**UCT513-MACHINE LEARNING**

Medical Insurance Cost Prediction Using different regression models

****

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY, PATIALA

**Under the Guidance of:**

Dr. Sushma Jain

Computer Science and Engineering Department

Submitted by: Aadish Mehta 101918024

Course: Computer Science and Business System

**Index**

|  |  |  |
| --- | --- | --- |
| **Serial No.** | **Section** | **Page Number** |
| **1** | **Overview** | **3** |
| **2** | **Dataset Description** | **4** |
| **3** | **Libraries and Data Collection** | **5** |
| **4** | **Data Analysis and Visualization** | **7** |
| **5** | **Data Pre-processing** | **14** |
| **6** | **Regression Models and Evaluation** | **15** |
| **7** | **Result** | **18** |

**Overview**

This project acts as an automated way of predicting medical insurance based on various factors like age, BMI, region, no. of children, sex, etc. This is done with the help of various regression models such as Linear Regression Model, Decision Tree Regression Model, Artificial Neural Networks and others.

The flow of the project follows as such:

DATA COLLECTION

DATA ANALYSIS

DATA PRE-PROCESSING

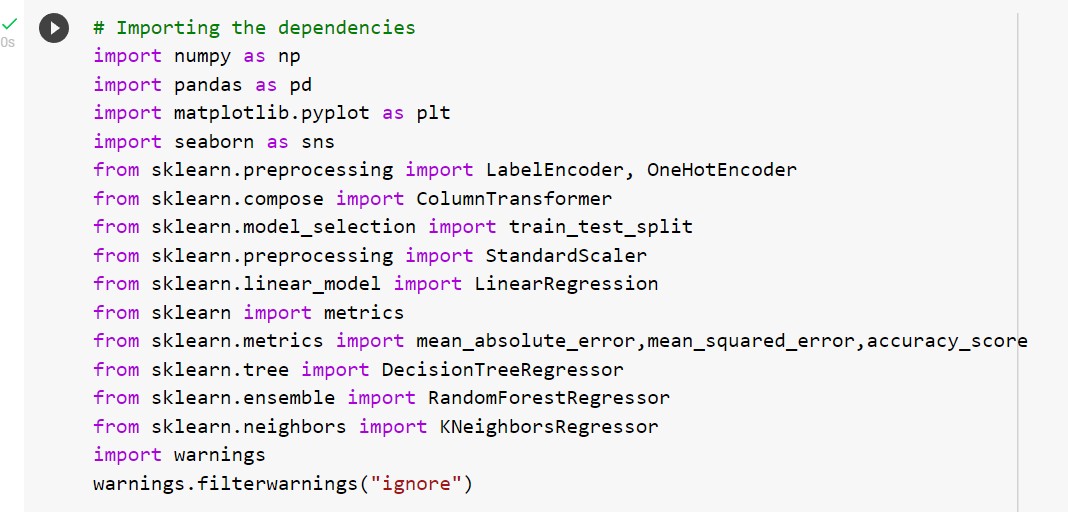
MODEL PREDICTIONS AND EVALUATION

RESULT

# Data Description

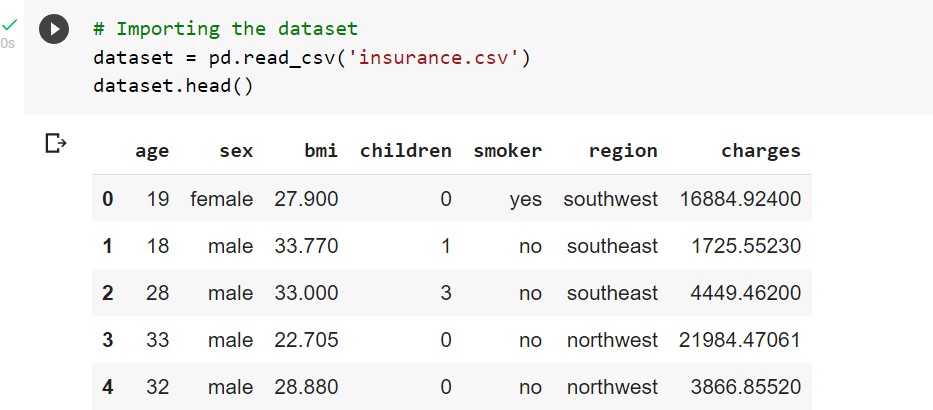
|  |  |
| --- | --- |
| **Dataset** | insurance.csv |
| **Number of attributes** | 6 |
| **Number of targets** | 1 |
| **Type of target** | Numerical |
| **Type of attributes** | Numerical & Categorical |
| **Type of Problem** | Regression |

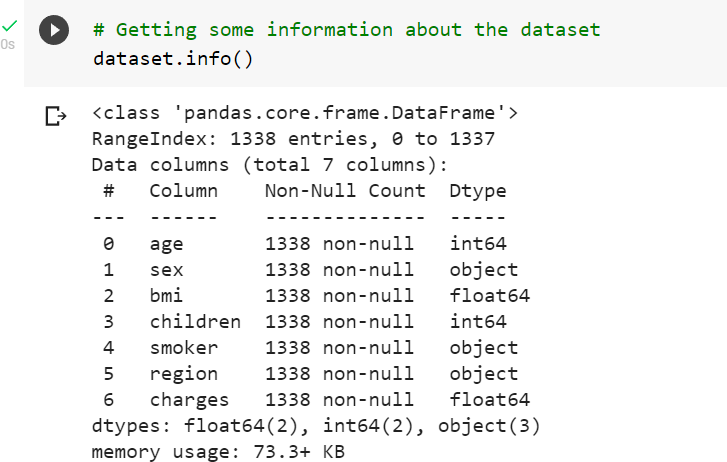
**Code**

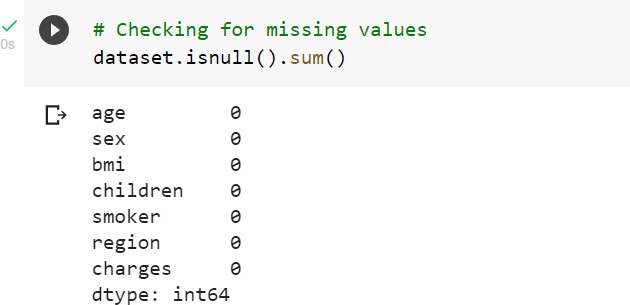


## DATA COLLECTION

The initial view of the data in a pandas data frame was as follows: -

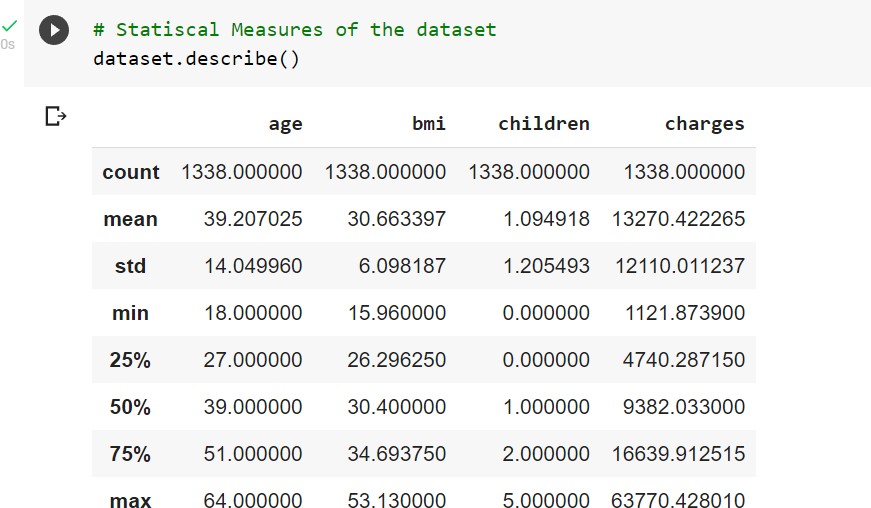


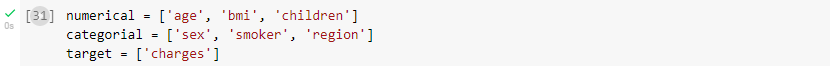


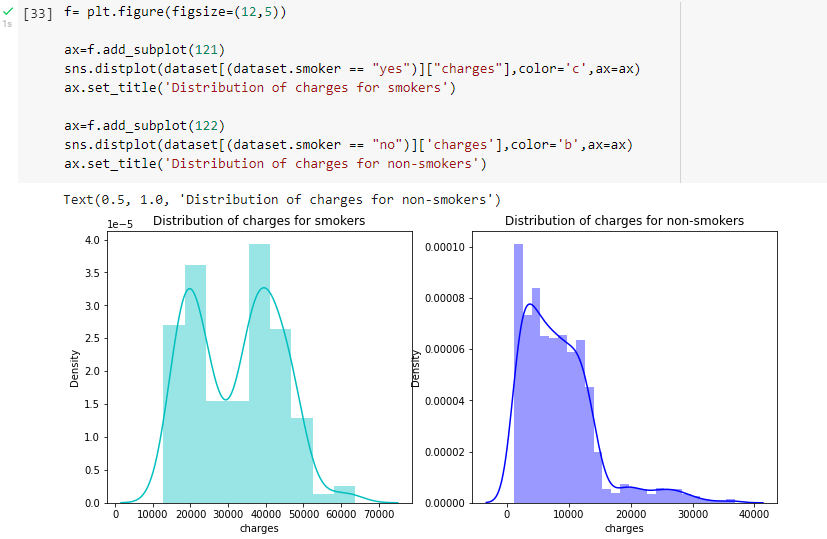


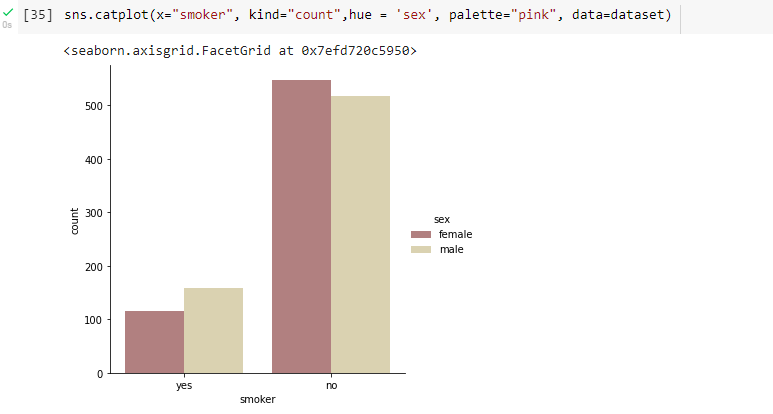
No NULL values were found in this dataset, so no need for any removal of them.

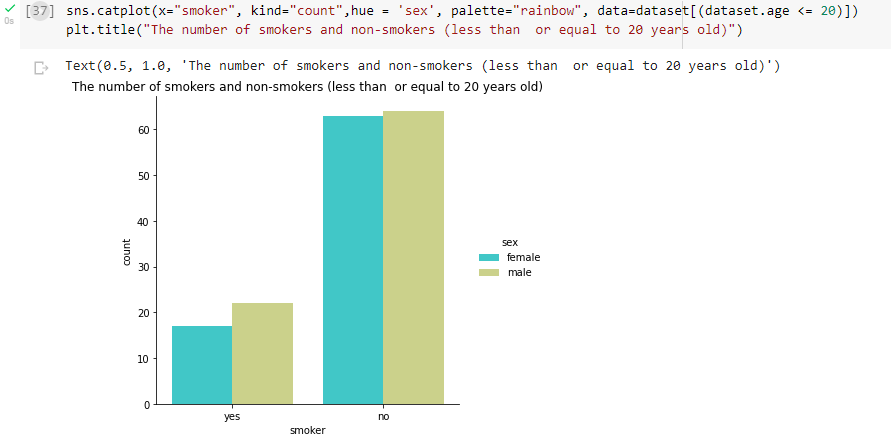
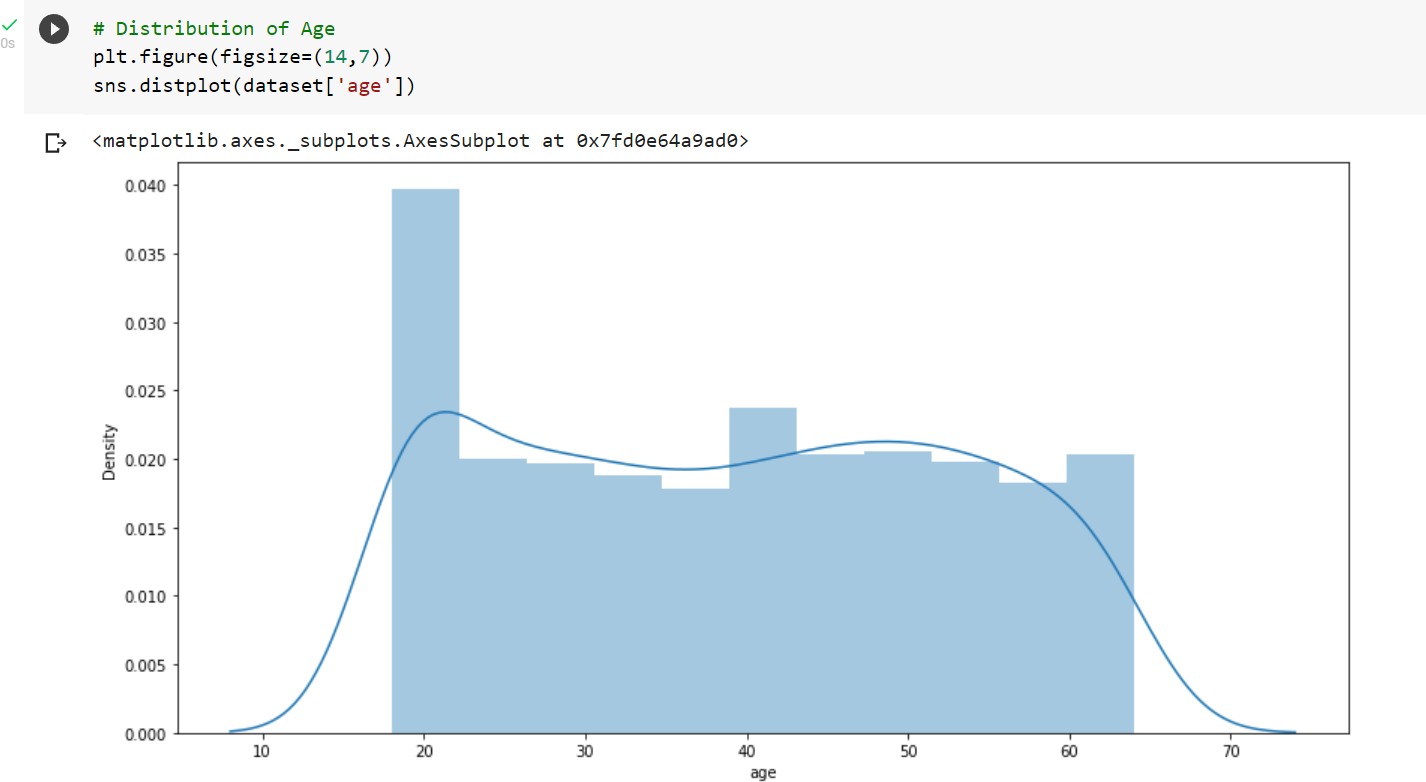
## DATA ANALYSIS AND VISUALIZATION

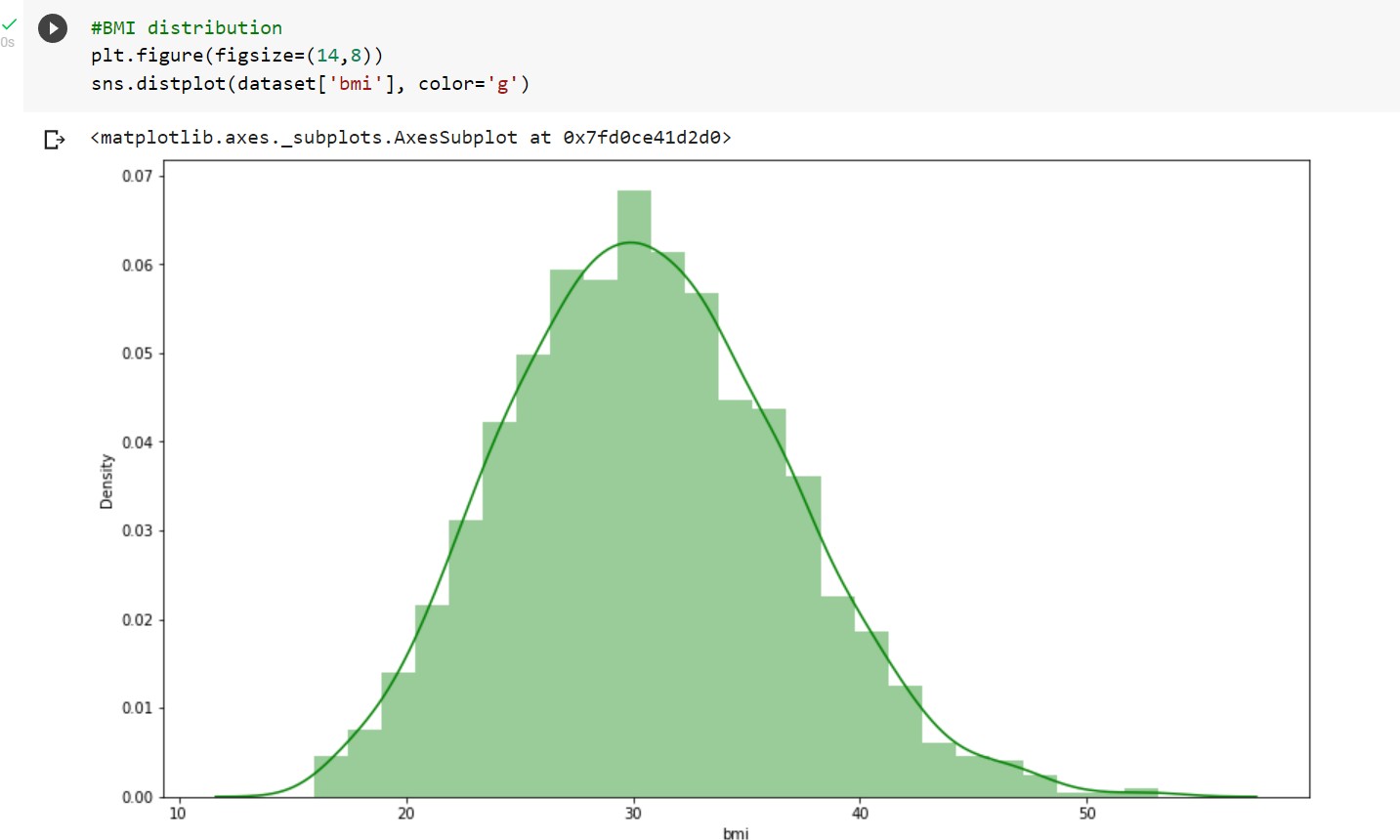
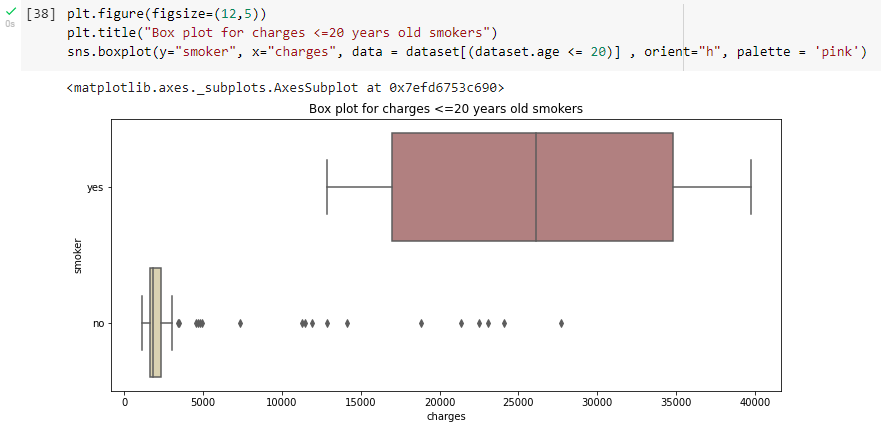


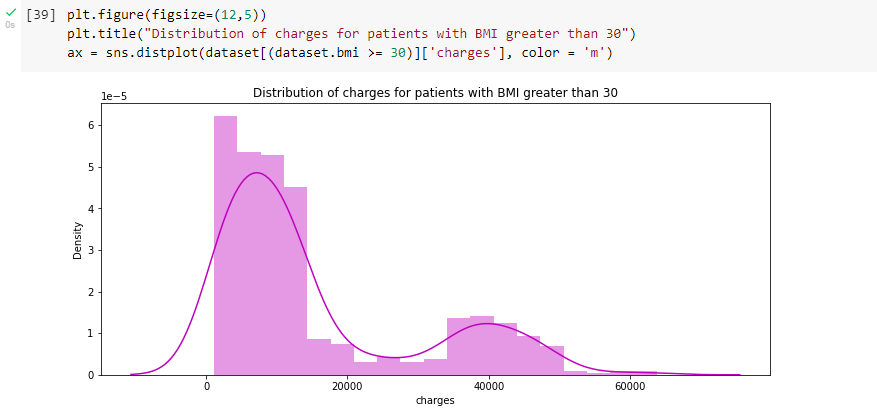
****

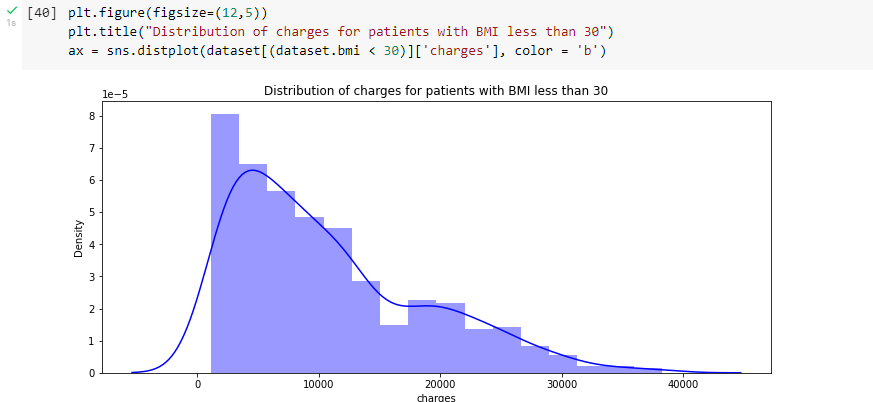
****

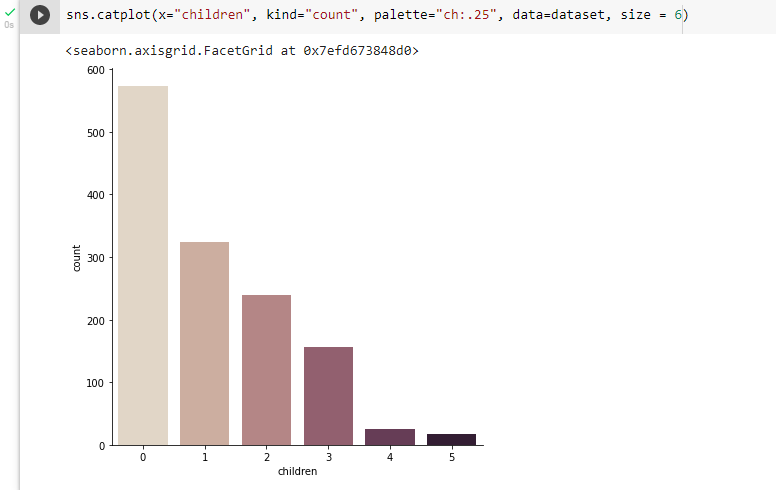
****

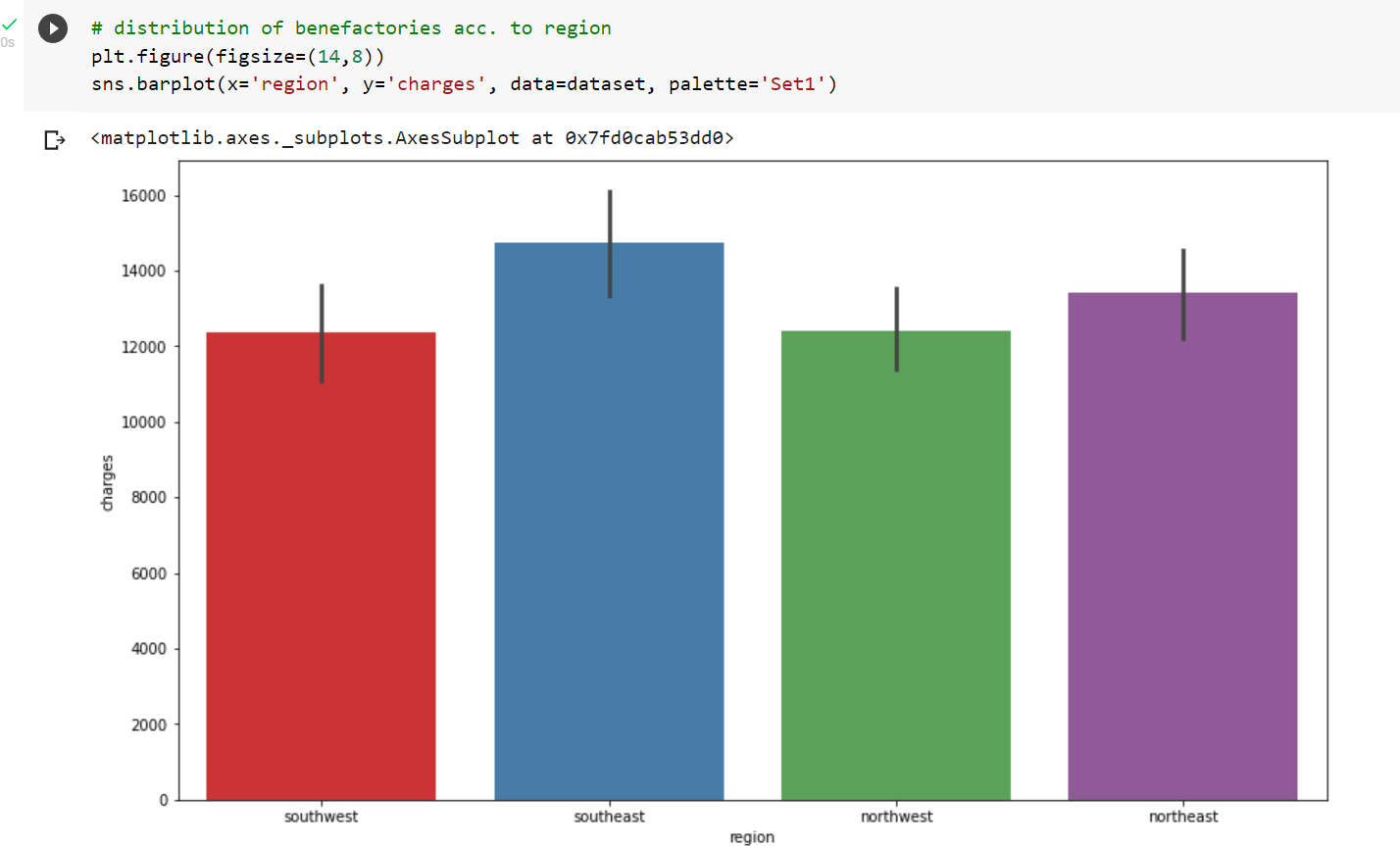
****

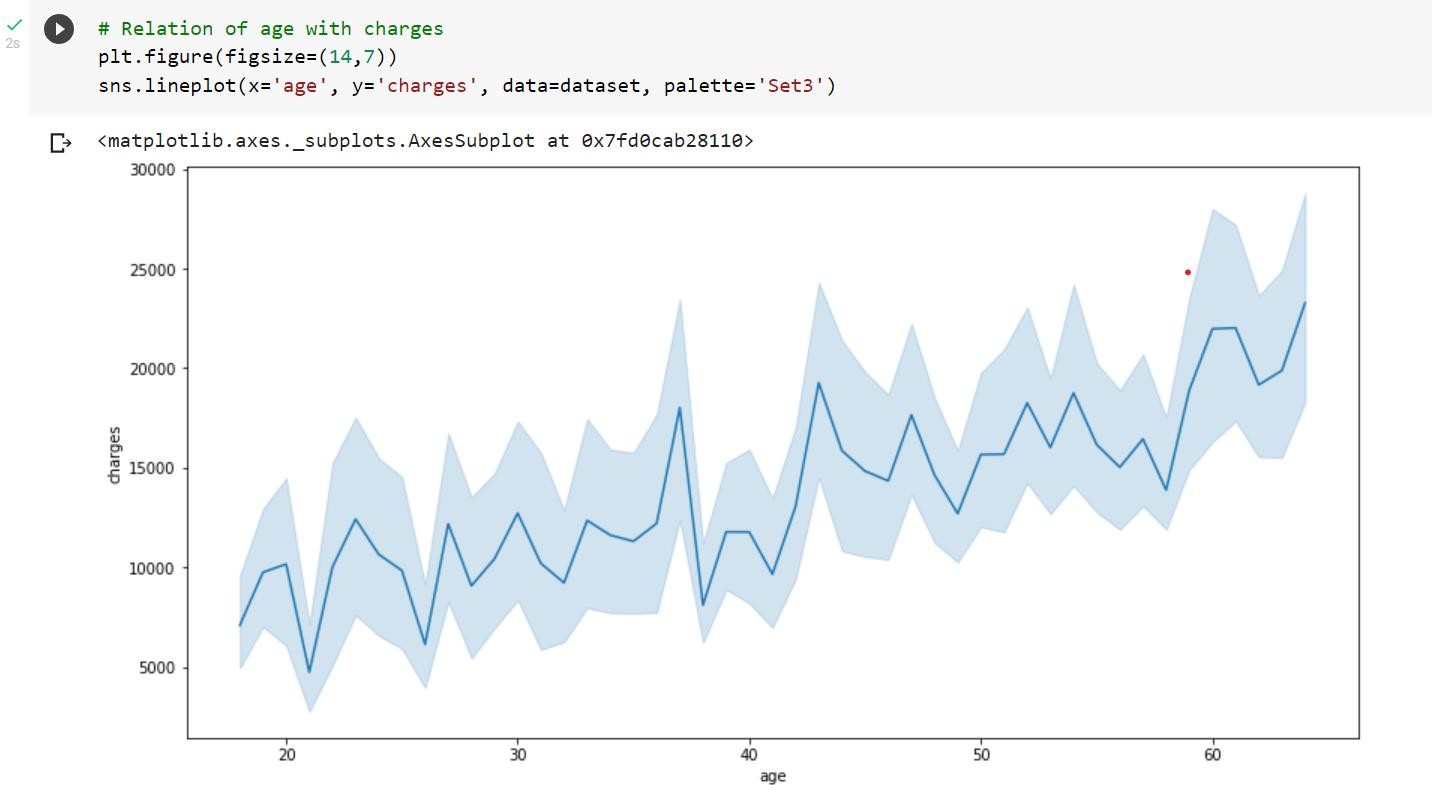


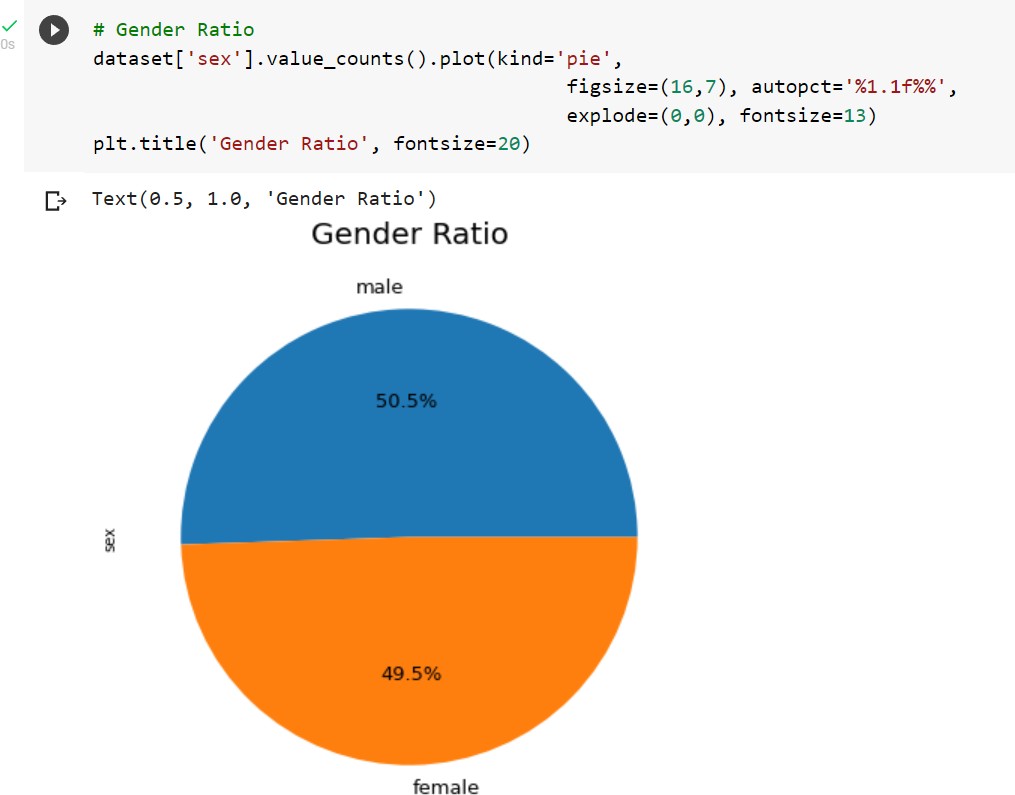






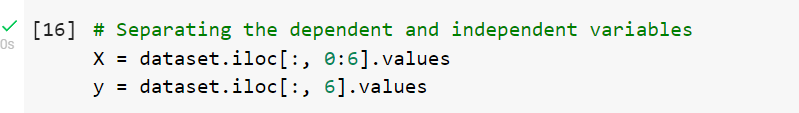




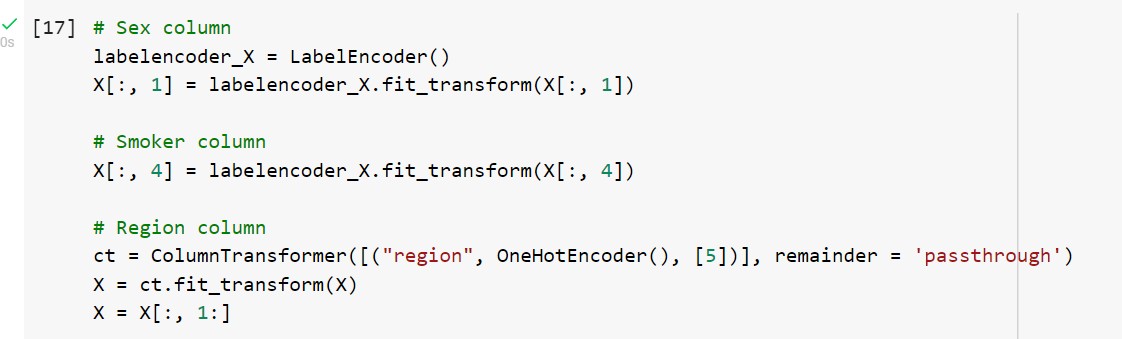


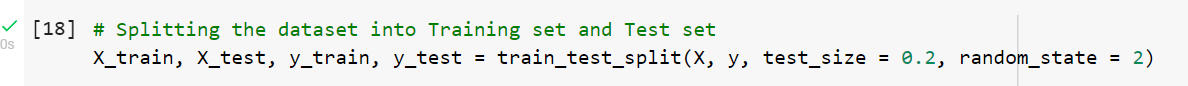


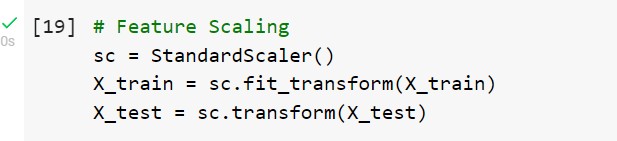
**Data Pre-Processing**



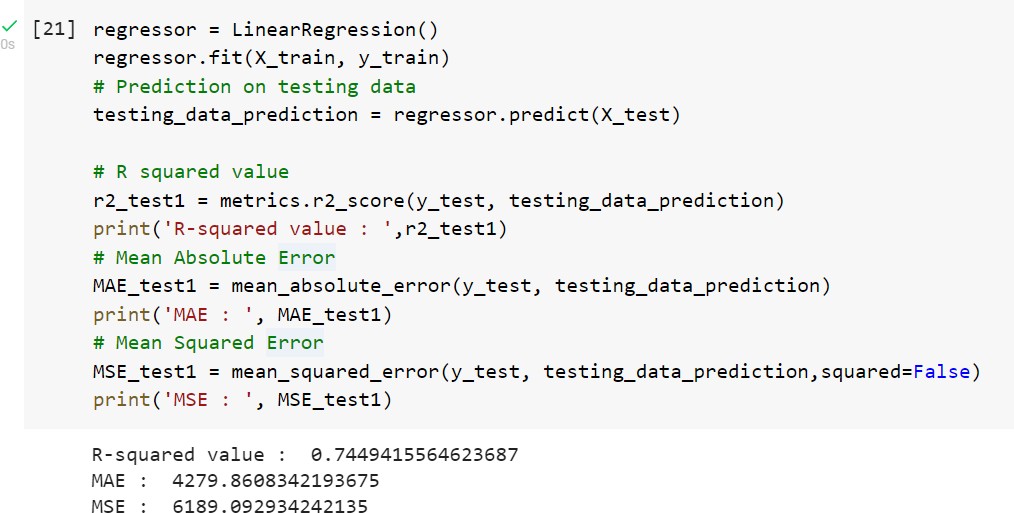
**ENCODING CATEGORICAL FEATURES**







**LINEAR REGRESSION MODEL**

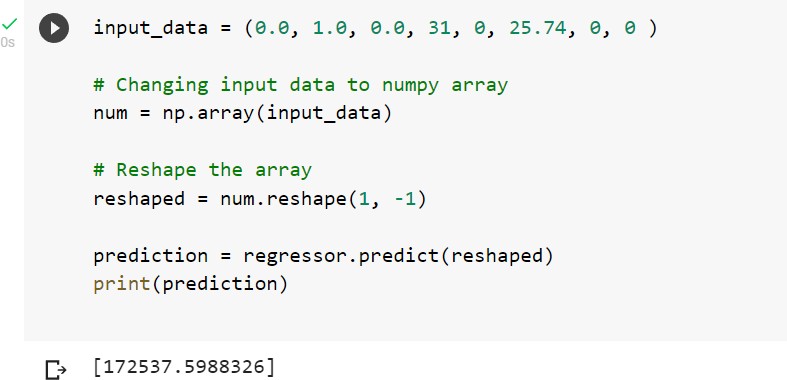


**PREDICTION FOR SINGLE TUPLE**

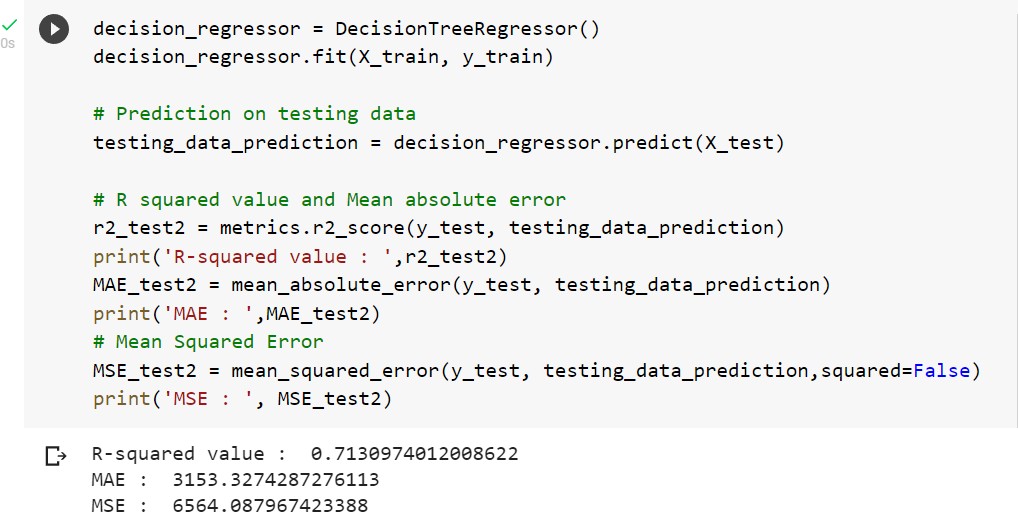
***AGE: 31***

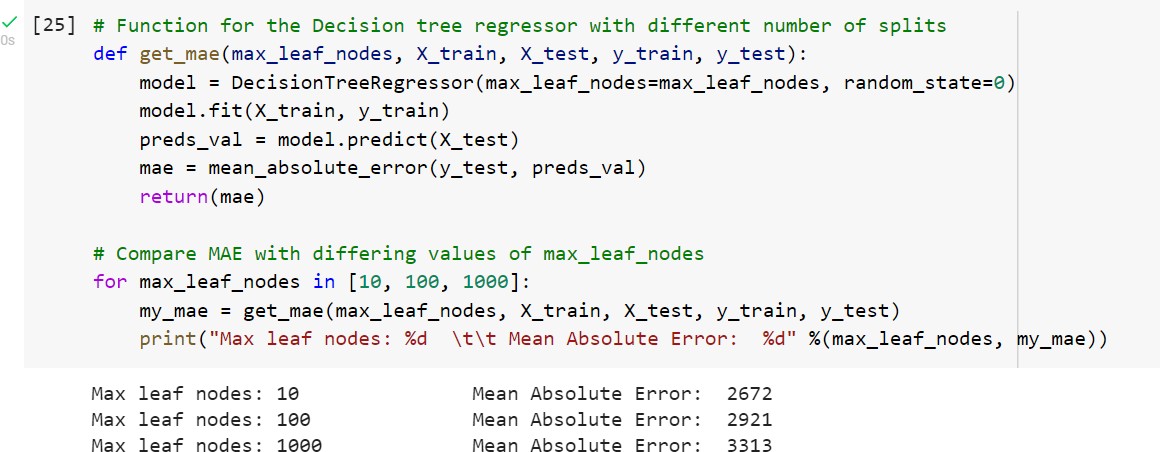
***SEX: FEMALE BMI: 25.74***

***CHILDREN: 0 SMOKER: NO REGION: SOUTHWEST***

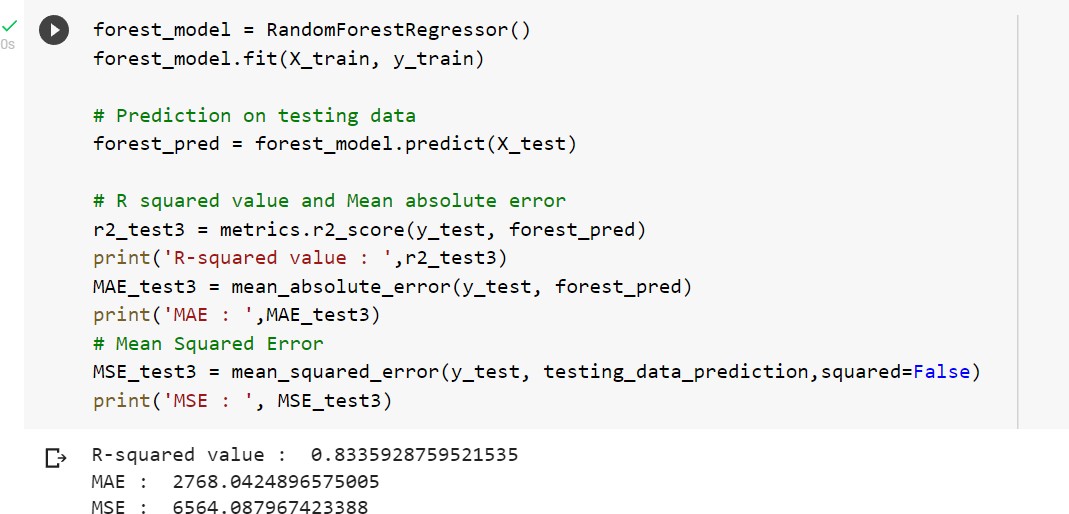


**DECISION TREE REGRESSION MODEL**

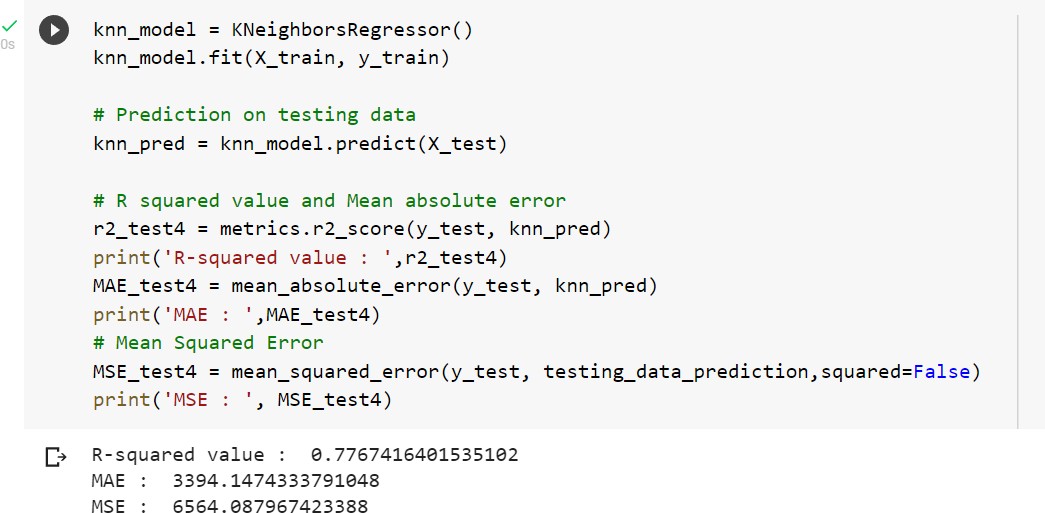




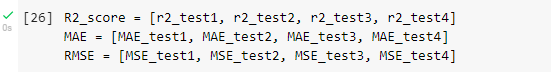
**RANDOM FOREST REGRESSION MODEL**



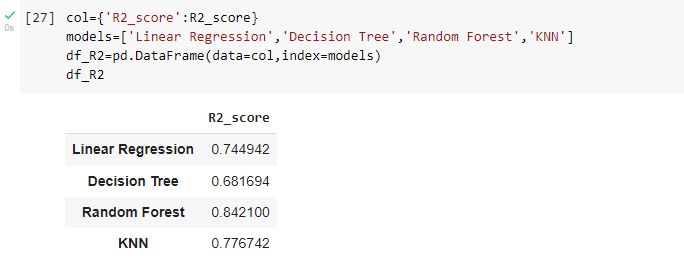
**K NEAREST NEIGHBOR REGRESSION MODEL**

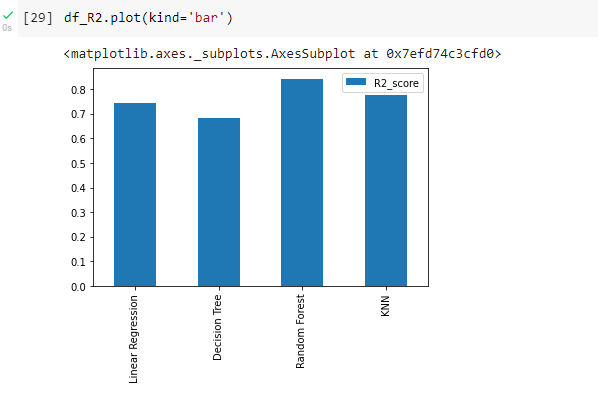


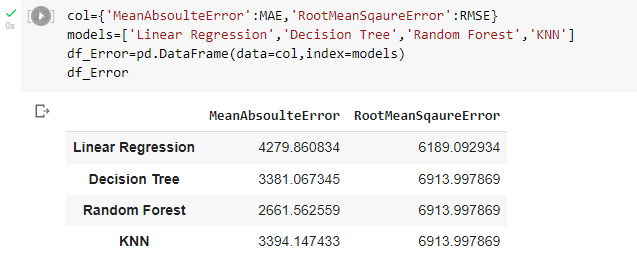
**RESULTS**

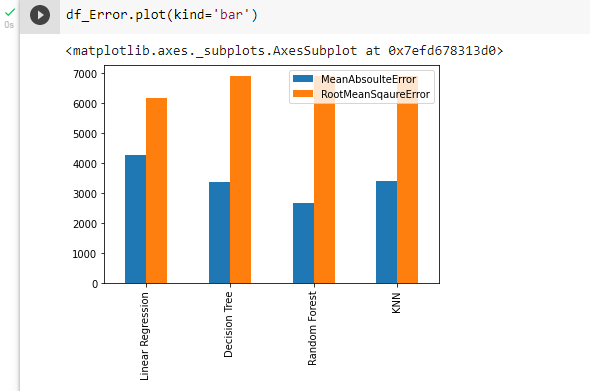


**R-SQUARED VALUE, MEAN ABSOLUTE ERROR, MEAN SQUARED ERROR MATRICS AND GRAPHS**









***From the result graphs it is clearly visible that Random Forest Regression Model is the most robust of all the models with a high R-squared value and low errors.***