**PRESENTED BY: 1. KIRTIKAA SHARMA (2115800013)**

**2.SHIVA SRIVASTAVA (2115800023)**

**3.ANANYA AGRAWAL (2115800002)**

**4.MUSKAN GARG (2115800018)**

****

Project Technical Document

~~~~

Contents:

[Project Description:](about:blank)

[Project Technical Details:](about:blank)

[Exploratory Data Analysis](about:blank)

[Data Visualization:](about:blank)

[Model Building:](about:blank)

[Deployment using Flask:](about:blank)

Boston House Price Prediction



**Project description:**

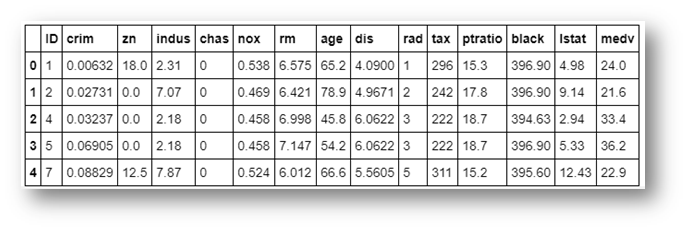
Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. Ask a home buyer to describe their dream house, and they probably won’t begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition’s data-set proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

**About the Dataset**

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. The goal of this project is to create a regression model that is able to accurately estimate the price of the house given the features.

In this dataset made for predicting the Boston House Price Prediction. Here I just show the all of the feature for each house separately. Such as Number of Rooms, Crime rate of the House’s Area and so on. We’ll show in the upcoming part.

**Data Overview**

****

1. **CRIM** per capital crime rate by town

2. **ZN** proportion of residential land zoned for lots over 25,000 sq.ft.

3. **INDUS** proportion of non-retail business acres per town

4. **CHAS** Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

5. **NOX** nitric oxides concentration (parts per 10 million)

6. **RM** average number of rooms per dwelling

7. **AGE** proportion of owner-occupied units built prior to 1940

8. **DIS** weighted distances to five Boston employment centers

9. **RAD** index of accessibility to radial highways

10.**TAX** full-value property-tax rate per 10,000 USD

11. **PTRATIO** pupil-teacher ratio by town

12. **Black** 1000(Bk — 0.63)² where Bk is the proportion of blacks by town

13. **LSTAT** % lower status of the population

**About the Algorithms used in**

The major aim of in this project is to predict the house prices based on the features using some of the regression techniques and algorithms.

**1.** **Linear Regression**

**2. Random Forest Regression Technique**

**About the Algorithms used in**

The major aim of in this project is to predict the house prices based on the features using some of the regression techniques and algorithms.

**Machine Learning Packages are used for in this Project**



**Data Collection**

I got the Dataset from [Kaggle.](https://www.kaggle.com/altavish/boston-housing-dataset) This Dataset consist several features such as Number of Rooms, Crime Rate, and Tax and so on. Let’s know about how to read the dataset into the Jupyter Notebook. You can download the dataset from [Kaggle](https://www.kaggle.com/altavish/boston-housing-dataset) in csv file format.

As well we can also able to get the dataset from the sklearn datasets. Yup! It’s available into the [sklearn Dataset](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html" \t "_blank).

# 2. Project Highlights

# This project is designed to get you acquainted to working with datasets in Python and applying basic machine learning techniques using NumPy and Scikit-Learn. Before being expected to use many of the available algorithms in the sklearn library, it will be helpful to first practice analyzing and interpreting the performance of your model.

# Things you will learn by completing this project:

# How to use NumPy to investigate the latent features of a dataset.

# How to analyze various learning performance plots for variance and bias.

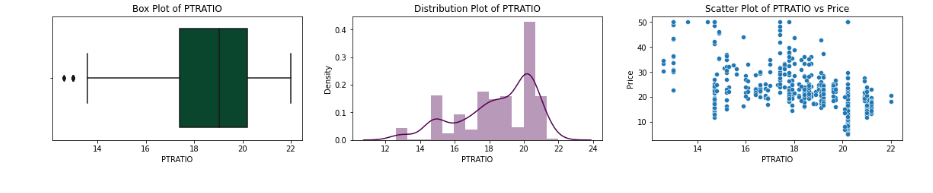
# How to determine the best-guess model for predictions from unseen data.

# How to evaluate a model's performance on unseen data using previous data.

# Exploratory Data Analysis:

Analysis were done on the basis of several charts which are shown below:

#Box Plot, Distribution Plot and Scatter Plot for PTRATIO



#### Observations:

- "PTRATIO" is normally distributed.  
- This shows that it is left skewed  
- It has no outliers

create a linear regression line

Chart, scatter chart

Description automatically generated

### Observation:

- This is a regression plot between RM and Price with the regression line.  
- Some variation is also there, but is also directly proportional aligned to the line.  
- Data is densely distributed showing high correlation.

This is a regression plot between CHAS and Price with the regression line.

Chart, scatter chart

Description automatically generated

Observation : Data is distributed only along the extreme boundary points.

Plot this residuals

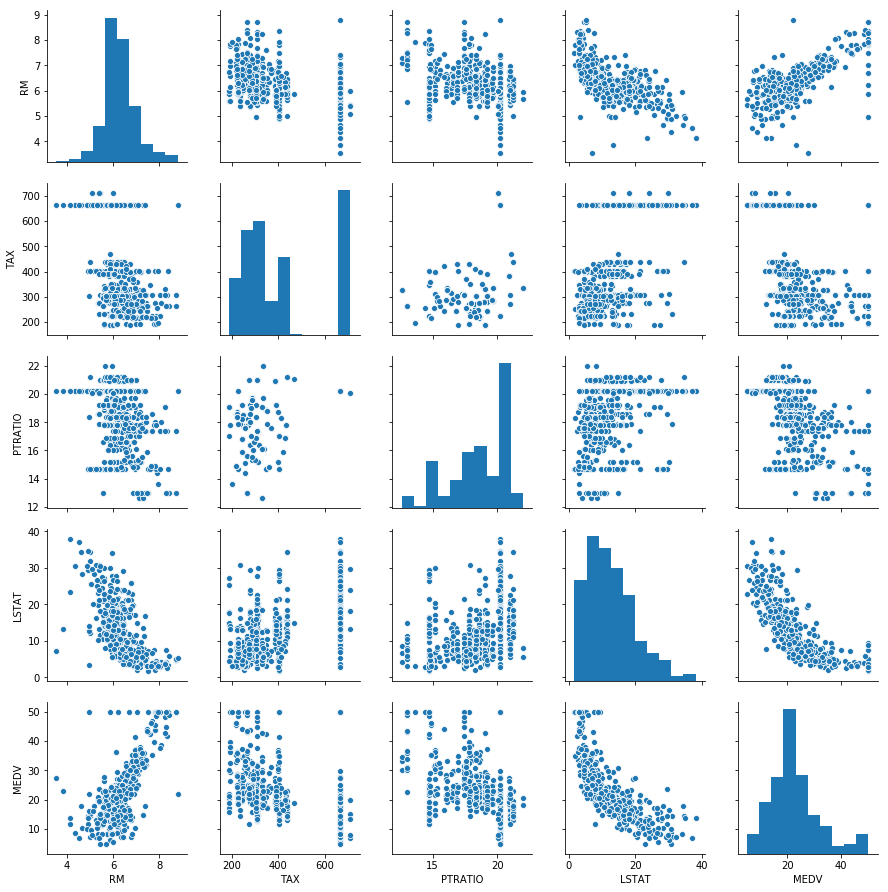
Chart, histogram

Description automatically generated

#### Observation:

- The given output plot has come out to be normally distributed so, our predicted model is fit till now.

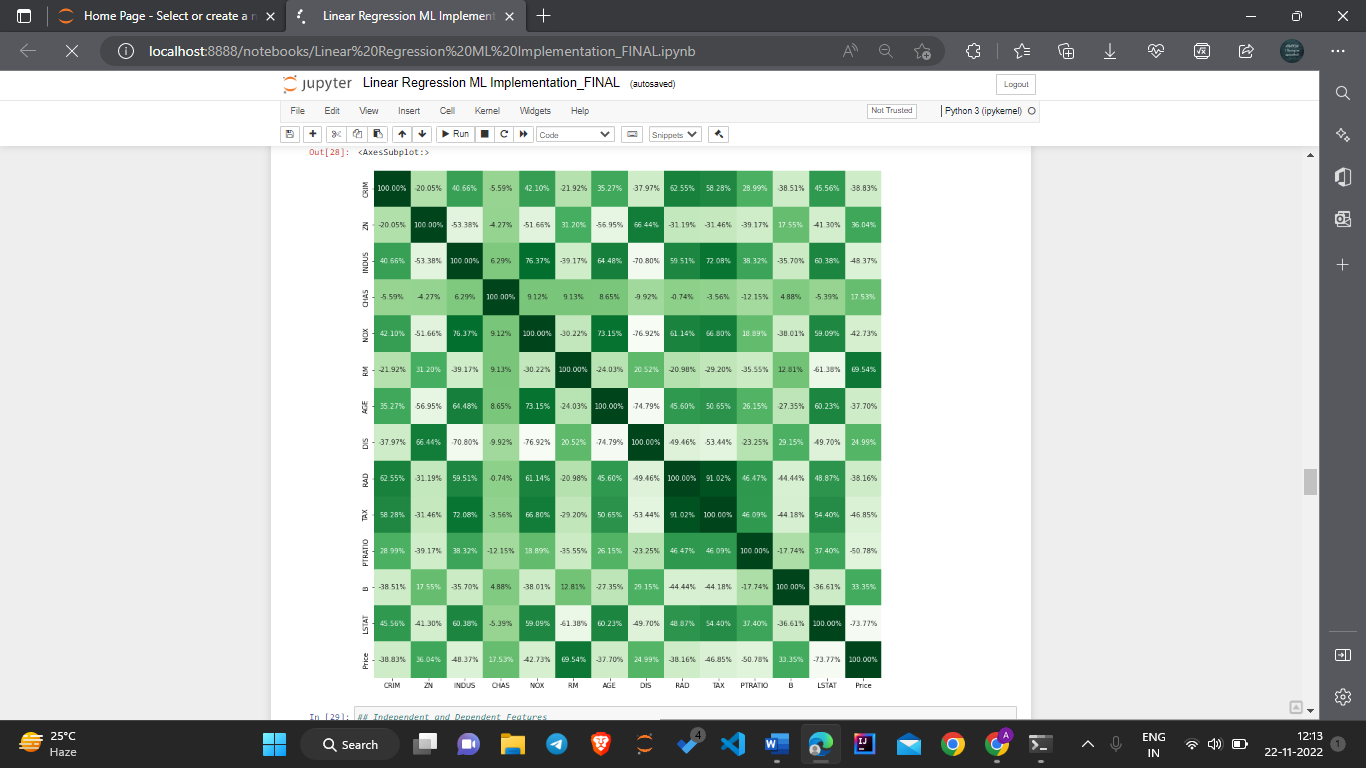
Create a pair plot



Observation:

It shows the relation between independent variables or independent and dependent variable through scatter plot and histogram.

**Heat map as described below:**



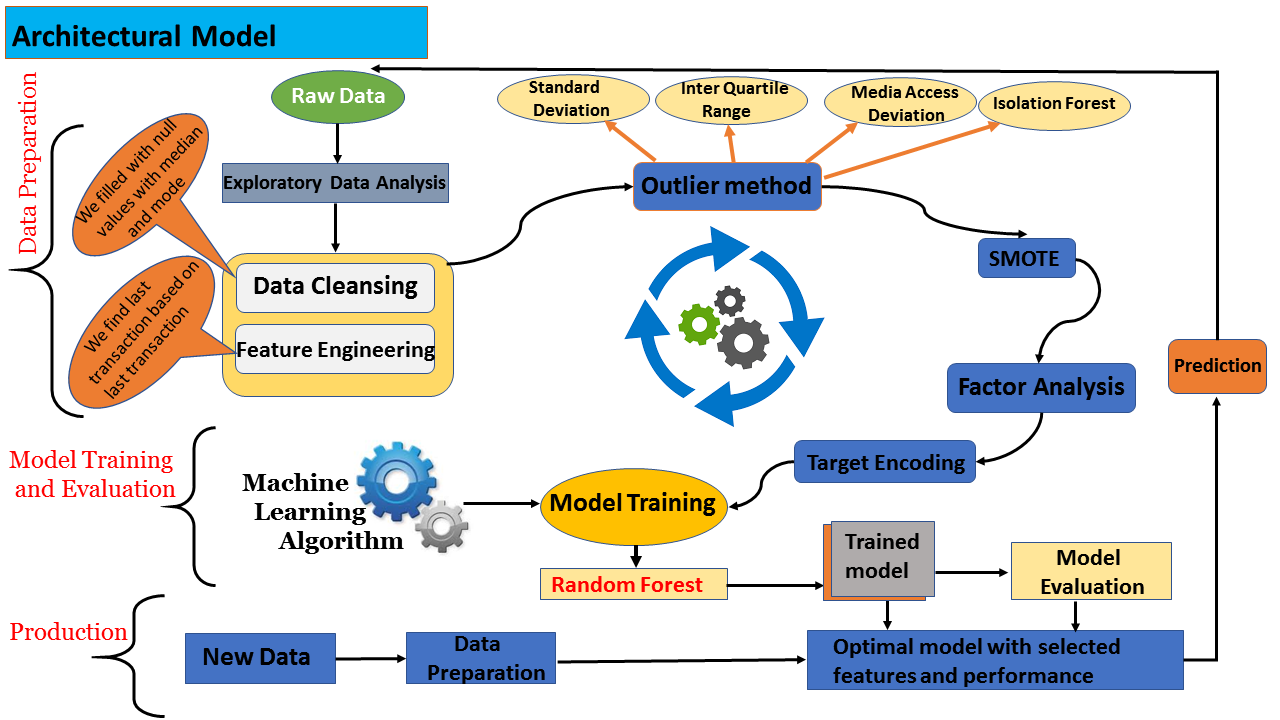
#### Observation:

- Positive sign implies positive correlation between two features whereas Negative sign implies negative correlation between two features.  
- To know which features have good correlation with our dependent variable PRICE and can help in having good predictions.  
- It is observed that INDUS, RM, TAX, PTRATIO and LSTAT shows some good correaltion with PRICE.

Results

After necessary data pre-processing is done various models are implemented and evaluated. For evaluation of models Train-Test split method is used in this paper. Data is split into 80% and 20%, 80% of the data is used as training data and the rest of the data is used as test data. The first partition which is the training data is used for fitting the model and the second partition which is the test data is used for testing the accuracy of the model. There are various ways of splitting the data like 70-30, 70% for training and 30% for testing or 75-25, 75% for training and 25% for testing etc .there is no hard and fast rule, but in this paper the dataset is split into the ratio mentioned above which is 80% for training and 20% for testing. After data set is split the models are implemented which are Simple Linear Regression, Multiple Linear Regressor and Random Forest Regressor Technique.

4. Model building:



**Observation**

### The Boston dataset used here contains only 506 instances which are very small however some important observations are made using this dataset. Firstly after splitting the data in 80-20, 80% for training and 20% for testing using Train-Test split method and using metrics like R-Squared, RMSE, Cross-Validation, it is seen that in all the cases Randon Forest Regression performs the best and the second best performance is given by Multiple Linear Regression. Since, adjusted R^2 value of the Random Forest Regression Technique is greater than the adjusted r^2 value of the Multiple linear Regression Technique. Hence, Random Forest Regression Technique is more efficient. So, We are using Random Forest Regression Model for further pickling and deployment.

.

**Future Works**

This model can be considered as the baseline for predicting house price. Further evaluation can be done here by increasing the data. More data can be collected and more attributes can be increased for getting a much better evaluation of the model. The data collected in Boston house dataset is from 1978 which is almost 50 years old and since then a lot of changes have occurred in house price due to inflation rate. Thus, new data can be collected and further evaluation can be made on the new collected data. In this paper three models are implemented which are Simple Linear Regression, Multiple Linear Regression and Random forest Linear Regressor on the Boston House dataset. More advanced models like Support Vector Machine, Decision

Tree, Random Forest, Multiple Linear Regression etc can be implemented and the results can be compared. Other ensemble learning techniques can be used like Adaboost, Xgboost etc and the results can be compared to the previous models. Feature selection techniques like Linear Discriminant Analysis, Principle Component Analysis, Independent component Analysis etc can be used before implementing the models and a study can be made on the performance of the models before applying feature selection methods and after implementing feature selection methods

**Conclusion**

It is very important to predict house price accurately. To accurately predict house price various variables must be taken into consideration like location of the house, the views that are visible from the house, crime rate around that area etc. A lot of time people pay overprice from the actual market price for a real estate property, similarly a lot of time sellers get very low price compare to the actual market price of the property. Not only people, various estate agencies also face the same problem where they are not sure whether to invest toward a certain property or not. They are confused as they are not able to predict what the price of the house can be in future. The main purpose of this paper is to help people who are facing these issues to predict the house price in future years. In this paper an intelligent system is made using the Regressor models which are Simple Linear Regression, Multiple Linear Regression,on the Boston House Dataset to predict the house price. In this paper it is observed that using the Boston house dataset, and implementing various data preprocessing techniques which are needed on the dataset.

**Deployment**

**Software And Tools Requirements in this project**

1. [**Github Account**](https://github.com/)
2. [**HerokuAccount**](https://heroku.com/)
3. [**VSCodeIDE**](https://code.visualstudio.com/)
4. [**GitCLI**](https://git-scm.com/book/en/v2/Getting-Started-The-Command-Line)

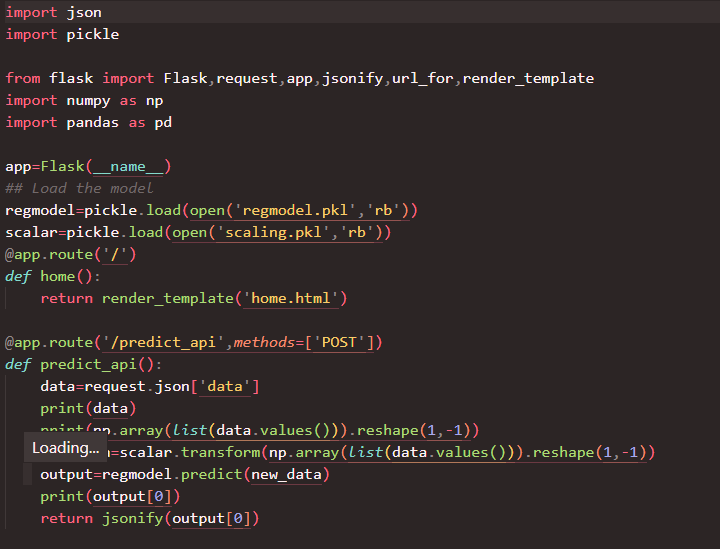
**Create a new environment**

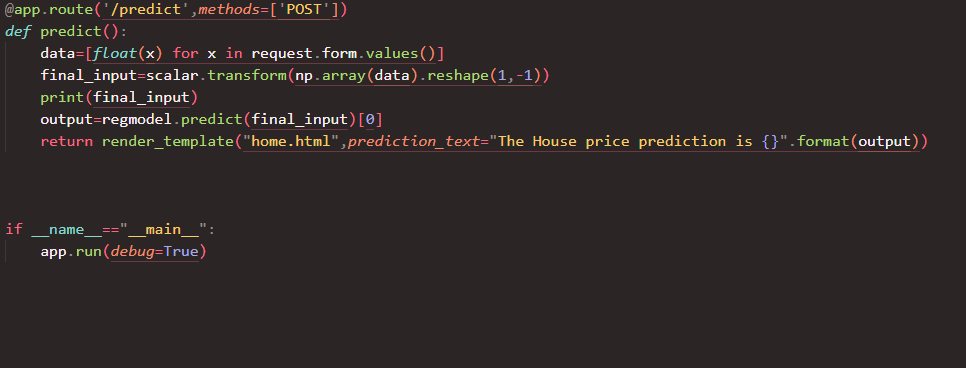
**python -m venv venv**

**venv/Scripts/Activate.bat**

**then select intrepreter 3.10.5**

**App.py**

****

****

**Boston House Price prediction uses the following packages and library from python:**

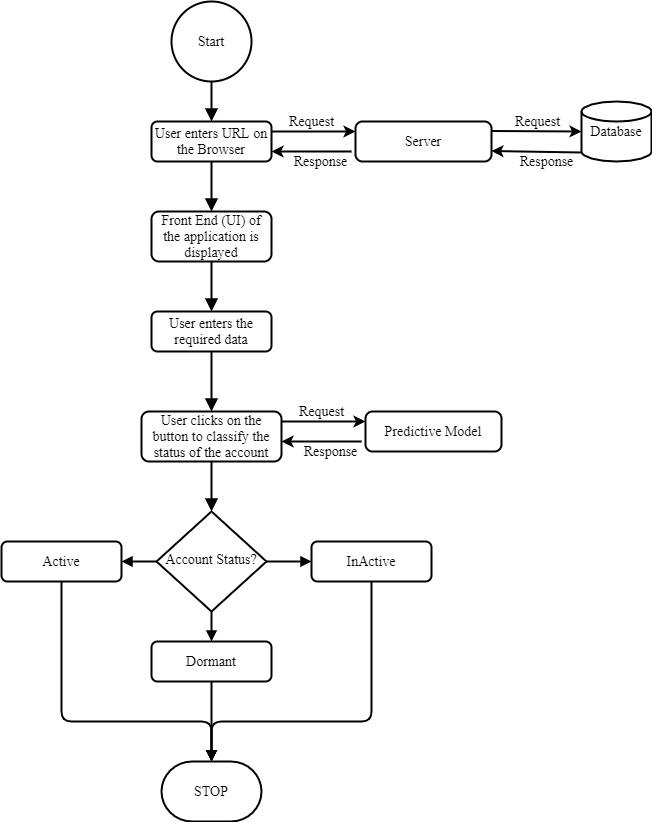
|  |
| --- |
| **import pandas as pd**  **import matplotlib.pyplot as plt**  **import seaborn as sns**  **import numpy as np**  **import sklearn**  **from sklearn.ensemble import RandomForestClassifier**  **from dataprep.eda import plot,plot\_correlation,plot\_missing,create\_report**  **from sklearn.preprocessing import StandardScaler** |

.

# 6. Deployment using Flask:

Deployment process was done using flask technique.

**Deployment Architecture**

****

Created app.py file to show prediction of House accounts

