**A PROJECT REPORT**

**ON**

**BIKE PRICE PREDICTION USING**

**K-NEAREST NEIGHBOR ALGORITHM**

Submitted in partial fulfillment for the requirement of the award of

TRAINING

IN

**Data Analytics, Machine Learning and AI using Python**



*Submitted By*

**Kartekay Kaushik(Ramanujan College,Delhi University)**

*Under the guidance of*

**Mr. Ashish Saini**

**INTRODUCTION**

In a world were transportation vehicles are ever increasing along with the users of such products.There needs to be a easier way to predict the selling price of various bikes so, that the seller could get a best or reasonable price for his vehicle and also help buyers determine if they are paying the right price for their vehicle.This model explores price prediction for used bikes according to various parameters given in the dataset.This problem will be treated as regression problem since we are predicting a continuous variable i.e. price.

**Problem statement**

To predict the price of a second hand bike using the dataset and available parameters.So,the user could get a reasonable price for his/her vehicle or compare the price of their item if they are buying second-hand to the reasonable price of the vehicle.

**Description of Dataset**

The dataset contains data of 32000 bikes scrapped from droom.in.The dataset was created by using webscraper.io and instant scraper tools by a Kaggle user named Sai Saathvik Domala.It consists of features like power,kilometers driven,Age of bike etc.

**Operations on the Data**

1. **Pre-processing:**

* Dropped column ‘owner’ which respresented wether the owner of the bike was the first owner or second owner and so on because the age of the bike was already given and it provided little to no extra information.
* Dropped column ‘city’ which represented in which city was the bike located,since the brands have a national price listing which doesn’t differ much from city to city.
* Dropped column ‘brand’ since bike name already consists of bike brand.
* Dropped column ‘bike\_name’ since it wasn’t able to get converted to float.

1. **Final Dataset:**

* Final dataset contains 32648 rows,4 columns.
* It will be spilt into train and test segments in the percentages 10% & 90% respectively.
* Another splitting of this dataset would be between X and y,where X is the features considered for estimating the price i.e age,kms\_driven,power columns,y is the price column.
* All values are non-null and of float datatype.

1. **Hyperparameter Tuning:**

* Test segment was decreased from 20% to 10% for increasing accuracy of the model and predicting more accurately as dataset has huge set of values to test even after reduction.
* The value of k in k-nearest neighbor was taken 5 at first but then increased to 7 to increase accuracy.
* Weights attribute of the k-nearest neighbor regressor was set to ‘distance’ instead of ‘uniform’ to increase the accuracy.
* Metric attribute of k-nearest neighbor reggreser was changed from ‘minkowski’ to ‘euclidean’ to increase accuracy.

**Algorithm Used *(K-nearest neighbor regressor)***

1. **Basic Description:**

* It is used because the task is for predicting a continuous variable which lies in the realm of regressor otherwise classifier would have been used.It is an example of supervised learning i.e. where both input and output data is provided for learning.It is non-parametric and instance-based algorithm i.e. it doesn’t make any assumption about the data and directly relies on the training data to make predictions on the given data points.

1. **Execution principles:**

* Data preparation: pre-processing the data i.e adding/deleteing columns,replacing string with float values,handling missing values etc.
* Distance calculation: calculate the distance between the input data point and other data points in the given dataset.
* Neighbour selection: specifying the value of k so that you can calculate the nearest distanced neighbours and use them to predict the data required.
* Regression:computing the predicted output using the algorithm.
* Generating output: create the final dataset and give it as output.

1. **Hyperparameter tuning:**

* Adjusting the value of k to get maximum accuracy.
* Adjusting the metric attribute to increase accuracy.
* Adjusting the weights attribute to increase accuracy.

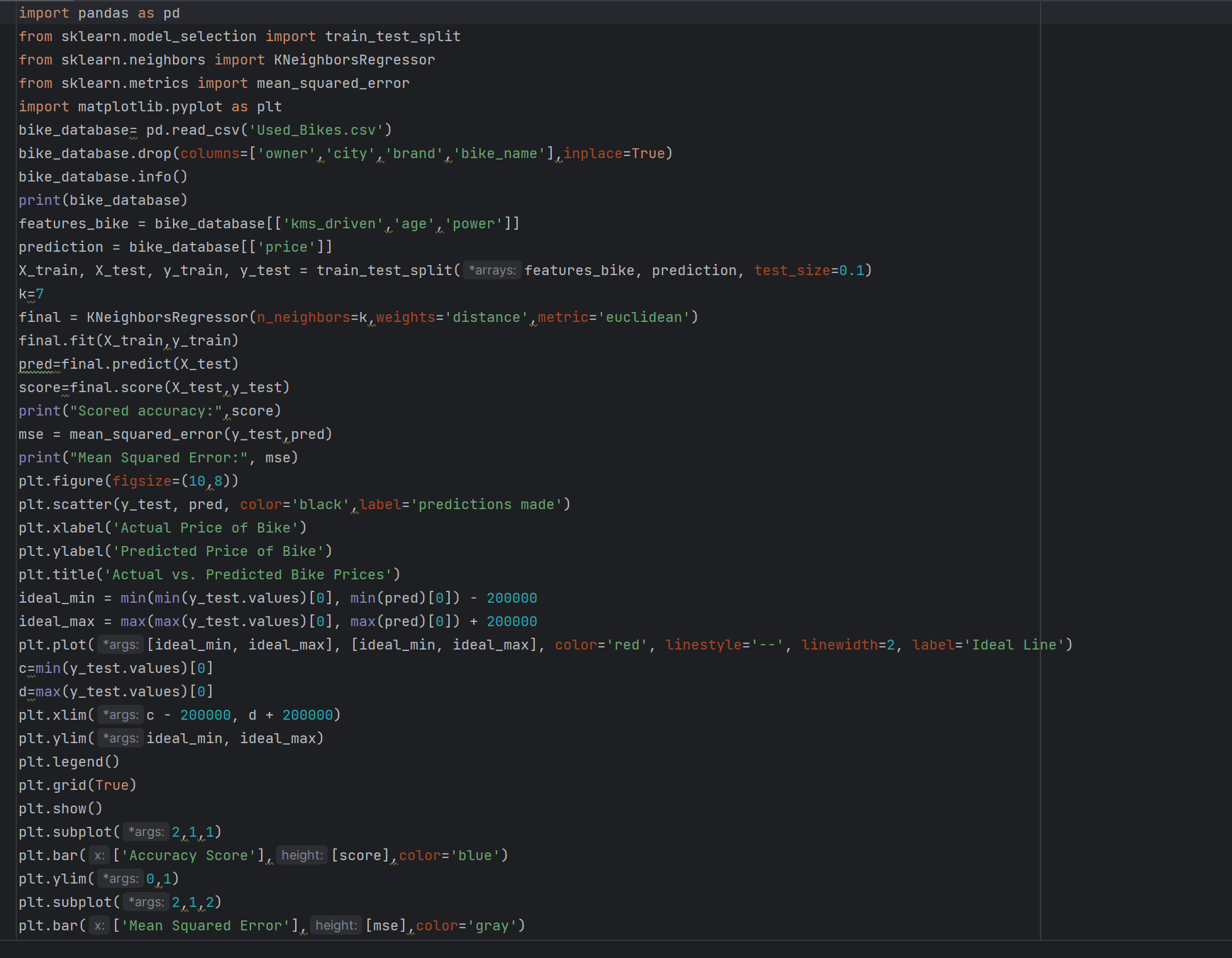
**Industrial Scope for this project**

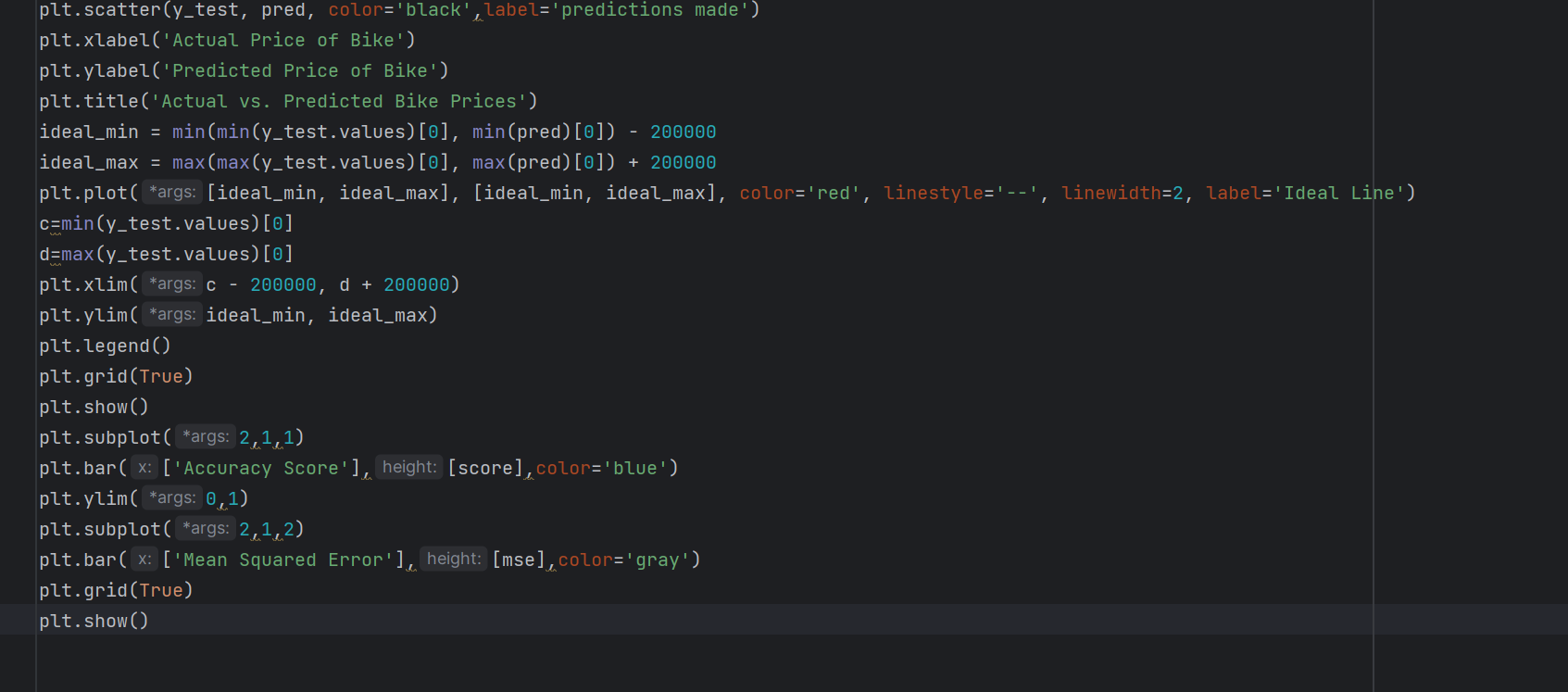
* Demand Forecasting : sellers can sell vehicles according to demand to avoid selling at lower rates.
* Pricing strategy : sellers could create competitive prices based on market trends which would create more demand hence,more profit for their vehicles.
* Consumer Empowerment : customers can make more logical decisions based on this algorithm hence increasing experience and savings.
* Market insights : analysts could determine what factors affecting the market and price trends.
* Market transparency : adds more transparency to market and connects sellers to buyers directly,promotes healthy competition inside the used bike market.

**Error Correction Procedure**

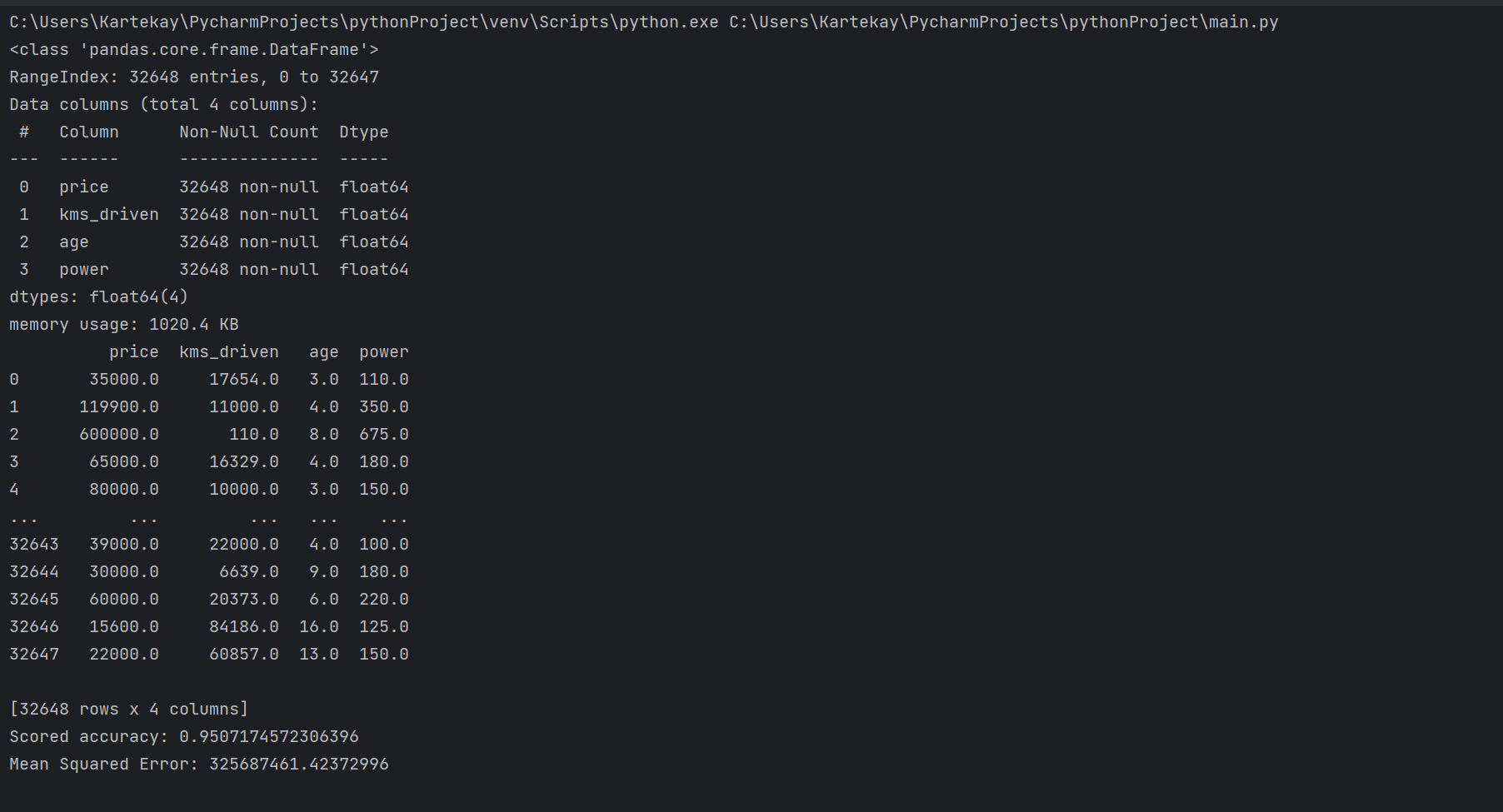
* Trial and error method: each runtime error was shown it was corrected by to tweaking the code and re running it.

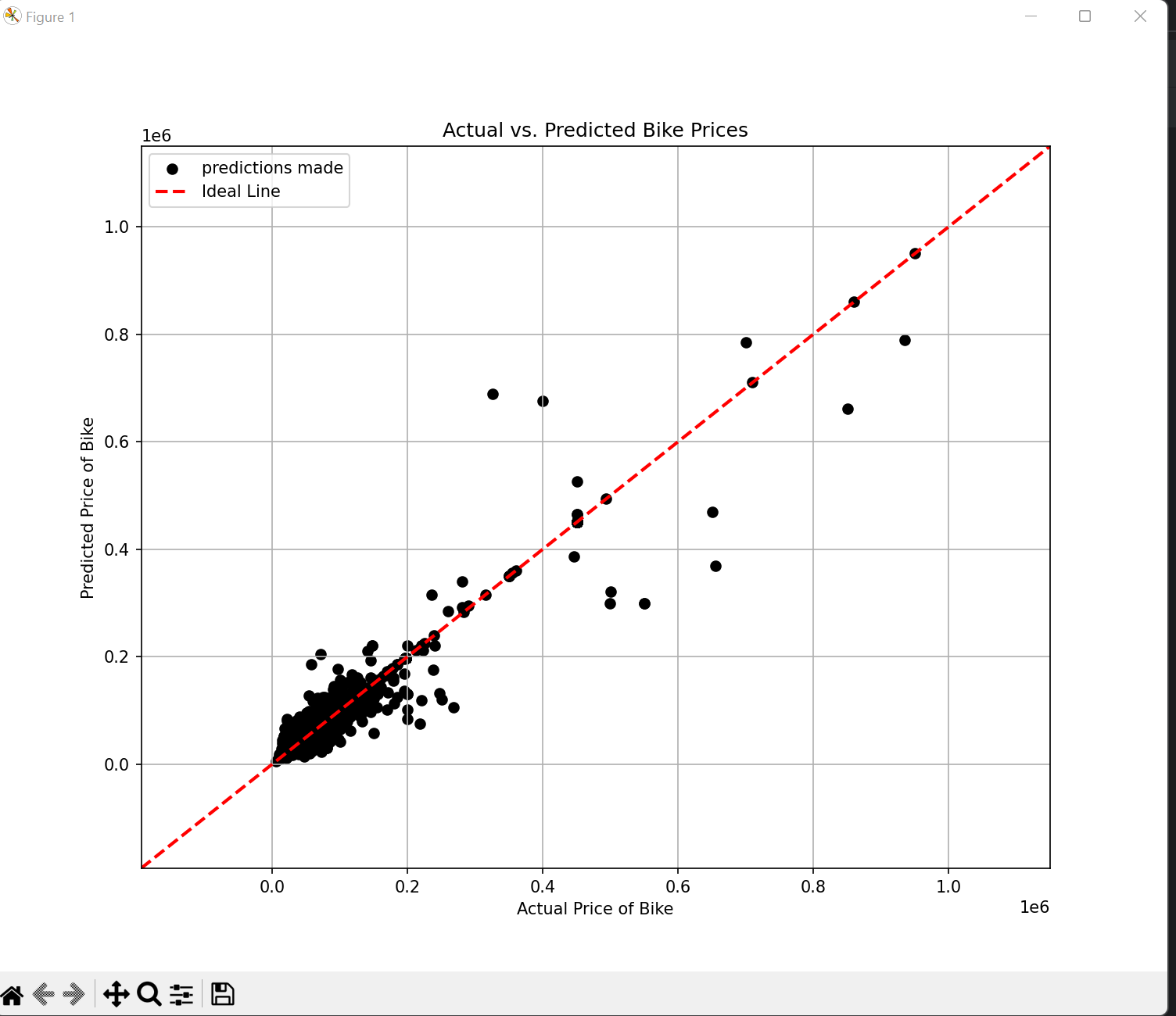
**Complete Code**

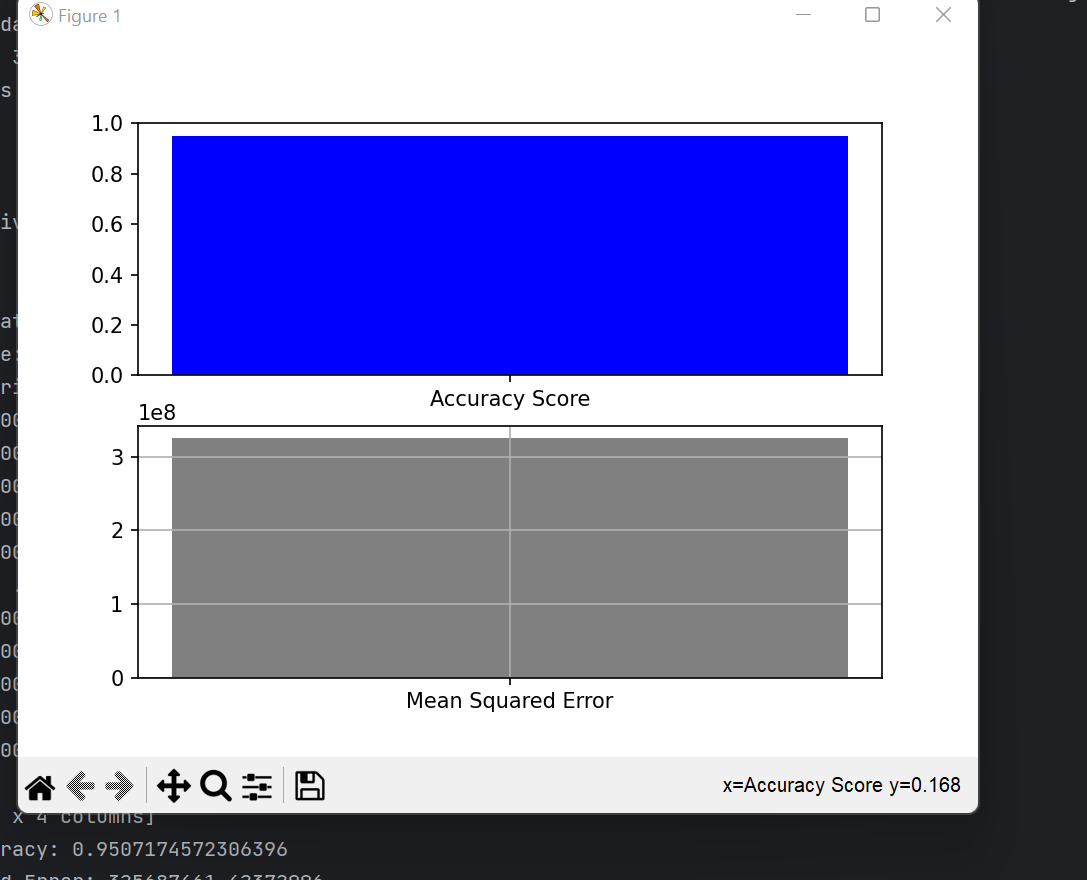




**Output Screenshots**

****

****

****

**Conclusion**

The project for predicting the price of used bikes using parameters such as age,kilometres driven and power using k nearest neighbour regression algorithm has 95% accuracy approximately.The project concludes although not perfect but highly accurate predictions could be made by setting k nearest neighbour to 7,distance attribute to Euclidean and weights attribute to distance would bring out the most accurate values and increase the prediction accuracy from 78% to 93%,we also noticed that the larger the train segment from the original file the larger will be the accuracy,after setting the test size to 10% and train size to 90% of the data, the accuracy increased even further to 95%.The mean squared error value was minimized as much as possible in the following procedure which indicates higher accuracy.

**Bibliography**

* <https://www.kaggle.com>
* <https://www.google.com>
* <https://www.youtube.com>