



PUC Minas

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
PÓS-GRADUAÇÃO *LATO SENSU* EM CIÊNCIA DE DADOS E BIG DATA
TRABALHO DE CONCLUSÃO DE CURSO

UM MODELO DE APRENDIZADO DE MÁQUINA SUPERVISIONADO PARA PREVISÃO DE QUANTIDADE DE PÚBLICO NOS JOGOS DO CAMPEONATO BRASILEIRO DE FUTEBOL

Salomão Fernandes de Freitas Júnior

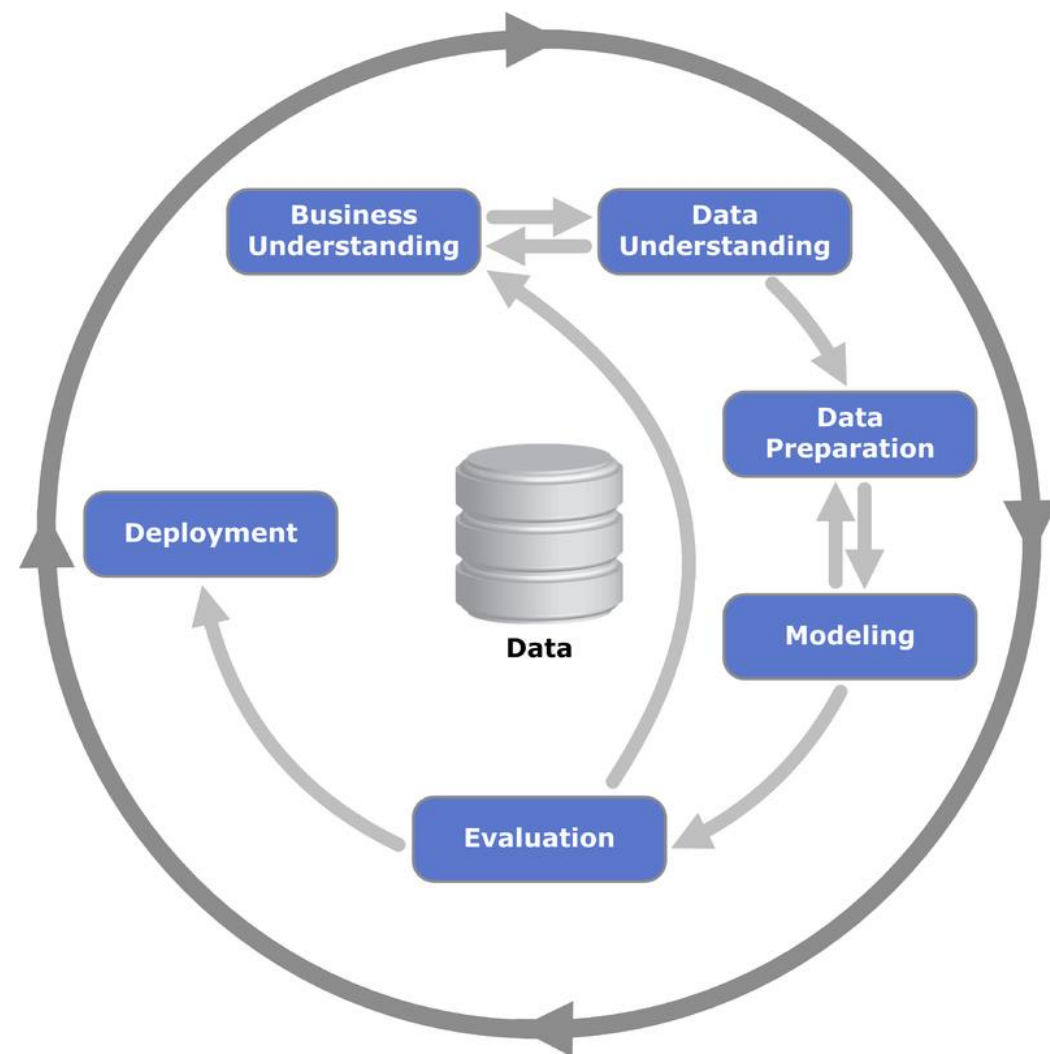
JUNHO/2024

INTRODUÇÃO

ETAPAS DO PROJETO

1. Entendimento do Domínio do Problema
2. Coleta de Dados
3. Análise Exploratória dos Dados
4. Preparação dos Dados
5. Criação de Modelos de Machine Learning
6. Apresentação dos Resultados
7. Implantação da Solução em Ambiente de Produção

MODELO CRISP-DM



AMBIENTE TECNOLÓGICO

ANÁLISE EXPLORATÓRIA / PREPARAÇÃO DE DADOS / MACHINE LEARNING



```
O2 - Preparação de Dados.ipynb
Arquivo  Editar  Ver  Inserir  Ambiente de execução  Ferramentas  Ajuda

+ Código  + Texto

ANÁLISE DE OUTLIERS

Definição de algumas funções que irão ajudar na análise dos outliers

[ ] #para calcular os limites de uma coluna na amostra
#valores acima do superior ou abaixo do inferior são considerados outliers
def limites(coluna):
    q1 = coluna.quantile(0.25) #1º Quartil da amostra
    q3 = coluna.quantile(0.75) #3º Quartil da amostra
    amplitude = q3-q1 # Distância Interquartil = Valor do 3º Quartil - Valor do 1º Quartil (Q3 - Q1)

    limite_inferior = q1-1.5*amplitude
    limite_superior = q3+1.5*amplitude

    return limite_inferior, limite_superior

# Para a exclusão de outliers de uma coluna (atributo)
def excluir_outliers(df, nome_coluna):
    qtd_linhas = df.shape[0]
    lim_inf, lim_sup = limites(df[nome_coluna])

    #filtrando somente as linhas que não representam outliers na coluna.
    df = df.loc[(df[nome_coluna] >= lim_inf) & (df[nome_coluna] <= lim_sup), :]

    linhas_removidas = qtd_linhas - df.shape[0]

    return df, linhas_removidas
```



APLICAÇÃO WEB DE PRODUÇÃO



```
preve_publico_brasileirao.py
requirements.txt

159 with st.container(border=True): # linha 2
160     st.markdown('##### Dados do Mandante')
161     col2_1, col2_2, col2_3 = st.columns([2, 4, 3, 3.7], gap='small') # cria duas colunas informando a proporção da largura
162     with col2_1:
163         with st.container(border=True, height=115):
164             # time_mandante (e grau_investimento_mandante)
165             mandante = st.selectbox('**Time Mandante:**', times_2024)
166             grau_mand = grau_investimento_times_2024[times_2024.index(mandante)]
167             if mandante not in ['América-MG']: # 'América-MG' saiu na dumização
168                 dicionario[f'time_mandante({mandante})'] = 1 # coluna dummie
169                 x_encoded['grau_investimento_mandante'] = grau_mand # coluna encoded
170
171     with col2_2:
172         with st.container(border=True, height=115):
173             # colocacao_mandante_antes
174             coloc_mand = st.slider(label = '**Colocação do mandante na tabela:**',
175                                   min_value = 1,
176                                   max_value = 20,
177                                   step = 1)
178             x_numericos['colocacao_mandante_antes'] = coloc_mand # atualiza o valor no dicionário
179
180     with col2_3:
181         with st.container(border=True, height=115):
182             # pontos_mand_last_5
183             pontos_mand = st.slider(label = '**Pontos conquistados pelo mandante nas últimas 5 rodadas:**',
184                                    min_value = 0,
185                                    max_value = 15,
186                                    step = 1)
187             x_numericos['points_mand_last_5'] = pontos_mand # atualiza o valor no dicionário
188
189 with st.container(border=True): # linha 3
190     st.markdown('##### Dados do Visitante')
```



ENTENDIMENTO DO DOMÍNIO DO PROBLEMA



Objetivo

Predição de Quantidade de Público em Jogos do Campeonato Brasileiro de Futebol

Características Relevantes

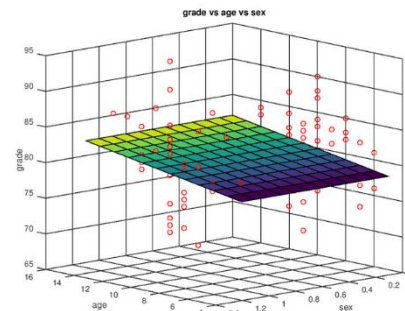
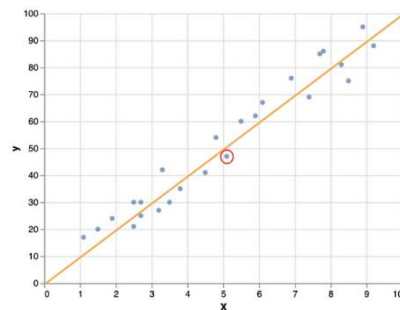
Etapa do campeonato, Desempenho do mandante, Importância do visitante, Desempenho recente do mandante, Dia da semana, Período do ano

Aplicabilidade

Apoio à Tomada de decisão em questões de Logística e Estratégias de marketing

Técnica utilizada

Aprendizado supervisionado com Regressão



ANÁLISE EXPLORATÓRIA DOS DADOS

Ações

1. Livre Exploração dos Datasets – Obtenção de “intimidade” com os dados
2. Aquisição de Conhecimento da Estrutura dos Dados: Atributos e Tipos de Dados
3. Exploração Visual dos Dados
4. Estatística Descritiva e Tabelas de Frequência dos Atributos
5. Identificação Inicial de Dados Inconsistentes, Valores Nulos e Outliers



<https://blog.unipar.br/analise-de-dados/>

ano_campeonato	data	rodada	publico	time_mandante	time_visitante	colocacao_mandante
2023	08/11/2023	33	18904	Athletico-PR	Fortaleza	7
2023	04/11/2023	32	37797	Grêmio	EC Bahia	4
2023	27/11/2023	35	6493	Goiás	Cruzeiro	18
2023	12/11/2023	34	25095	Atlético-MG	Goiás	6
2023	24/11/2023	35	39133	Corinthians	EC Bahia	14
2023	29/11/2023	36	59921	Flamengo	Atlético-MG	4
2023	02/12/2023	37	40338	Corinthians	Internacional	13
2023	03/12/2023	37	29986	Palmeiras	Fluminense	1
2023	23/11/2023	30	54804	Flamengo	RB Bragantino	3
2023	06/12/2023	38	0	Coritiba FC	Corinthians	19



PREPARAÇÃO DOS DADOS



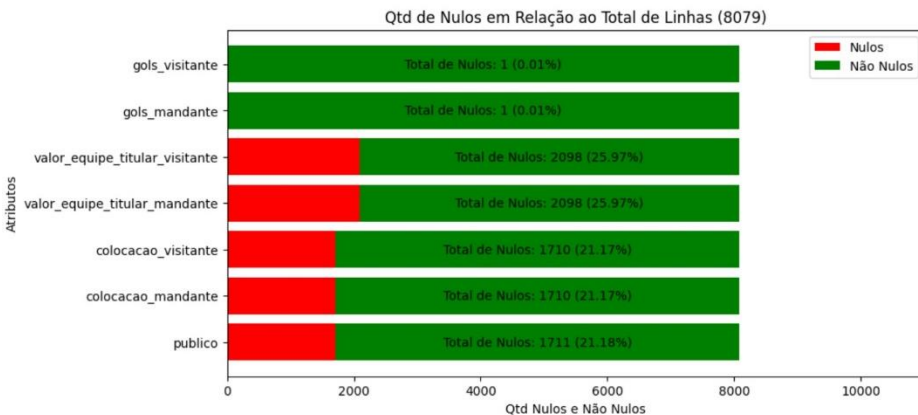
<https://pt.linkedin.com/pulse/s%C3%A9rie-analytics-cap%C3%ADtulo-iii-prepara%C3%A7%C3%A3o-dos-dados-marcelo-fernandes>

Seleção Inicial de Atributos (Feature Selection)

[ano_campeonato, data, rodada, publico, time_mandante, time_visitante, colocacao_mandante, colocação_visitante, valor_equipe_titular_mandante, valor_equipe_titular_visitante, gols_mandante e gols_visitante]

Ajustes de Valores

Uniformização dos nomes dos clubes: Goiás EC -> Goiás



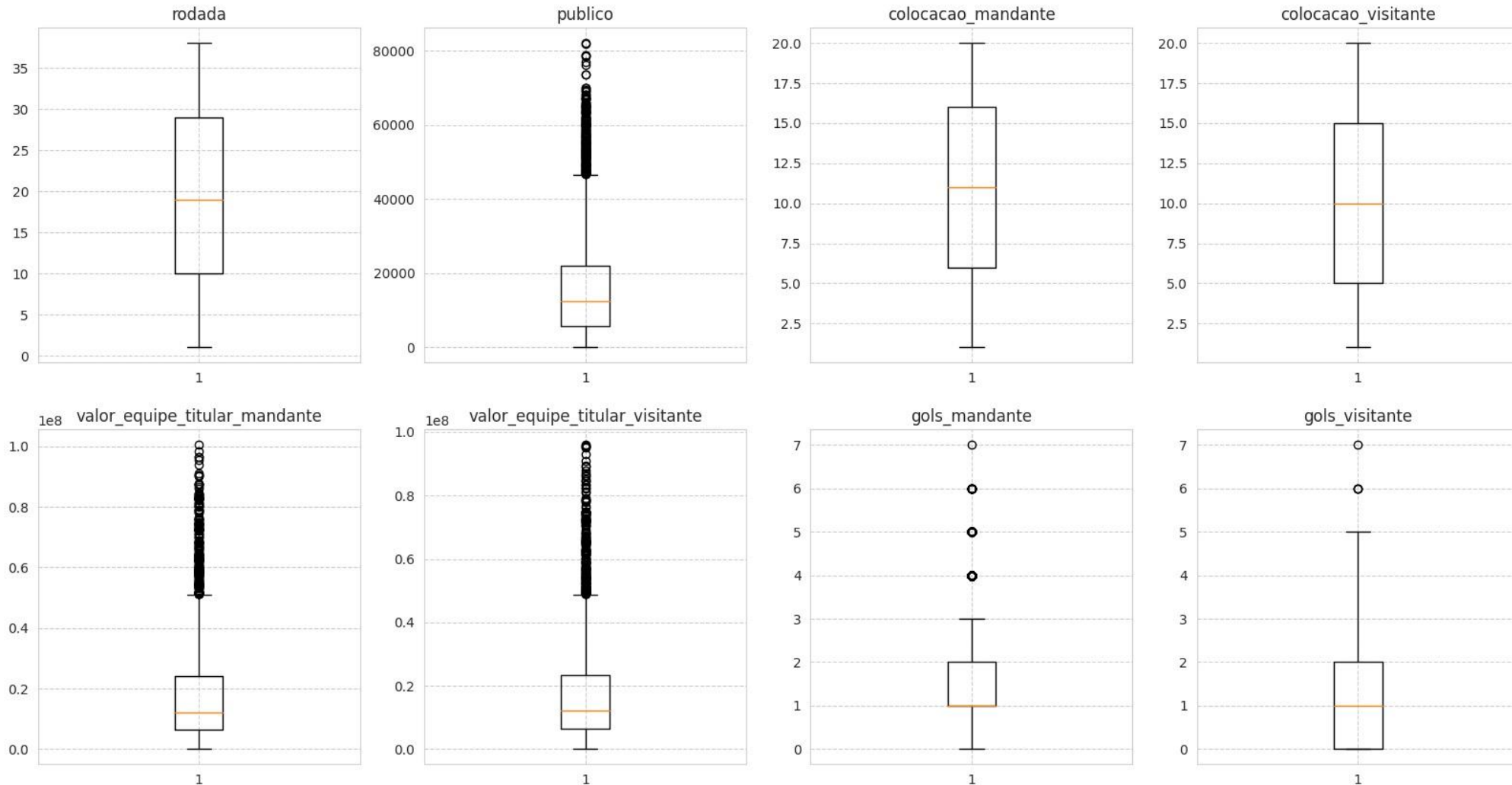
Tratamento de Valores Nulos

Atributos de valor de equipe: preenchimento com a média do clube no ano
Exclusão de algumas linhas onde não foi possível estimar (anos 2003 a 2006)

PREPARAÇÃO DOS DADOS

Análise de Outliers

Outliers identificados tratavam-se de valores reais: gols em grandes goleadas, jogos com grandes públicos, equipes com alto investimento



PREPARAÇÃO DOS DADOS



Engenharia de Atributos (Feature Engineering)

Criados os atributos dia_semana e trimestre, a partir do atributo data

Criados os atributos de pontuação recente dos clubes, a partir dos atributos de gols

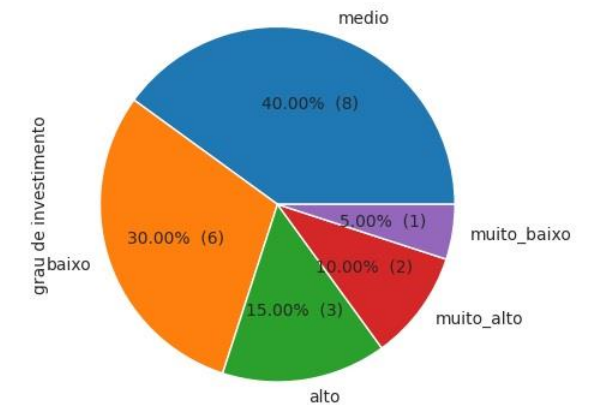
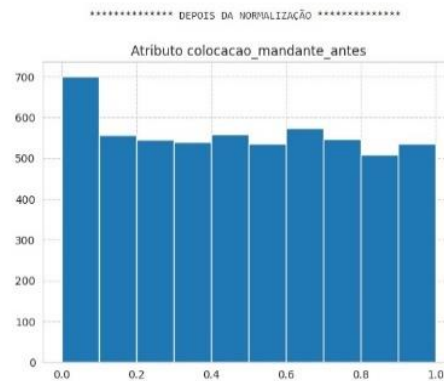
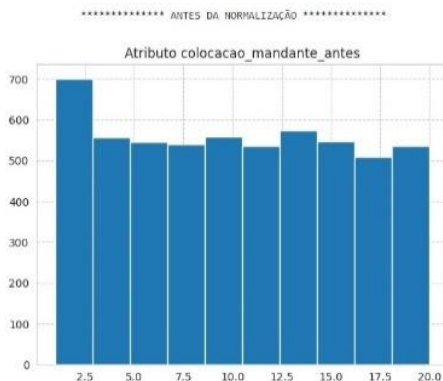
Transformações de Valores Numéricos

Foi aplicada a normalização por escala nos atributos numéricos

Codificação de Variáveis Categóricas

Aplicado o *Label Encoding* para as variáveis de grau de investimento

Aplicado o *Dummy Encoding* para as demais variáveis categóricas



CRIAÇÃO DE MODELOS DE MACHINE LEARNING



<https://blog.cefis.com.br/machine-learning-aprendizado-de-maquina/>

Algoritmos

Selecionados diversos algoritmos de vários vieses indutivos

LinearRegression, Kneighboor, DecisionTree, RandomForest, ExtraTrees, Bagging, GradienteBoosting, SVR, GaussianNB, MLP (Redes Neurais Artificiais)

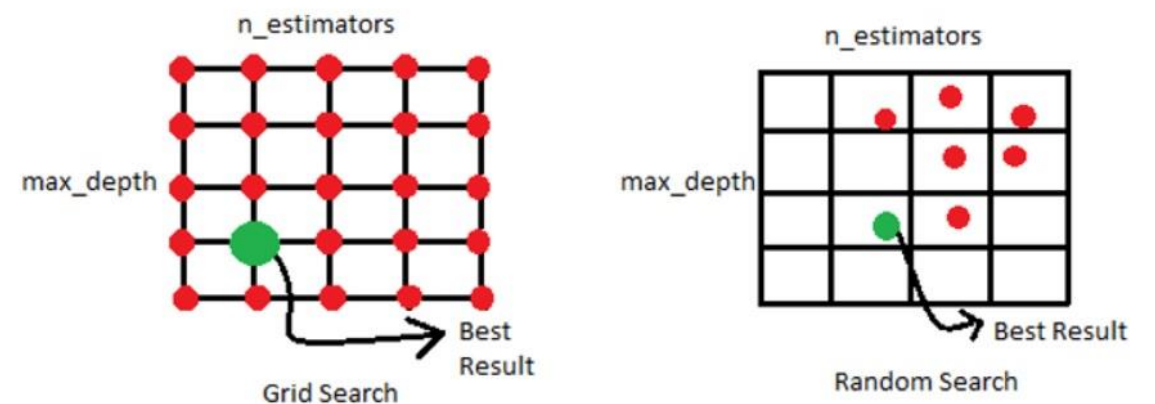
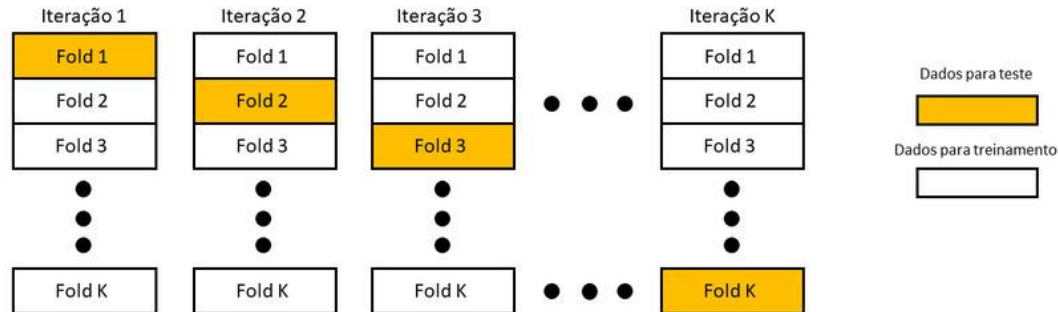
Estratégias de Validação

CrossValidation, com KFold Repetido $k = 10$

GridSearch e RandomSearch

Métricas

MSE, RMSE, MAE, R2



APRESENTAÇÃO DOS RESULTADOS



Ranking de Modelos

Modelos rankeados pelo Scorer (métrica MSE)

	Melhores Parâmetros	Scorer	MSE	RMSE	MAE	R2
VotingRegressor()	{}	86944338.1229	86944338.1229	9308.3293	6667.5721	0.5107
GradientBoostingRegressor()	{'n_estimators': 2000, 'min_samples_split': 55...	87560233.1613	87560233.1613	9352.2293	6751.4998	0.5053
BaggingRegressor()	{'n_estimators': 600, 'max_features': 1.0, 'bo...	90532539.3209	90532539.3209	9505.6257	6793.0617	0.4859
RandomForrestRegressor()	{}	92084897.3344	92084897.3344	9584.6729	6865.0558	0.4812
MLPRegressor()	{'learning_rate_init': 0.00021544346900318823,...	92891723.5119	92891723.5119	9618.0417	6923.0146	0.4776
ExtraTreesRegressor()	{}	94917758.7600	94917758.7600	9727.5549	6761.5418	0.4640
LinearRegression()	{}	96729969.1116	96729969.1116	9825.2441	7202.7539	0.4535
DecisionTreeRegressor()	{'min_samples_split': 110, 'max_depth': 40}	110594541.1342	110594541.1342	10502.2186	7648.8327	0.3753
KNeighborsRegressor()	{'n_neighbors': 25}	112760278.4747	112760278.4747	10611.8786	8055.1902	0.3643
SVR()	{'kernel': 'linear', 'degree': 2}	187200890.8312	187200890.8312	13661.5880	9777.7021	0.0535
GaussianNB()	{'var_smoothing': 1e-08}	210647944.1080	210647944.1080	14504.2932	10287.3673	0.1860
LinearSVR()	{'fit_intercept': True}	273265094.3289	273265094.3289	16510.7133	11416.7974	0.5398

APRESENTAÇÃO DOS RESULTADOS



Testes do Melhor Modelo com Dados Novos

***** MELHOR MODELO *****

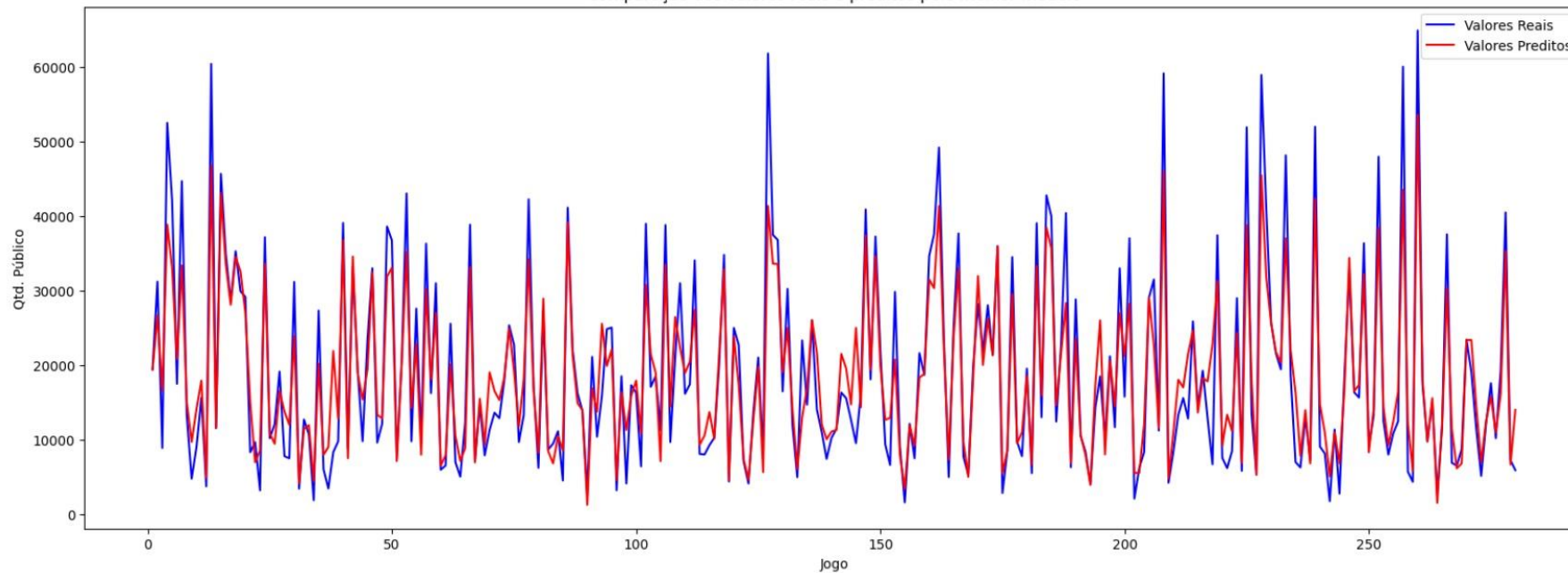
```
VotingRegressor(estimators=[('Modelo_1',  
                             GradientBoostingRegressor(max_depth=2,  
                                                         min_samples_leaf=2,  
                                                         min_samples_split=55,  
                                                         n_estimators=2000)),  
                             ('Modelo_2', BaggingRegressor(n_estimators=600)),  
                             ('Modelo_3', RandomForestRegressor()),  
                             ('Modelo_4',  
                              MLPRegressor(batch_size=256, early_stopping=True,  
                                             max_iter=1000)),  
                             ('Modelo_5', ExtraTreesRegressor())])
```

***** MÉTRICAS DO MODELO *****

MSE: 24182072.16 RMSE: 4917.53 MAE: 3503.07 R2: 0.87


***** Comparando os valores reais e preditos em um gráfico *****


Comparação dos valores reais e preditos pelo melhor modelo



IMPLANTAÇÃO EM AMBIENTE DE PRODUÇÃO

Aplicação WEB – Entrada Individual de Dados



Previsão de Público em Jogos do Campeonato Brasileiro de Futebol 

Previsão de Público - Jogo Individual - Jogos em lote - Sobre

Forneça os dados do jogo para predição do público esperado

Dados do Jogo

Data do Jogo:

28/04/2024

Rodada do Campeonato:

1

4

38

Dados do Mandante

Time Mandante:

Cruzeiro

Colocação do mandante na tabela:

1

12

20

Pontos conquistados pelo mandante nas últimas 5 rodadas:

0

4

15

Dados do Visitante

Time Visitante:

EC Vitória

Colocação do visitante na tabela:

1

17

20

Pontos conquistados pelo visitante nas últimas 5 rodadas:

0

1

15

Prever público para o jogo

Público estimado: 20.130

Salomão Freitas Jr.
salomaofreitasjr@gmail.com

Obrigado