

MACHINE LEARNING PROJECT – 2 – BUILDING A MODEL USING RANDOM FOREST REGRESSOR TO PREDICT MEDIAN HOUSE VALUE BY ANALYSING CALIFORNIA HOUSING DATASET

```
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error
```

```
cal_housing_data = pd.read_csv("D:/6.Data Analytics/Machine learning - Kaggle/Project1/housing.csv")
pd.set_option('display.max_column', None)
```

```
print(cal_housing_data.shape)
```

```
=====
(20640, 10)
|
```

```
print(cal_housing_data.head())
```

```
===== RESTART: D:\coding files so far\python\a2.py =====
   longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
0    -122.23    37.88             41.0         880.0         129.0
1    -122.22    37.86             21.0        7099.0        1106.0
2    -122.24    37.85             52.0        1467.0         190.0
3    -122.25    37.85             52.0        1274.0         235.0
4    -122.25    37.85             52.0        1627.0         280.0

   population  households  median_income  median_house_value  ocean_proximity
0         322.0         126.0         8.3252         452600.0        NEAR BAY
1        2401.0        1138.0         8.3014        358500.0        NEAR BAY
2         496.0         177.0         7.2574        352100.0        NEAR BAY
3         558.0         219.0         5.6431        341300.0        NEAR BAY
4         565.0         259.0         3.8462        342200.0        NEAR BAY
>
```

```
print(cal_housing_data.isna().any().any())
print("\n")
print(cal_housing_data.isna().sum())
```

```
===== RESTART: D:\c
True

longitude          0
latitude           0
housing_median_age  0
total_rooms         0
total_bedrooms      207
population          0
households          0
median_income       0
median_house_value  0
ocean_proximity    0
dtype: int64
```

```
print(cal_housing_data['total_bedrooms'].tail())
```

```
===== RESTART: D:\coding fil
20635    374.0
20636    150.0
20637    485.0
20638    409.0
20639    616.0
Name: total_bedrooms, dtype: float64
```

```
cal_housing_data.info()
```

```
===== RESTART: D:\coding files so far\p
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   longitude              20640 non-null  float64
1   latitude               20640 non-null  float64
2   housing_median_age     20640 non-null  float64
3   total_rooms            20640 non-null  float64
4   total_bedrooms         20433 non-null  float64
5   population             20640 non-null  float64
6   households             20640 non-null  float64
7   median_income          20640 non-null  float64
8   median_house_value     20640 non-null  float64
9   ocean_proximity        20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

```
print(cal_housing_data['ocean_proximity'].head())
```

```
===== RESTART: D:\coding f
0    NEAR BAY
1    NEAR BAY
2    NEAR BAY
3    NEAR BAY
4    NEAR BAY
Name: ocean_proximity, dtype: object
```

```
filtered_cal_housing_data = cal_housing_data.dropna(axis = 0) #removed all the rows that contains NA values
print(filtered_cal_housing_data.isna().sum())
print("\n")
print(filtered_cal_housing_data.columns)
```

```
===== RESTART: D:\coding files so far\python\az.py =====
longitude      0
latitude       0
housing_median_age  0
total_rooms    0
total_bedrooms  0
population     0
households     0
median_income  0
median_house_value  0
ocean_proximity  0
dtype: int64

Index(['longitude', 'latitude', 'housing_median_age', 'total_rooms',
      'total_bedrooms', 'population', 'households', 'median_income',
      'median_house_value', 'ocean_proximity'],
      dtype='object')
```

```
y = filtered_cal_housing_data.median_house_value
print(y.head())
```

```
===== RESTART: D:\coding files s
0    452600.0
1    358500.0
2    352100.0
3    341300.0
4    342200.0
Name: median_house_value, dtype: float64
```

```
cal_housing_features = ['longitude', 'latitude', 'housing_median_age', 'total_rooms', 'population', 'households',
                        'median_income', 'ocean_proximity', 'total_bedrooms']
```

```
X = filtered_cal_housing_data[cal_housing_features]
X = pd.get_dummies(X, columns = ['ocean_proximity'])
print(X.head())
```

```
===== RESTART: D:\coding files so far\python\a2.py =====
   longitude  latitude  housing_median_age  total_rooms  population  \
0   -122.23    37.88           41.0         880.0        322.0
1   -122.22    37.86           21.0       7099.0       2401.0
2   -122.24    37.85           52.0       1467.0        496.0
3   -122.25    37.85           52.0       1274.0        558.0
4   -122.25    37.85           52.0       1627.0        565.0

   households  median_income  total_bedrooms  ocean_proximity_<1H OCEAN  \
0         126.0         8.3252           129.0             False      True
1        1138.0         8.3014          1106.0             False      True
2         177.0         7.2574           190.0             False      True
3         219.0         5.6431           235.0             False      True
4         259.0         3.8462           280.0             False      True

   ocean_proximity_INLAND  ocean_proximity_ISLAND  ocean_proximity_NEAR BAY  \
0                False                False                True
1                False                False                True
2                False                False                True
3                False                False                True
4                False                False                True

   ocean_proximity_NEAR OCEAN
0                False
1                False
2                False
3                False
4                False
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2)
```



```
cal_housing_model = RandomForestRegressor()  
cal_housing_model.fit(X_train,y_train)  
prediction = cal_housing_model.predict(X_test)  
mae = mean_absolute_error(y_test, prediction)  
print("The Mean Absolute Error: ",round(mae,2))
```

```
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 31807.44  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 31110.02  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 31493.82  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 33390.85  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 32051.39  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 32113.03  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 31672.06  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 30978.71  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 31005.44  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 32377.82  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 31050.66  
  
===== RESTART: D:\coding files so far\python\a2.py =====  
The Mean Absolute Error: 32087.48
```