



APLICAÇÃO DE ALGORITMOS DE MACHINE LEARNING PARA PROBLEMAS REAIS



Predição de odds em apostas esportivas

Fernando de Abreu e Lima Filho

Recife - PE

CONTEXTO/PROBLEMÁTICA



- As apostas esportivas movimentam bilhões de dólares no mundo todo.
- Bookmakers calculam probabilidades para definir odds, influenciando lucros e decisões dos apostadores.

CONTEXTO/PROBLEMÁTICA

Dataset: <https://github.com/xgabora/Club-Football-Match-Data-2000-2025>

- *Contém milhares de jogos de clubes ao redor do mundo*
- *Fácil compreensão*
- *Atualizado até os dias atuais*
- *Variáveis chaves:*
 - *Elo (Rating)*
 - *Forms*
 - *Home, Away, Division*



OBJETIVO GERAL

Desenvolver um modelo preditivo capaz de estimar as odds de um jogo de futebol com base em dados históricos da partida.

1

OddHome Bet365

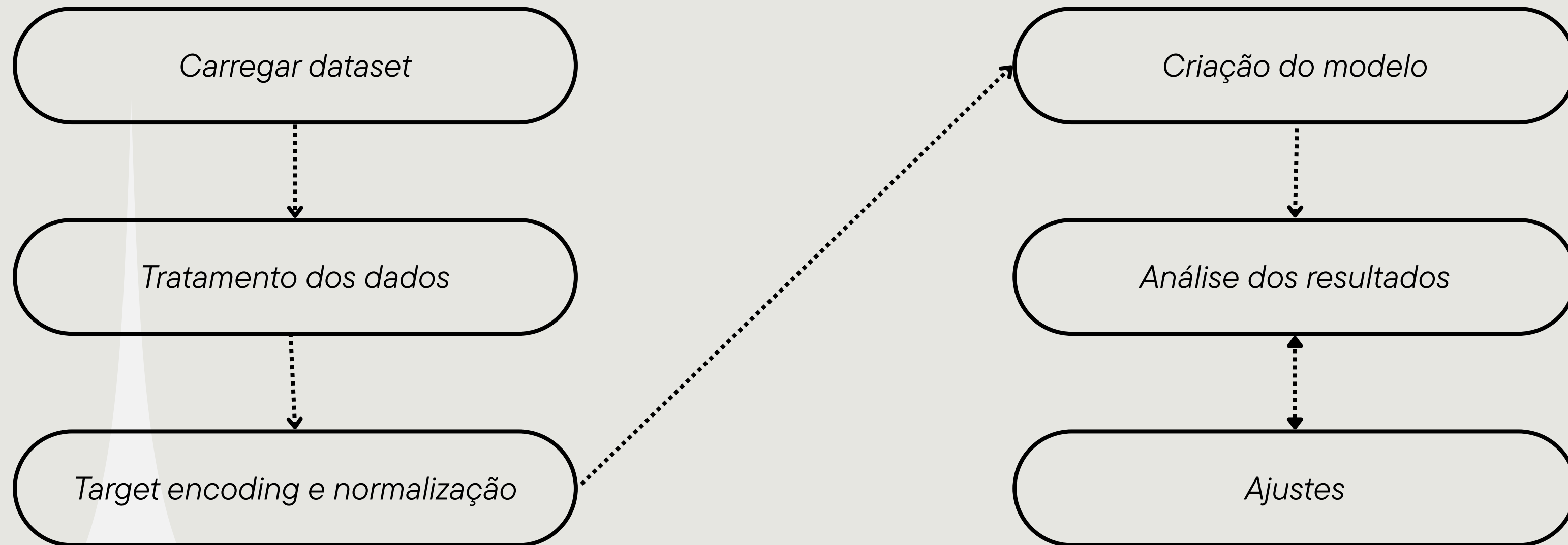
2

OddAway Bet365

3

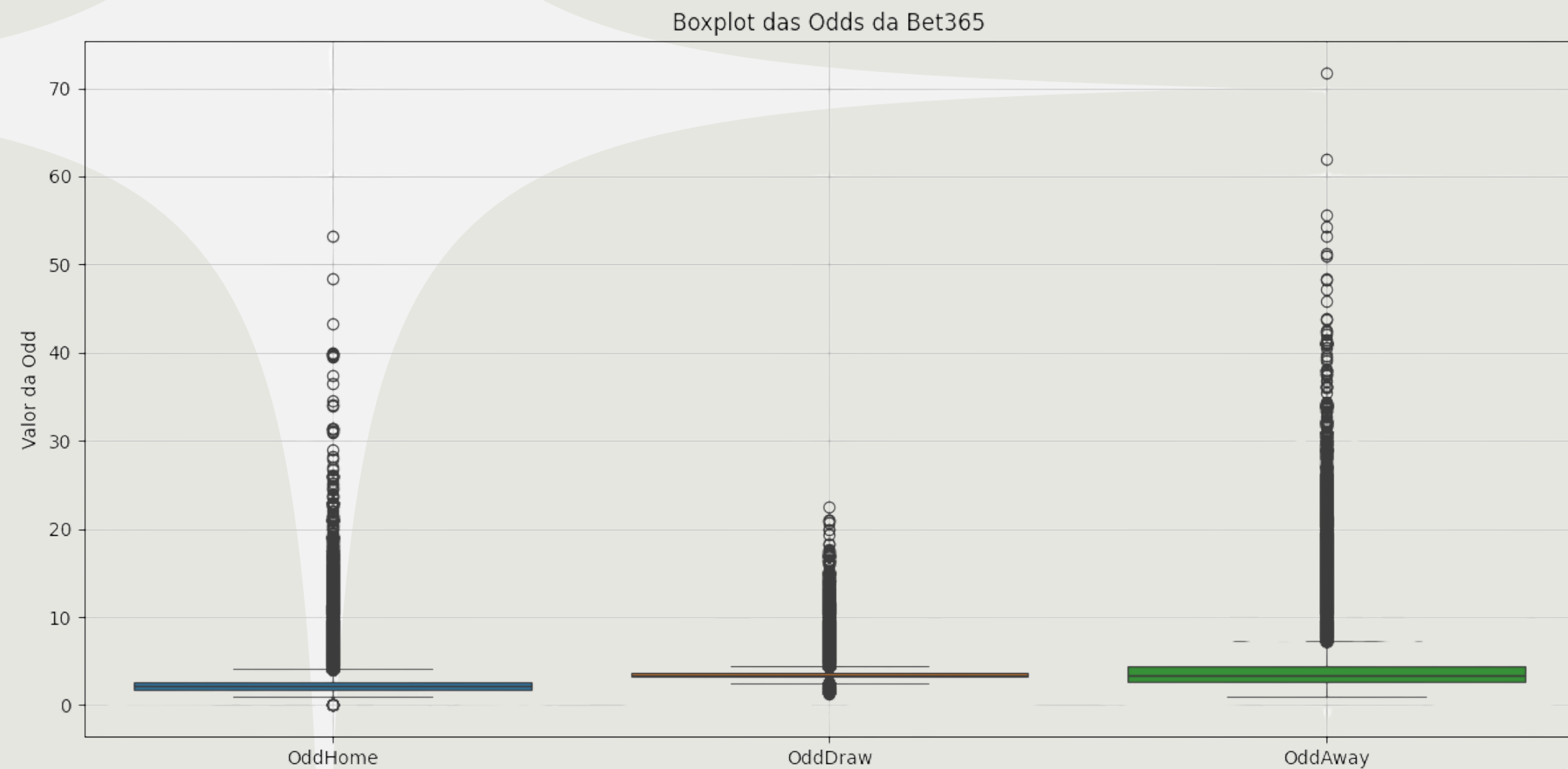
OddDraw Bet365

METODOLOGIA



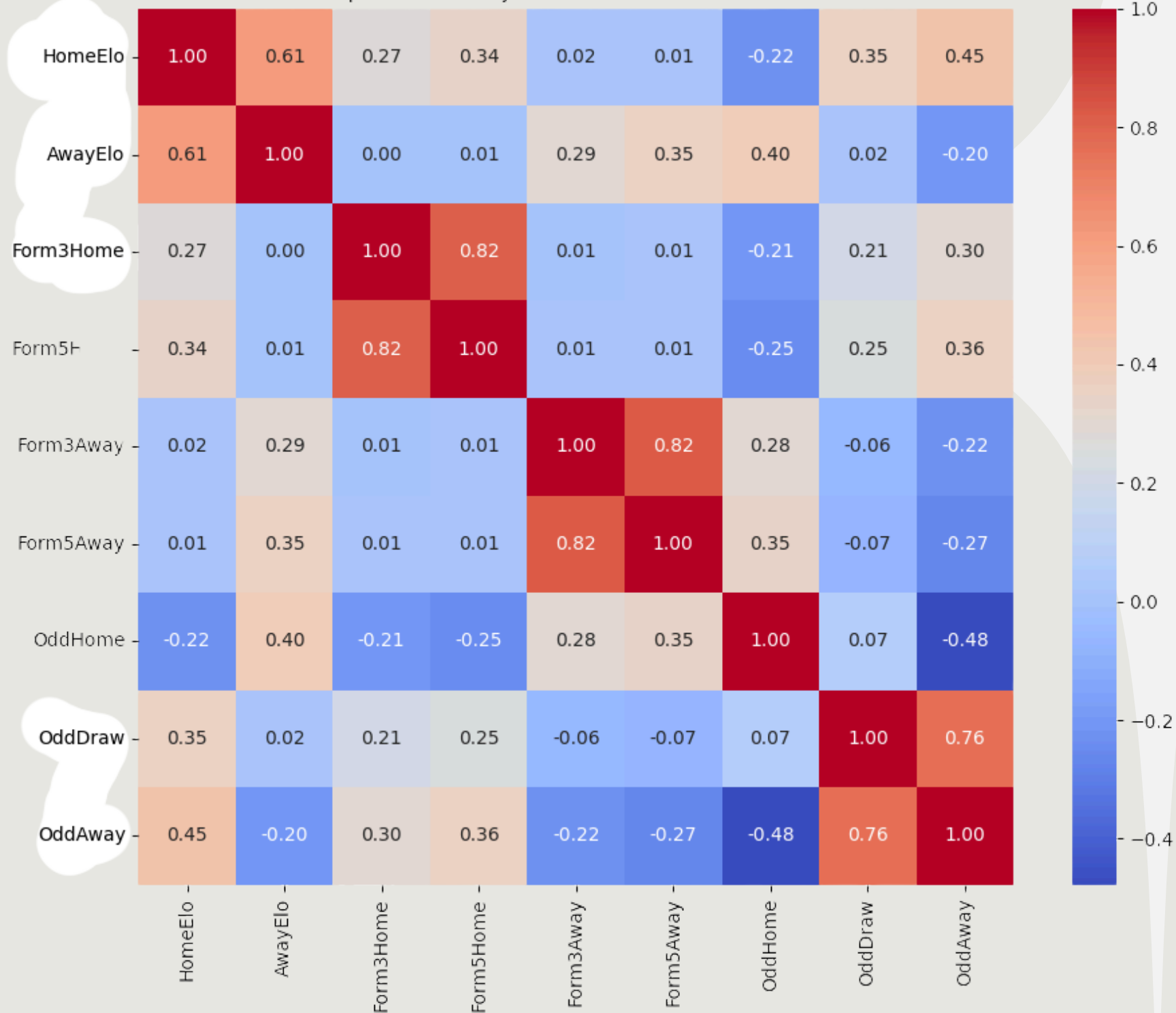
TRATAMENTO DOS DADOS

```
AwayElo      89029
HomeElo      88960
OddAway      3030
OddDraw      3030
OddHome      3030
Form3Away    1500
Form3Home    1500
Form5Home    1500
Form5Away    1500
dtype: int64
```

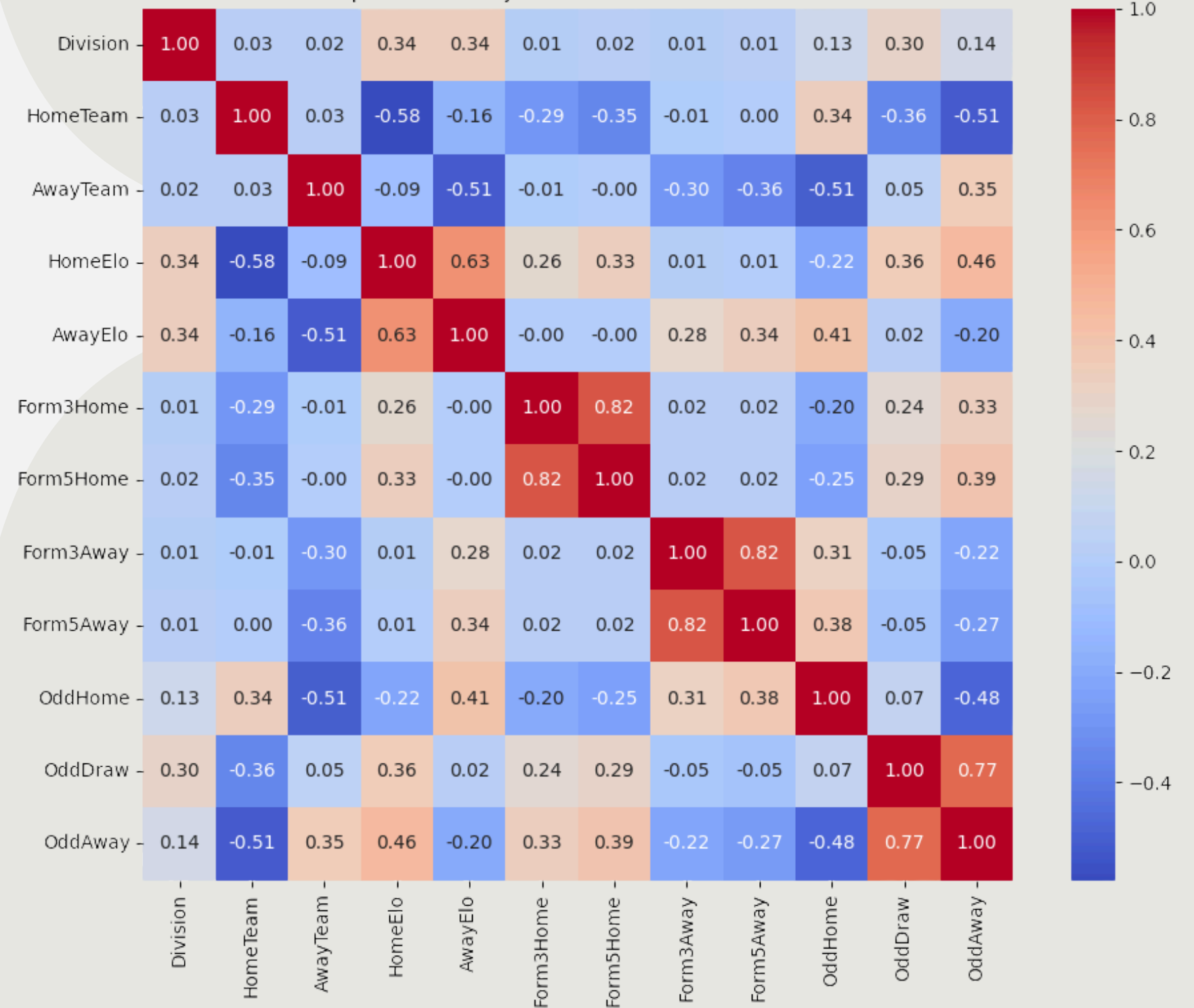


TRATAMENTO DOS DADOS

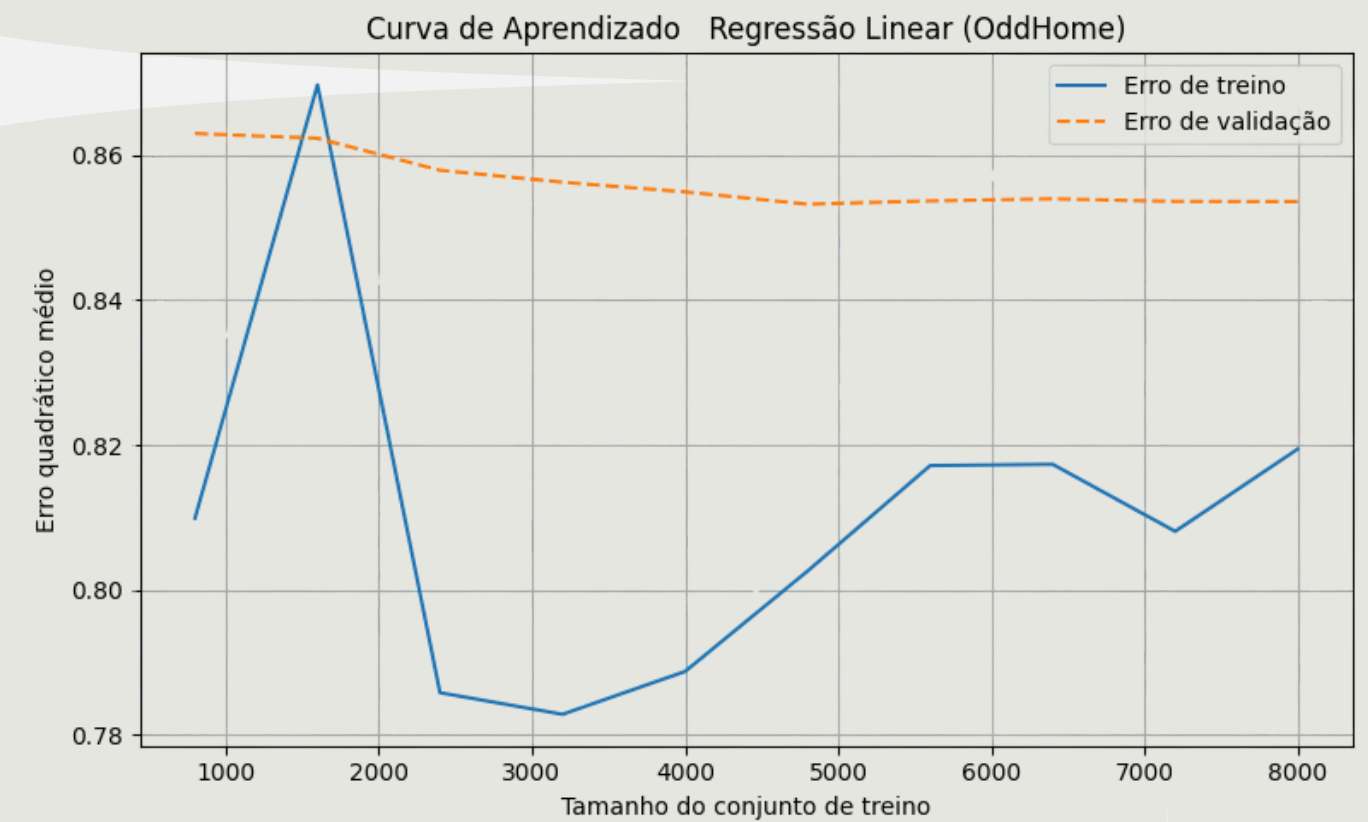
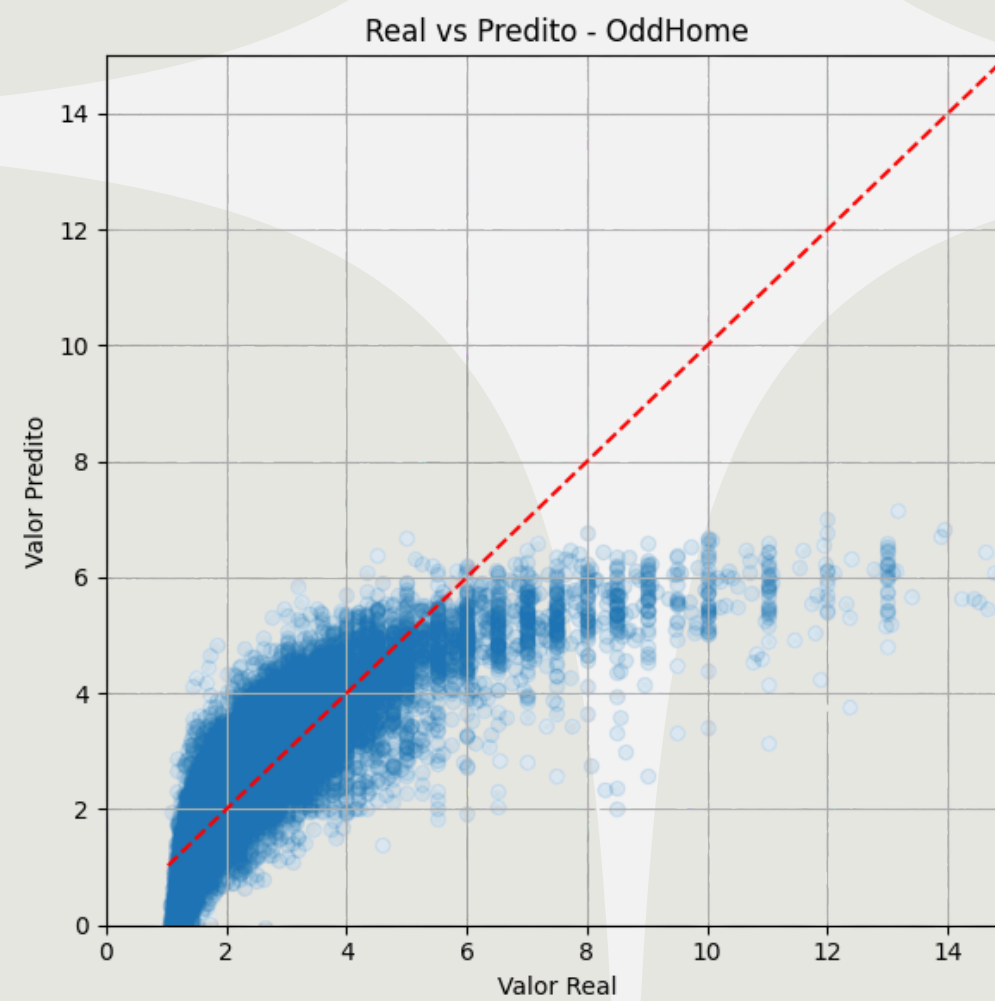
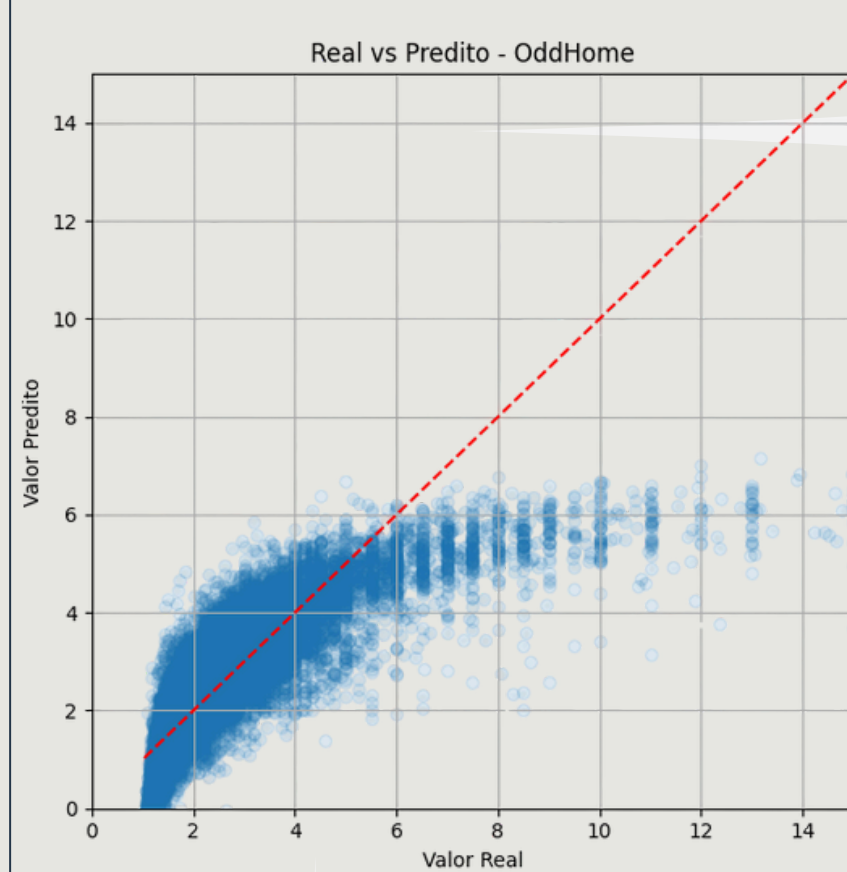
Mapa de Correlação entre Variáveis Numéricas



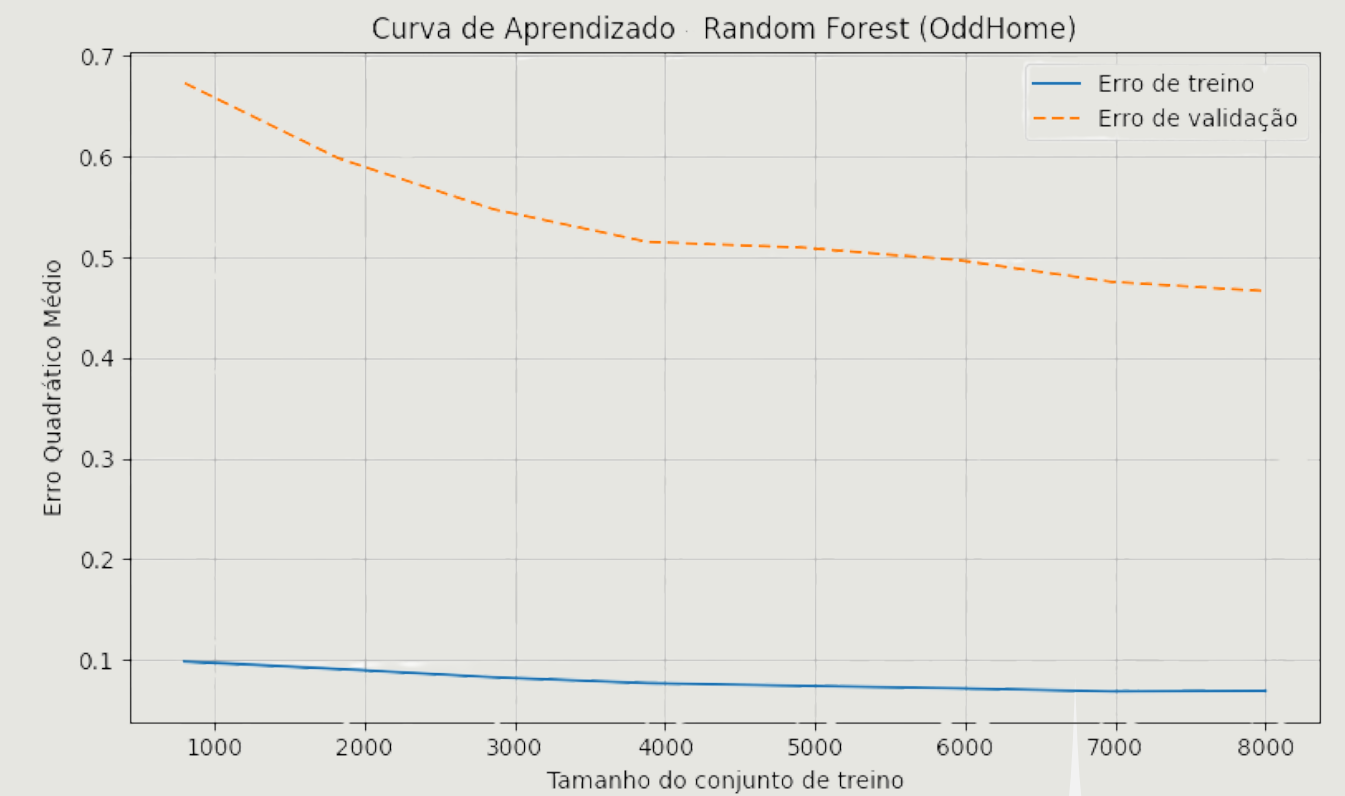
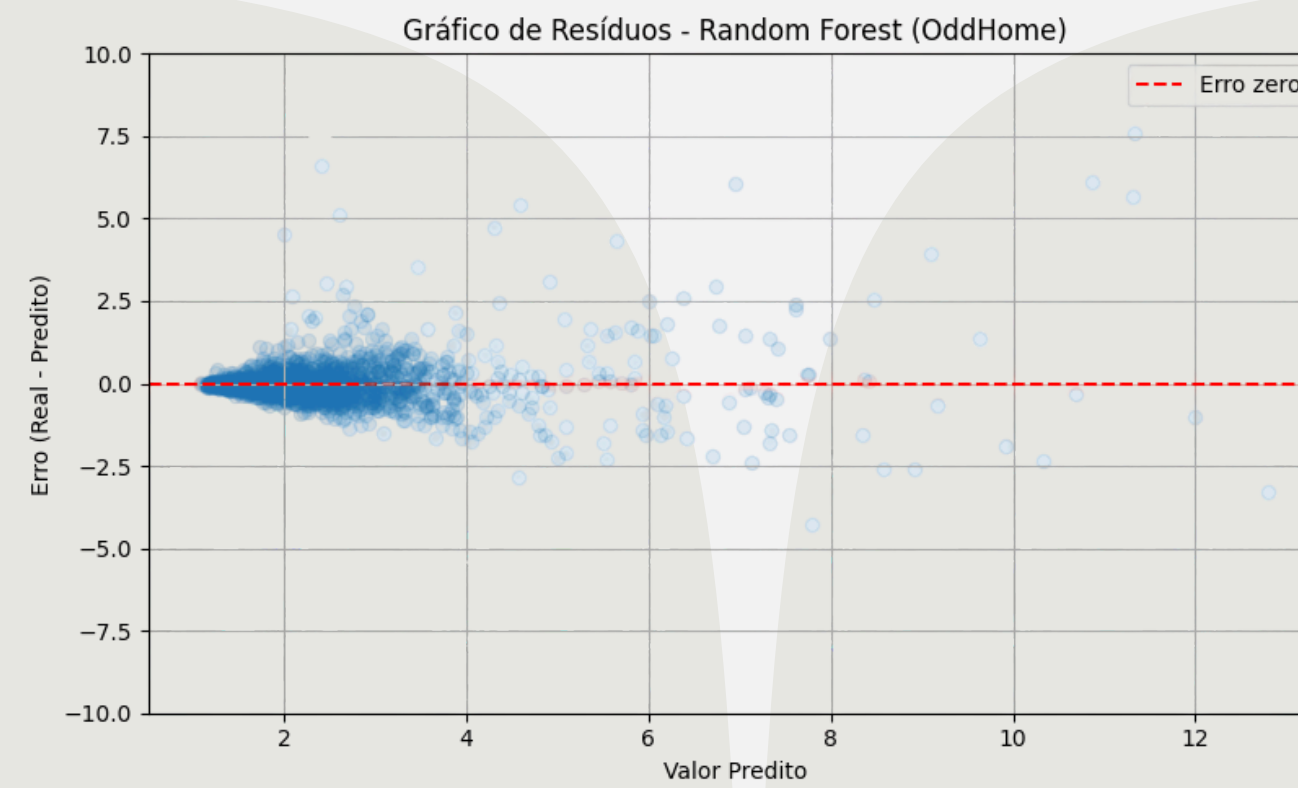
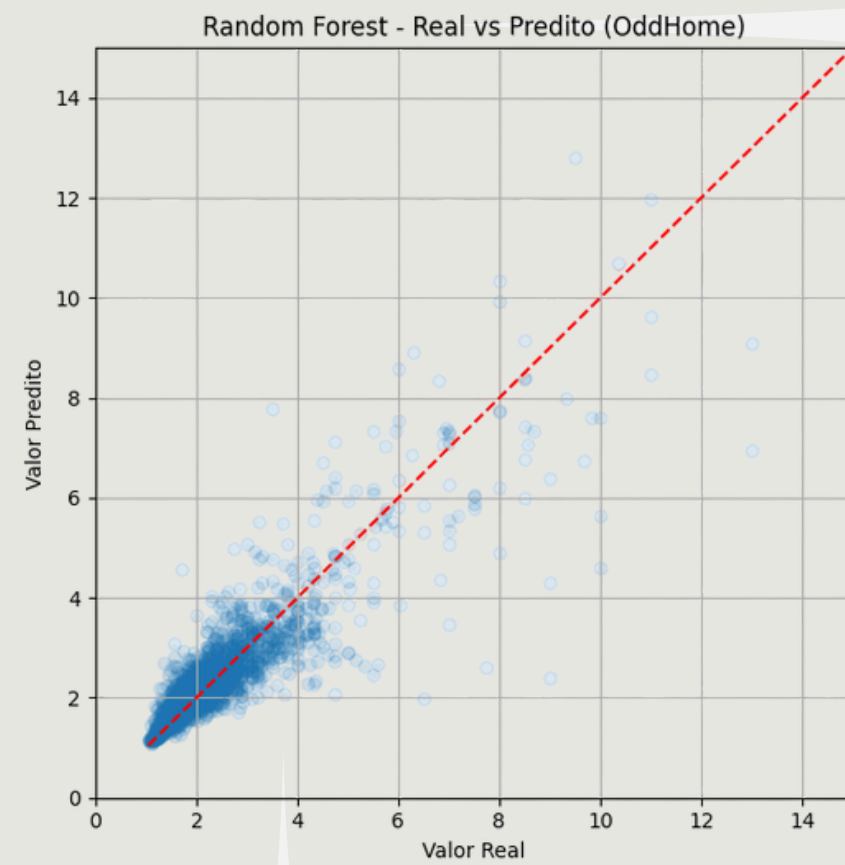
Mapa de Correlação entre Variáveis Numéricas



REGRESSÃO LINEAR



RANDOM FOREST



XGBOOST REGRESSOR

XGBoost - Real vs Predito (OddHome)

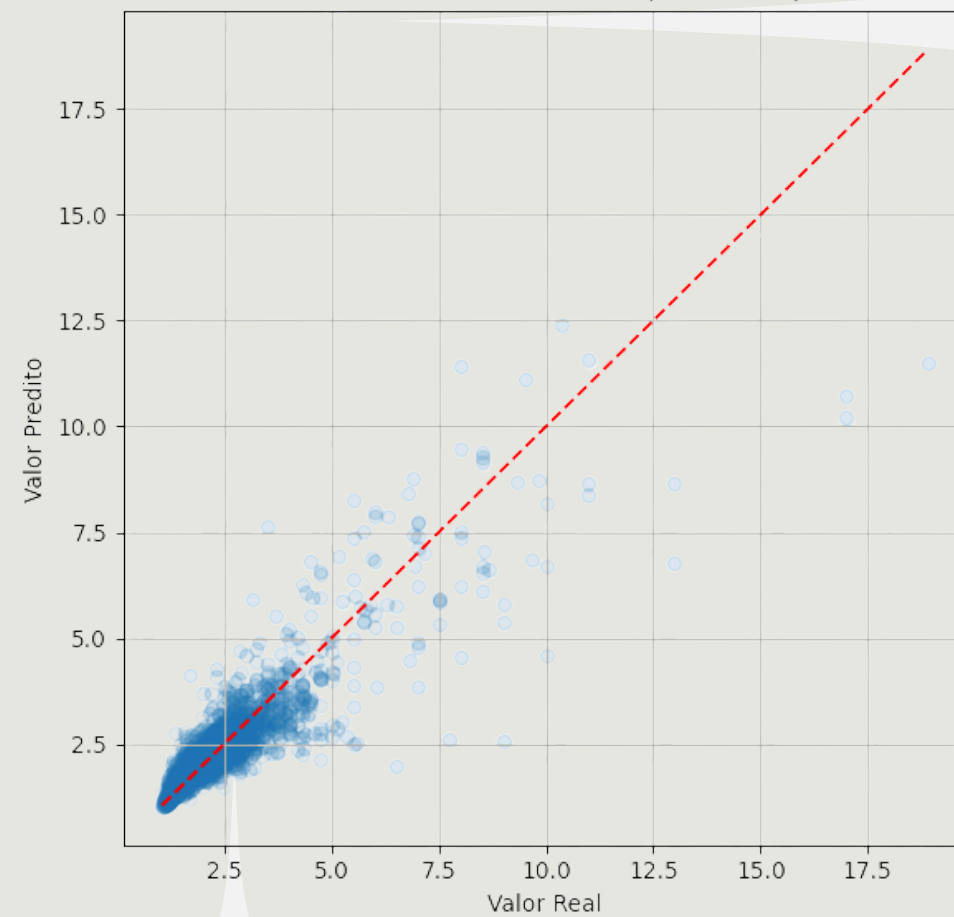
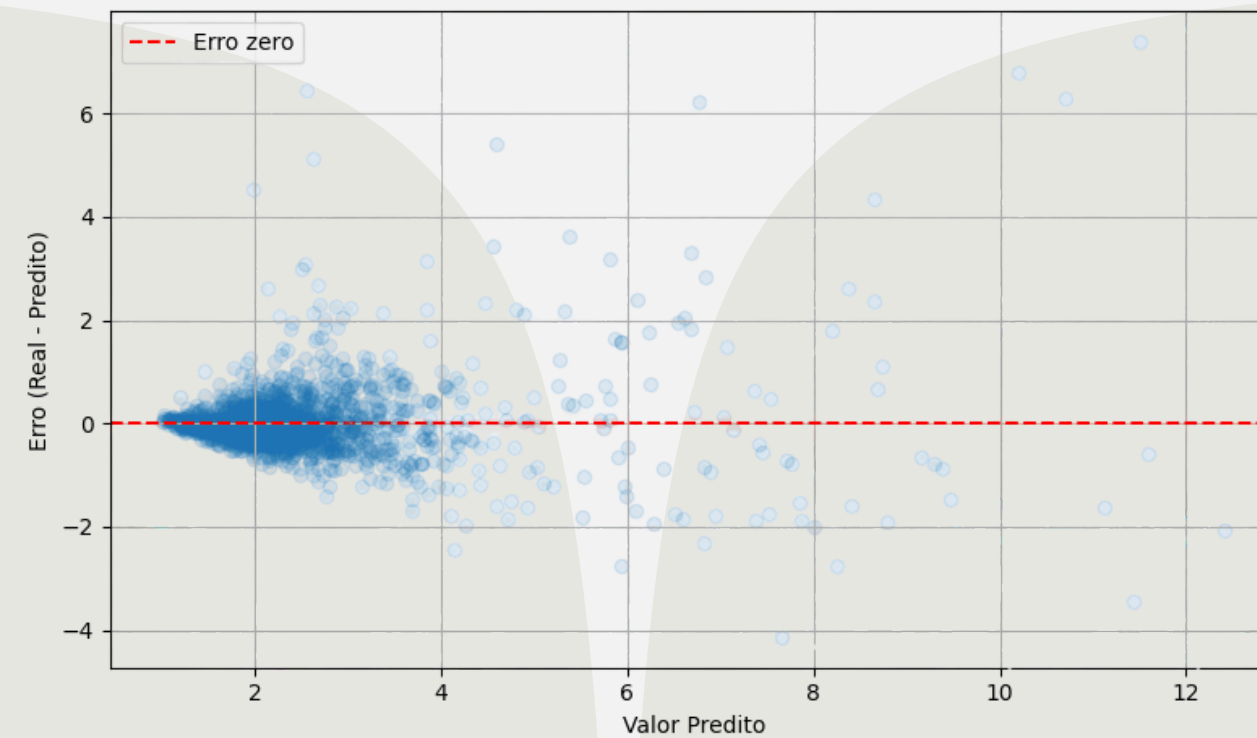
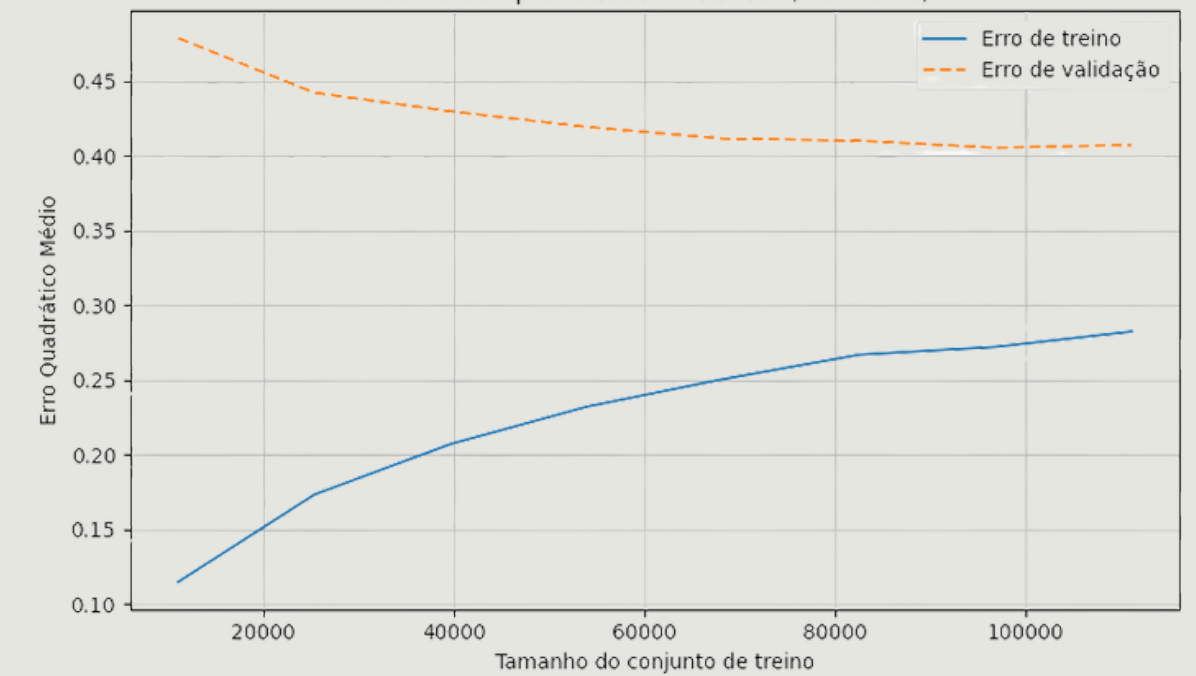


Gráfico de Resíduos - OddHome - XGBoost



Curva de Aprendizado - XGBoost (OddHome)



COMPARAÇÃO

Modelo	MAE	RMSE	R ²	Tempo de treino
LinearRegressor	0.535888	0.807781	0.582028	Rápido
RandomForest	0.3721	0.4743	0.7722	Lento
XGBoost	0.3731	0.4799	0.7695	Rápido

PCA - XGBOOST













	Com PCA	Sem PCA
MAE	<i>0.339535</i>	<i>0.427902</i>
RMSE	<i>0.415940</i>	<i>0.613407</i>
R ²	<i>0.784779</i>	<i>0.682604</i>

CONCLUSÃO

Melhorias futuras:

- Explorar outros algoritmos de regressão não linear, como redes neurais.
- Considerar possíveis viéses temporais nos dados
- Alimentação e criação de novas colunas no dataset

Modelo	Tempo de treino	Desempenho
LinearRegression	✓	✗
RandomForest	✗	✓
XGBoost	✓	✓

FEATURE ENGINEERING		
LAGGED FEATURES		
Given that the dataset includes a time data, it offers a possibility to create lagged features. One of those features (Form) is already included in the dataset, but others might include things like goal potency, points gathered during the season, table position, streaks or Elo shifts. These are planned to be added into the dataset in the future as they might contain valuable information for match prediction:		
Column	Data Type	Description
 GF3Home	int	Goals scored by the Home team in last 3 matches.
 GF3Away	int	Goals scored by the Away team in last 3 matches.
 GA3Home	int	Goals conceded by the Home team in last 3 matches.
 GA3Away	int	Goals conceded by the Away team in last 3 matches.
 GF5Home	int	Goals scored by the Home team in last 5 matches.
 GF5Away	int	Goals scored by the Away team in last 5 matches.
 GA5Home	int	Goals conceded by the Home team in last 5 matches.
 GA5Away	int	Goals conceded by the Away team in last 5 matches.
 GA5Away	int	Goals conceded by the Away team in last 5 matches.
 PointsHome	int	Points gathered in the respective season in the league by Home team.
 PointsAway	int	Points gathered in the respective season in the league by Away team.
 PositionHome	int	Home team's current position in the league.



*OBRIGADA
PELA ATENÇÃO!*