**TESLA STOCK PRICE PREDICTION**

***A SUMMER INTERNSHIP Report Submitted to***

**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY-KAKINADA, KAKINADA**

***Submitted in partial fulfillment of the requirements for the award of degree***

#### BACHELOR OF TECHNOLOGY

**In**

#### INFORMATION TECHNOLOGY

**III B. Tech – I Semester Submitted by**

#### K. PRADEEP (20KN1A1225)

**V. LUKE SUNDAR (20KN1A1264)**

#### A. DEVI PRIYA (20KN1A1201)

***Under the esteemed Guidance of***

## Dr. M. CHAITANYA KISHORE REDDY,Ph.D.

**Professor & HOD**

**Department of Information Technology**

NRI INSTITUTE OF TECHNOLOGY

(AUTONOMOUS)

**Approved by AICTE, New Delhi :: Permanent Affiliation to JNTUK, Kakinada**

**Accredited by NBA (CSE,ECE& EEE), Accredited by NAAC with A-Grade An ISO 9001:2015 Certified Institution**

**Pothavarappadu (V), Via Nunna, Agiripalli (M), PIN-521 212. Ph : 0866 – 2469666 | e-mail :** [**nrihitech@rediffmail.com**](mailto:nrihitech@rediffmail.com)

**Website :[www.nriit.edu.in](http://www.nriit.edu.in/)**

(2020-2024)

**TESLA STOCK PRICE PREDICTION**

***A SUMMER INTERNSHIP Report submitted to***

**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY-KAKINADA, KAKINADA**

***Submitted in partial fulfillment of the requirements for the award of degree***

**BACHELOR OF TECHNOLOGY**

**In**

**INFORMATION TECHNOLOGY**

**III B. Tech – I Semester Submitted by**

**K. PRADEEP (20KN1A1225)**

**V. LUKE SUNDAR (20KN1A1264)**

**A. DEVI PRIYA (20KN1A1201)**

***Under the esteemed Guidance of***

### Dr. M. CHAITANYA KISHORE REDDY, Ph.D.

**Professor & HOD**

**Department of Information Technology**

(AUTONOMOUS)

NRI INSTITUTE OF TECHNOLOGY

**Approved by AICTE, New Delhi: Permanent Affiliation to JNTUK, Kakinada**

**Accredited by NBA (CSE, ECE & EEE), Accredited by NAAC with A-Grade An ISO 9001:2015 Certified Institution**

**Pothavarappadu (V), Via Nunna, Agiripalli (M), PIN-521 212. Ph : 0866 – 2469666 | e-mail :** [**nrihitech@rediffmail.com**](mailto:nrihitech@rediffmail.com)

**Website :** [**www.nriit.edu.in**](http://www.nriit.edu.in/)

**i**

NRI INSTITUTE OF TECHNOLOGY

(AUTONOMOUS)

**Approved by AICTE, New Delhi: Permanent Affiliation to JNTUK, Kakinada**

**Accredited by NBA (CSE, ECE & EEE), Accredited by NAAC with A-Grade An ISO 9001:2015 Certified Institution**

**Pothavarappadu (V), Via Nunna, Agiripalli (M), PIN-521 212. Ph : 0866 – 2469666 | e-mail :** [**nrihitech@rediffmail.com**](mailto:nrihitech@rediffmail.com)

**Website :** [**www.nriit.edu.in**](http://www.nriit.edu.in/)

**Department of Information Technology**

Certificate

**This is to certify that the project work entitled**

**“TESLA STOCK PRICE PREDICTION”**

**Is a bonafide record work done by**

**K. PRADEEP (20KN1A1225)**

**V. LUKE SUNDAR (20KN1A1264)**

**A. DEVI PRIYA (20KN1A1201)**

**In the Department of Information Technology, NRI Institute of Technology, is affiliated to JNTU-Kakinada in partial fulfillment of the requirements for the award of Bachelor of Technology in Information Technology during 2020- 2024. This work has been carried out under my guidance and supervision.**

**The results embodied in this Project report have not been submitted in any University or Organization for the award of any degree or diploma.**

**INTERNAL GUIDE: HEAD OF DEPARTMENT:**

**Dr. M. CHAITANYA KISHORE REDDY, Dr. M. CHAITANYA KISHORE REDDY,**

**Professor & HOD Professor & HOD**

**External Examiner**

**ii**

# ACKNOWLEDGEMENT

Before getting into the thickest of things, we would like to thank the personalities who were part of my project in numerous ways, those who gave me outstanding support from birth of the project.

We are extremely thankful to our beloved **chairman,** **Dr.**

**R. Venkat Rao** for providing necessary infrastructure and resources for the accomplishment of our project at **NRI Institute of Technology, Agiripalli**.

We are highly indebted to **Dr. C. Naga Bhaskar, Principal** of **NRI Institute of Technology,** for his support during the tenure of the project.

We hereby wish to express our deep sense of gratitude to **Dr.**

**M. Chaitanya Kishore Reddy, Prof & HOD**, Department of Information Technology, NRI Institute of Technology for the esteemed guidance, moral support and invaluable advice provided by them for the success of the project.

With grateful thanks, I express my deep sense of gratitude and respect towards my internal guide **Dr. M. CHAITANYA KISHORE REDDY*,* Prof & HOD** for his valuable suggestions, guidance and encouragement throughout the project.

We are also thankful to all the staff members of Information Technology department who have cooperated in making our project a success. We would like to thank all our parents and friends who extended their help, encouragement and moral support either directly or indirectly in our project work.

Thanks for Your Valuable Guidance and kind support.

**iii**

# DECLARATION

We hereby declare that project report entitled **“TESLA STOCK PRICE PREDICTION”** is a genuine project work carried out by us, in **B. Tech (Information Technology)** degree course of **JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY,**

**KAKINADA** and has not been submitted to any other courses or University for award of any degree by us.

### Signature of the Students

|  |  |
| --- | --- |
| K. PRADEEP | 20KN1A1225 |
| V. LUKE SUNDAR | 20KN1A1264 |
| A. DEVI PRIYA | 20KN1A1201 |

**iv**

# ABSTRACT

Prediction of stock prices has been an important area of research for a long time. In recent years’ stock price prediction has been one of the most significant concern. Stock market trading is an important economic activity of the society which helps people to earn more money. Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on a financial exchange. The successful prediction of a stock's future price will maximize investor’s gains. Machine Learning methods are often used for the prediction of stock prices. We use several machine learning approaches (Supervised or Unsupervised) and methods through which the investors get to know the stock prices increase or decrease. It was done in five phases, such as data acquired, pre-processing of dataset, extraction of features, prediction of stock price using different techniques and display the result.

The Tesla vehicles became very popular in the car industry as it was affordable in the consumer market and it left no carbo foot print. Due to large decline in the stock prices of Tesla Inc. at the beginning of 2019, Tesla owners started selling their vehicles in the used car market. This used car prices depended on attributes such as, model of the vehicle, year of production, miles driven and the battery used for the vehicle. Prices were different for a specific vehicle in different months. In this project, it is discussed how a machine learning technique is being implemented in order to predict the stock prices. To reach this goal, different machine learning techniques were used.

**v**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.No** | **CHAPTERS** | **PAGE NO** |
|  | **TITLE PAGE** | **i** |
| **CERTIFICATE** | **ii** |
| **ACKNOWLEDGEMENT** | **iii** |
| **DECLARATION** | **iv** |
| **ABSTRACT** | **v** |
| **LIST OF FIGURES** | **viii** |
| **LIST OF ABBREVIATIONS** | **ix** |
| **LIST OF SCREENS** | **x** |
| **1** | **INTRODUCTION** | **1** |
| **2** | **LITERATURE SURVEY** | **2** |
| **3** | **SOFTWARE REQUIREMENT ANALYSIS** | **3-5** |
|  | 3.1 Functional Requirements | 4 |
|  | 3.2 Non-functional Requirements | 5 |
| **4** | **SOFTWARE DESIGN** | **6** |
| **5** | **PROPOSED SYSTEM** | **7-23** |
|  | 5.1 Technology used | 7 |
|  | 5.1.1 Machine Learning | **7** |
|  | 5.1.2 Supervised Learning | 8 |
|  | 5.1.3 Unsupervised Learning | 8 |
|  | 5.2 Approaches | 9 |
|  | 5.3 Predictive Modelling | 9 |
|  | 5.4 Common Predictive Algorithms | 10 |
|  | 5.4.1 Decision Tree Model | 10 |
|  | 5.4.2 Random Forest | 11 |
|  | 5.4.3 XG Boost | 11 |
|  | 5.5 Applications of Machine Learning | 12 |

**vi**

|  |  |  |
| --- | --- | --- |
|  | 5.6 Algorithm used in our project | 13 |
|  | 5.7 Modules involved | 14 |
|  | 5.8 Data Processing in Machine Learning | 14-16 |
|  | 5.9 Datasets | 17 |
|  | 5.10 Kaggle Dataset | 17 |
|  | 5.11 Dataset Splitting | 18 |
|  | 5.11.1 Train Dataset | 19 |
|  | 5.11.2 Test Dataset | 20-21 |
|  | 5.12 Tools Used | 22 |
|  | 5.12.1 Python | 22 |
|  | 5.12.2 Numpy | 22 |
|  | 5.12.3 Scikit Learn | 23 |
|  | 5.12.4 Matplotlib | 23 |
|  | 5.13 Jupyter Notebook | 23 |
| **6** | **CODING** | **24-25** |
| **7** | **OUTPUT SCREENS** | **26 - 34** |
| **8** | **CONCLUSION AND FURTHER WORK** | **35** |
| **9** | **REFERENCES** | **36** |

**vii**

# LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **FIG.NO.** | **NAME OF THE FIGURE** | **PAGE NO.** |
| 4.1 | System Design . | 6 |
| 5.1 | Linear Regression Model | 13 |
| 5.2 | Data Processing | 15 |
| 5.3 | Kaggle | 18 |
| 5.4 | Data Splitting | 21 |

**viii**

# LIST OF ABBREVIATIONS

|  |  |  |
| --- | --- | --- |
| **S.NO.** | **MNEMONICS** | **ABBREVIATION NAME** |
| 1 | CSV | Common Separated values |
| 2 | UX | User Experience |
| 3 | PHP | Hypertext preprocessor |
| 4 | MNIST | Modified National Of Standards And Technology Database |

**ix**

# LIST OF SCREENS

|  |  |  |
| --- | --- | --- |
| **SCREEN NO.** | **SCREEN NAME** | **PAGE NO.** |
| 7.1 | Libraries used and attributes of Dataset | 26 |
| 7.2 | Using the methods isnull() and isna() | 27 |
| 7.3 | Displaying complete information about Dataset | 28 |
| 7.4 | Graph | 29 |
| 7.5 | Displaying the shape of the train and test variables | 30 |
| 7.6 | Displaying the Data frame | 31 |
| 7.7 | Displaying the Actual and Predicted values | 32 |
| 7.8 | Checking the Accuracy | 33 |
| 7.9 | Bar chart | 34 |

**x**



CHAPTER-1

**INTRODUCTION**

An American car manufacturing company Tesla Inc was founded on 2003. It is one of the companies that developed electric cars to a manufacturing standard and made it available for the consumer market . In the beginning of 2019, the stock price of Tesla started to drop. This made Tesla vehicle owners feared and most of them started selling of their vehicle. A huge demand increased in the used car market for Tesla vehicles. But the price for Tesla vehicle depended on the stock price of the vehicle and it impacted largely in the used Tesla vehicle market. Tesla vehicles highly depends on Tesla company’s post sells services. In order to keep the vehicles usable in the market and also decline of the stock prices the accurate prediction of used car prices was very important for the new and existing customers and the company itself.

Over the last two years few stocks have been most exciting and controversial as tesla. The company aims to change the world fundamentally by disrupting one of the essential industries. Not only did it make lot of money for its investors tesla made Elon Musk into one of the most powerful CEO’s in the world. However, market participants are rightfully skeptical about how high the tesla stock can go. Tesla just reported its production and delivery numbers for Q3 2020.the company produced 237,000 vehicles while it has delivered 241,000 vehicles these numbers bring it close to the annual production volume of 1 million units. Meanwhile domestic brands like ford produce over 4 million vehicles per year, and the international leader Volkswagen group makes 9.3 million.

**1**



CHAPTER-2

### LITERATURE SURVEY

**TESLA STOCK PRICE PREDICTION USING SENTIMENT ANALYSIS**

The stock market is a very volatile component of the financial domain. Accurate predictions of various stocks are a highly active area of research and analysis. Following the previous ML prediction techniques using Artificial Neural networks and fuzzy-based techniques, this research aims to extend the accurate prediction results. Since multiple qualitative factors go into the decision-making of a buy-sell of stock, a blend of algorithmic trading is at the cornerstone of the research. This research work aims to look into the unique relationship between Elon Musk's Tweets and Tesla's stock value. Exploratory Data Analysis was employed as the primary analysis method to better differentiate patterns within our dataset, which had been pre-processed to remove any stop words. Combining these methodologies and elements yielded a decisive conclusion with a clear correlation: an increase in the number of tweets/engagements corresponded to an increase in Tesla's closing price and vice versa.

**SENTIMENT ANALYSIS**

Sentiment Analysis, also known as Opinion Mining, is a branch of humanities and linguistics combined with advanced data science and (AI) artificial intelligence. Sentiment Analysis' basic assumption is to take a set of linguistic data (in this case, Elon Musk's Twitter) and determine the emotions and subjective feelings behind those words. Sentiment Analysis is done in several ways. For example, a machine might learn to discern human emotions by reviewing enormous datasets of text. Sentiment Analysis was utilized because we predicted that emotions expressed in tweets could be linked to Tesla stock prices. Sentiment Analysis is employed in almost every aspect of our lives. Sentiment Analysis is used in various ways in advertisements, campaigns, and commercials, all of which include pathos. Sentiment Analysis was used extensively in our study, with the subjective state behind Elon Musk's tweets/replies being applied to Tesla's stock, revealing novel correlations between the two.

Each tweet on Twitter was divided into three categories based on its expected emotional language: positive, neutral, and negative. This was accomplished with Python and the Scikitlearn program. This machine learning was carried out using pre-processed tweets from the text column of the CSV file.

2



CHAPTER-3

## SOFTWARE REQUIREMENT ANALYSIS

## HARDWARE REQUIREMENTS:

* Processor: Intel P4 Processor
* RAM: 4 GB
* Processor Speed: 2GHZ
* Hard Disk: 500MB

## SOFTWARE REQUIREMENTS:

* Operating System: Windows 10(64-bit Operating System)
* Coding Language: Python
* Tools: Pandas, Numpy, Sklearn, Matplotlib
* Platform/Web Server: Jupyter Notebook
* Technology: Machine Learning



* 

### Functional Requirements

Functional requirements deal with the functionality of the software in the engineering view. The component flow and the structural flow of the same is enhanced and described by it. The functional statement deals with the raw datasets that are categorized and learning from the same dataset. Later the datasets are categorized into clusters and the impairment of the same is checked for the efficiency purpose. After the dataset cleaning the data are cleansed and the machine learns and finds the pattern set for the same it undergoes various iteration and produce output.

Functional requirements describe what the software should do (the functions). Think about the core operations. Because the “functions” are established before development, functional requirements should be written in the future tense. In developing the software for Stock Price Prediction, some of the functional requirements could include:

* + The software shall accept the tw\_spydata\_raw.csv dataset as input.
  + The software should shall do pre-processing (like verifying for missing data values) on input for model training.
  + The software shall LINEAR REGRESSION as main component of the software.
  + It processes the given input data by producing the most possible outcomes of a CLOSING STOCK PRICE.

### Non-functional Requirements:

Nonfunctional requirement deals with the external factors which are nonfunctional in nature It is used for analysis purpose. Under the same the judgment of the operations is carried out for its performance. Stock is feasible and is ever changing so these extra effects and the requirements helps it to get the latest updates and integrate in a one goes where the technicians can work on and solve a bug or a draft if any. The non-functional requirements followed are its efficiency and hit gain ratio. The usability of the code for the further effectiveness and to implement and look for the security console. The System is reliable and the performance is maintained with the support of integration and portability of the same.

4

## Product properties:

* **Usability:** It defines the user interface of the software in terms of simplicity of understanding the user interface of stock prediction software, for any kind of stock trader and other stakeholders in stock market.
* **Efficiency:** maintaining the possible highest accuracy in the closing stock prices in shortest time with available data.
* **Performance:** It is a quality attribute of the stock prediction software that describes the responsiveness to various user interactions with it.

5

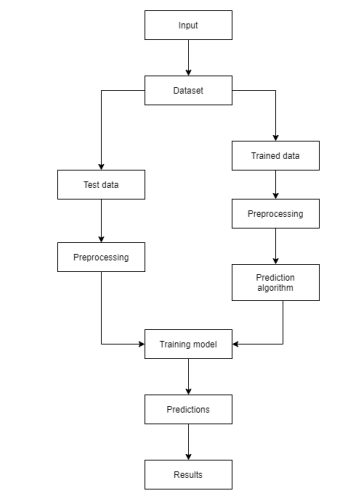


CHAPTER-4

**SOFTWARE DESIGN**

## Structure Chart:

A structure chart in software engineering and organizational theory is a chart which shows the breakdown of a system to its lowest manageable levels. They are used in structured programming to arrange program modules into a tree. Each module is represented by a box, which contains the module's name.



**Fig. 4.1 System Design**

**6**



CHAPTER-5

### PROPOSED SYSTEM

In this project an attempt is made to predict the stock market trend using a culmination of prediction modelling and regression algorithms. This model takes the current stock values from the data sets gathered. The data gathered is modelled into various sub parts or data sets which is used to train and test the algorithm. Then, regression model in python or R is used to model the data. This model runs a comprehensive search algorithm on the data sets and create a summary table based on the output. The values are plot on a chart and regression and clustering techniques are applied to find out the increase or decrease in price of that stock. Based on the calculation this model extrapolates the current stock prices to generate a prediction after a given time. The models are built and trained using supervised machine learning algorithms. The output will be in graphical form and will change with change in dataset. This model expects up to 69% of in-sample accuracy and 37% or out-of-sample accuracy using supervised machine learning algorithms.

## 5.1 TECHNOLOGY USED 5.1.1 MACHINE LEARNING:

Machine Learningis the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: *The ability to learn*. Machine learning is actively being used today, perhaps in many more places than one would expect.

Two of the most widely adopted machine learning methods are Supervised leaningwhich trains algorithms based on example input and output data that is labelled by humans, and Unsupervised learningwhich provides the algorithm with no labelled data in order to allow it to find structure within its input data. Let’s explore these methods in more detail.

7

**5.1.2 Supervised learning**:

The model or algorithm is presented with example inputs and their desired outputs and then finding patterns and connections between the input and the output. The goal is to learn a general rule that maps inputs to outputs. The training process continues until the model achieves the desired level of accuracy on the training data. Some real-life examples are:

* **Image Classification:** You train with images/labels. Then in the future you give a new image expecting that the computer will recognize the new object.
* **Market Prediction/Regression:** You train the computer with historical market data and ask the computer to predict the new price in the future.

**5.1.3 Unsupervised learning**: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. It is used for clustering population in different groups. Unsupervised learning can be a goal in itself (discovering hidden patterns in data).

* **Clustering:** You ask the computer to separate similar data into clusters, this is essential in research and science.
* **High Dimension Visualization:** Use the computer to help us visualize high dimension data.
* **Generative Models:** After a model captures the probability distribution of your input data, it will be able to generate more data. This can be very useful to make your classifier more robust.

The goal of unsupervised learning may be as straightforward ads discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify the raw data.

8

## 5.2 APPROACHES:

As a field, machine learning is closely related to computational statistics, so having a background knowledge in statistics is useful for understanding and leveraging machine learning algorithms.

For those not have studied statistics, it can be helpful to first define correlation and regression, as they are commonly used techniques for investigating the relationship among quantitative variables. Correlation is a measure of association between two variables that are not designated as either dependent or independent. Regression at a basic level is used to examine the relationship between one independent and one independent variable. Because regression statistics can be used to anticipate the dependent variable when the independent variable is known, regression enables prediction capabilities.

Approaches to machine learning are continuously being developed. For our purposes, we’ll go through a few of the popular approaches that are being used in machine learning at the time of writing.

## 5.3 PREDICTIVE MODELLING:

In short, predictive modelling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes. Predictive modelling can be used to predict just about anything, from TV ratings and a customer’s next purchase to credit risks and corporate earnings.

A predictive model is not fixed; it is validated or revised regularly to incorporate changes in the underlying data. In other words, it’s not a one-and-done prediction. Predictive models make assumptions based on what has happened in the past and what is happening now. If incoming, new data shows changes in what is happening now, the impact on the likely future outcome must be recalculated, too. For example, a software company could model historical sales data against marketing expenditures across multiple regions to create a model for future revenue based on the impact of the marketing spend.

9

Most predictive models work fast and often complete their calculations in real time. That’s why banks and retailers can, for example, calculate the risk of an online mortgage or credit card application and accept or decline the request almost instantly based on that prediction.

Some predictive models are more complex, such as those used in Computational biology and Quantum computing the resulting outputs take longer to compute than a credit card application but are done much more quickly than was possible in the past thanks to advances in technological capabilities, including computing power.

## 5.4 COMMON PREDICTIVE ALGORITHMS:

* Linear Regression
* Decision Tree Model
* Random Forest
* XG Boost Model

## 5.4.1 DECISION TREE MODEL:

A decision tree is an algorithm that displays the likely outcomes of various actions by graphing structured or unstructured data into a tree-like structure. Decision trees divide different decisions into branches and then list alternative outcomes beneath each one. It examines the training data and chooses the independent variable that separates it into the most diverse logical categories. The popularity of decision trees stems from the fact that they are simple to understand and interpret.

Decision trees also work well with incomplete datasets and are helpful in selecting relevant input variables. Businesses generally leverage decision trees to detect the essential target variable in a dataset. They may also employ them because the model may generate potential outcomes from incomplete datasets.

10

## 5.4.2 RANDOM FOREST:

A random forest is a vast collection of decision trees, each making its prediction. Random forests can perform both classification and regression. The values of a random vector sampled randomly with the same distribution for all trees in the random forest determine the shape of each tree. The power of this model comes from the ability to create several trees with various sub-features from the features. Random forest uses the bagging approach, i.e., it generates data subsets from training samples that you can randomly choose with replacement.

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.

## 5.4.3 XG BOOST:

XG Boost is an implementation of Gradient Boosted decision trees. XG Boost models majorly dominate in many Kaggle Competitions.

In this algorithm, decision trees are created in sequential form. Weights play an important role in XG Boost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

11

## 5.5 APPLICATIONS OF MACHINE LEARNING:

* **Web Search Engine:** One of the reasons why search engines like google, Bing etc. work so well is because the system has learnt how to rank pages through a complex learning algorithm.
* **Photo tagging Applications:** Be it Facebook or any other photo tagging application, the ability to tag friends makes it even more happening. It is all possible because of a face recognition algorithm that runs behind the application.
* **Spam Detector:** Our mail agent like Gmail or Hotmail does a lot of hard work for us in classifying the mails and moving the spam mails to spam folder. This is again achieved by a spam classifier running in the back end of mail application.
* **Database Mining for growth of automation:** Typical applications include Web- click data for better UX (User experience), Medical records for better automation in healthcare, biological data and many more.
* **Applications that cannot be programmed:** There are some tasks that cannot be programmed as the computers we use are not modelled that way. Examples include Autonomous Driving, Recognition tasks from unordered data (Face Recognition/ Handwriting Recognition), Natural language Processing, computer Vision etc.
* **Understanding Human Learning:** This is the closest we have understood and mimicked the human brain. It is the start of a new revolution, The real AI. Now, after a brief insight lets come to a more formal definition of Machine Learning.
* **Stock Market Trading:** Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's **long short term memory neural network** is used for the prediction of stock market trends.
* **Medical Diagnosis:** In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain. It helps in finding brain tumors and other brain-related diseases easily.

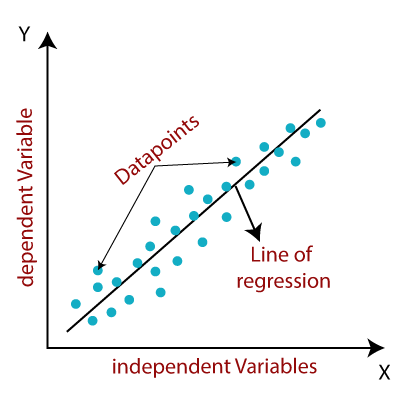


## 5.6 ALGORITHM USED IN OUR PROJECT:

**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 getting 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.



**FIG 5.1 LINEAR REGRESSION MODEL**

13

## 5.7 MODULES INVOLVED:

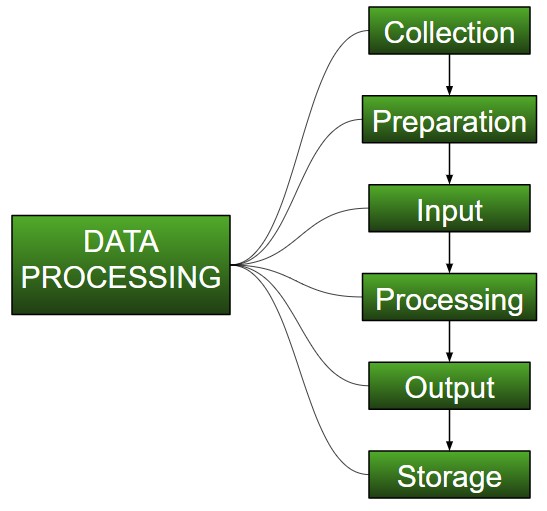
1. Input as Dataset.
2. Preprocessing.
3. Data splitting.
4. Training and Testing.
5. Output as Predicted Result.

Attribute such as: price of open, high, low, close, adjusted close price taken from huge dataset are fed as input to the models for training to pre-process the data techniques like normalization & one hot encoding in applied on dataset. After this data is divided in two sets namely training & testing which are ratio of 80:20 respectively. Then, this set are used to train a model using a linear regression model. Finally, all these modules are evaluated using root mean square error.

## 5.8 DATA PROCESSING IN MACHINE LEARNING:

Data Processing is the task of converting data from a given form to a much more usable and desired form i.e. making it more meaningful and informative. Using Machine Learning algorithms, mathematical modeling, and statistical knowledge, this entire process can be automated. The output of this complete process can be in any desired form like graphs, videos, charts, tables, images, and many more, depending on the task we are performing and the requirements of the machine. This might seem to be simple but when it comes to massive organizations like Twitter, Facebook, Administrative bodies like Parliament, UNESCO, and health sector organizations, this entire process needs to be performed in a very structured manner. So, the steps to perform are as follows:

14



**FIG 5.2 DATA PROCESSING**

## COLLECTION:

The most crucial step when starting with ML is to have data of good quality and accuracy. Data can be collected from any authenticated source like data.gov.kaggle.in or UCI dataset repository. For example, while preparing for a competitive exam, students study from the best study material that they can access so that they learn the best to obtain the best results. In the same way, high-quality and accurate data will make the learning process of the model easier and better and at the time of testing, the model would yield state-of-the-art results. A huge amount of capital, time and resources are consumed in collecting data.

Example: Working on the Facial Expression Recognizer, needs numerous images having a variety of human expressions. Good data ensures that the results of the model are valid and can be trusted upon.

15

## PREPARATION:

The collected data can be in a raw form which can’t be directly fed to the machine. So, this is a process of collecting datasets from different sources, analyzing these datasets and then constructing a new dataset for further processing and exploration. This preparation can be performed either manually or from the automatic approach. Data can also be prepared in numeric forms also which would fasten the model’s learning. **Example:** An image can be converted to a matrix of N X N dimensions, the value of each cell will indicate the image pixel.

## INPUT:

Now the prepared data can be in the form that may not be machine-readable, so to convert this data to the readable form, some conversion algorithms are needed. For this task to be executed, high computation and accuracy is needed. Example: Data can be collected through the sources like MNIST Digit data(images), Twitter comments, audio files, video clips.

## PROCESSING:

This is the stage where algorithms and ML techniques are required to perform the instructions provided over a large volume of data with accuracy and optimal computation.

## OUTPUT:

In this stage, results are procured by the machine in a meaningful manner which can be inferred easily by the user. Output can be in the form of reports, graphs, videos, etc.

## STORAGE:

This is the final step in which the obtained output and the data model data and all the useful information are saved for future use.

16

## 5.9 DATASETS:

The key to success in the field of machine learning or to become a great data scientist is to practice with different types of datasets. But discovering a suitable dataset for each kind of machine learning project is a difficult task. **A dataset** is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table. A tabular dataset can be understood as a database table or matrix, where each column corresponds to a **particular variable,** and each row corresponds to the **fields of the dataset.** The most supported file type for a tabular dataset is **"Comma Separated File,"** or **CSV.** But to store a "tree-like data," we can use the JSON file more efficiently.

## TYPES OF DATASETS:

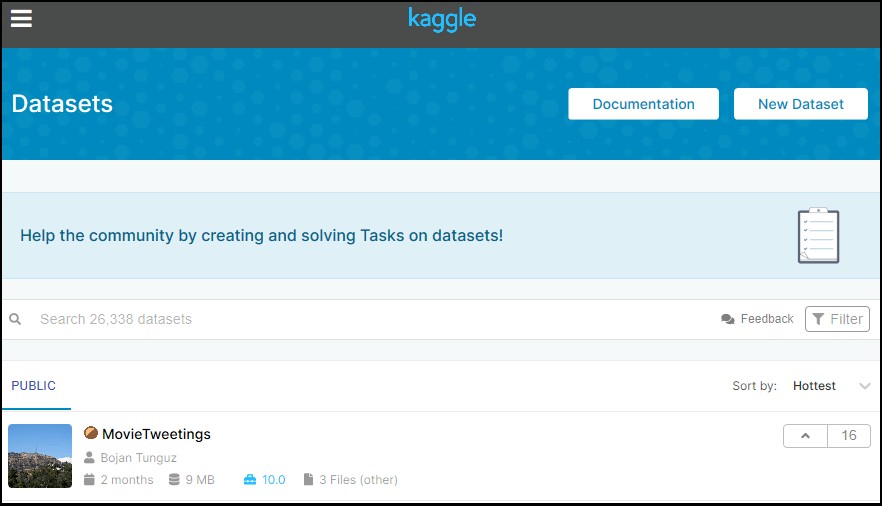
* **Numerical data:** Such as house price, temperature, etc.
* **Categorical data:** Such as Yes/No, True/False, Blue/green, etc.
* **Ordinal data:** These data are similar to categorical data but can be measured on the basis of comparison.

## 5.10 KAGGLE DATASETS:

Kaggle is one of the best sources for providing datasets for Data Scientists and Machine Learners. It allows users to find, download, and publish datasets in an easy way. It also provides the opportunity to work with other machine learning engineers and solve difficult Data Science related tasks.

Kaggle provides a high-quality dataset in different formats that we can easily find and download.

17



## FIG 5.3 KAGGLE

## 5.11 DATA SPLITTING:

Data is at the heart of every ML problem. Without proper data, ML models are just like bodies without soul. But in today’s world of ‘big data’ collecting data is not a major problem anymore. We are knowingly (or unknowingly) generating huge datasets every day. However, having surplus data at hand still does not solve the problem. Though making sense out of raw data is an art in itself and requires good feature engineering skills and domain knowledge (in special cases), the quality data is of no use until it is properly used. The major problem which ML/DL practitioners face is how to divide the data for training and testing. Though it seems like a simple problem at first, its complexity can be gauged only by diving deep into it. Poor training and testing sets can lead to unpredictable effects on the output of the model. It may lead to overfitting or under fitting of the data and our model may end up giving biased results.

18

#### How to divide the data?

The data should ideally be divided into 3 sets – namely, train, test, and holdout cross- validation or development (dev) set. Let’s first understand in brief what these sets mean and what type of data they should have.

#### Train Set:

The train set would contain the data which will be fed into the model. In simple terms, our model would learn from this data. For instance, a Regression model would use the examples in this data to find gradients in order to reduce the cost function. Then these gradients will be used to reduce the cost and predict data effectively.

#### Dev Set:

The development set is used to validate the trained model. This is the most important setting as it will form the basis of our model evaluation. If the difference between error on the training set and error on the dev set is huge, it means the model as high variance and hence, a case of over-fitting.

#### Test Set:

The test set contains the data on which we test the trained and validated model. It tells us how efficient our overall model is and how likely is it going to predict something which does not make sense. There are a plethora of evaluation metrics (like precision, recall, accuracy, etc.) which can be used to measure the performance of our model.

## 5.11.1 TRAINING DATASET:

The training dataset is biggest (in-size) subset of the original dataset, which is used to train or fit the machine learning model. Firstly, the training data is fed to the ML algorithms, which lets them learn how to make predictions for the given task.

The training data varies depending on whether we are using Supervised Learning or Unsupervised Learning Algorithms.

For **Unsupervised learning**, the training data contains unlabeled data points, i.e., inputs are not tagged with the corresponding outputs. Models are required to find the patterns from the given training datasets in order to make predictions.

19

On the other hand, for supervised learning, the training data contains labels in order to train the model and make predictions.

The type of training data that we provide to the model is highly responsible for the model's accuracy and prediction ability. It means that the better the quality of the training data, the better will be the performance of the model. Training data is approximately more than or equal to 60% of the total data for an ML project.

## 5.11.2 TEST DATASET:

Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. The test dataset is another subset of original data, which is independent of training dataset.However, it has some similar types of features and class probability distribution and uses it as a benchmark for model evaluation once the model training is completed. Test data is a well-organized dataset that contains data for each type of scenario for a given problem that the model would be facing when used in the real world. Usually, the test dataset is approximately 20-25% of the total original data for an ML project.

At this stage, we can also check and compare the testing accuracy with the training accuracy, which means how accurate our model is with the test dataset against the training dataset. If the accuracy of the model on training data is greater than that on testing data, then the model is said to have overfitting.

The testing data should:

* Represent or part of the original dataset.
* It should be large enough to give meaningful predictions.

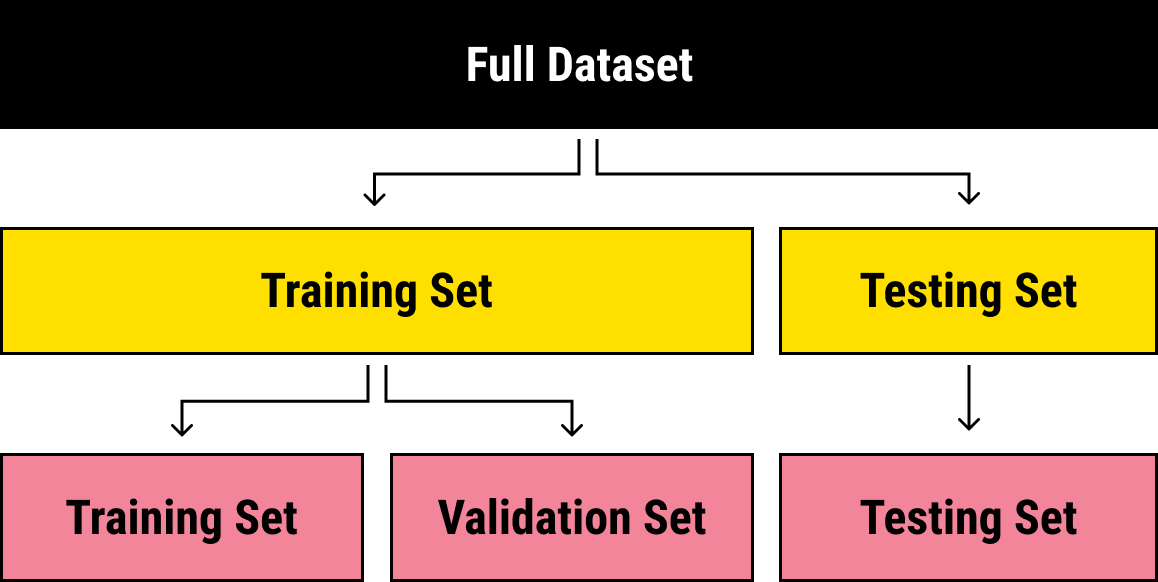
20

**x\_train**: It is used to represent features for the training data

**x\_test**: It is used to represent features for testing data

**y\_train:** It is used to represent dependent variables for training data

**y\_test:** It is used to represent independent variable for testing data.



## FIG 5.4 DATA SPLITTING

21

**5.12 TOOLS USED:**

## 5.12.1 PYTHON :

The language that we have selected for this project was Python. This was a straightforward call for many reasons.

1. Python as a language has a vast community behind it. Any problems which may be faced is simply resolved with visit to Stack Overflow. Python is the foremost standard language on the positioning that makes it is very straight answer to any question.
2. Python is an abundance of powerful tools ready for scientific computing Packages. The packages like NumPy, Pandas and SciPy area unit freely available and well documented. These Packages will intensely scale back, and variation the code necessary to write a given program. This makes repetition fast.
3. Python is a language as forgiving and permits for the program that appear as if pseudo code. This can be helpful once pseudo code give in tutorial papers should be required and verified. Using python this step is sometimes fairly trivial. However, Python is not without its errors. The python is dynamically written language and packages are area unit infamous for Duck writing. This may be frustrating once a package technique returns one thing that, for instance, looks like an array instead of being an actual array. Plus the standard Python documentation did not clearly state the return type of a method, this can’t lead without a lot of trials and error testing otherwise happen in a powerfully written language. This is a problem that produces learning to use a replacement Python package or library more difficult than it otherwise may be.

## 5.12.2 NUMPY:

Numpy is python package which provide scientific and higher level mathematical abstractions wrapped in python. It is [20] the core library for scientific computing, that contains a provide tools for integrating C, strong n-dimensional array object, C++ etc. It is also useful in random number capability, linear algebra etc. Numpy’s array type augments the Python language with an efficient data structure used for numerical work, e.g., manipulating matrices. Numpy additionally provides basic numerical routines, like tools for locating Eigenvectors

22

## 5.12.3 SCIKIT LEARN:

Scikit-learn could be a free machine learning library for Python. It features numerous classification, clustering and regression algorithms like random forests, k-neighbours, support vector machine, and it furthermore supports Python scientific and numerical libraries like SciPy and NumPy. In Python Scikit-learn is specifically written, with the core algorithms written in Cython to get the performance. Support vector machines are enforced by a Cython wrapper around LIBSVM .i.e., linear support vector machines and logistic regression by a similar wrapper around LIBLINEAR.

## 5.12.4 MATPLOTLIB:

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

## 5.13 JUPYTER NOTEBOOK:

The Jupyter Notebook is an open-source web application that enables to making and sharing documents that contain visualizations, narrative text, live code and equations. Uses include: data , data visualization, data transformation, statistical modelling, machine learning, numerical simulation, data cleaning and much more .

Jupyter Notebook is an open-source, web-based interactive environment**,** which allows you to create and share documents that contain live code, mathematical equations**,** graphics, maps, plots, visualizations, and narrative text. It integrates with many programming languages like Python, PHP, R, C#,etc.

23



CHAPTER-6

### CODING

**import pandas as pd**

**import numpy as np**

**from sklearn import metrics**

**%matplotlib inline**

**import matplotlib.pyplot as plt**

**dataset=pd.read\_csv('D:\Tesla-stock-analysis-and-prediction- master/tesla\_dataset.csv')**

**dataset.head()**

**dataset.isnull().sum()**

**dataset.isna().any()**

**dataset.info()**

**dataset.describe()**

**print(len(dataset))**

**dataset['Open'].plot(figsize=(16,6))**

**X = dataset[['Open','High','Low','Volume']]**

**y = dataset['Close']**

**from sklearn.model\_selection import train\_test\_split**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,random state=0)**

**X\_train.shape**

**X\_test.shape**

**from sklearn.linear\_model import LinearRegression**

**from sklearn.metrics import confusion\_matrix, accuracy\_score**

**regressor = LinearRegression()**

**regressor.fit(X\_train,y\_train)**

**print(regressor.coef\_)**

**print(regressor.intercept\_)**

24

**predicted=regressor.predict(X\_test)**

**print(X\_test)**

**predicted.shape**

**dframe=pd.DataFrame(y\_test,predicted)**

**dfr=pd.DataFrame({'Actual':y\_test,'Predicted':predicted})**

**print(dfr)**

**dfr.head(25)**

**from sklearn.metrics import confusion\_matrix, accuracy\_score**

**from sklearn.metrics import confusion\_matrix, accuracy\_score**

**import math**

**print('Mean Absolute Error:',metrics.mean\_absolute\_error(y\_test,predicted))**

**print('Mean Squared Error:',metrics.mean\_squared\_error(y\_test,predicted))**

**print('Root Mean squaredError :,math.sqrt(metrics.mean squared error(y\_test,predicted)))**

**graph=dfr.head(20)**

**graph.plot(kind='bar')**

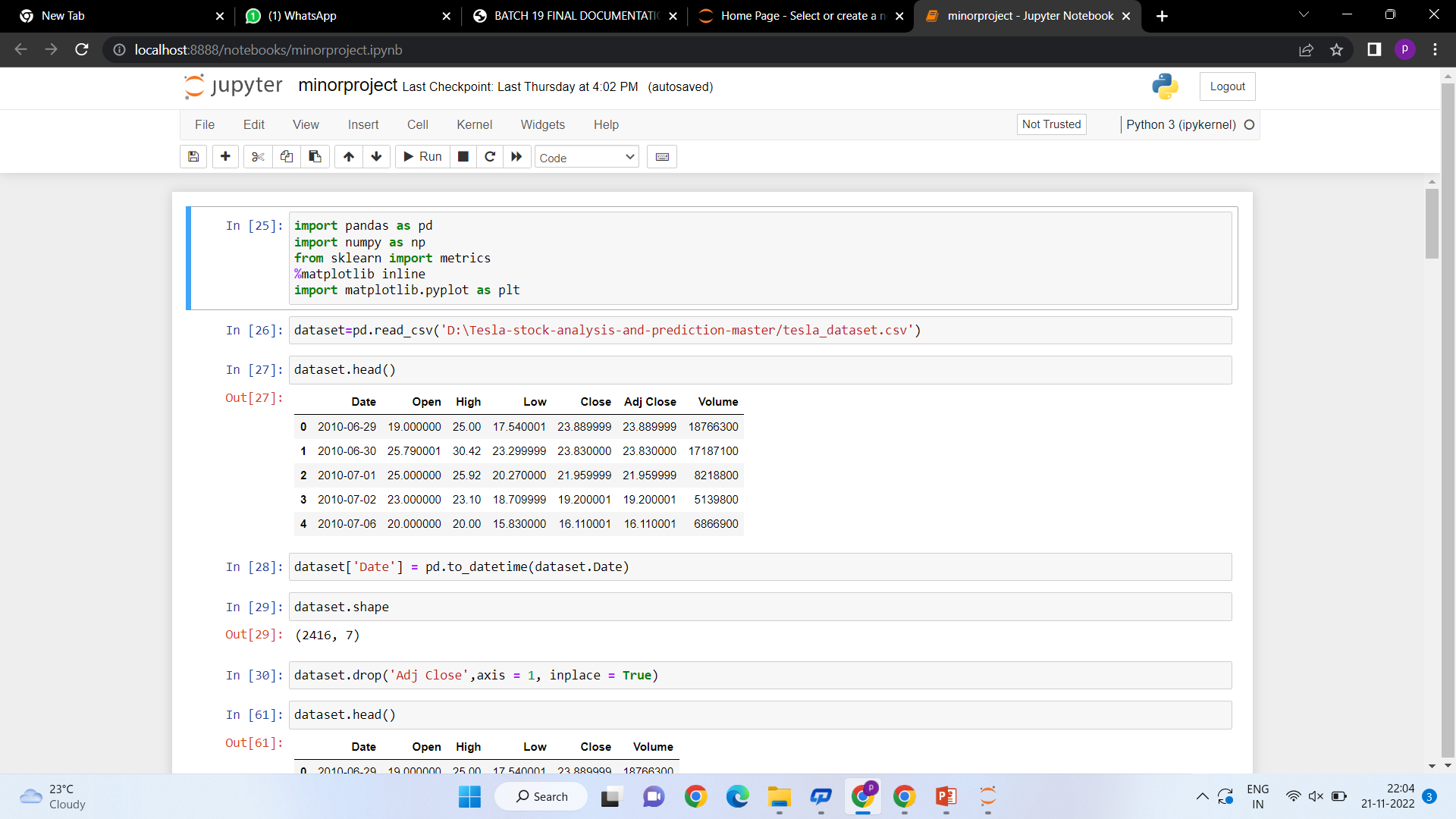
### 25



CHAPTER-7

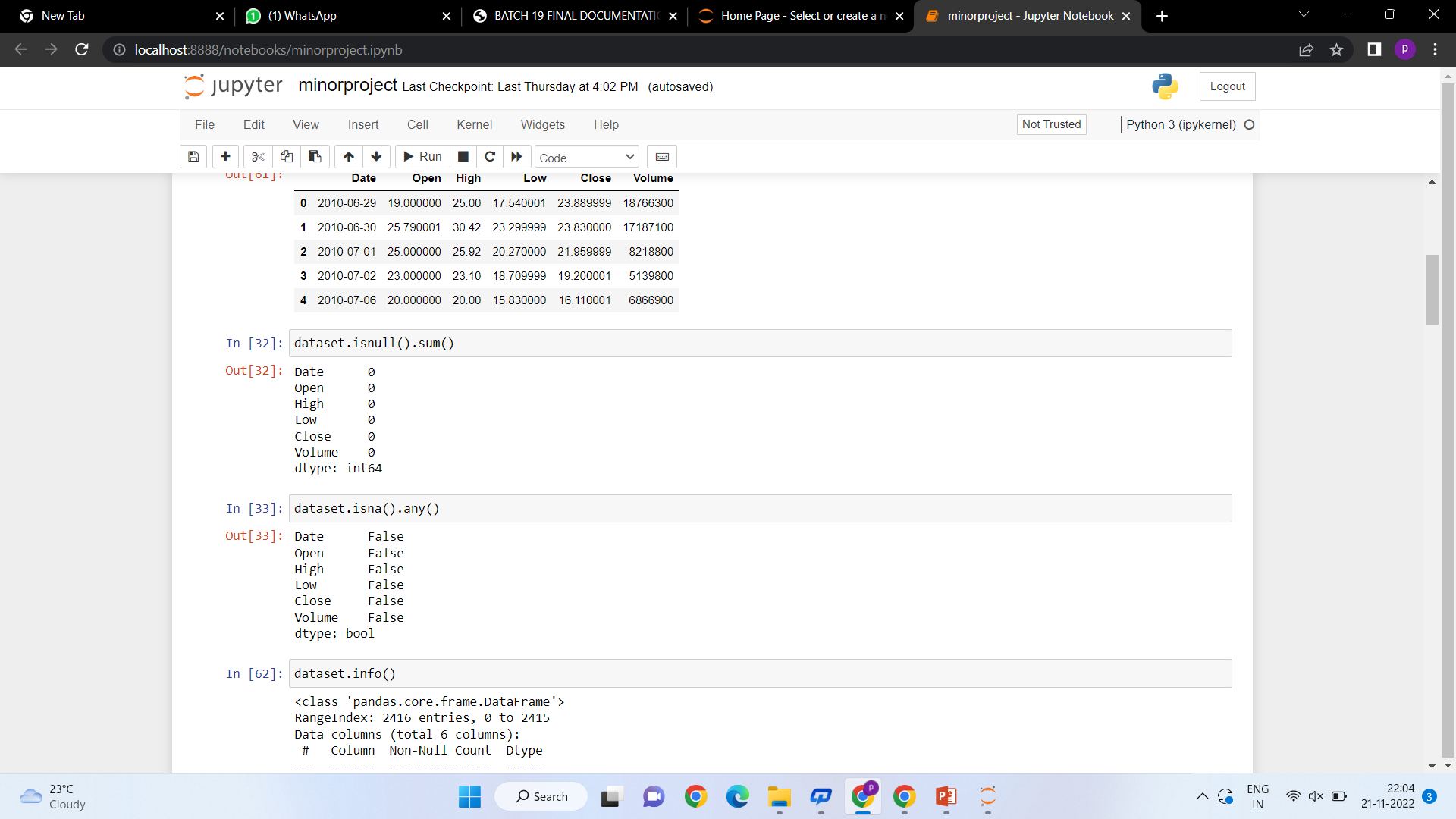
**OUTPUT SCREENS**

**­­­­­**

****

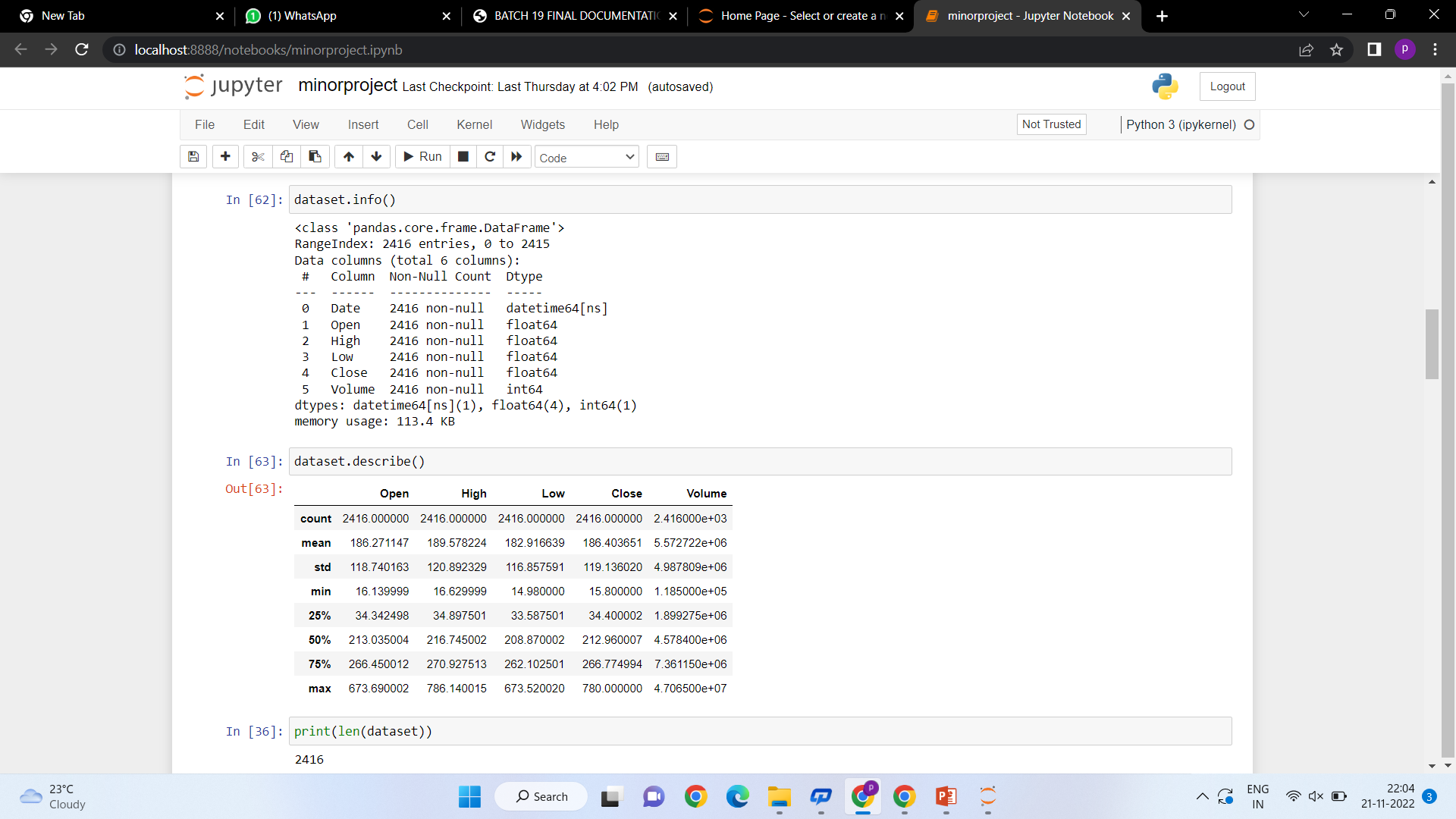
**Fig 7.1 Libraries Used and Attributes of Data Set**

26



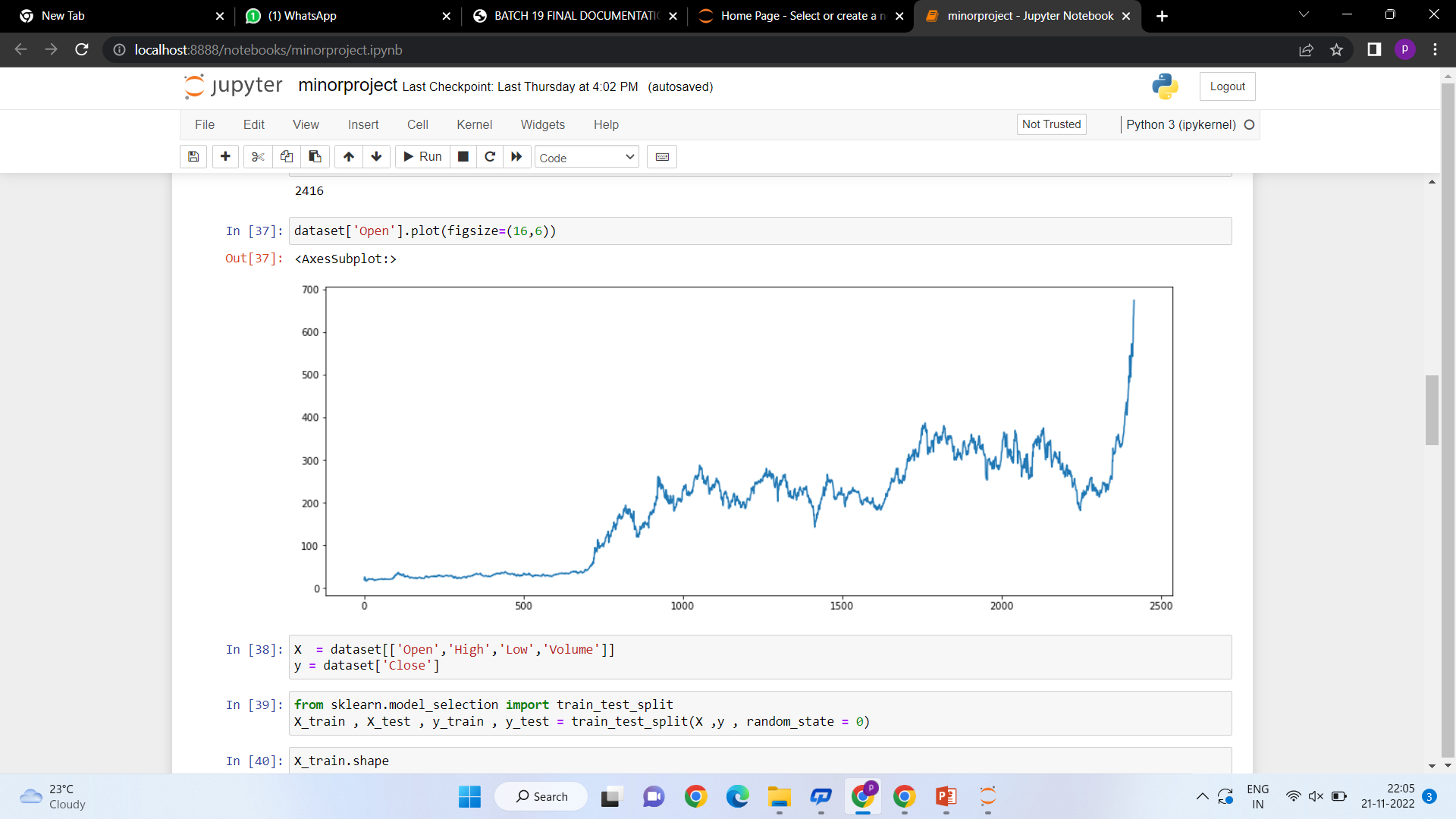
**Fig 7.2 Using Methods isnull () And isna ()**

27



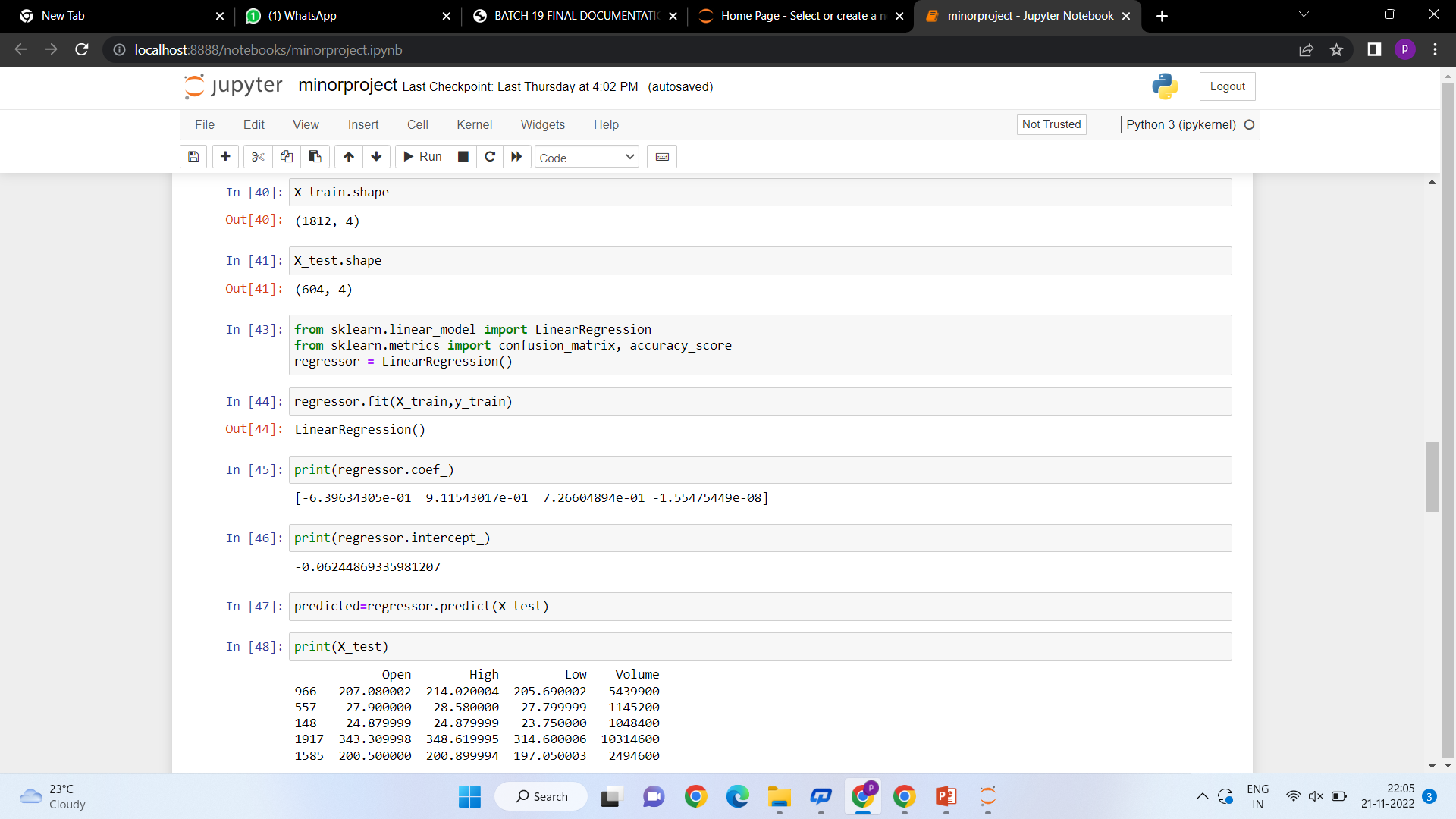
**Fig 7.3 Displaying Complete Information About Dataset**

28



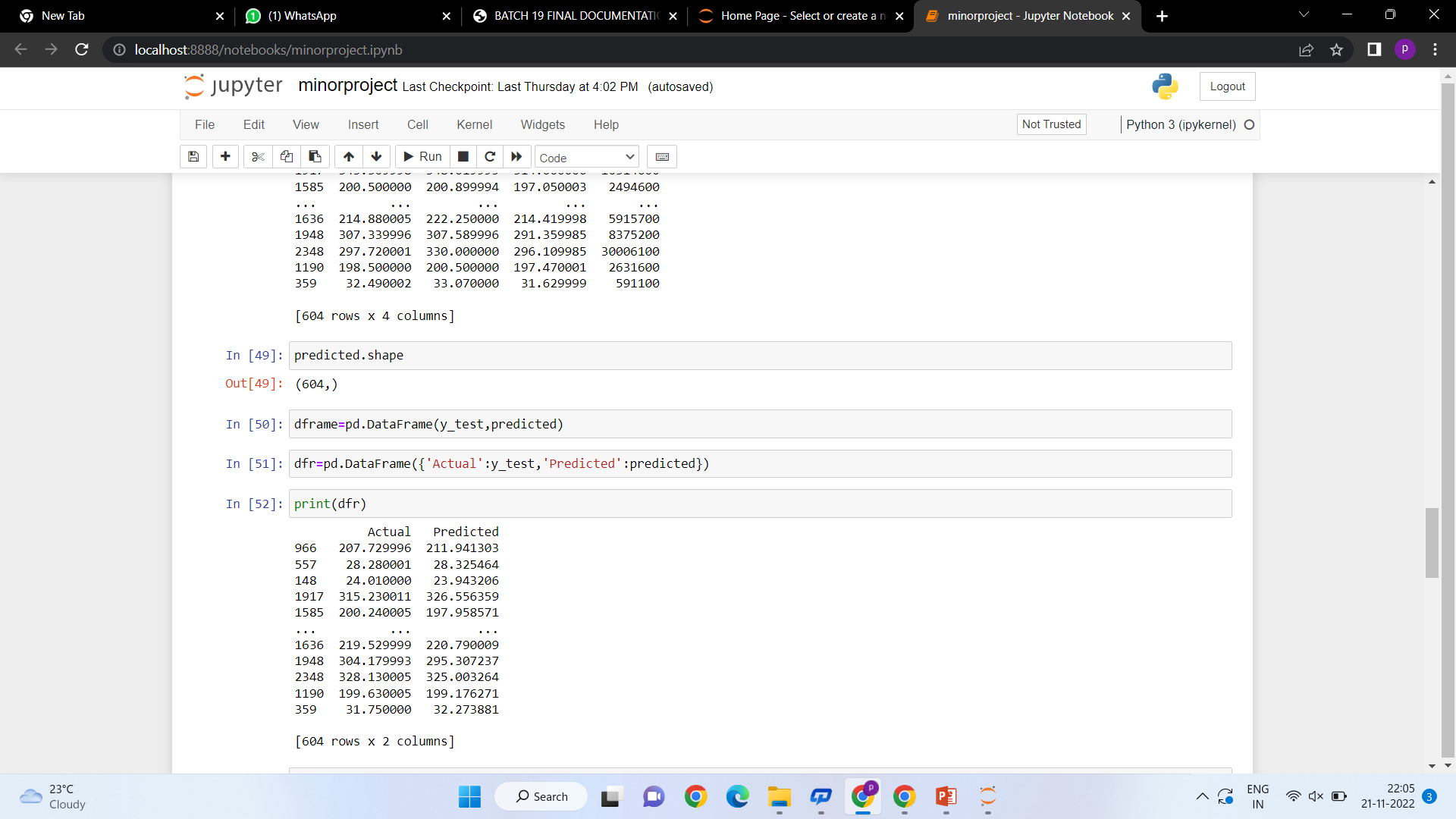
**Fig 7.4 Graph**

29

****

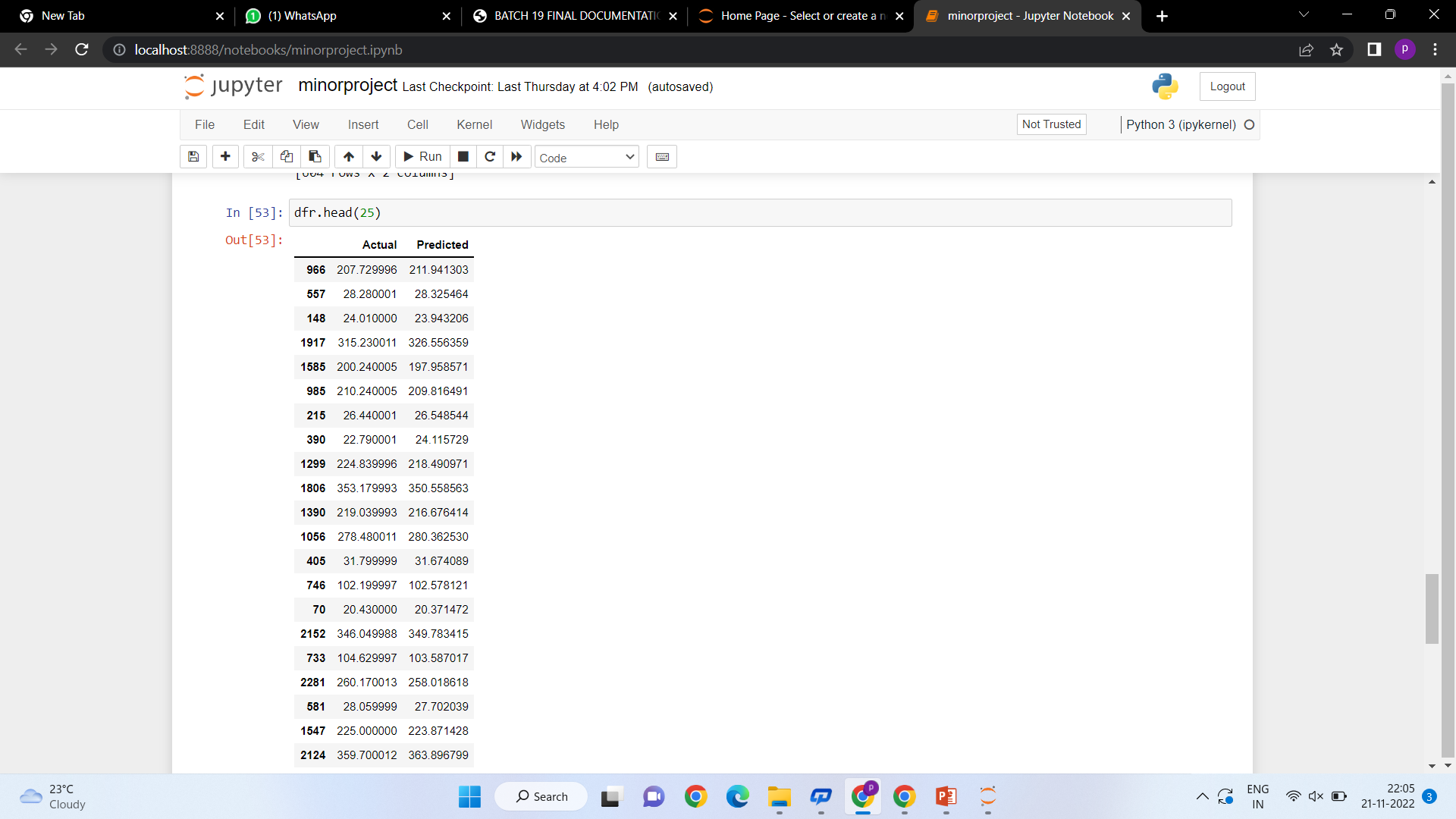
**Fig 7.5 Displaying Shape of the Train and Test Variables**

30

****

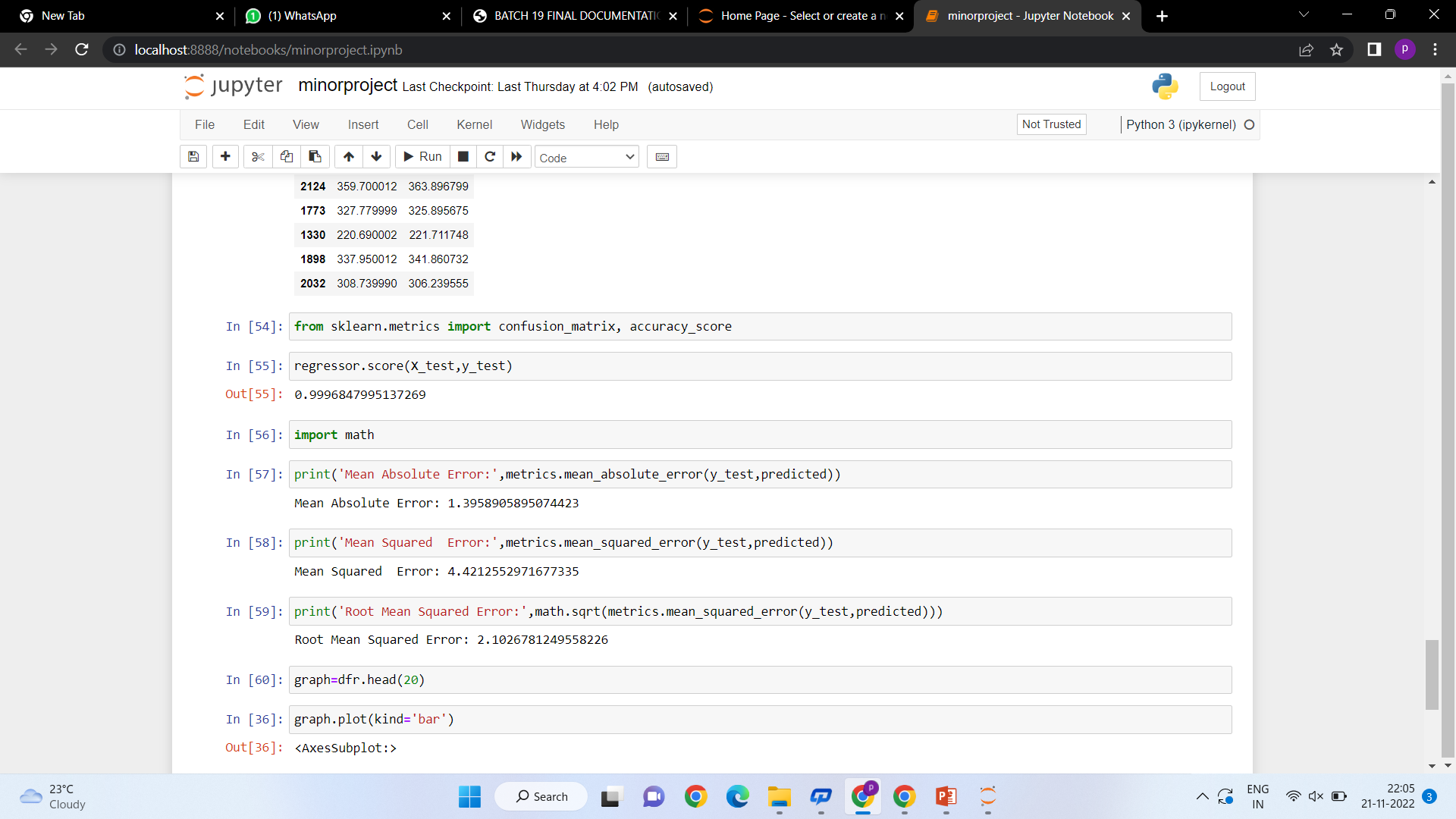
**Fig 7.6 Displaying the Data Frame**

31

****

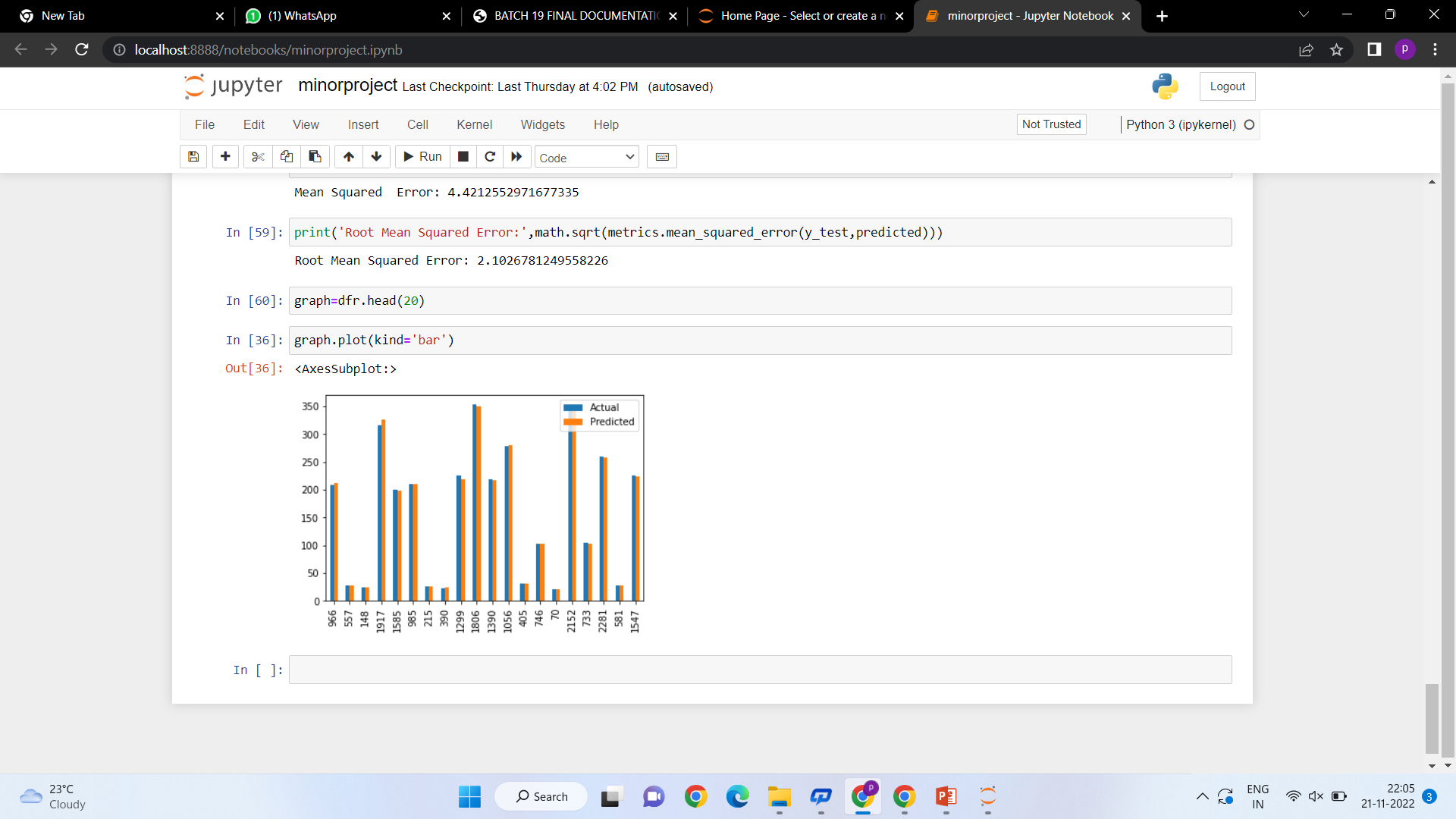
**Fig 7.7 Displaying the Actual and Predicted Values**

32

****

**Fig 7.8 Checking Accuracy**

33

****

**Fig 7.9 Bar chart**

34



CHAPTER-8

### CONCLUSION

### In this project, we have compared the actual value with the predicted value by applying

### the linear regression algorithm for the datasets belonging to kaggle to obtain the accuracy result. The changes in the tesla stock market is not always be in a regular pattern or not always follow the continuous cycle.It seems like Tesla’s stock prices will decrease in the coming future if they don’t come up with a new idea of representing their vision. This may be possible as other companies have also started manufacturing electric vehicles at a very low price as compared to Tesla.The successful prediction of a stock's future price will maximize investor’s gains.

**FUTURE WORK:**

In the future, for better accuracy model can be trained with more varied and detailed data.Also, other algorithms along with proposed can be used to create a new hybrid model

35



CHAPTER-9

### REFERENCES

### [1] Zhen Hu, Jibe Zhu, and Ken Tse “Stocks Market Prediction Using Support Vector Machine”, 6th International Conference on Information Management, Innovation Management and Industrial Engineering, 2013.M.

### [2] Wei Huang, Yoshiteru Nakamori, Shou-Yang Wang, “Forecasting stock market movement direction with support vector machine”, Computers & Operations Research, Volume 32, Issue 10, October 2005, Pages 2513– 2522.

### [3] N. Ancona, Classification Properties of Support Vector Machines for Regression, Technical Report, RIIESI/CNRNr. 02/99.

### [4] K. jae Kim, “Financial time series forecasting using support vector machines,” Neurocomputing, vol. 55, 2003.

### [5] Debashish Das and Mohammad shorif Uddin data mining and neural network techniques in stock market prediction: a methodological review, international journal of artificial intelligence & applications, vol.4, no.1, January 2013 [2] Pradeep K Sinha, Vrushali Y Kulkarni “Efficient Learning of Random Forest Classifier using Disjoint Partitioning Approach” Proceedings of the World Congress on Engineering 2013 Vol II, WCE 2013, July 3 - 5, 2013, London, U.K [3] Grahn H, Lavesson N, Lapajne M, Slat D, A CUDA implementation of Random Forest – Early Results, Master Thesis Software Engineering, School of Computing, Blekinge Institute of Technology, Sweden.

36