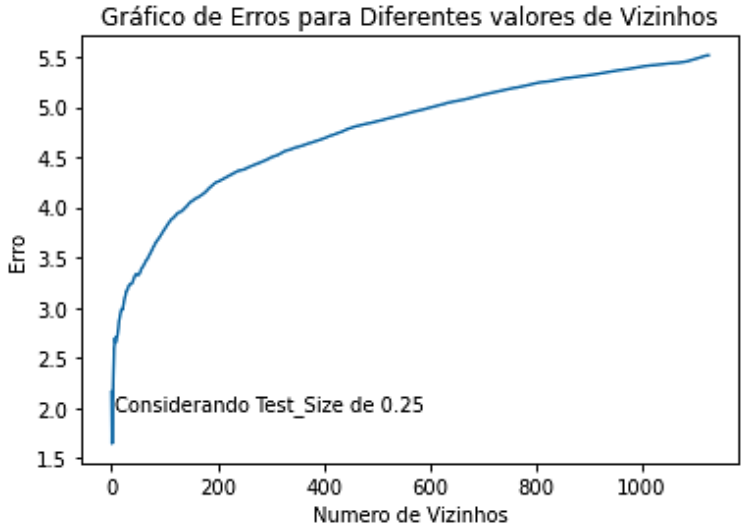
```python
# -*- coding: utf-8 -*-
"""
@author: Laura
"""
import pandas as pd
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt

base_dados = pd.read_csv('bd_tarefa4.CSV')
estatisticas = base_dados.describe()

previsores = base_dados.iloc[:,0:5].values
saida = base_dados.iloc[:,5].values

scaler = StandardScaler()
previsores = scaler.fit_transform(previsores)

x = 0.25
(prev_treino, prev_teste, saida_treino, saida_teste) = train_test_split(
        previsores, saida, test_size=x, random_state=0)

"""
Grafico - Evolução do erro
"""

erro = []
for i in range(1, len(prev_treino)):
    regressao = KNeighborsRegressor(n_neighbors=i)
    regressao.fit(prev_treino, saida_treino)
    regressaoFinal = regressao.predict(prev_teste)

    erro.append(mean_absolute_error(saida_teste, regressaoFinal))

print('Menor erro é {} e ocorre com {} vizinhos'.format(
    round(min(erro), 3), erro.index( min(erro) ) ))
print('Maior erro é {} e ocorre com {} vizinhos'.format(
    round(max(erro), 3), erro.index( max(erro) ) ))

plt.plot(erro)
plt.title("Gráfico de Erros para Diferentes valores de Vizinhos")
plt.annotate('Considerando Test_Size de {}'.format(x), xy=(5, 2))
plt.xlabel("Numero de Vizinhos")
plt.ylabel("Erro")
plt.show()

>> Menor erro é 1.645 e ocorre com 1 vizinhos
>> Maior erro é 5.515 e ocorre com 1125 vizinhos
```

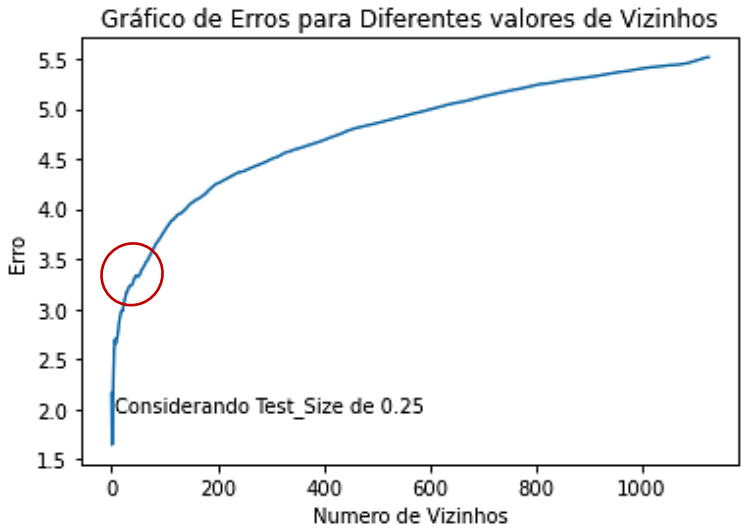
Inicialmente, na própria base de dados acrescentou-se uma linha com os títulos das colunas.



O gráfico gerado refere-se a valores de erro, para valores de vizinhos (*n_neighbors*) diferentes, para comparação e otimização de valores.

Também foram simulados diferentes "*test_sizes*" (entre 10% e 75%) porém os resultados se mostraram extremamente semelhantes.

Percebe-se que, quanto mais vizinhos, maior o número de erros (como esperado), porém pode-se considerar um ponto ideal na região entre 0 e 100 vizinhos (próximo a um "pico").



Define-se este ponto pois é o que apresenta menos erros, considerando um número razoável de vizinhos (aproximado em 50 com erro de 3.32873).

Sabendo que o maior erro registrado é de 5.515 e que o menor erro é 1.645, pode-se definir uma boa aproximação, já que a média entre estes valores é de 3,58 (o erro definido é menor).