K.Narayanapura, Kothanur, Bengaluru, Karnataka 560077

**AFFILIATED TO BENGALURU NORTH UNIVERSITY**

**DEPARTMENT OF COMPUTER SCIENCE (PG)**

**MCA**

**“MACHINE LEARNING”**

**RESEARCH PAPER PRESENTATION**

**“CRUDE OIL PRICE PREDECTION USING MACHINE LEARNING”**

**SUBMITTED BY**

|  |  |
| --- | --- |
| **NAME** | **REGISTER NUMBER** |
| **AREF BEGH** | **22MCAB30** |
| **HEMASAGAR T G** | **22MCAB36** |

|  |
| --- |
| **UNDER THE GUIDENCE OF:** |
| Dr. Ranjitha,  Department of Computer Science (PG),  Kristu Jayanti College, Autonomous,  Bengaluru. | |

|  |
| --- |
| Dr. Kumar R,  Head of Department,  Department of Computer Science (PG),  Kristu Jayanti College, Autonomous,  Bengaluru. |

2022-2023

**1. ABSTRACT**

The most popular fuel in the world is crude oil. Depending on the nature of the datapoints, certain machine learning models successfully fit the dataset. Finding models that effectively fit the datapoints and anticipate the price of fuel using machine learning is the major goal of this research. This study seeks to assess the effectiveness of several supervised learning models by comparing them. 5 supervised learning models were employed.

We tested these algorithms, which are often adaptable to many situations, to determine which offers the best results in terms of accuracy and performance: SVR (linear, RBF, polynomial), Random Forest Regression, and Linear Regression. As a result of factors like global inflation, the price of oil has been rising dramatically recently. In India, where the cost of LPG (Liquid Petroleum Gas), petrol and diesel has been rising, this has grown to be a significant issue. These are therefore made from or extracted from crude oil, which India obtains from nearby nations like Dubai and Saudi Arabia. We decided to employ machine learning algorithms to forecast the prices of petrol and diesel in the near future, and after considering a number of techniques, we settled on the linear regression model since it produced the most accurate findings.

*Keywords: Prediction, Oil Prices, Machine Learning Models.*

**2. INTRODUCTION**

The most popular fuel in the world is crude oil. The key benefits of crude oil are its high density and accessibility. Almost all industry depend on oil. A reliable source of power is oil. When compared to other types of energy like solar and wind energy, oil energy is highly dependable.Depending on the sort of data points offered, certain machine learning algorithms match the dataset effectively. The major goal of this research is to use machine learning models to anticipate the price of fuel by identifying several models that effectively fit the data points. This project compares several supervised learning models and makes a judgement call based on effectiveness. To determine which provides the best results in terms of accuracy and performance, we used 3 supervised learning models, namely Random Forest Regression, Linear Regression, and Decision Tree Regression. These algorithms output a numerical value. We may thus contrast the results of these models with the original models. Due to several factors, such as global inflation, the price of oil has been rising rapidly recently. As a result, they are obtained from petroleum. Crude oil for India is imported from nearby nations like Saudi Arabia and Dubai. We have chosen to employ machine learning techniques and ensemble learning to forecast the prices of petroleum products like petrol and diesel in the future. In the approach of ensemble learning, many algorithms or a single algorithm are used repeatedly. On this basis, we may evaluate many algorithms and choose the most suitable one for our given task.

**3. METHODOLOGY**

By automatically estimating their latent connection from data, regression analysis is a machine learning technique that seeks to effectively predict the value of continuous output variables from a set of independent input variables.

**Algorithms:**

The algorithms used in this project are as following:

**3.1 Linear Regression:**

Multiple linear regression model will be expressed as followed:

y = a0 + a1x + a2x2+....+e

y is the dependent variable and x is the independent variable, a0 is the constant term, is the intercept of the regression line on the vertical axis and a1 is the regression coefficient that is the slope of the regression line. e is the random error which will be used to express the effect of random factors on dependent variable. Step wise algorithm is as follows:

***STEP 1: IMPORTING LIBRARIES AND LOADING THE DATA.***

Import any libraries that may be necessary to construct our model. We first imported Matplotlib, NumPy, and Pandas. The next step after importing the libraries is to retrieve the dataset and load our data. The data should be in (.csv/.xls) format.

***STEP 2: VISUALISING THE DATA***

It's crucial to visualise the data in order to identify any relationships between the various factors. The great package Matplotlib may be used to plot our data in a variety of ways.

***STEP 3: FEATURE ENGINEERING***

We discovered that there is a significant link between the two criteria, date and price, when we visualised our data. Thus, these parameters will be used to develop our model.

***STEP 4: FITTING THE LINEAR REGRESSION MODEL***

Then import the sklearn library's train\_test\_split function. Our data are divided into training and testing data using this method. The training dataset is typically made up of 70–80 percent of the data, while the remaining data makes up the testing dataset. Following that, we can determine our model's intercept and coefficient.

**4. IMPLEMENTATION RESULTS**

The whole project is based on python, machine learning.

***4.1 Libraries used:*** sklearn. numpy, pandas, matplotlib.

**sklearn:**

The most effective and reliable Python machine learning package is called Skearn (Skit-Learn). Through a Python consistency interface, it offers a variety of effective methods for statistical modelling and machine learning, including classification, regression, clustering, and dimensionality reduction.

**numpy:**

Two 64-bit floats representing a complex number with real and fictitious components Each instance of a dtype (data-type) object, including the numerical types in NumPy, has certain properties. The supported dtypes are np. bool\_, np. float32, etc.

**pandas:**

Working with "relational" or "labelled" data may be simple and straightforward thanks to the Python module pandas, which offers quick, adaptable, and expressive data structures. It seeks to serve as the essential, high-level building block for using Python for actual, useful data analysis.

**matplotlib:**

For Python and its numerical extension NumPy, Matplotlib is a cross-platform data visualisation and graphical charting package (histograms, scatter plots, bar charts, etc.). As a result, it presents a strong open-source substitute for MATLAB.

***4.2 Data Collection:***

We gathered information on crude oil and its pricing from repositories on Kaggle, Google, and Git Hub.

***4.3 Data Preprocessing:***

The model cannot be trained using the raw data directly. As a result, we preprocess the raw data. The libraries that are often utilised in our project are first imported. Here, we've utilised NumPy, which mostly focuses on operations on arrays, as well as pandas to interact with data and graphs.

We import our data collection after loading the libraries. The dataset could contain some missing data. Missing data may cause our results to differ. We utilise the Simple Inputter class to substitute missing data with mean in order to prevent this. We use mean, median, constant number, etc. to replace missing values based on the data set and dependent variables. We disregard the rows with missing values if the data set is too large and there are just 1% of the required data. Now, we use the One Hot Encoder class to encrypt the category data. By using this technique, the category variable is converted into a collection of binary variables, also referred to as dummy variables. To display N labels, it used N-1 characteristics. This enhances the machine's ability to comprehend the data. The data was then divided into a training set and a testing set. On the training set, we apply machine learning techniques. We employed basic linear regression, decision tree regression, and random forest regression in our study. We train the models using the train set as the input.

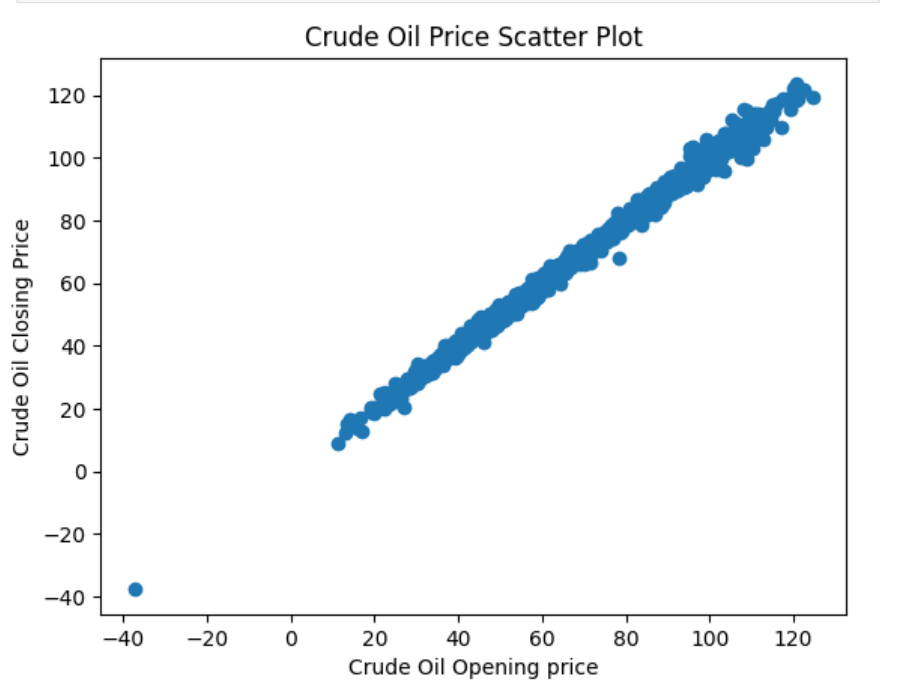


Fig 4.1 CRUDE OIL PRICE SCATTER PLOT ON OPEN AND CLOSING PRICE.

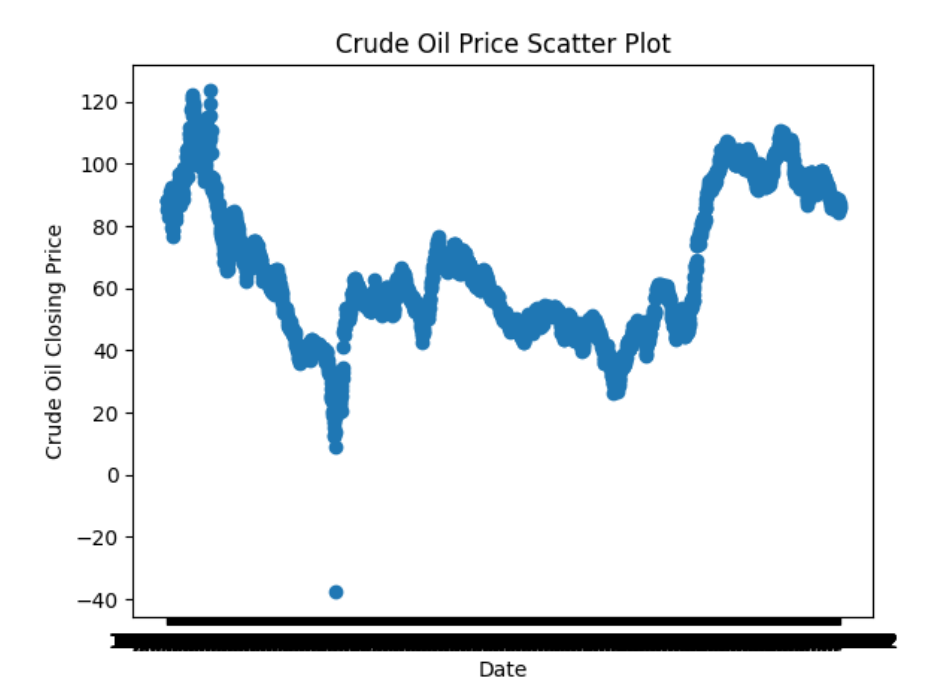


Fig 4.2 CRUDE OIL PRICE SCATTER PLOT ON ACTUAL CLOSING PRICE.

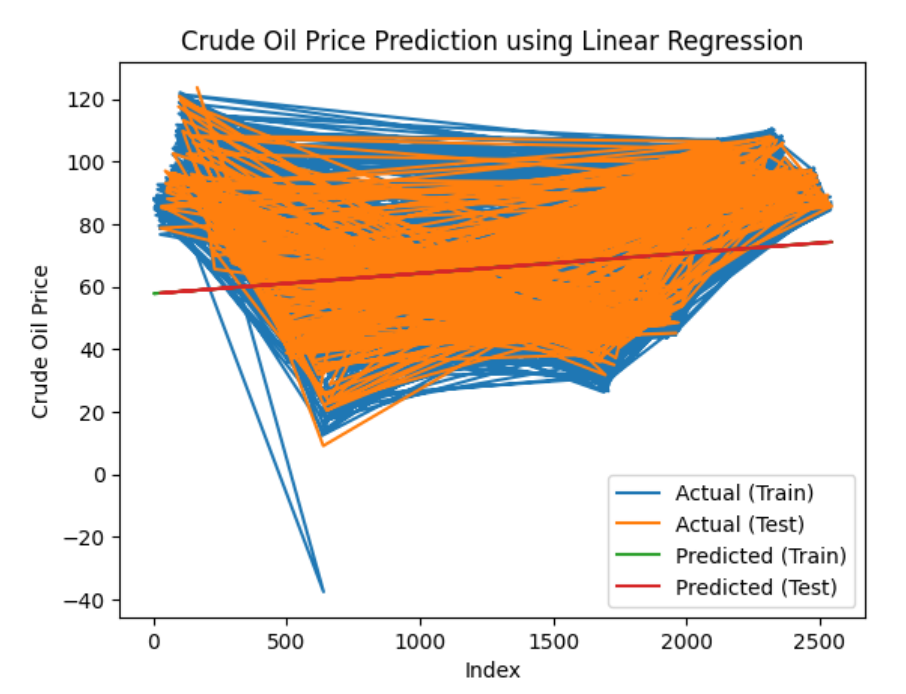
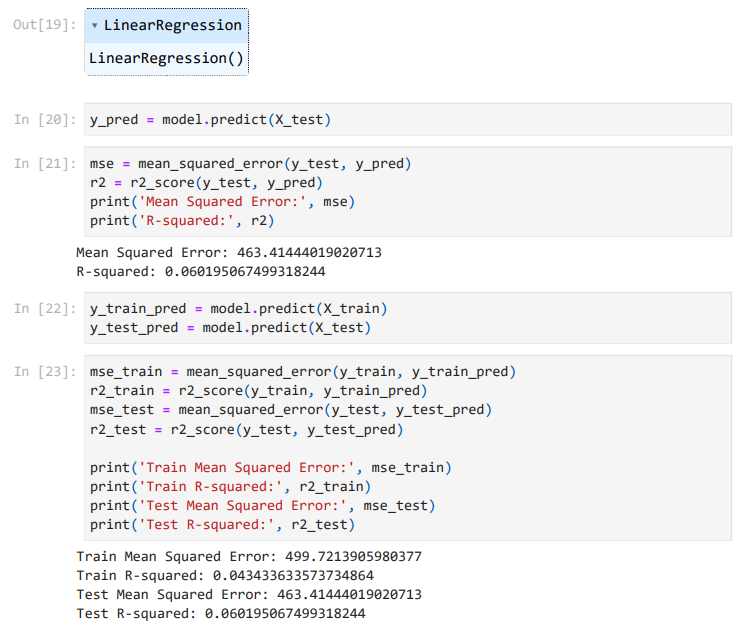


Fig 4.3 PREDECTION GRAPH USIGN LINEAR REGRESSION.

Fig 4.4 OUTPUT OF THE MODEL.

***4.4 Data output:***

Train Mean Squared Error: 499.7213905980377

Train R-squared: 0.043433633573734864

Test Mean Squared Error: 463.41444019020713

Test R-squared: 0.060195067499318244

**5. CONCLUSION**

As a result, we may contrast these models' output with their real models. As a result of factors like global inflation, the price of oil has been rising dramatically recently. So, things come from petroleum or are extracted from it. India obtains its crude oil from nations that are nearby, such Saudi Arabia and Dubai. We've made the decision to deploy machine learning algorithms and employ ensemble learning to forecast the future prices of petroleum products like petrol and diesel. Using several instances of one algorithm or several distinct algorithms is a technique called ensemble learning. By comparing many algorithms, we may determine which is optimum for our issue statement.