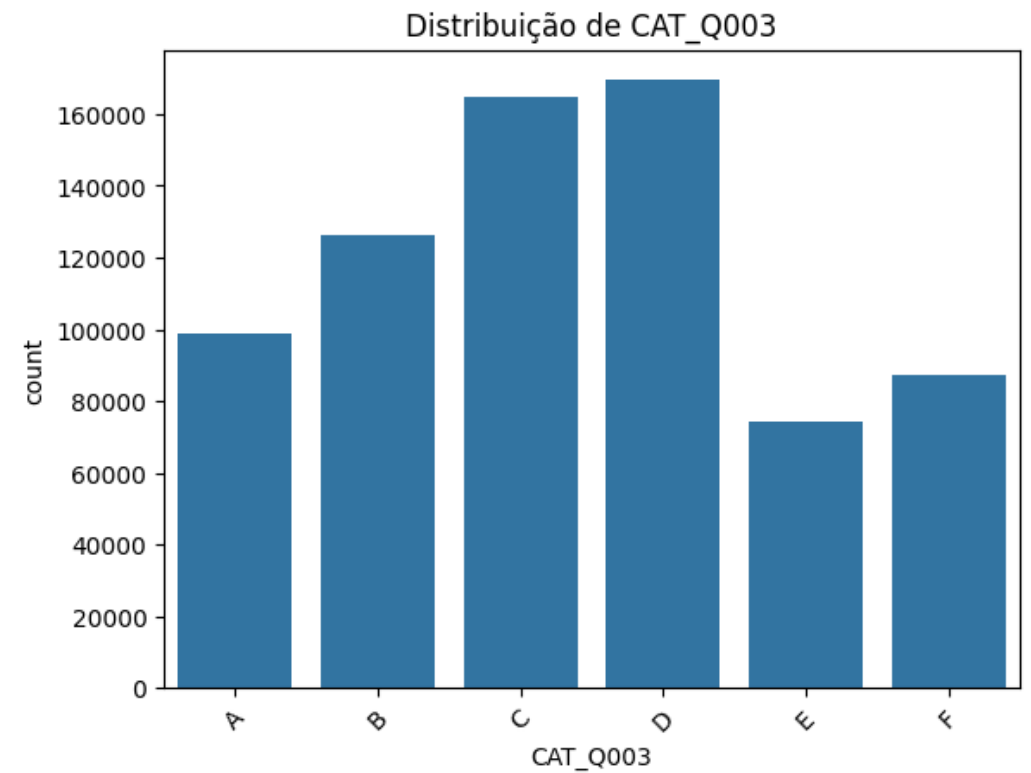
CATEGÓRICOS

Classes desbalanceadas e padrões claros

Contagem sexo



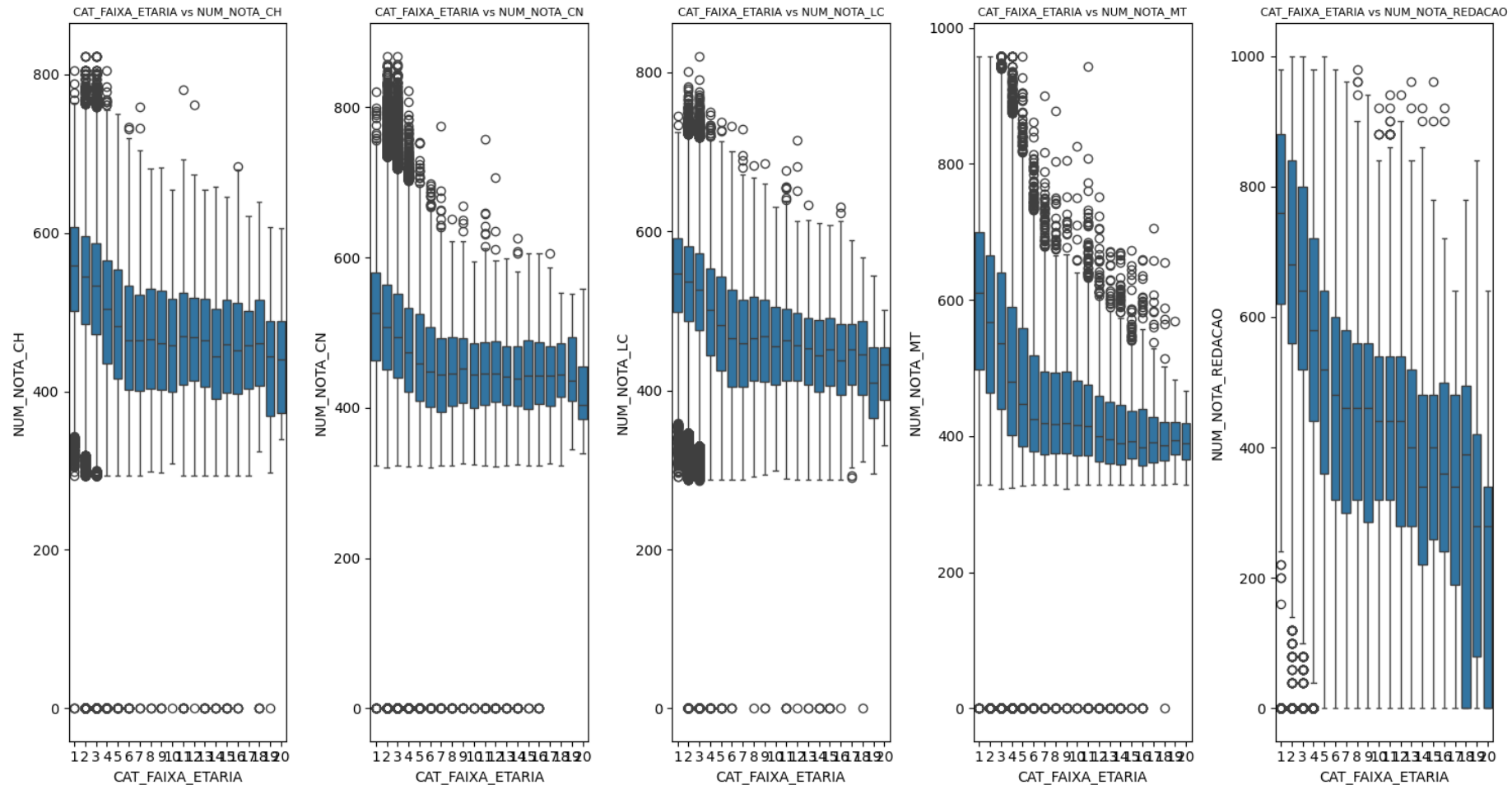
Contagem Q003 (ocupação pai)



CATEGÓRICOS

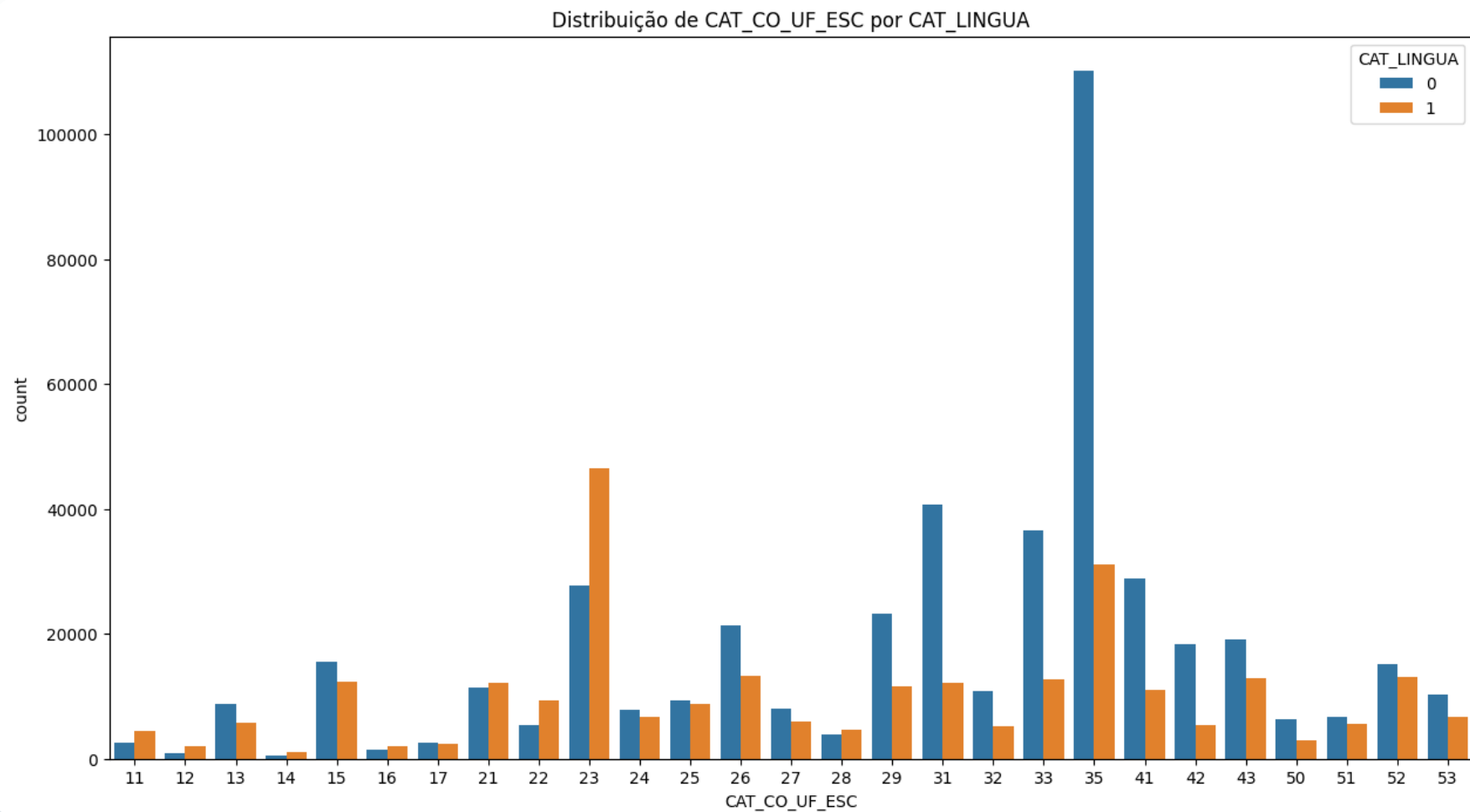
Relação entre alvo e variáveis

Relação notas ENEM e faixa etária



CATEGÓRICOS

Crosstab



Preparação dos dados

01

APLICAÇÃO DE FILTROS

- Manter alunos com código de município da escola preenchido
 - Apenas alunos do terceiro ano do ensino médio estão entre esses, porém não necessariamente todos
- Alunos que possuem registro em todas as cinco provas: remover nulos
- Remover alunos que zeraram alguma prova exceto redação

VOLUME DE DADOS

3.933.955



958.506



721.429



716.944

Preparação dos dados

— Análise de alunos que zeraram alguma prova

- ✓ Significado nota zero: ausência, eliminação e critério redação (branco, fuga do tema etc.)
- ✓ Dia 1: LC, CH, RED e Dia 2: CN e MT
- ✓ CONCLUSÃO:
 - Poucos alunos faltam no primeiro dia zerando as 3 disciplinas: 65
 - Muitos alunos deixam de ir no segundo dia: 3.572
 - Maior parte das notas zero do segundo dia é por ausência
 - Possivelmente maior parte das notas zero no primeiro dia é por não dar tempo de fazer a disciplina ou desistência/eliminação

PROVA	CONTAGEM	%
CH	864	0,12
CN	3572	0,49
LC	223	0,03
MT	3595	0,49
RED	28.763	3,98

PROVA	CONTAGEM ZEROS
Apenas LC	31
Apenas CH	531
Apenas RED	28.763
LC, CH e RED	65

PROVA	CONTAGEM ZEROS
Apenas CN	32
Apenas MT	35
CN e MT	3572

Preparação dos dados

02

LIMPEZA DOS DADOS

- Remover colunas desnecessárias – não agregam valor
- Remoção de n/a ou nulos
 - TP_ENSINO: única variável com nulos
 - Outras variáveis da base possuem o campo "Não informado"
 - Salvar como 0 (Não informado)

TP_ENSINO

CATEGORIA	CONTAGEM	%
1 (Regular)	702.073	97,92
2 (Especial)	2.334	0,33
Nulos	12.537	1,75

Preparação dos dados

03

AJUSTES DE TIPOS

- 'floats' que podem ser 'ints' (ocupa menos espaço e simplifica base)
- Variáveis de categorias como tipo 'category'

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 716944 entries, 0 to 716943
Data columns (total 43 columns):
#   Column                Non-Null Count  Dtype
---  -
0   TP_FAIXA_ETARIA        716944 non-null  category
1   TP_SEXO                716944 non-null  category
2   TP_ESTADO_CIVIL        716944 non-null  category
3   TP_COR_RACA            716944 non-null  category
4   TP_NACIONALIDADE       716944 non-null  category
5   TP_ESCOLA              716944 non-null  category
6   NU_NOTA_CN             716944 non-null  float64
7   NU_NOTA_CH             716944 non-null  float64
8   NU_NOTA_LC             716944 non-null  float64
9   NU_NOTA_MT             716944 non-null  float64
10  TP_LINGUA              716944 non-null  category
11  NU_NOTA_REDACAO        716944 non-null  float64
12  Q001                   716944 non-null  category
13  Q002                   716944 non-null  category
14  Q003                   716944 non-null  category
15  Q004                   716944 non-null  category
16  Q005                   716944 non-null  int64
17  Q006                   716944 non-null  category
```

Preparação dos dados



Remover ou manter uma categoria?



Remover a categoria quando:

- há proporção baixa no grupo
- marcar como 'não informado' ou 'ausente' não possuir um motivo aparente para omissão da informação
- é claramente um outlier ou ruído



Manter a categoria quando:

- apresenta comportamento relevante: é um grupo claramente significativo, não aleatório
- grande proporção (remover pode viciar o conjunto)
- categoria possui estatísticas diferentes das outras categorias da variável, pode agregar informação



Não é possível 'inputar' dados nestas variáveis para substituir a categoria

Preparação dos dados

04

REMOÇÃO DE CATEGORIAS

4.1 Análise e ajuste de campos que podem trazer viés aleatório

- TP_ESTADO_CIVIL: 0 (Não informado)
- TP_COR_RACA: 0 (Não declarado)
- TP_NACIONALIDADE: 0 (Não informado)
- TP_ESCOLA: 1 (Não Respondeu)
- TP_ENSINO: 0 (Não informado)

VARIÁVEL	CONTAGEM	%
TP_ESTADO_CIVIL	25.896	3,61
TP_COR_RACA	8.056	1,12
TP_NACIONALIDADE	372	0,05
TP_ESCOLA	2	0,00
TP_ENSINO	12.537	1,75

Preparação dos dados

04

REMOÇÃO DE CATEGORIAS

4.1.1 TP_ESTADO_CIVIL: 0 (Não informado) --> **MANTER**

- Possui uma quantidade representativa de dados (3,61%)
- Média de notas da categoria: parece haver uma diferença, podendo indicar algum significado.

Amostra	Média CH	Dif. Média	Desv P.
Base	527	0	84
-----	-----	-----	-----
Cat 0	515	-12	83
Cat 1	528	+1	84
Cat 2	488	-39	83
Cat 3	513	-14	84
Cat 4	480	-47	83

```
TP_ESTADO_CIVIL
0      25896
1      681625
2        4119
3        5116
4         188
```


Preparação dos dados

04

REMOÇÃO DE CATEGORIAS

4.1.2 TP_COR_RACA: 0 (Não declarado) --> **MANTER**

- Possui uma quantidade relativa de dados (1,12%)
- Média de notas da categoria: parece haver uma diferença, podendo indicar algum significado.

Amostra	Média CH	Dif. Média	Desv P.
Base	527	0	84
-----	-----	-----	-----
Cat 0	519	-8	90
Cat 1	549	+22	80
Cat 2	506	-21	81
Cat 3	507	-20	83
Cat 4	523	- 4	88
Cat 5	477	- 50	78

```
TP_COR_RACA
0      8056
1    347156
2    68341
3    279669
4    10409
5     3313
```

Obs.: há uma opção 6 de "Não dispõe", porém não aparece nesse banco de dados filtrado.

Deve-se unir os campos 0 ("Não informado") e 6 ("Não dispo"), transformar ambos em 0.

Preparação dos dados

04

REMOÇÃO DE CATEGORIAS

4.1.3 TP_NACIONALIDADE: 0 (Não declarado) --> **REMOVER**

- Representa apenas 0,05% da base.
- Considerando a baixa representatividade e o potencial ruído que pode causar, optou-se por remover as linhas com esta categoria.

TP_NACIONALIDADE	
0	372
1	701524
2	11606
3	1488
4	1954

Preparação dos dados

04

REMOÇÃO DE CATEGORIAS

4.1.4 TP_ESCOLA: 1 (Não Respondeu) --> REMOVER

- Representa menos de 0,001% da base.
- Considerando sua baixa representatividade e o potencial de ruído, optou-se por remover essas linhas do conjunto de dados.

TP_ESCOLA	
1	2
2	523156
3	193786

Preparação dos dados

04

REMOÇÃO DE CATEGORIAS

4.1.5 TP_ENSINO: 0 (Não informado) --> **MANTER**

- Possui uma quantidade relativa de dados (1,75%)
- Média de notas da categoria: parece haver uma diferença, podendo indicar algum significado.

Amostra	Média CH	Dif. Média	Desv P.
Base	527	0	84
-----	-----	-----	-----
Cat 0	476	-51	78
Cat 1	528	+1	84
Cat 2	498	-39	81

```
TP_ENSINO
0      12537
1      702073
2       2334
```

Preparação dos dados

05

FEATURE ENGINEERING

5.1 Variáveis categóricas que podem ser numéricas

- Campos efetivamente com significados numéricos são as notas e a questão 5
- Vários podem ser transformados em numéricos usando quantificação dos itens da variável



MOTIVO

- A variável é ordinal (existe uma ordem lógica entre as categorias).
- A conversão mantém essa ordem, transformando-a em variável numérica contínua discreta.



PERMITE

- Usar modelos que se beneficiam de relações numéricas diretas, como regressões lineares e redes neurais.
- Evitar codificação densa (como OneHot) que pode gerar alta dimensionalidade e desperdiçar a ordem natural.

Preparação dos dados

05

Variáveis categóricas que podem ser numéricas: perguntas de quantidade

- 'Q008' Quantidade de banheiro (código A a E)
- 'Q009' Quantidade de quartos (código A a E)
- 'Q010' Quantidade de carros (código A a E)
- 'Q011' Quantidade de motos (código A a E)
- 'Q012' Quantidade de geladeira (código A a E)
- 'Q013' Quantidade de freezer (código A a E)
- 'Q014' Quantidade de máquina de lavar roupa (código A a E)
- 'Q015' Quantidade de máquina de secar roupa (código A a E)
- 'Q016' Quantidade de micro-ondas (código A a E)
- 'Q017' Quantidade de máquina de lavar louça (código A a E)
- 'Q019' Quantidade de televisores (código A a E)
- 'Q022' Quantidade de celulares (código A a E)
- 'Q024' Quantidade de computadores (código A a E)

A. 0 itens
B. 1 item
C. 2 itens
D. 3 itens
E. 4 ou mais

Categoria
A, B, C, D, E



Valor
0, 1, 2, 3, 5

- Analisada proporção de cada categoria
 - Q008 (banheiro): 5% "4 ou mais"
 - Q022 (celular): 28% "4 ou mais"
- Estimativa conservadora de 5 por baixa porcentagem em geral (busca média provável)

QTD_Q008	716570	non-null	int64
QTD_Q009	716570	non-null	int64
QTD_Q010	716570	non-null	int64
QTD_Q011	716570	non-null	int64
QTD_Q012	716570	non-null	int64
QTD_Q013	716570	non-null	int64
QTD_Q014	716570	non-null	int64
QTD_Q015	716570	non-null	int64
QTD_Q016	716570	non-null	int64
QTD_Q017	716570	non-null	int64
QTD_Q019	716570	non-null	int64
QTD_Q022	716570	non-null	int64
QTD_Q024	716570	non-null	int64

Preparação dos dados

05 5.2 Variáveis categóricas que podem ser numéricas + dummy

- 'Q001' Grau de escolaridade do pai
- 'Q002' Grau de escolaridade da mãe

- A.** Nunca estudou.
- B.** Não completou a 4ª série/5º ano do Ensino Fundamental.
- C.** Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
- D.** Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
- E.** Completou o Ensino Médio, mas não completou a Faculdade.
- F.** Completou a Faculdade, mas não completou a Pós-graduação.
- G.** Completou a Pós-graduação.
- H.** Não sei.

Q001

A	0.023233
B	0.107423
C	0.112053
D	0.115831
E	0.320373
F	0.119138
G	0.097870
H	0.104078

Q002

A	0.012056
B	0.067653
C	0.085895
D	0.114562
E	0.369499
F	0.153181
G	0.157638
H	0.039515

- 10% não sabe a escolaridade do pai
- 4% não sabe a escolaridade da mãe

Preparação dos dados

05

SOLUÇÃO

- coluna numérica + uma coluna dummy

1. Conversão ordinal para numérico

- As categorias A a G representam níveis crescentes de escolaridade, então foram transformadas em valores numéricos de 0 a 6, respeitando essa ordem.

2. Tratamento da categoria "H" (Não sei)

- Como "H" não representa um nível de escolaridade real, ela não foi incluída na escala ordinal, sendo criada uma variável binária (dummy) indicando se a resposta foi "H" (1) ou não (0).

3. Imputação dos valores "H" na variável numérica

- Os casos com "H" foram preenchidos com a mediana dos demais valores numéricos válidos, para evitar distorções nos modelos.

Essa abordagem ajuda o modelo a entender tanto a escolaridade quanto a ausência dessa informação.



Q001_DUMMY_H	716570	non-null	bool
Q001_NUM	716570	non-null	int64
Q002_DUMMY_H	716570	non-null	bool
Q002_NUM	716570	non-null	int64

Preparação dos dados

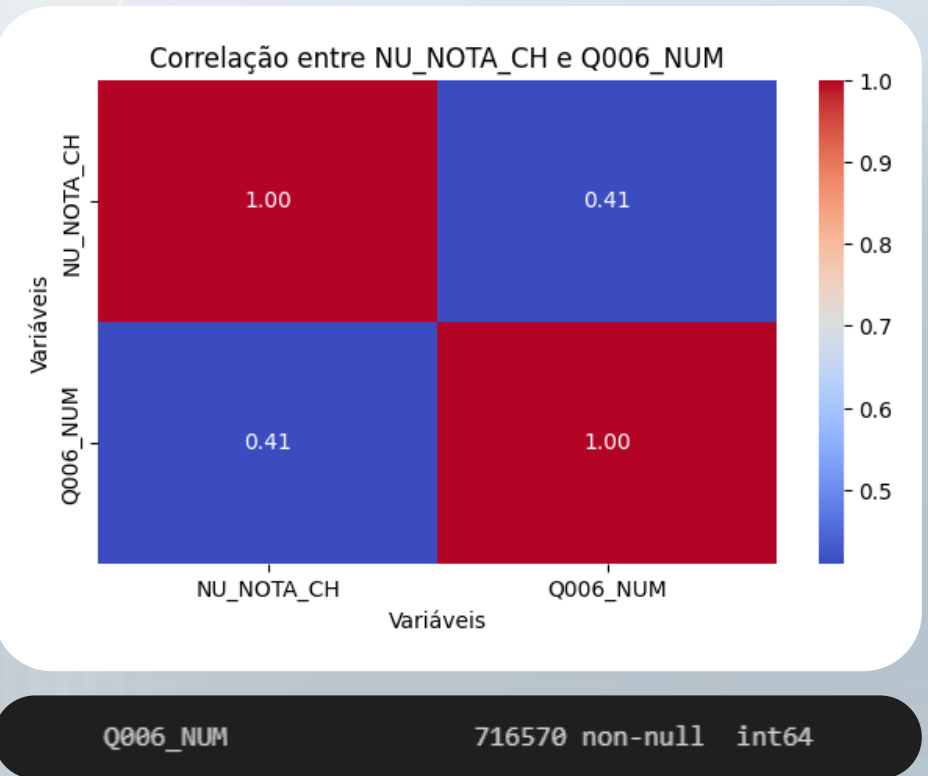
05

5.3 Conversão para numérica colunas com ordem lógica: renda

- A. Nenhuma Renda
- B. Até R\$ 1.320,00
- C. De R\$ 1.320,01 até R\$ 1.980,00.
- D. De R\$ 1.980,01 até R\$ 2.640,00.
- E. De R\$ 2.640,01 até R\$ 3.300,00.
- F. De R\$ 3.300,01 até R\$ 3.960,00.
- G. De R\$ 3.960,01 até R\$ 5.280,00.
- H. De R\$ 5.280,01 até R\$ 6.600,00.
- I. De R\$ 6.600,01 até R\$ 7.920,00.
- J. De R\$ 7.920,01 até R\$ 9.240,00.
- K. De R\$ 9.240,01 até R\$ 10.560,00.
- L. De R\$ 10.560,01 até R\$ 11.880,00.
- M. De R\$ 11.880,01 até R\$ 13.200,00.
- N. De R\$ 13.200,01 até R\$ 15.840,00.
- O. De R\$ 15.840,01 até R\$ 19.800,00.
- P. De R\$ 19.800,01 até R\$ 26.400,00.
- Q. Acima de R\$ 26.400,00.

	CH_MEDIA	CH_DESVP
Q006		
A	471.987819	75.815146
B	487.590602	78.221274
C	512.013869	77.623855
D	524.863413	77.103934
E	534.117018	76.564661
F	542.872168	76.577475
G	551.354456	75.650048
H	561.115961	75.900236
I	566.928161	75.504468
J	572.064128	75.232491
K	576.701892	73.895251
L	580.241201	73.626003
M	588.239723	72.123003
N	589.535600	71.846202
O	594.494189	72.454061
P	602.992898	71.269680
Q	607.385957	70.669476

- As médias das notas aumentam conforme a renda sobe (forte correlação)
- Evita a explosão de variáveis com dummies



Preparação dos dados

05

5.3 Conversão para numérica colunas com ordem lógica: empregado

- o 'Q007' Em sua residência trabalha empregado(a) doméstico(a)?

- A. Não.
- B. Sim, um ou dois dias por semana.
- C. Sim, três ou quatro dias por semana.
- D. Sim, pelo menos cinco dias por semana.

Q007	
A	0.888261
B	0.061109
C	0.013972
D	0.036658

	CH_MEDIA	CH_DESVP
Q007		
A	522.320705	82.962979
B	571.012899	83.180977
C	572.019101	86.678891
D	574.705616	85.874905

Há uma relação crescente entre a frequência de trabalho doméstico e as médias das notas.

Q007_NUM	716570 non-null	int64
----------	-----------------	-------

Preparação dos dados

05

FEATURE ENGINEERING

5.4 Variáveis categóricas que podem ser binárias



MOTIVO

- Reflete bem o significado semântico
- Melhora compatibilidade com algoritmos



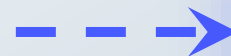
PERMITE

- Análise estatística direta (Correlações, médias condicionais etc.)
- Evita One-Hot desnecessário

Preparação dos dados

- 'Q018' Existência de aspirador de pó (A ou B)
- 'Q020' Existência de DVD (A ou B)
- 'Q021' Existência de TV por assinatura (A ou B)
- 'Q023' Existência de telefone fixo (A ou B)
- 'Q025' Existência de internet (A ou B)

Categoria
'Não', 'Sim'



Valor
0 ou 1

```
BIN_Q018      716570 non-null bool
BIN_Q020      716570 non-null bool
BIN_Q021      716570 non-null bool
BIN_Q023      716570 non-null bool
BIN_Q025      716570 non-null bool
```

- As variáveis com apenas duas categorias (ex: possui ou não um bem)
- Convertidas diretamente para valores binários (0 = não possui, 1 = possui).

Preparação dos dados

05

5.5 Conversão para numérica colunas com ordem lógica: ocupação

- 'Q003' Ocupação pai
- 'Q004' Ocupação mãe

- A.** Grupo 1 ex: lavrador
- B.** Grupo 2 ex: diarista
- C.** Grupo 3 ex: padeiro
- D.** Grupo 4 ex: professor
- E.** Grupo 5 ex: médico
- F.** Não sei

Q003	
A	0.137084
B	0.174984
C	0.228163
D	0.235666
E	0.103376
F	0.120728

	CH_MEDIA	CH_DESVP
Q003		
A	483.384661	80.208439
B	512.610479	79.556114
C	521.908523	78.070217
D	558.600366	77.887702
E	583.014806	77.762306
F	504.943209	81.669708

Q004	
A	0.107740
B	0.363494
C	0.065280
D	0.291415
E	0.077854
F	0.094217

	CH_MEDIA	CH_DESVP
Q004		
A	476.282000	78.783898
B	515.747708	79.305178
C	519.811190	78.918916
D	554.916415	79.134444
E	583.056140	79.314432
F	510.445088	83.771192

- Nota-se uma média crescente da nota entre as categorias A e E, com F de exceção
- Será mantida F por trazer uma informação relevante e ter o desvio padrão próximo aos outros
- Será mantida como categórica

Preparação dos dados

06

AJUSTES FINAIS E EXPORTAÇÃO DA BASE

- Realizada padronização dos nomes por tipo
- Reset no index

ENTRADAS

```
'NUM_Q001', 'NUM_Q002',  
'NUM_Q005', 'NUM_Q006',  
'NUM_Q007', 'NUM_Q008',  
'NUM_Q009', 'NUM_Q010',  
'NUM_Q011', 'NUM_Q012',  
'NUM_Q013', 'NUM_Q014',  
'NUM_Q015', 'NUM_Q016',  
'NUM_Q017', 'NUM_Q019',  
'NUM_Q022', 'NUM_Q024'
```

```
'BIN_Q001_DUMMY_H', 'BIN_Q002_DUMMY_H',  
'BIN_Q018', 'BIN_Q020', 'BIN_Q021',  
'BIN_Q023', 'BIN_Q025',
```

```
'CAT_COR_RACA', 'CAT_CO_MUNICIPIO_ESC',  
'CAT_CO_UF_ESC', 'CAT_DEPENDENCIA_ADM_ESC',  
'CAT_ENSINO', 'CAT_ESCOLA', 'CAT_ESTADO_CIVIL',  
'CAT_FAIXA_ETARIA', 'CAT_LINGUA',  
'CAT_LOCALIZACAO_ESC', 'CAT_NACIONALIDADE',  
'CAT_Q003', 'CAT_Q004', 'CAT_SEXO',  
'CAT_SIT_FUNC_ESC',
```

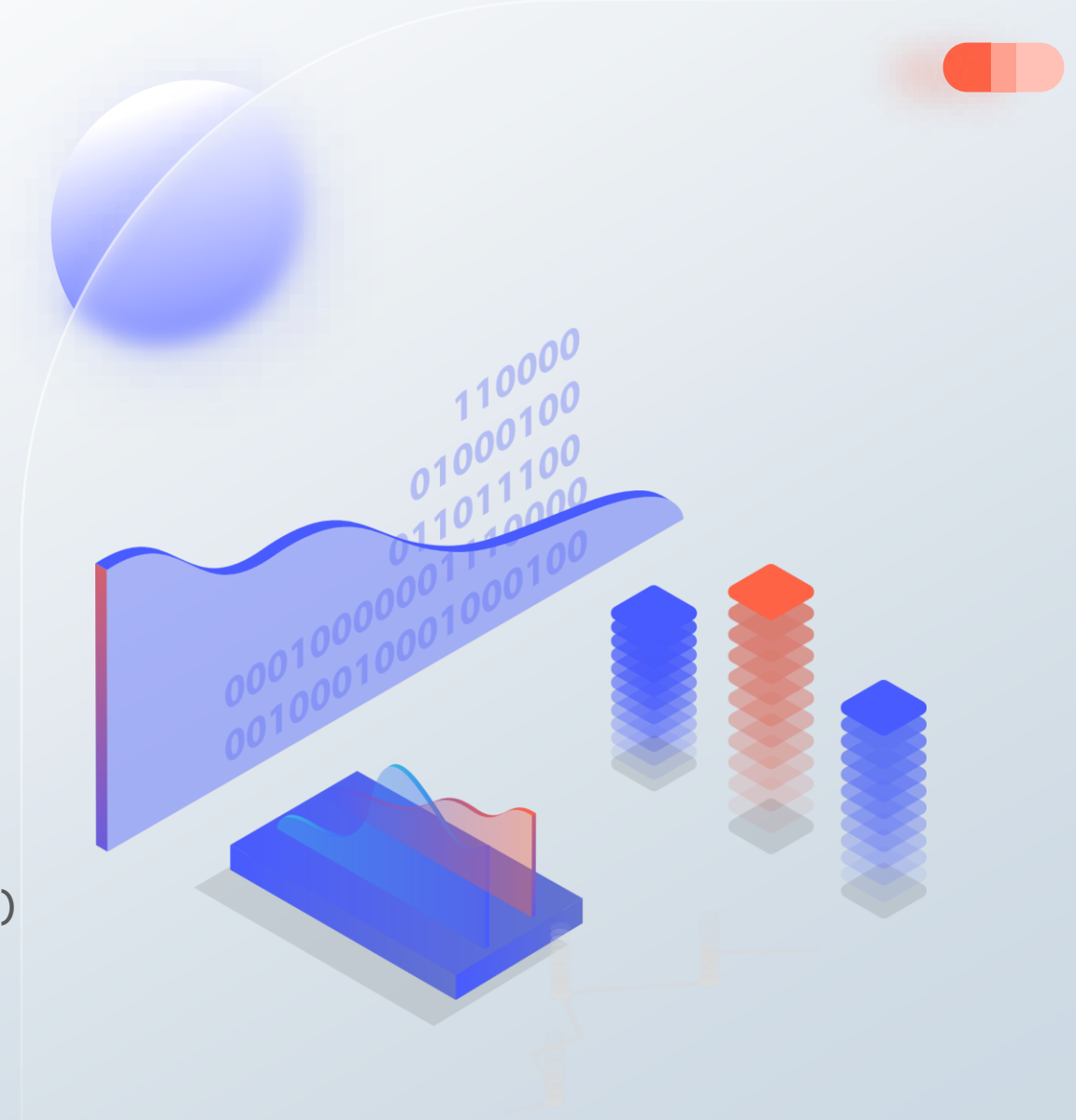
SAÍDAS

```
'NUM_NOTA_CH',  
'NUM_NOTA_CN',  
'NUM_NOTA_LC',  
'NUM_NOTA_MT',  
'NUM_NOTA_REDACAO',
```

Modelagem

ESTRATÉGIA

- Regressão
- Selecionar métricas
- Leitura da base tratada e encoding se necessário
- Seleção de grupo de treino e teste (20%)
- Treinamento rápido (volume de dados e tempo disponível)
- Modelos que lidam bem com dados categóricos
- Aplicar modelo baseline simples
- Aplicar mais de uma técnica (Árvore de Decisão e LightGBM)
- Comparar resultados (MLFlow)
- Realizar Tuning (Grid e Bayes Search)
- Selecionar melhores modelos e exportar (.pkl)



Modelagem

MÉTRICAS

R^2

**Coeficiente de
Determinação**

mede a proporção da
variabilidade explicada,
métrica generalista

MAE

Erro Absoluto Médio

menor penalização para
outliers, reduzindo a distorção
por grandes erros, mantém a
mesma unidade de medida
original dos dados

RMSE

**Raiz do Erro Quadrático
Médio**

penaliza erros maiores,
sendo sensível a outliers e
mantendo a unidade de
medida original dos dados

Modelagem

Árvore de Decisão

Estrutura-se em divisões sucessivas dos dados com base em perguntas binárias, formando subconjuntos mais homogêneos

* **Label Encoding:**
converte colunas
categóricas em
numéricas. Necessário
para modelos de árvore

TRÊS MODELOS TREINADOS

Baseline

- Hiperparâmetros mínimos
- Treinamento rápido

Melhor CCP Alpha (Cost Complexity Pruning)

- Controla a poda da árvore, removendo nós com impacto mínimo no erro

Grid Search

- Realiza busca exaustiva na lista de hiperparâmetros definidos

Modelagem

Light GBM

É uma variante do GBM (Gradient Boosting Machine) que gera múltiplas árvores de forma sequencial, onde cada nova árvore busca corrigir os erros das anteriores, a versão Light melhora a eficiência computacional, principalmente com grandes bases e variáveis categóricas.

* Especificar colunas categóricas para o modelo

DOIS MODELOS TREINADOS

Baseline

- Hiperparâmetros mínimos
- Treinamento rápido

Bayes Search

- Realiza busca inteligente e eficiente na lista de hiperparâmetros definidos por meio aprendizagem com a variação dos erros

Avaliação

Árvore de Decisão

TREINO	BASE	MELHOR CCP	GRID SEARCH	TEST	BASE	MELHOR CCP	GRID SEARCH
R ²	0,26	0,26	0,29	R ²	0,26	0,26	0,28
MAE	57,71	57,71	56,53	MAE	57,73	57,73	56,73
RMSE	72,78	72,78	71,49	RMSE	72,85	72,85	71,76

LightGBM

TREINO	BASE	BEYES SEARCH
R ²	0,35	0,36
MAE	53,68	53,50
RMSE	68,10	67,91

TESTE	BASE	BEYES SEARCH
R ²	0,31	0,32
MAE	55,34	55,02
RMSE	70,14	69,80

Modelo pronto para uso: métricas melhoradas e prevendo a nota com sua margem de erro

Registro

Uso do MLFlow para armazenamento, controle, versionamento e comparação dos modelos

Notas CH ENEM 2023 ⓘ Provide Feedback ⓘ Add Description

Runs Evaluation Experimental Traces

					Metrics		Tags	
					mae	r2 ↕	rmse	model_type
<input type="checkbox"/>	Run Name	Created	Duration	Models				
<input type="checkbox"/>	magnificent-bear-137	✓ 1 minute ago	7.0s	modelo_lgbm_bayes v10	55.0236357...	0.3183606553...	69.7973999...	LGBMRegressor - BayesSearchCV
<input type="checkbox"/>	fearless-fowl-934	✓ 2 minutes ago	5.2s	modelo_lgbm_bayes v9	55.0550289...	0.3173666377...	69.8482732...	LGBMRegressor - BayesSearchCV
<input type="checkbox"/>	casual-mouse-499	✓ 9 hours ago	45.0s	modelo_lgbm_bayes v4	55.1407230...	0.3160904496...	69.9135337...	LGBMRegressor - BayesSearchCV
<input type="checkbox"/>	salty-perch-318	✓ 1 day ago	2.2min	modelo_lgbm_base v6	55.3376167...	0.3117378022...	70.1356583...	LGBMRegressor
<input type="checkbox"/>	treasured-dove-153	✓ 8 hours ago	21.2s	modelo_arvore_decisao_grid v3	56.7338174...	0.2794840562...	71.7602126...	Decision Tree Regressor com GridSearchCV
<input type="checkbox"/>	big-rook-716	✓ 9 hours ago	18.7s	modelo_arvore_decisao_alpha v5	57.7268636...	0.2575139981...	72.8460579...	Decision Tree Regressor com alpha otimizado
<input type="checkbox"/>	hilarious-pug-382	✓ 1 day ago	51.0s	modelo_arvore_decisao_alpha v4	57.7268636...	0.2575139981...	72.8460579...	Decision Tree Regressor com alpha otimizado
<input type="checkbox"/>	suave-croc-548	✓ 1 day ago	41.4s	modelo_arvore_decisao_alpha v3	57.7268636...	0.2575139981...	72.8460579...	Decision Tree Regressor com alpha otimizado
<input type="checkbox"/>	sneaky-moth-594	✓ 9 hours ago	29.5s	modelo_arvore_decisao_base v6	57.7284860...	0.2574738657...	72.8480265...	Decision Tree Regressor
<input type="checkbox"/>	stylish-bee-532	✓ 1 day ago	1.2min	modelo_arvore_decisao_base v5	57.7284860...	0.2574738657...	72.8480265...	Decision Tree Regressor
<input type="checkbox"/>	debonair-skunk-2	✓ 1 day ago	54.2s	modelo_arvore_decisao_base v4	57.7284860...	0.2574738657...	72.8480265...	Decision Tree Regressor

Registro

Armazenamento dos hiperparâmetros e características de cada modelo

Notas CH ENEM 2023 >

magnificent-bear-137

Overview

Model metrics

System metrics

Traces

Artifacts

Run ID	aabfdc4c4a6c4baeb8a9838fc40fce5
Duration	7.0s
Datasets used	—
Tags	model_type: LGBMRegressor - BayesSearchCV
Source	d:\Armazenamento\MBA\TCC\Codigos_Iniciais\venv\Lib\site-packages\ipykernel_launcher.py
Logged models	sklearn
Registered models	modelo_lgbm_bayes v10
Registered prompts	—

Parameters (8)

Search parameters

Parameter	Value
colsample_bytree	0.319934457252814
learning_rate	0.005
max_depth	54
num_leaves	49
n_estimators	5000
reg_alpha	0.15096028361393093
reg_lambda	0.020875891601585508
subsample	0.31695966954573485

Metrics (3)

Search metrics

Metric	Value
mae	55.023635767301606
r2	0.31836065534070224
rmse	69.79739992005827

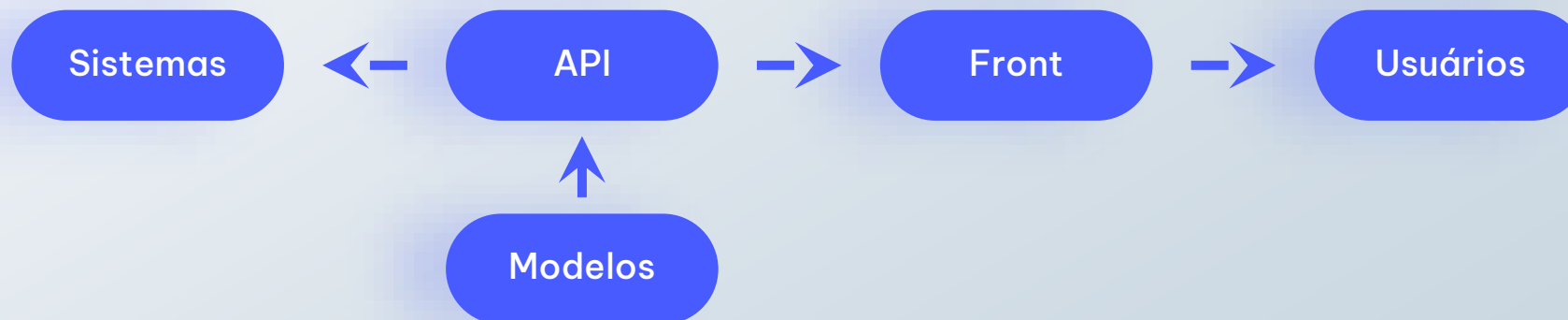
Implantação

DISPONIBILIZAÇÃO DO MODELO

- Serviço de API (FastAPI)
 - Consumo pode ser realizado por diversos sistemas
 - Permite escalabilidade
- Serviço via Aplicação Web (Streamlit)
 - Interface amigável para o usuário final não técnico
 - Possibilita testes rápidos

MOTIVAÇÕES

- Dois serviços permitem a separação de responsabilidades: modelo, API, UI
- Flexibilidade de consumo

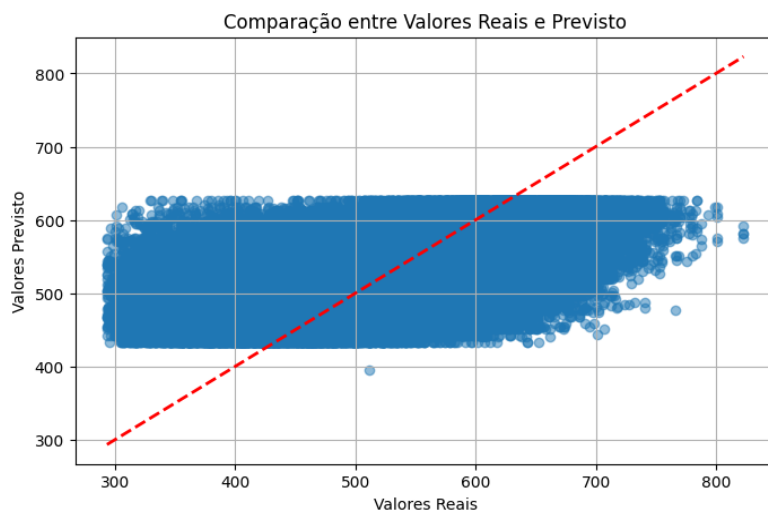


Resultados Modelo

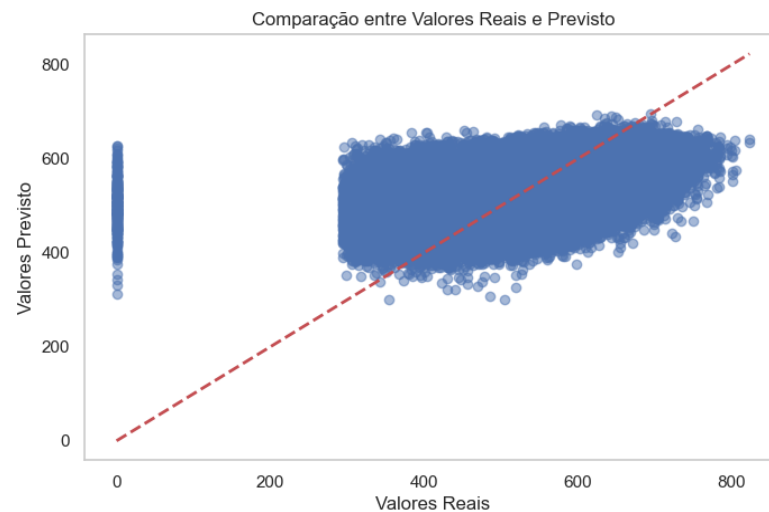
REAL VS PREVISTO

- Reconhece a tendencia geral
- Prevê bem entre 350 e 650, dificuldade com notas maiores e menores
- Subestimação das notas maiores, superestimação das menores

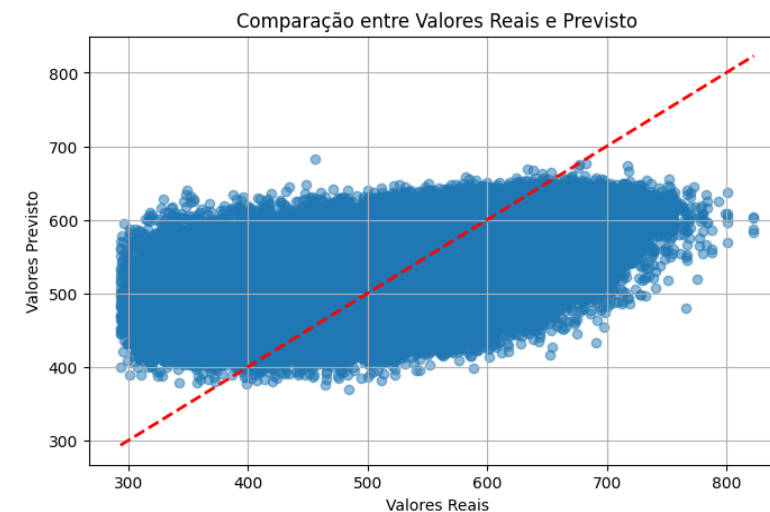
Árvore de Decisão (sem outliers zero)



LightGBM (com outliers zero)



LightGBM (sem outliers zero)





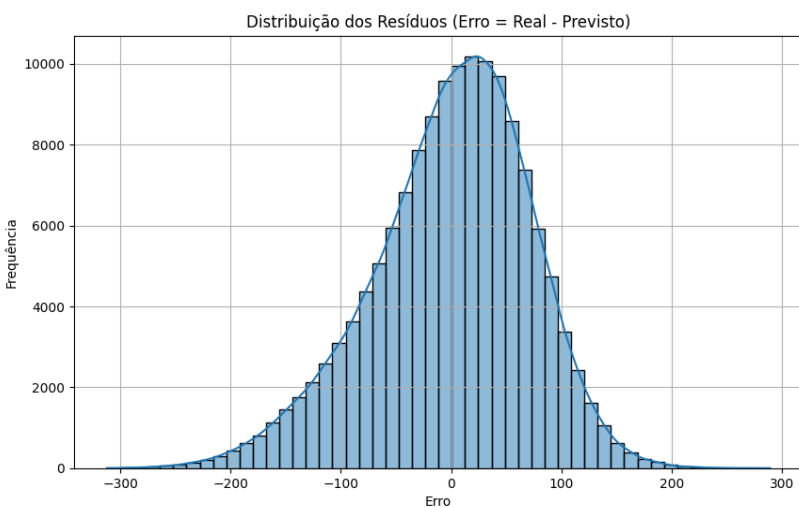
Resultados Modelo

CÁLCULO DE RESÍDUOS

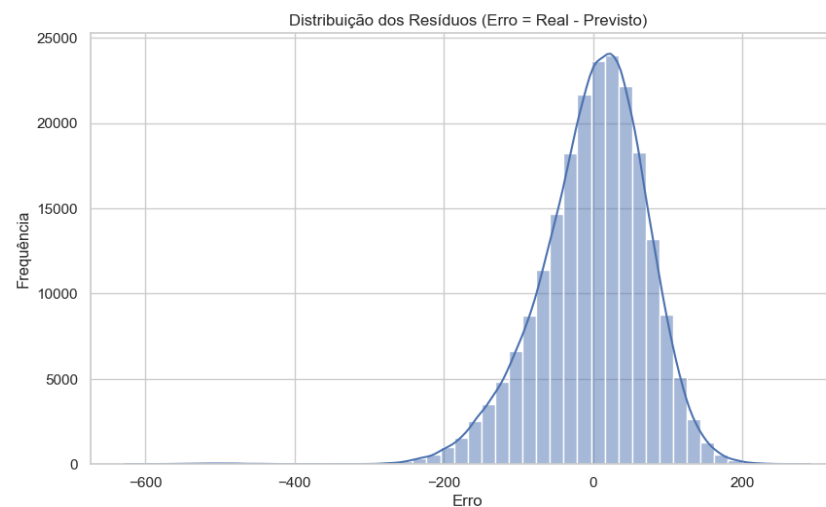
- Distribuição normal centrada em zero
- Baixo viés de modo geral



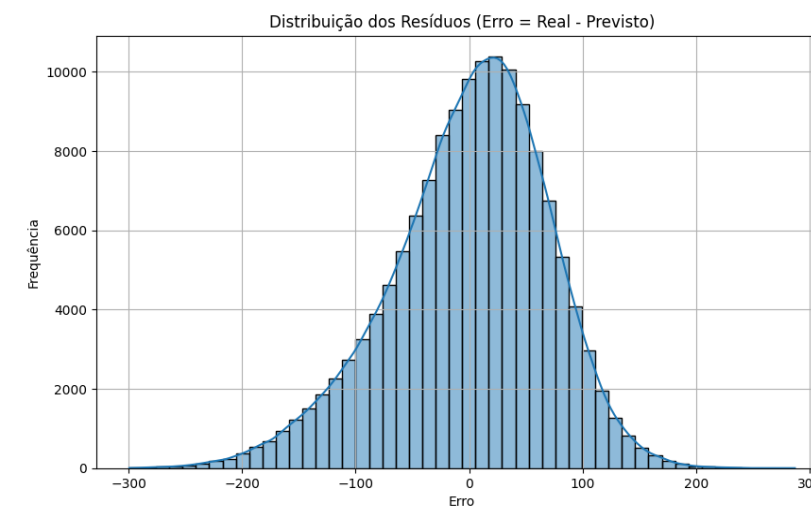
Árvore de Decisão
(sem outliers zero)



LightGBM
(com outliers zero)



LightGBM
(sem outliers zero)



Algumas conclusões dos resultados

- Aplicar modelos melhores e com mais treinamento
- Explorar melhor notas maiores e menores
- Aplicar modelos por faixas
- Base desequilibrada (SMOTE ou subamostragem para extremidades)

MAE por faixa

faixa	
(0, 300]	201.545286
(300, 500]	69.132923
(500, 700]	46.181468
(700, 1000]	127.771426

Monitoramento

Modelo (qualidade preditiva):

- Data Drift: mudança nos dados de entrada ao longo do tempo
- Concept Drift: mudanças nas saídas mesmo com entradas semelhantes
- Acompanhar métrica definidas: R^2 , MAE e RMSE
- Uso de logs, comparações periódicas com distribuição original

API

- Latência e tempo de resposta
- Taxa de erros
- Disponibilidade
- Volume de requisições
- Uso de logs, middleware, alertas

Apresentação prática

Chamada de API

```
# Fazendo a requisição POST
response = requests.post(url, json=dados_nome_front)

# Exibindo a resposta da API
print("Status Code:", response.status_code)

if response.status_code == 200:
    print("Resposta:", response.json())
    print('')
else:
    print("Erro:", response.text)
```

```
Status Code: 200
Resposta: {'mensagem': 479.47074971315425}
```

Utilização UI

Previsão com Modelo de Machine Learning

Seleção de Modelo

Escolha o modelo de Machine Learning

LightGBM

Insira os dados do modelo

- ☐ Não sabe o grau de estudo do pai
- ☐ Não sabe o grau de estudo da mãe
- ☐ Presença de aspirador
- ☐ Presença de DVD
- ☐ Presença de TV por assinatura
- ☐ Presença de telefone fixo
- ☐ Presença de internet

Cor/Raça

Não declarado

Código do município da escola

Alta Floresta DOeste

Calcular Previsão

Resultado da previsão: 583.8771725170914