4

Impactos de Beber e fumar

Felipe Martins & Caio Righetto

Sobre a base de dados

- A base de dados utilizada é a "Smoking and Drinking Dataset";

- Ela contém informações gerais sobre os organismos de indivíduos que bebem e fumam (ou não);

- Os dados foram coletados a partir do serviço Nacional de Saúde Pública da Coreia do Sul.

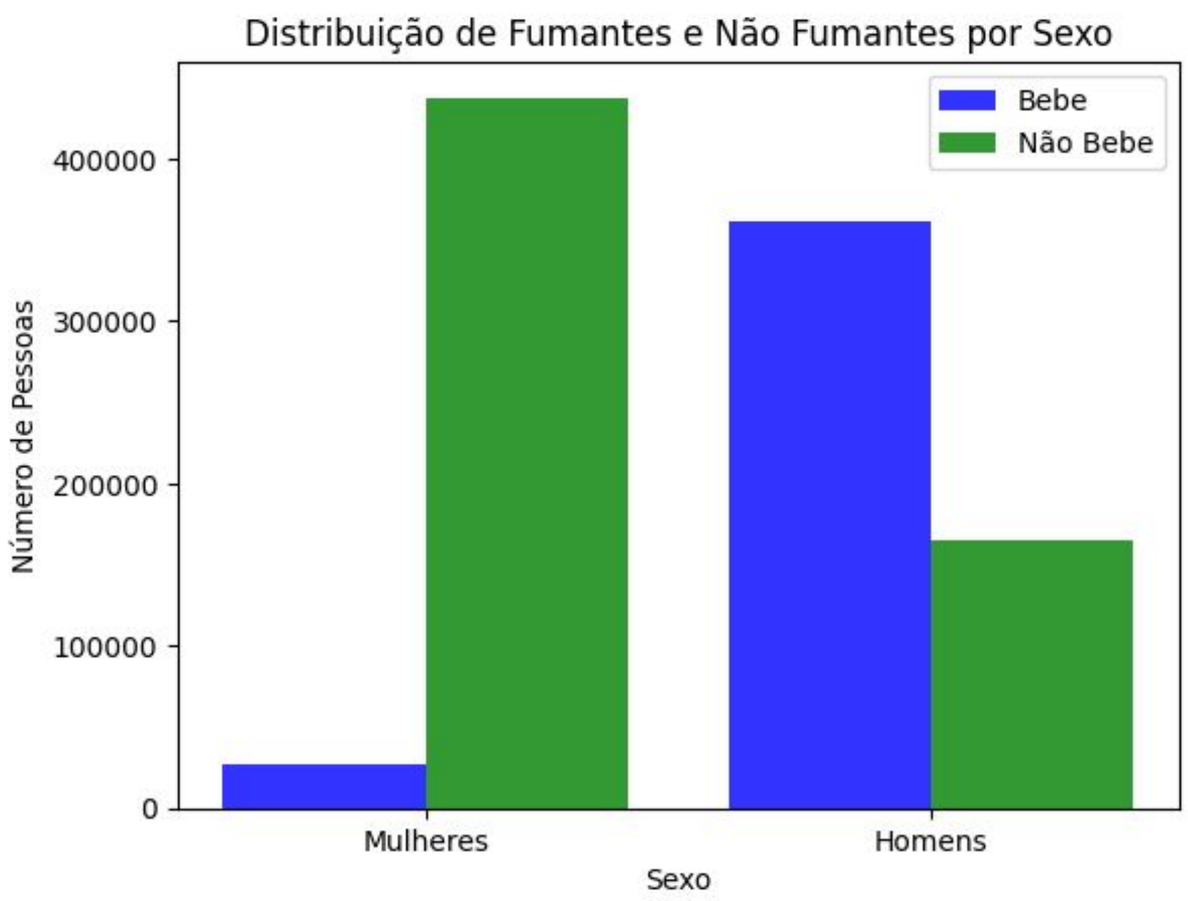
São 24 colunas ao todo, com mais de 900.000 registros

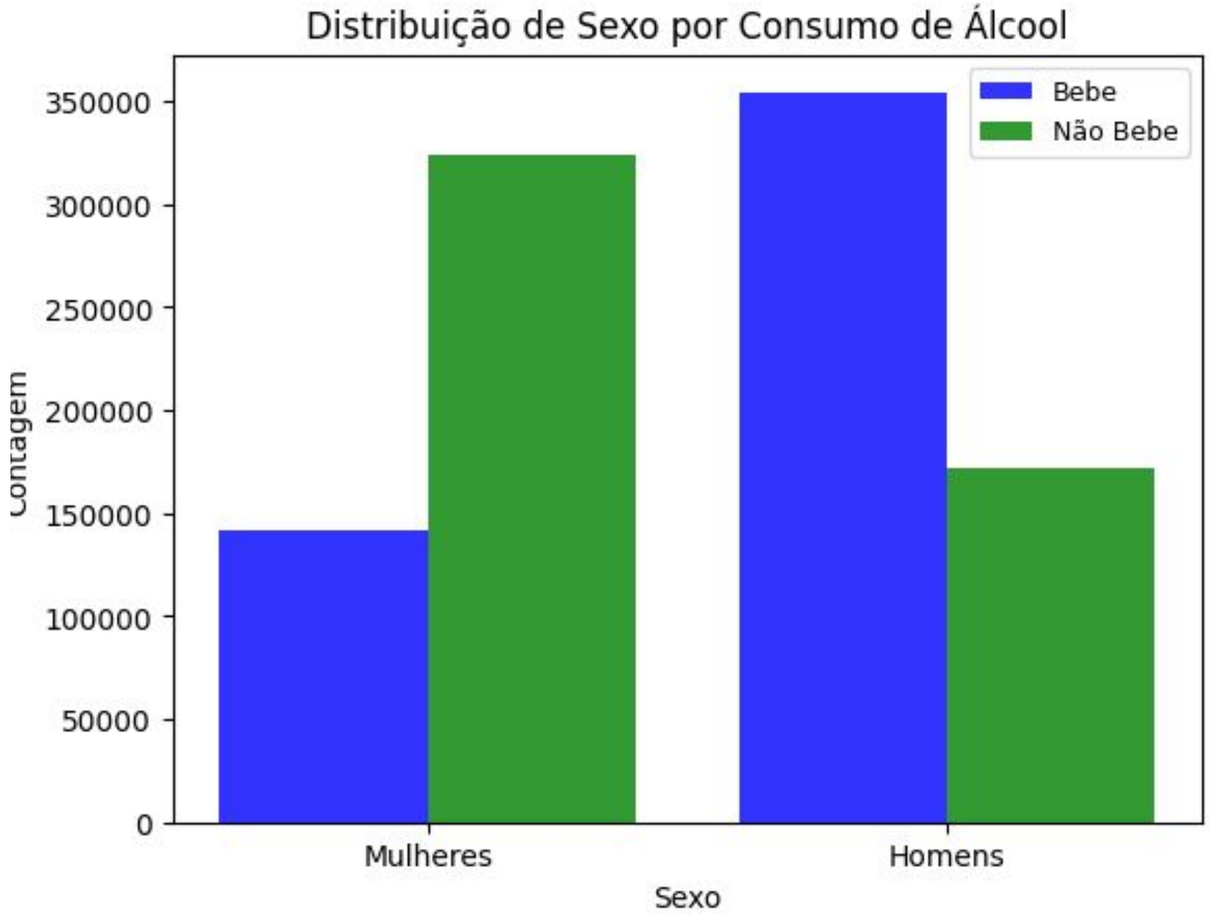
- sex
- age
- height
- weight
- sight_left
- sight_right
- hear_left
- hear_right
- SBP
- DBP
- BLDS
- tot_chole
- HDL_chole
- LDL_chole
- triglyceride
- hemoglobin
- urine_protein
- serum_creatinine
- SGOT_AST
- SGOT_ALT
- gamma_GTP
- **SMK_stat_type_cd**
- **DRK_YN**

Referente ao status de fumante do indivíduo

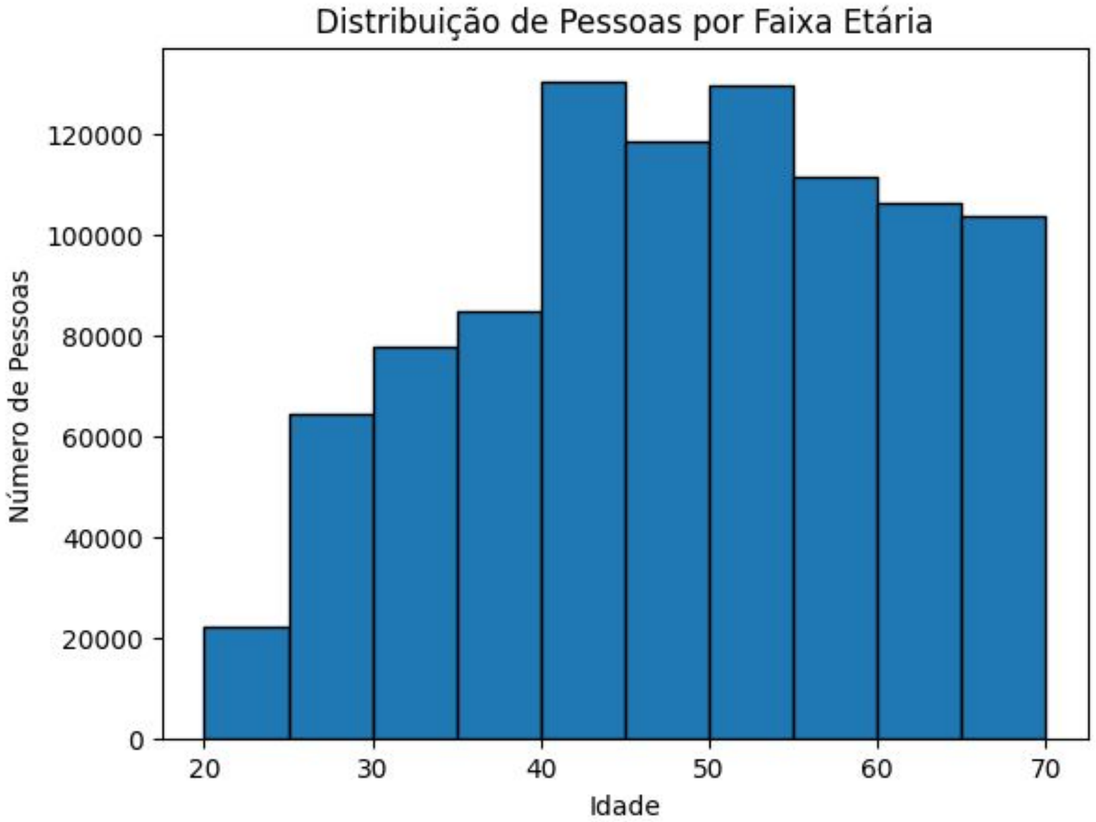
Referente ao status de bebida (se bebe ou não)

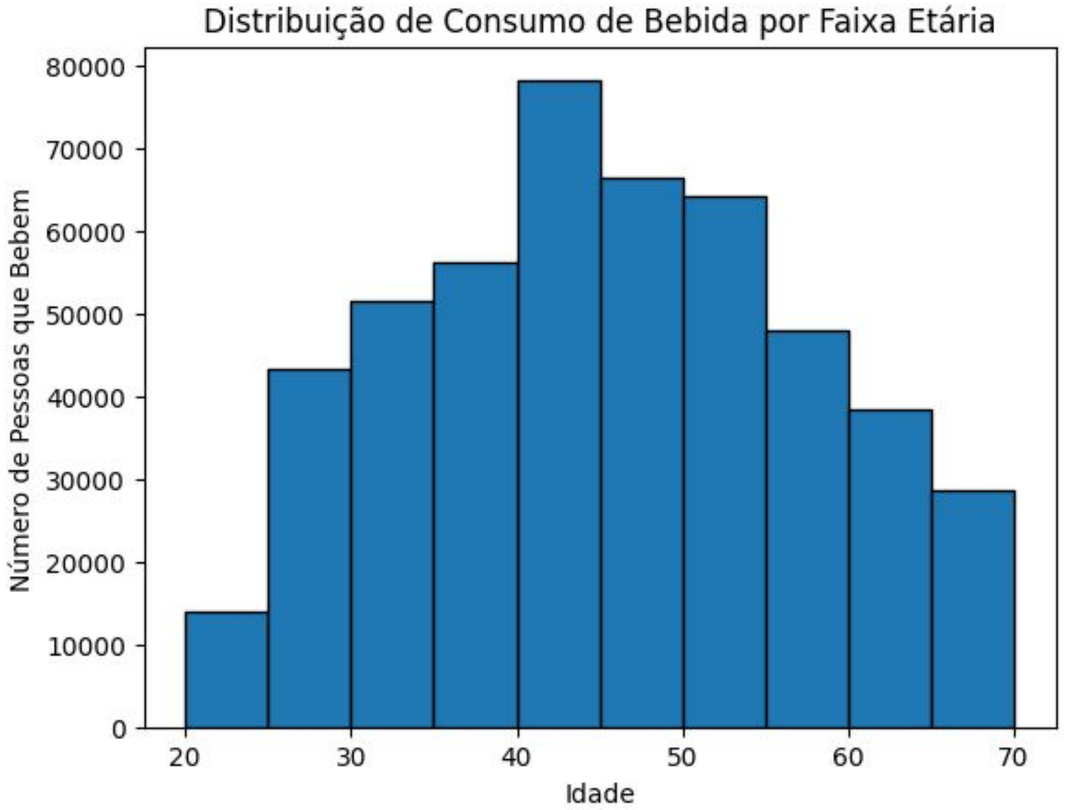
EDA





EDA





Modelos usados

Objetivo

Treinar modelos para perceberem e identificarem quais as marcas mais significativas que estão naqueles que fumam ou bebem

Regressão Logística

Para prever se a pessoa bebe ou não

CatBoost Classifier

Para prever se a pessoa bebe ou não

Comparar os resultados

Determinar qual modelo se saiu melhor

Tratamento dos Dados

Coluna ''DRK_YN''

DRK_YN
Y
N
N
N
N

→

DRK_YN
1
0
0
0
0

Coluna ''SMK_stat_type_cd''

SMK_stat_type_cd
1.0
3.0
1.0
1.0
1.0

→

SMK_stat_type_cd
0
1
0
0
0

Coluna ''sex''

sex
Male
Male
Male
Male
Male

→

sex
1
1
1
1
1

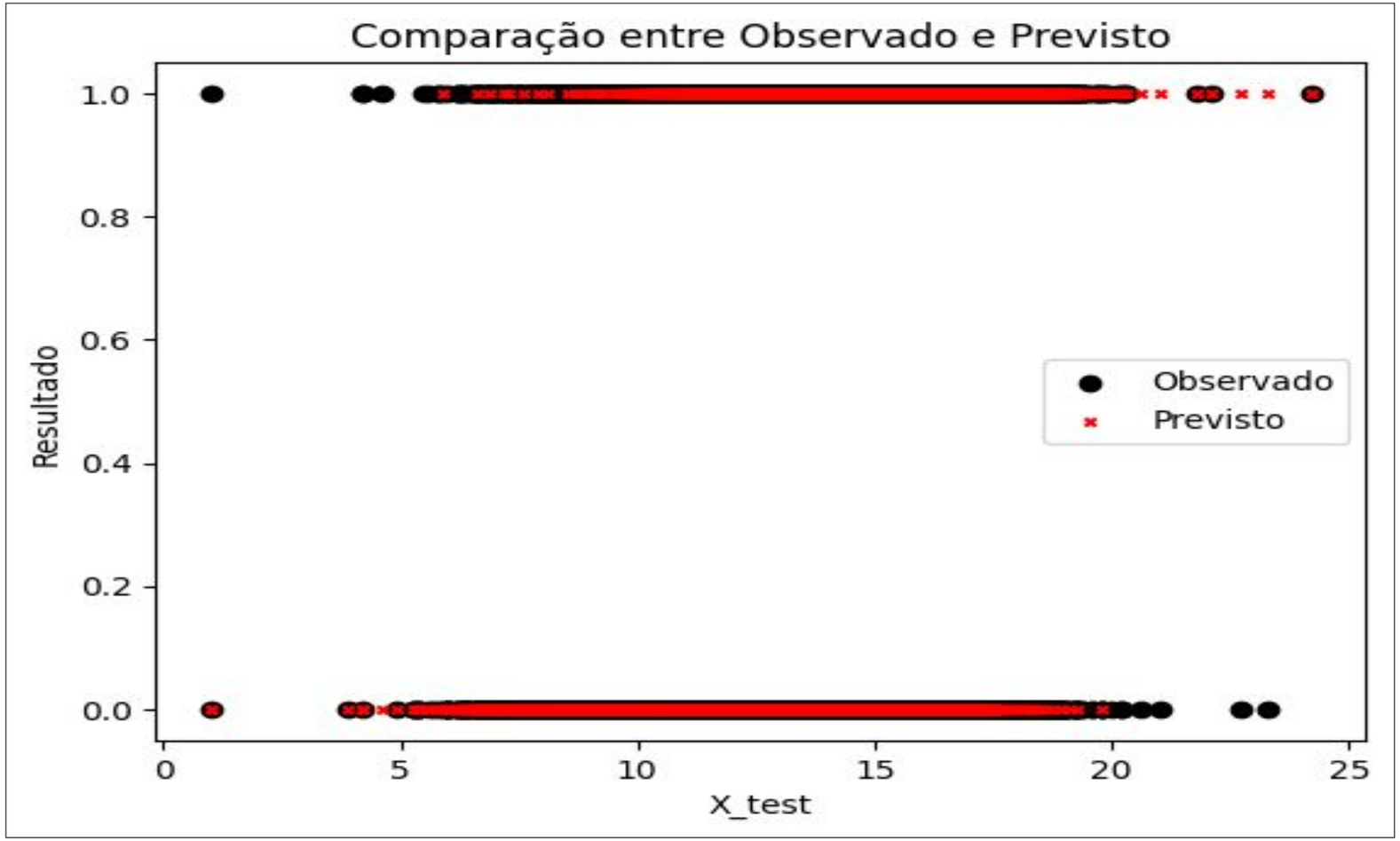
Regressão Logística

O modelo de Regressão Logística foi usado para classificar se uma pessoa bebe ou não;

Foi usada a biblioteca Logistic Regression do Scikit Learn;

```
Pontuacao obtida com X e Y de treino:  0.7153728015978292
Pontuacao obtida com X e Y de teste:  0.7143743380239068
Pontuacao obtida com Y da previsão e o Y real:  0.7143743380239068
```

Regressão Logística



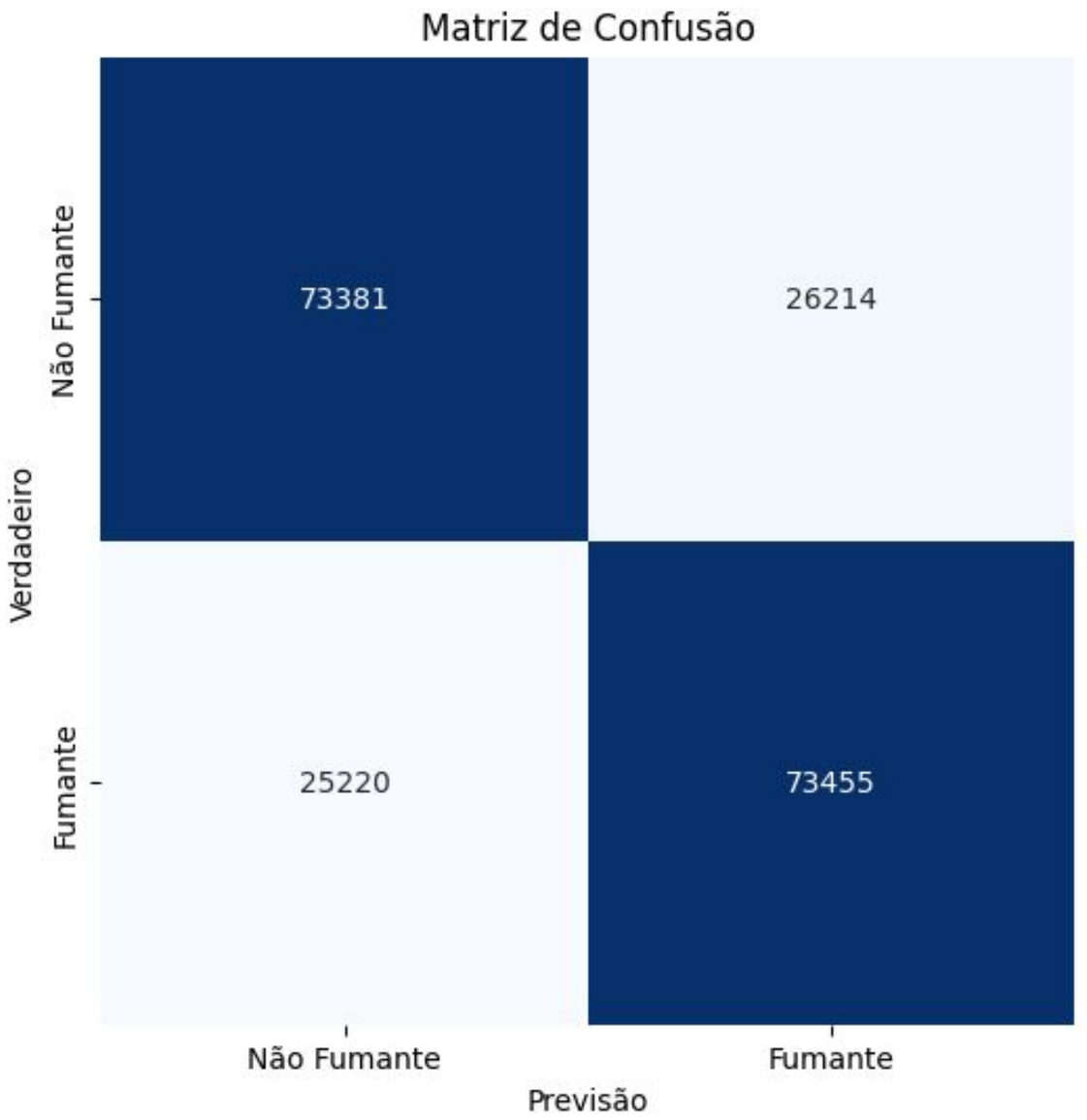
	Feature	Importance
22	SMK_stat_type_cd	0.386025
16	hemoglobin	0.266964
0	sex	0.244508
3	weight	0.038258
13	HDL_chole	0.020269

Catboost Classifier

O catboost classifier é baseado em árvores de decisão impulsionadas por gradientes.

Durante o treinamento, são construídas consecutivas árvores de decisão, cada árvore é construída com uma perda menor do que a anterior.

```
Precisão do modelo: 0.74
```



Catboost Classifier

	Feature	Importância
1	age	22.527525
21	gamma_GTP	19.550073
0	sex	12.703390
13	HDL_chole	9.785088
22	SMK_stat_type_cd	9.500303
20	SGOT_ALT	7.680877
2	height	3.037309

A Gama-GT é uma enzima que é muito presente no fígado, quem consome muito álcool têm níveis mais altos dela

Um valor elevado de colesterol alto pode ser associado ao consumo regular de álcool

A ALT é é outra enzima que é muito presente no fígado, principalmente em quem tem o fígado gorduroso

Comparação dos resultados

Regressão Logística

- Acurácia: 71%
- Features importantes menos coerentes
- Executa mais rápido

```
modelo_LR = LogisticRegression()
modelo_LR.fit(X_train,y_train)
y_pred = modelo_LR.predict(X_test)
```

✓ 4.0s

VS

Catboost

- Acurácia: 74%
- Features importantes mais coerentes
- Executa mais devagar

```
from catboost import CatBoostClassifier

model = CatBoostClassifier(iterations=500

model.fit(X_train, y_train, eval_set=(X_t
```

✓ 30.4s