**MODELLING AND SIMULATION**

**( IT – 205 )**



**PROJECT REPORT**

**SUBMITTED BY : - Under the Supervision of : -**

*VARUN KUMAR (2K19 / IT / 140) Mr . Ankit Yadav*

*YASHIT KUMAR (2K19 / IT / 149) Asst.Prof.*

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

**CERTIFICATE**

I hereby certify that the Project titled “**House Price Prediction** ” which is submitted by Varun Kumar ; Roll No – 2K19/IT/140 ; and Yashit Kumar ; Roll No – 2K19/IT/149 INFORMATION TECHNOLOGY, Delhi Technological University, Delhi in fulfillment of the requirement for the 3rd semester of Bachelor of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

**Place : Delhi Mr. Ankit Yadav**

**Date : 15-11-20 Supervisor**

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

**ACKNOWLEDGEMENT**

We would like to convey our heartfelt thanks to our supervisor Mr. Ankit Yadav for his ingenious ideas, tremendous help and cooperation. We are extremely grateful to our friends who gave valuable suggestions and guidance for completion of our project. The cooperation and healthy criticism came handy and useful with them.

**OBJECTIVE**

Thousands of houses are sold everyday. There are some questions every buyer asks himself like: What is the actual price that this house deserves? Am I paying a fair price?

**In this Project a machine learning model is proposed to predict a house price based on data sets (features) related to the house .**

**Our model will be used to predict house prices in given area and invest in that area.**

While learning about machine learning it is best to actually work with real world data .

The main objectives of this Project are as follows:

* To apply data preprocessing and preparation techniques in order to obtain clean data
* To build machine learning models able to predict house price based on house features
* To analyze and compare models performance in order to choose the best model

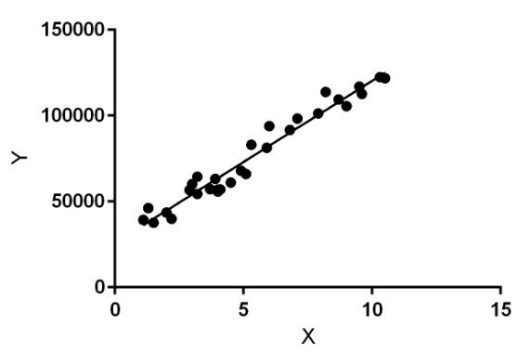


**ALGORITHMS USED**

We are trying to predict the House price using the machine learning techniques with the help of the previous works. We have used the Simple Linear Regression , Decision Tree Regression and Random Forest Regression So, it would be helpful for the people and it may avoid them in making mistakes.

1. **Linear Regression**

**Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.  
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

1. **Decision Tree Regression**

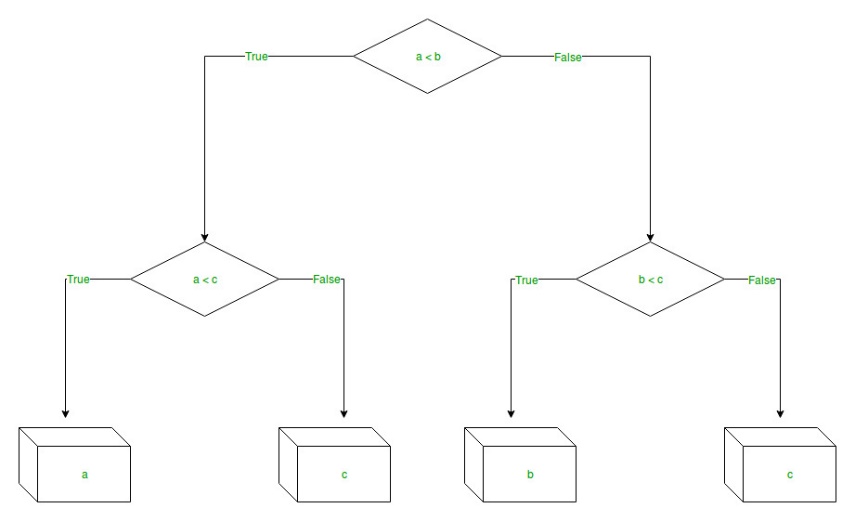
**Decision Tree** is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs and utility.

Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

The branches/edges represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]
2. Result [End Nodes]

The branches/edges represent the truth/falsity of the statement and takes makes a decision based on that in the example below which shows a decision tree that evaluates the smallest of three numbers:

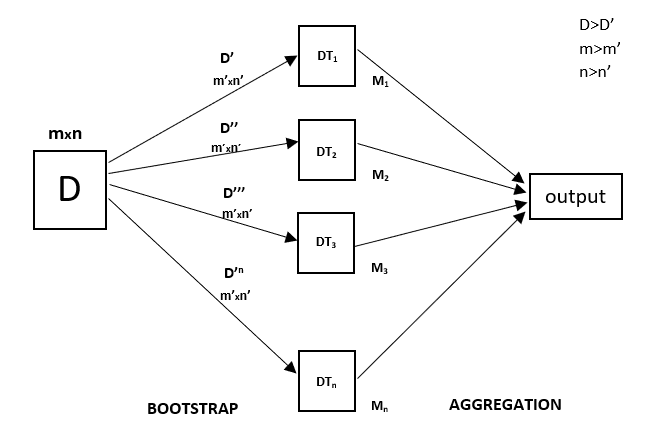


Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

**Discrete output example:** A weather prediction model that predicts whether or not there’ll be rain in a particular day.  
**Continuous output example:** A profit prediction model that states the probable profit that can be generated from the sale of a product

1. **Random Forest Regression**

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.  
Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.



**DATASETS**

1. Title: Boston Housing Data

2. Sources:

(a) Origin: This dataset was taken from the StatLib library which is

maintained at Carnegie Mellon University.

(b) Creator: Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the

demand for clean air', J. Environ. Economics & Management,

vol.5, 81-102, 1978.

(c) Date: July 7, 1993

3. Past Usage:

- Used in Belsley, Kuh & Welsch, 'Regression diagnostics ...', Wiley,

1980. N.B. Various transformations are used in the table on

pages 244-261.

- Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning.

In Proceedings on the Tenth International Conference of Machine

Learning, 236-243, University of Massachusetts, Amherst. Morgan

Kaufmann.

4. Relevant Information:

Concerns housing values in suburbs of Boston.

5. Number of Instances: 506

6. Number of Attributes: 13 continuous attributes (including "class"

attribute "MEDV"), 1 binary-valued attribute.

7. Attribute Information:

1. CRIM per capita 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 centres

9. RAD index of accessibility to radial highways

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

11. PTRATIO pupil-teacher ratio by town

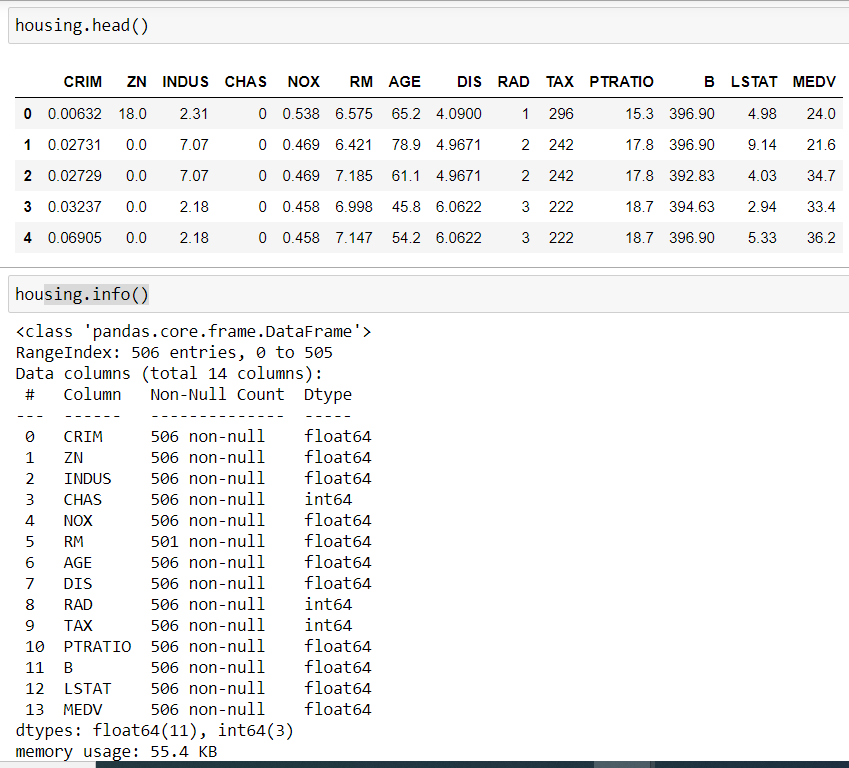
12. B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks

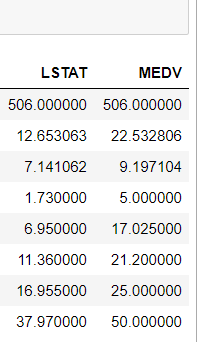
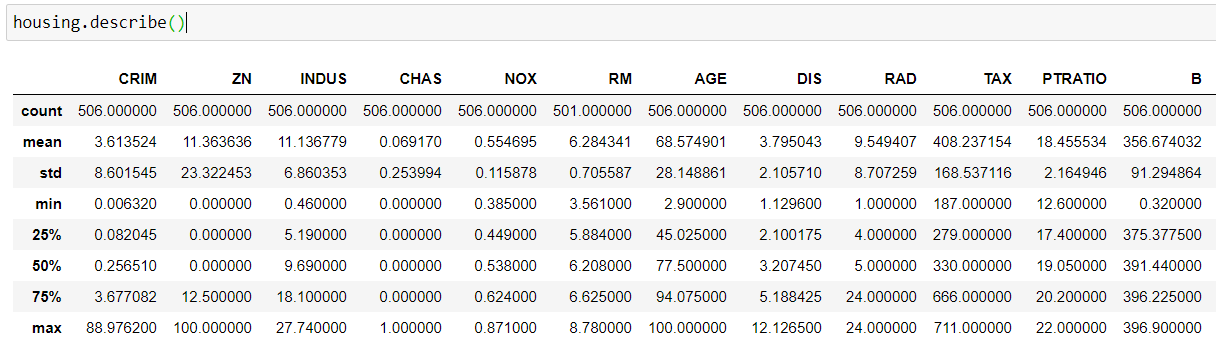
by town

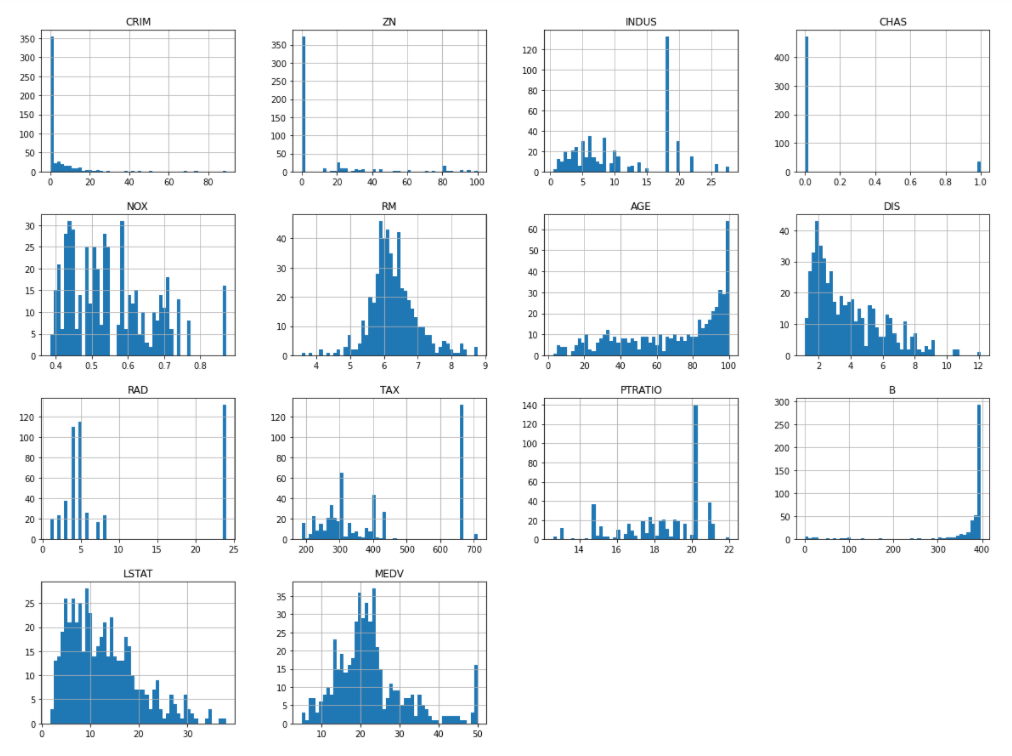
13. LSTAT % lower status of the population

14. MEDV Median value of owner-occupied homes in $1000's

8. Missing Attribute Values: None.

 **DATASET ANALYSIS**



**FEATURES GRAPH ANALYSIS** 

**FEATURES OBSERVATION**

Data Science is the process of making some assumptions and hypothesis on the data, and testing them by performing some tasks. Initially we could make the following intuitive assumptions for each feature:

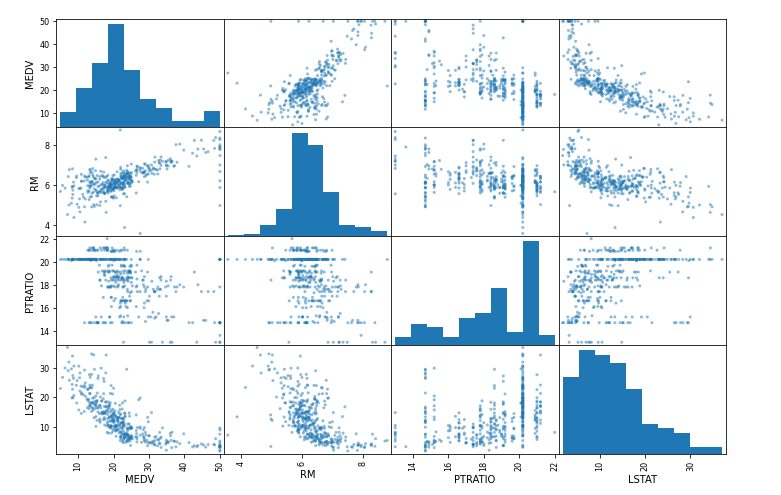
* Houses with more rooms (higher ‘RM’ value) will worth more. Usually houses with more rooms are bigger and can fit more people, so it is reasonable that they cost more money. They are directly proportional variables.
* Neighborhoods with more lower class workers (higher ‘LSTAT’ value) will worth less. If the percentage of lower working class people is higher, it is likely that they have low purchasing power and therefore, they houses will cost less. They are inversely proportional variables.
* Neighborhoods with more students to teachers ratio (higher ‘PTRATIO’ value) will be worth less. If the percentage of students to teachers ratio people is higher, it is likely that in the neighborhood there are less schools, this could be because there is less tax income which could be because in that neighborhood people earn less money. If people earn less money it is likely that their houses are worth less. They are inversely proportional variables.

We’ll find out if these assumptions are correct through the project.

## **LOOKING FOR CORRELATIONS**

We will start by creating a scatterplot matrix that will allow us to visualize the pair-wise relationships and correlations between the different features.

It is also quite useful to have a quick overview of how the data is distributed and wheter it cointains or not outliers.



We can spot a linear relationship between ‘RM’ and House prices ‘MEDV’. In addition, we can infer from the histogram that the ‘MEDV’ variable seems to be normally distributed but contain several outliers

**SHUFFLE AND SPLIT DATA**

For this section we will take the Boston housing dataset and split the data into training and testing subsets. Typically, the data is also shuffled into a random order when creating the training and testing subsets to remove any bias in the ordering of the dataset.

. **TRAINING AND TESTING**

It is useful to evaluate our model once it is trained. We want to know if it has learned properly from a training split of the data. There can be 3 different situations:

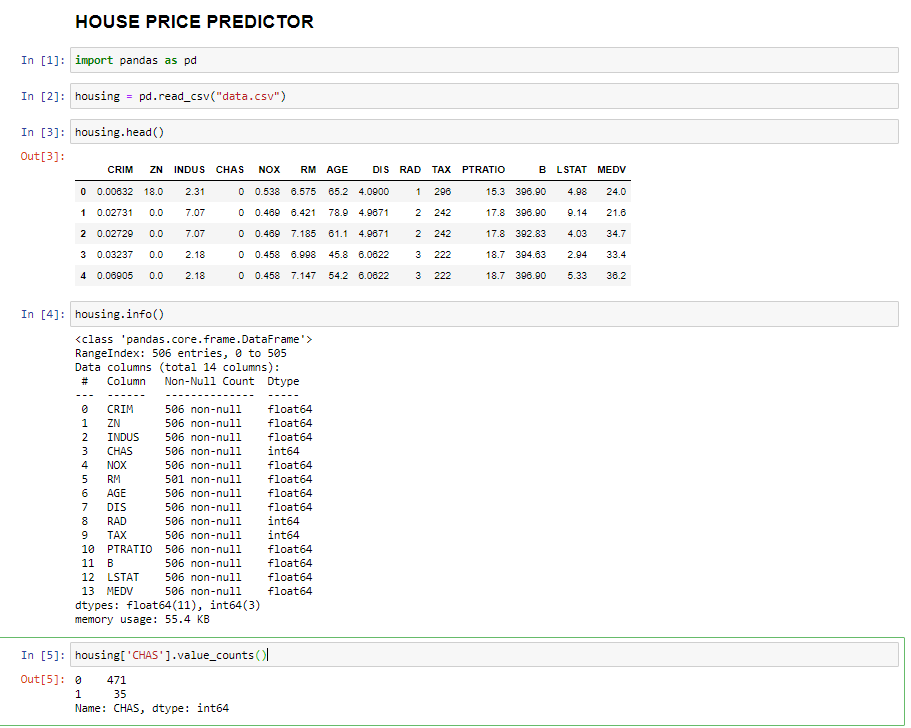
1) The model didn´t learn well on the data, and can’t predict even the outcomes of the training set, this is called underfitting and it is caused because a high bias.

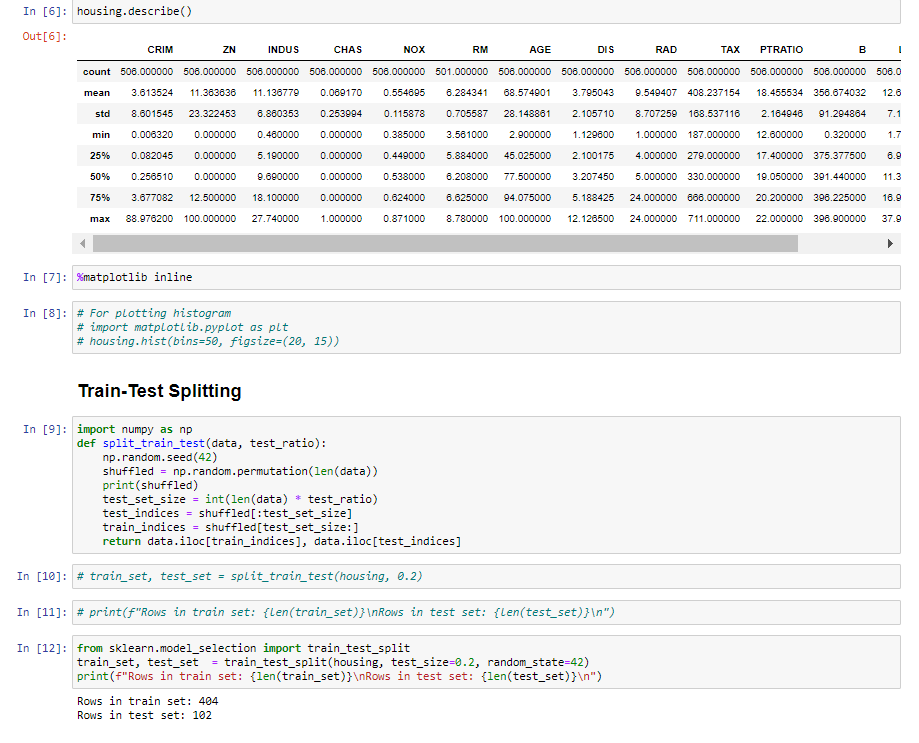
2) The model learn too well the training data, up to the point that it memorized it and is not able to generalize on new data, this is called overfitting, it is caused because high variance.

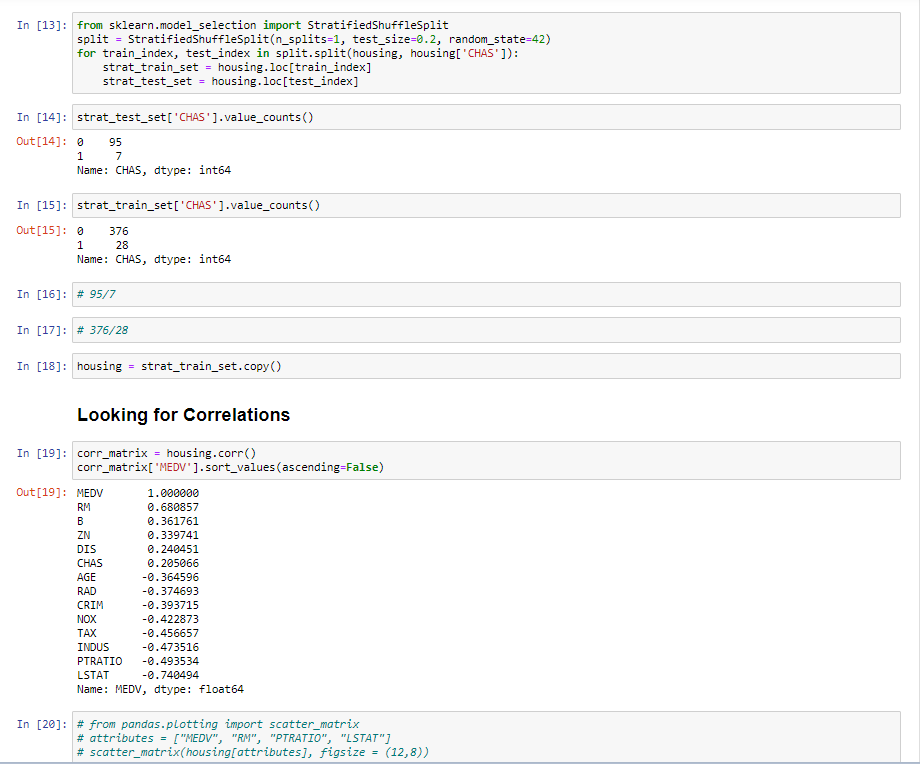
3) The model just had the right balance between bias and variance, it learned well and is able predict correctly the outcomes on new data.

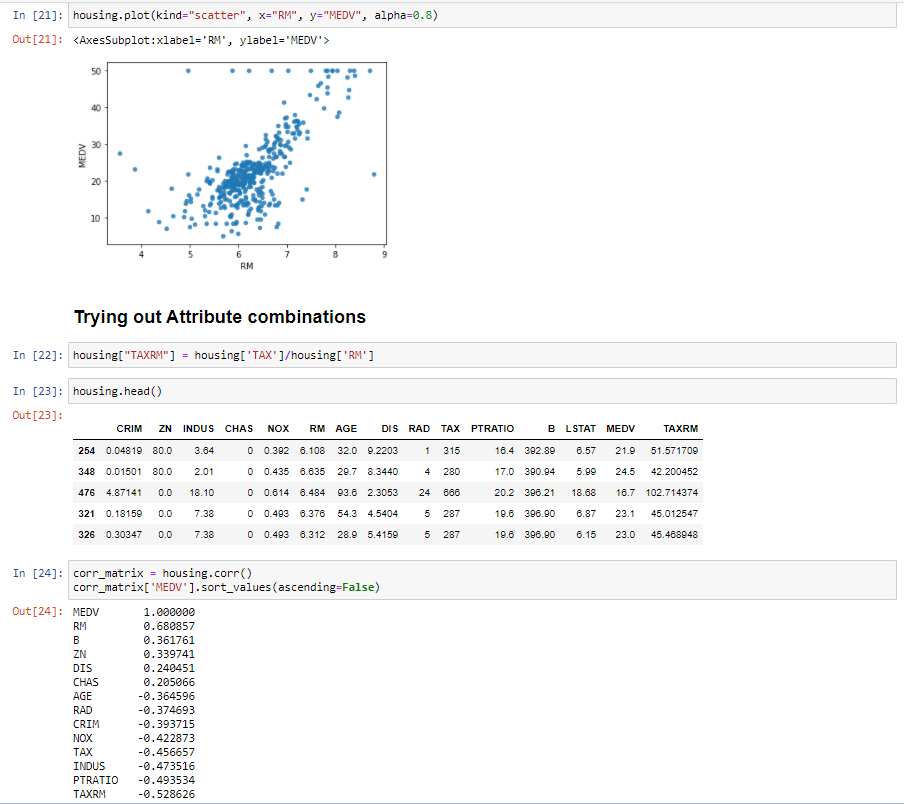
**SOURCE CODE**

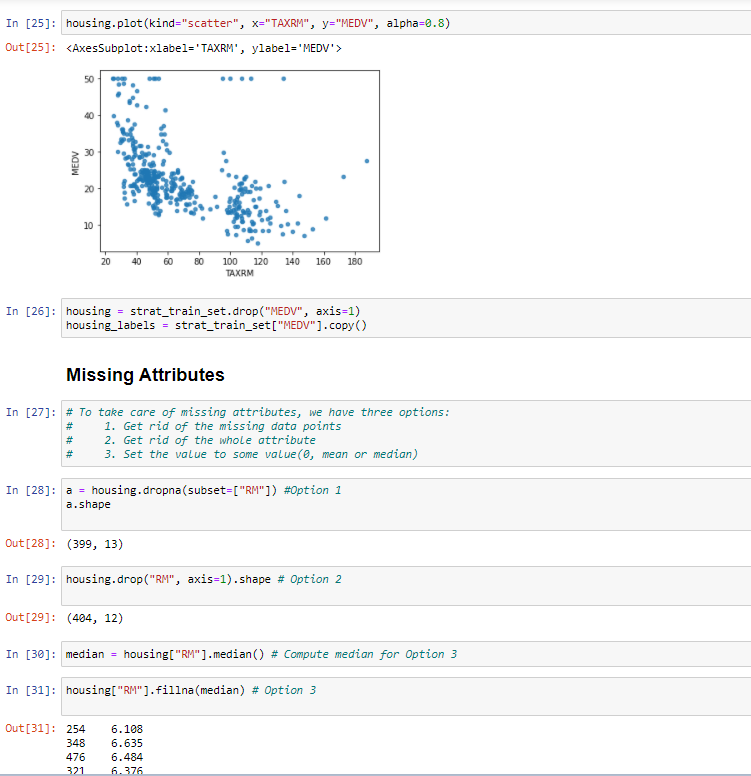
**Github link to code - https://github.com/varunkmr038/MS-PROJECT-**

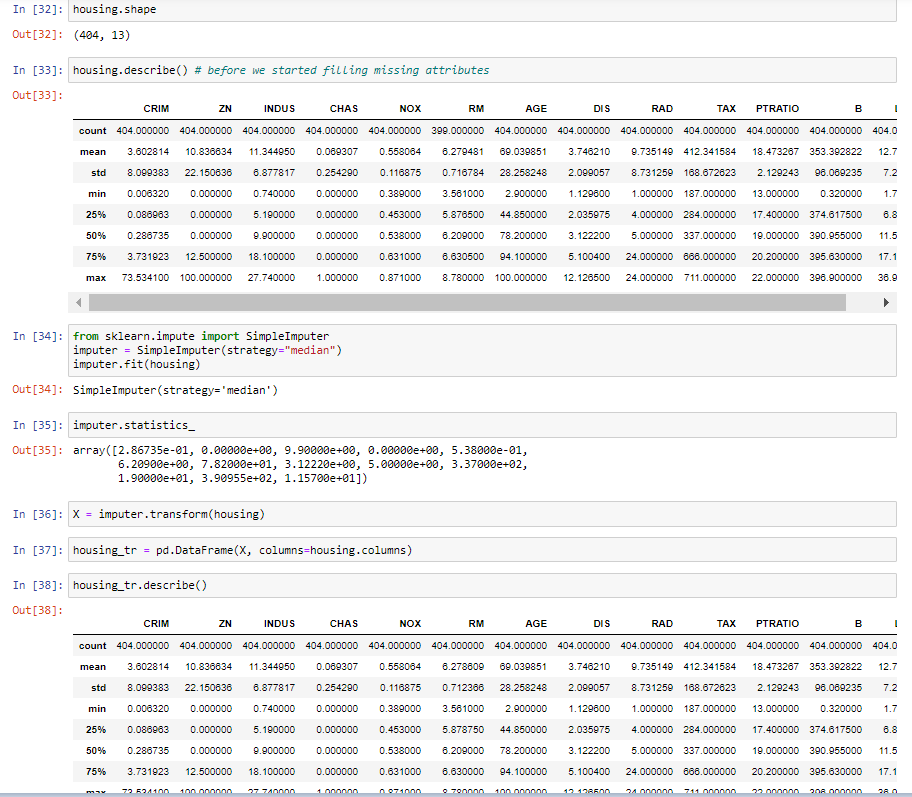


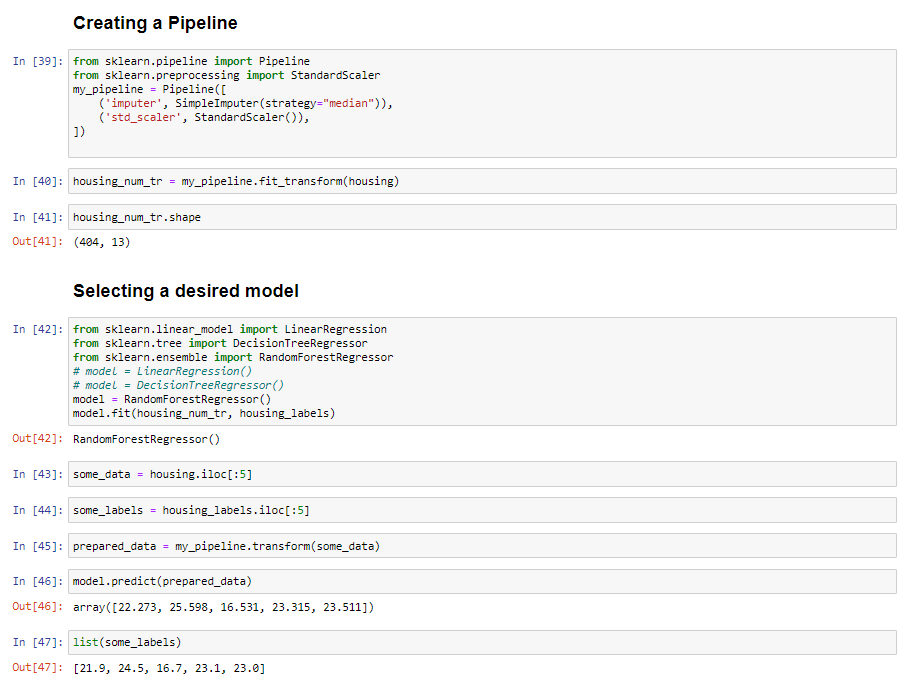


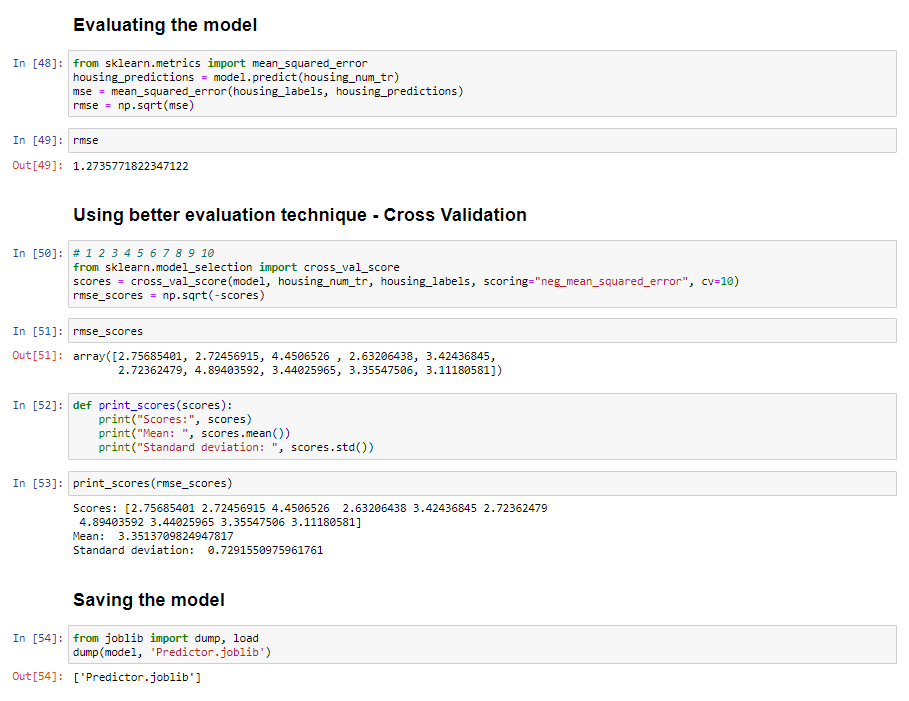
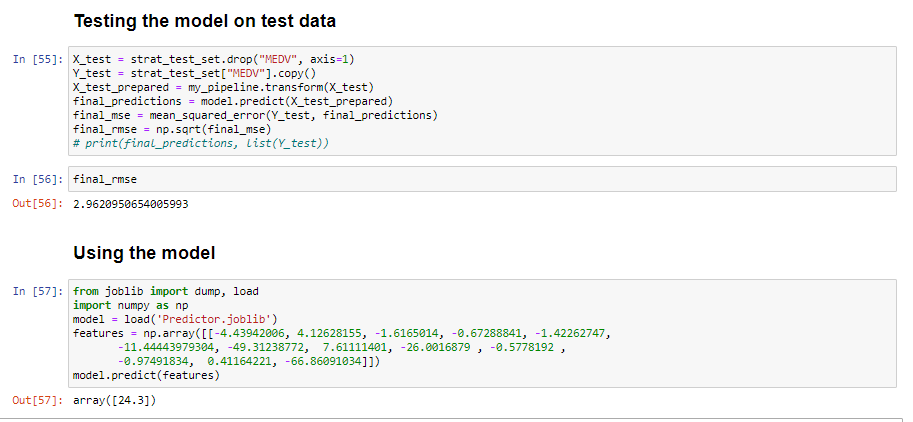












**CONCLUSION**

The main goal of this project is to determine the prediction for prices of houses for which we

tried three different machine learning algorithms namely Linear regression, Decision

tree regression and Random forest regression , And we got the following results : ---

Model Outputs

1. Linear Regression:

Mean (rmse error): 5.037482786117751

Standard deviation: 1.0594382405606948

2. Decision Tree Regression:

Mean (rmse error) : 4.220181728238616

Standard deviation: 0.7451258431327811

3. Random Forest Regression

Mean (rmse error) : 3.3513709824947817

Standard deviation: 0.7291550975961761

**So it’s clear from the outputs that the Random forest have more accuracy in prediction when compared to the others** . Hence , we have used this model to predict better house prices .

**BIBLIOGRAPHY**

* <https://archive.ics.uci.edu/ml/machine-learning-databases/housing/?C=N;O=D>
* <https://www.kaggle.com/shreayan98c/boston-house-price-prediction>
* <https://www.geeksforgeeks.org/random-forest-regression-in-python/>
* <https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/>
* <https://www.knowledgehut.com/blog/data-science/linear-regression-for-machine-learning>