

DETAILED OVERVIEW OF MACHINE LEARNING

ALGORITHMS

Major project report submitted in partial fulfilment of the requirement for the degree of

Bachelor of Technology

in

Computer Science and Engineering

By

HONEY GUPTA (181257)

UNDER THE SUPERVISION OF



Dr. VIVEK KUMAR SEHGAL

Department of Computer Science & Engineering and Information

Technology

Jaypee University of Information Technology, Wagnaghat, 173234, Himachal Pradesh,

INDIA

DECLARATION

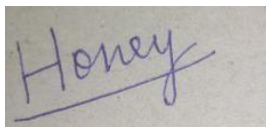
I hereby declare that, this “Detailed overview of Machine Learning algorithms” project has been done by me under the supervision of **(Dr Vivek Kumar Sehgal, Associate Professor (CSE/IT))**, Jaypee University of Information Technology.

I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

A small, handwritten signature in blue ink that reads "Vivek Sehgal". The signature is written in a cursive style and is underlined with two parallel lines.

Supervised by:
(Dr Vivek Kumar Sehgal)
Associate Professor (CSE/IT)

Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology

A rectangular stamp containing a handwritten signature in blue ink that reads "Honey". The signature is written in a cursive style and is underlined with a single line.

Submitted by:
(Honey Gupta)
(181257)

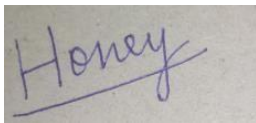
Computer Science & Engineering Department
Jaypee University of Information Technology

CERTIFICATE

This is to certify that the work which is being presented in the project report titled
“DETAILED OVERVIEW OF MACHINE LEARNING ALGORITHMS”

in partial

fulfilment of the requirements for the award of the degree of B Tech in Computer Science and Engineering submitted to the Department of Computer Science and Engineering, Jaypee University of Information Technology, Wagnaghat is an authentic record of work carried out by “Honey Gupta (181257)” during the period from January 2022 to May 2022 under the supervision of **Dr Vivek Kumar Sehgal**, Department of Computer Science and Engineering, Jaypee University of Information Technology, Wagnaghat.

A rectangular box containing a handwritten signature in blue ink that reads "Honey".

Honey Gupta
(181257)

The above statement made is correct to the best of my knowledge.

(Dr Vivek Kumar Sehgal)
Associate Professor
(CS/IT)

Computer Science & Engineering and Information Technology
Jaypee University of Information Technology, Wagnaghat,

AKCNOWLEDGEMENT

Firstly, I express my heartiest thanks and gratefulness to almighty God for his divine blessing makes us possible to complete the project work successfully.

I am really grateful and wish my profound my indebtedness to Supervisor Dr Vivek Kumar Sehgal, **Associate Professor (CSE/IT)** Department of CSE Jaypee University of Information Technology, Waknaghat. Deep Knowledge & keen interest of my supervisor in the field of “**DETAILED OVERVIEW OF MACHINE LEARNING ALGORITHMS**” to carry out this project.

His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to, Dr Vivek Kumar Sehgal Department of CSE, for his kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straight forwardly or in a roundabout way in making this project a win.

In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

TABLE OF CONTENT

INDEX

CONTENT	Page
LIST OF FIGURES	
LIST OF TABLES	
ABSTRACT	
1. INRODUCTION...	1
DATA SCIENCE	2
DATA SCIENCE PROCESS	3
STATISTICS	5
METHODOLOGY	22
2. LITERATURE SURVEY.	30
FEASIBILITY STUDY ON THE MAJOR PROJECT ...	30
REQUIREMENTS.....	31
TOOLS AND TECHNOLOGIES	31
3. SYSTEM DEVELOPMENT... ..	32
BENCH MARK DATA SETS	32
DATA SET FEATURES AND DEFINITIONS	33
THE PROPOSED APPROACH.....	34
EXPLORATORY DATA ANALYSIS	34

4. PERFORMANCE ANALYSIS.....	38
SUPPORT VECTOR MACHINE... ..	38
K NEAREST NEIGHBOR... ..	40
SCREEN SHOTS OF THE VARIOUS STAGES.....	43
DISCOVERING/ANALYZING THE DATA.....	43
VALIDATION AND PREPROCESSING.....	44
EXPLORATION AND VISUAL ANALYSIS	44
DATA SCALING	45
TRAIN-TEST SPLIT... ..	45
5. CONCLUSION.....	46
RESULTS AND DISCUSSIONS	46
FUTURE SCOPE	46

ABSTRACT

Data science is an indispensable area of interests when it comes on Information Technology sector. Currently, there is a huge boom in IT and finance market. Billions of data are being circulated, accumulated, utilized by different IT sectors. Data science is the scientific process of transforming data into meaningful insights in order to draw better decisions.

Different components of data science are:

Statistics or business statistics

Data visualization

Machine learning.

Thus, this project deals around machine learning algorithms and Exploratory Data Analysis in a much more depth. By seeing its implementation, we will get the meaningful insights about various algorithms and its functions. Hence machine learning is a subset of Data Science as its name suggests, “Machine is Learning”. And its various techniques are as follows, like Decision trees, Random forest, K nearest neighbor, Support Vector Machine, Neural networks using Keras. All of these features required applying data mining and machine learning techniques to real-world datasets and that's what this is all about. Data science, artificial intelligence, and machine learning are the most valuable technical skills to have right now. These fields are exploding with progress and new opportunities.

Later this project, there we ought to see detailed implementation and explanation of various machine learning techniques on different datasets such as: Zomato, Titanic etc. Coming onto various algorithms, we will see which one is the best suited for different datasets and much more. And for every algorithm there is a practical applied standpoint and some actual Python code to run so we can see it in action and refer to it later.

Machine Learning and artificial intelligence (AI) is everywhere; if we want to know how companies like Google, Amazon, and even Udemmy extract meaning and insights from massive data sets, the Data Science will give us the fundamentals we need. Data Scientists enjoy one of the top-paying jobs, with an average salary of \$120,000 according to Glassdoor and Indeed.

Keywords: Support vector machine, exploratory data analysis, k nearest neighbor, principal component analysis.

Supervised and Unsupervised machine learning algorithms.

CHAPTER 01

INTRODUCTION

Data Science

Data Science or Data Driven Science is one of the most essential tools in today's IT sector. Thus, it is used in financial systems, bank systems, architecture, industries, automobiles sectors, health and sanitation sectors, social media platforms such as Instagram, Facebook, LinkedIn, Twitter mostly everywhere in the world as far as data is concern we ought to study the data science in detail. Data Science is secured that makes it impossible to be counterfeited but most difficult to handle as every year, the world produces tons and tons of data across the globe.

Thus, data are used by Apple's Siri and Google's voice search and much more platforms. Data analysis or Analysis or Business analysis performs on past business performances and get insights for future business performances. Thus, 3V's of BIG data are Volume, Velocity, Variety. Data Science supports and encourages shifting between deductive which is hypothesis-based reasoning and inductive which is pattern based reasoning.

Data Science and machine learning algorithms built a vast complex in computer science fields. It's algorithms that remembers its input due to its internal memory, which makes the algorithm perfectly suited for solving machine learning problems involving sequential data. It is one of the algorithms that have great results also in deep learning.

Thus, deep learning is another area of implementing machine learning algorithms. Majorly used for Artificial Intelligence applications which can be used to solve any classification or regression problem world wide. It is most prominent uses neural networks which is hence used to build deep models which thus understands complex relations among most of the variables.

Data Science Process

Since globally looking, there is ton and tons of data being floating in the internet across the planet. Thus, data is collected from the internet and various sites and also from top MNCs such as Amazon, Flipkart, Tesla, Google, Microsoft, Apple and thousands of more likewise. The spread of data in the internet, measured by its rapid market capitalization growth and immense emerges the appreciation in the Data Science fields, led to the emergence of a large number of other forms of it.

So, this data in turn is called the raw data and it is therefore called raw data as it is blindly or eventually or completely taken from the internet source or Wikipedia or any other search engines such as Google chrome, Microsoft Edge, Yahoo, Internet Explorer.

Then this raw data is formatted or utilized or is being processed. Then moving onto further the next step is cleaning dataset and all these Exploratory Data Analysis. And cleaning dataset and exploratory Data Analysis which in turns get into models or algorithms. Hence the models or algorithms , in turn focuses on the next step in doing communication visualize reports and in data products which again gets dissolved in the internet source.

In turn the communicate visualize report will hence make effective and better decisions which will thus be helpful and effective for the world and in the future for future purposes.

Hence machine learning also plays a major role and a very crucial part of data science and so in its process as one of the fundamentals behind machine learning concepts are test and training of the data, and thus machine learning is basically an algorithm that can learn from observational data. And can make predictions based on it.

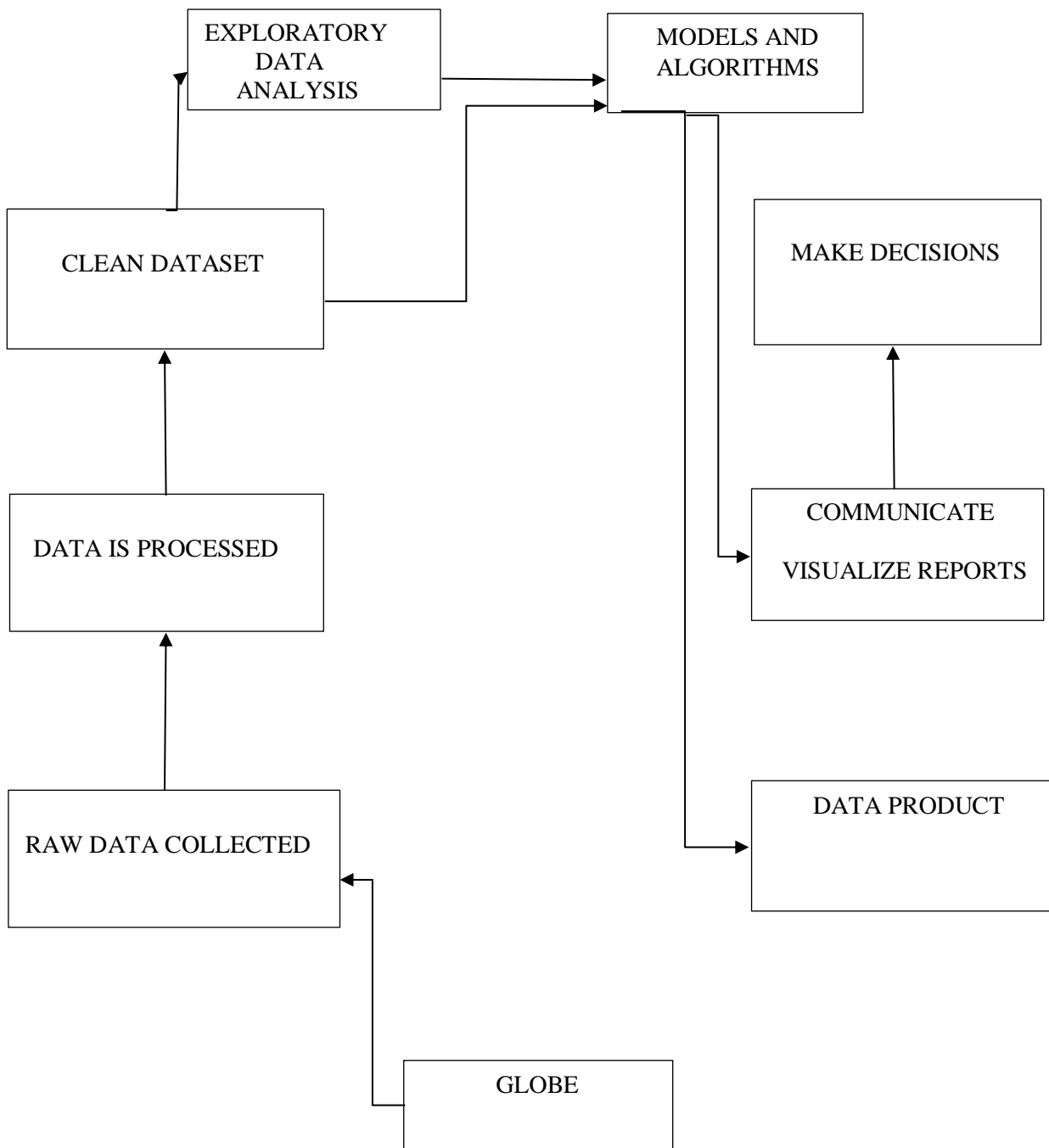


Fig 1: Process of Data Science

Since we know that there are three main components of Data Science and they are as follows:

- Statistics or Business Statistics
- Data Visualization
- Machine Learning

- **Statistics**

Why do we need to be aware of Traditional Statistics?

Firstly introducing term statistics as it is very crucial topic when coming onto data science and doing machine learning algorithms on various datasets. To understand the shape and distribution of our data in the datasets to choose the appropriate statistical analysis methods. Dispersion in data analysis and explains these measures in the business or process context. Hence calculating measures of dispersions such as range, interquartile range, standard deviation and variance.

Thus, Statistics is a collection, organization, analysis and interpretation of data. Hence statistics or business statistics includes

- Design of experiments
- Sampling
- Descriptive statistics
- Inferential statistics
- Probability theory

Thus, we therefore prefer Inferential statistics over Traditional Statistics because as Inferential statistics tells us the relationship between variables and helps us to get outcome or come out on evidence based approach and this approach is better than approximate approach i.e. Traditional Statistics and also the main reason is because of Hypothesis.

And knowledge of statistics helps us to make decisions based on the data collected, also read, evaluate, and interpret the results. Develop systematic approach to analyze data and also evaluate the informative information we have been given. As discussed, it is the collection or bundle or interpretation of data. And thus, the knowledge of statistics will help us to make effective and better some data driven decisions and also helps us to evaluate the information being given to us. Statistics helps us and decision makers to make decisions and also for various purposes like it is to present and describe business data and information accurately, properly and effectively. Secondly it is used to draw attentions or conclusions about large sets of data or datasets using information collected from subsets. Thirdly it is used to make reliable forecasts about a business activity and hence also improves business processes.

Types Of Statistics are Descriptive Statistics and Inferential Statistics, thus statistics is the branch of mathematics that transforms data into useful information for decision makers. Descriptive Statistics is used to collect, organize and characterize the present raw data. And Inferential Statistics used to make inferences and much more such as hypothesis testing and also determine relationships and also make predictions.

Descriptive Statistics

are methods for organizing and summarizing data. Example: Tables or graphs are used and descriptive values such as the average score are used to summarize data. A descriptive value for a population is called a parameter and a descriptive value for a sample is called a statistic.

Inferential Statistics are methods for using sample data to make general conclusions or inferences about populations. Sample is typically only a part of the whole population, sample data provide only limited information about the population. Thus, as a result, sample statistics Are generally imperfect representatives of the corresponding population parameters.

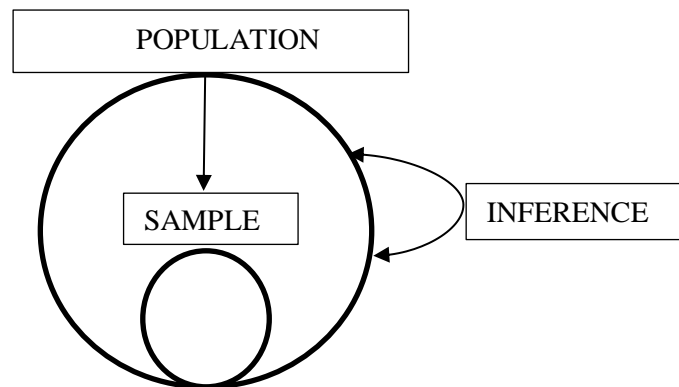


Fig 2: Vienn Diagram Between Population and sample

Population Vs Sample:

Population consists of all the items or individuals about which we want to draw a conclusion. And thus, population is entire group having studied. Whereas the random selection from population is called sample. Hence a Sample is the portion of population selected for analysis and thus, is a subset of the population that is being surveyed.

Classification of data or variables or the level of data analysis, thus, variables are of two types Categorical and other one is Numerical or continuous. Categorical variables are those whose values that can only be placed into categories such as “yes”, “no”, that’s it.

Whereas numerical variables have values that represent quantities such as discrete and continuous. In discrete we have countable elements that can be counted or be numbered like age of children or farmers surveyed in the report or number of working days of an employee. Whereas in the continuous columns it is those datasets that are continuous or uncountable like weight, height, blood sugar rate, something measurable characteristics.

Thus, a variable is a characteristics or condition that can change or take on different values. There are some descriptive statistics key concepts as location or central tendency, dispersion or spread, shape or distribution. Hence descriptive statistics helps in finding potential data sets problems such as mean value errors or mean squared errors or outliers etc. And also helps is finding process issues. Moreover, selecting of appropriate statistical test for understanding the underlying relationship or patterns. Descriptive statistics involves measures of central tendency or called summary statistics which contains mean, median, mode. Next we have measures of spread which involves range, quartiles, percentiles, absolute deviation, variance and standard deviation. Measures of symmetry involves skewness. Measures of Peaked ness involves kurtosis. Measures of central tendency, thus the goal of measures of the central tendency is to come up with one single number that best describes a distribution of scores. There are three basic measures of central tendency, and are mode, median, mean. The scale of measurement used, so that summary makes sense given the nature of the scores. The shape of the frequency distribution so that it measures accurate and hence nicely summarize the distribution.

Mean is the average of all the datasets taken into consideration and it is being calculated by summing up each and every data being encountered during calculation of it mean. And there are various types of mean like simple mean, arithmetic mean, geometric mean, harmonic mean weighted mean, thus in weighted mean choice of weights depends on the applications and the weighted mean computations, quantities ex dollars or volumes are hence thus frequently used.

In weighted mean we have numerator as sum of weighted data values and in denominator we have sum of the weights. Next example of mean or type of mean we have Trimmed mean which is used when extremes values are present is called the trimmed mean. It is generated by eliminating a percentage of the smallest and largest values from a data set and then hence computing and calculating the mean of the remaining values. For let's say 12% trimmed mean is evolved by removing the smallest 12% and the biggest 12% of the data values and then computing the mean of the remaining values. And then we have geometric mean which is again computed by getting the m th root of the product of m values.

It is often used in analyzing growth rates in financial data where using the arithmetic mean will provide misleading results. Hence it should be applied anytime when we want to determine the mean rate of change over several successive ranges. Its more usages are that in contains populations of species, crop yields, pollution levels and birth and death rates.

Median is the finding out the middle most element from the dataset. The median is the midpoint of the ordered dataset. It is not as popular as the mean, but is often used in academia and data science. That is since it is not affected by outliers. In an ordered dataset, the median is the number at position $\frac{n+1}{2}$. If this position is not a whole number, it, the median is the simple average of the two numbers at positions closest to the calculated value. In Excel, the median is calculated by: =MEDIAN(). The mode is the value that occurs most often. A dataset can have 0 modes, 1 mode or multiple modes. The mode is calculated simply by finding the value with the highest frequency. In Excel, the mode is calculated by: =MODE.SNGL() -> returns one mode =MODE.MULT() -> returns an array with the modes. It is used when we have more than 1 mode.

Variance and standard deviation measure the dispersion of a set of data points around its mean value. There are different formulas for population and sample variance & standard deviation. This is due to the fact that the sample formulas are the unbiased estimators of the population formulas.

Covariance is a measure of the joint variability of two variables.

- A positive covariance means that the two variables move together.
 - A covariance of 0 means that the two variables are independent.
 - A negative covariance means that the two variables move in opposite directions.
- Covariance can take on values from $-\infty$ to $+\infty$. This is a problem as it is very hard to put such numbers into perspective.

Correlation is a measure of the joint variability of two variables. Unlike covariance, correlation could be thought of as a standardized measure. It takes on values between -1 and 1, thus it is easy for us to interpret the result.

- A correlation of 1, known as perfect positive correlation, means that one variable is perfectly explained by the other.
- A correlation of 0 means that the variables are independent.
- A correlation of -1, known as perfect negative correlation, means that one variable is explaining the other one perfectly, but they move in opposite directions.

Skewness is a measure of asymmetry that indicates whether the observations in a dataset are concentrated on one side. Right (positive) skewness looks like the one in the graph. It means that the outliers are to the right (long tail to the right). Left (negative) skewness means that the outliers are to the left. Usually, we will use software to calculate skewness.

Graphs and tables for relationships between variables (Scatter plots)

When we want to represent two numerical variables on the same graph, we usually use a scatter plot. Scatter plots are useful especially later on, when we talk about regression analysis, as they help us detect patterns (linearity, homoscedasticity). Scatter plots usually represent lots and lots of data. Typically, we are not interested in single observations, but rather in the structure of the dataset.

A scatter plot that looks in the following way (down) represents data that doesn't have a pattern. Completely vertical 'forms' show no association. Conversely, the plot above shows a linear pattern, meaning that the observations move together.



Fig 3: Scatter Plot

Graphs and tables for relationships between variables (Cross tables)

Cross tables (or contingency tables) are used to represent categorical variables. One set of categories is labeling the rows and another is labeling the columns. We then fill in the table with the applicable data. It is a good idea to calculate the totals.

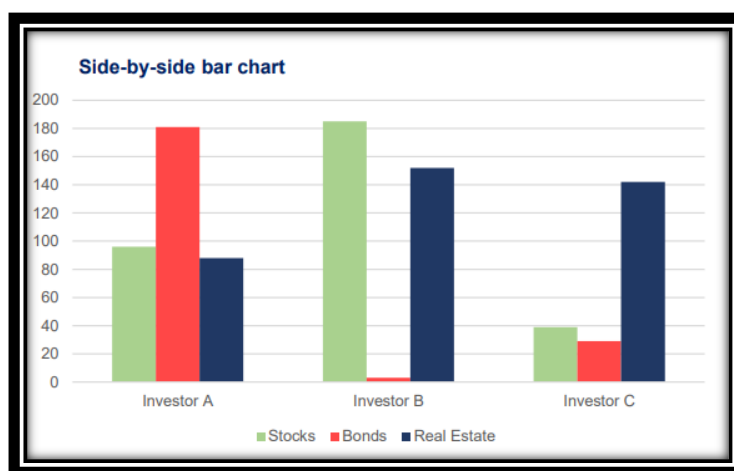


Fig 4: **Bar Chart**

A common way to represent the data from a cross table is by using a side-by-side bar chart.

Creating a side-by-side chart in Excel:

1. Choose your data
2. Insert -> Charts -> Clustered Column Selecting more than one series (groups of data) will automatically prompt Excel to create a side-by-side bar (column) chart. Thus, Frequency distribution tables for numerical variables are different than the ones for categorical. Usually, they are divided into intervals of equal (or unequal) length. The tables show the interval, the absolute frequency and sometimes it is useful to also include the relative (and cumulative) frequencies.

INFERENCEAL STATISTICS

In statistics, when we talk about distributions, we usually mean probability distributions. Definition (informal): A distribution is a function that shows the possible values for a variable and how often they occur. Definition (Wikipedia): In probability theory and statistics, a probability distribution is a mathematical function that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment. Examples: Normal distribution, Student's T distribution, Poisson distribution, Uniform distribution, Binomial distribution. It is a common mistake to believe that the distribution is the graph. In fact, the distribution is the 'rule' that determines how values are positioned in relation to each other. Very often, we use a graph to visualize the data. Since different distributions have a particular graphical representation, statisticians like to plot them.

The Normal distribution is also known as Gaussian distribution or the Bell curve. It is one of the most common distributions due to the following reasons:

- It approximates a wide variety of random variables
- Distributions of sample means with large enough samples sizes could be approximated to normal
- All computable statistics are elegant
- Heavily used in regression analysis
- Good track record.

The Standard Normal distribution is a particular case of the Normal distribution. It has a mean of 0 and a standard deviation of 1. Standardization allows us to compare different normally distributed datasets, detect normality, detect outliers, create confidence interval, test hypotheses • perform regression analysis. The Central Limit Theorem (CLT) is one of the greatest statistical insights. It states that no matter the underlying distribution of the dataset, the sampling distribution of the means would approximate a normal distribution.

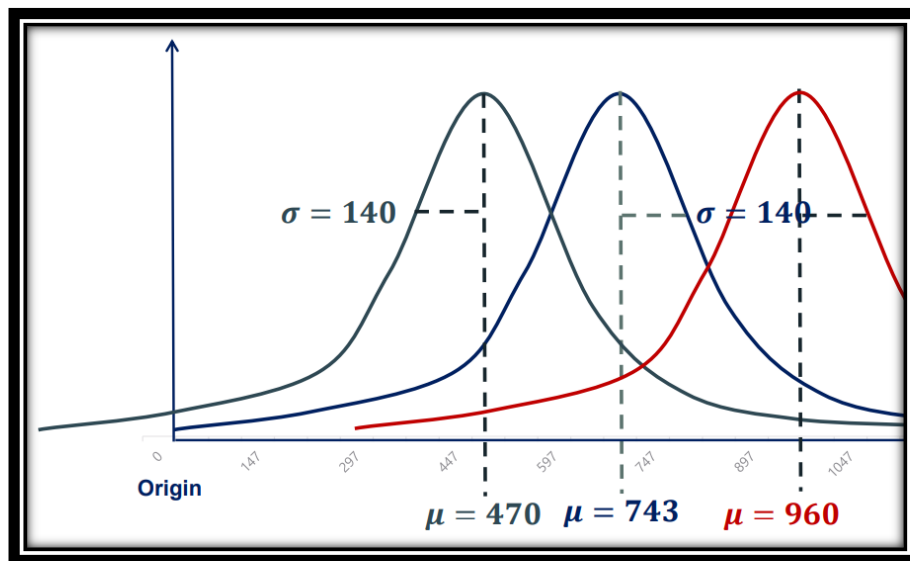


Fig 5: Normal Distribution

Moreover, the mean of the sampling distribution would be equal to the mean of the original distribution and the variance would be n times smaller, where n is the size of the samples.

The CLT applies whenever we have a sum or an average of many variables (example: sum of rolled numbers when rolling dice). Broadly, an estimator is a mathematical function that approximates a population parameter depending only on sample information. An estimate is the output that you get from the estimator (when you apply the formula). There are two types of estimates: point estimates and confidence interval estimates.

A confidence interval is an interval within which we are confident (with a certain percentage of confidence) the population parameter will fall. We build the confidence interval around the point estimate. The Student's T distribution is used predominantly for creating confidence intervals and testing hypotheses with normally distributed populations when the sample sizes are small. It is particularly useful when we don't have enough information or it is too costly to obtain it.

All else equal, the Student's T distribution has fatter tails than the Normal distribution and a lower peak. This is to reflect the higher level of uncertainty, caused by the small sample size. We can obtain the student's T distribution for a variable with a Normally distributed population.

Data Visualization

Data visualization is basically the pictorial representation of the data and datasets. As it has charts and data types and there is never only one visualization. There are many of its types which are as follows:

- Numerical and categorical
- Numerical and numerical
- Time series

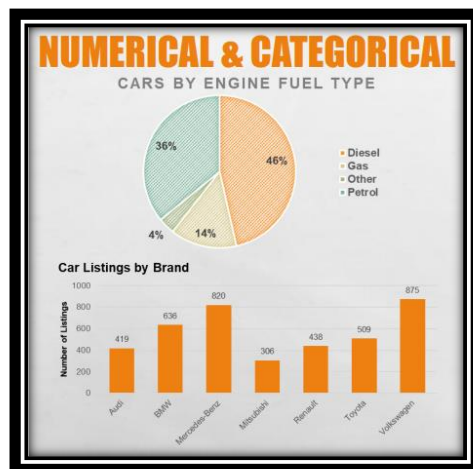


Fig 6: Numerical and categorical variables

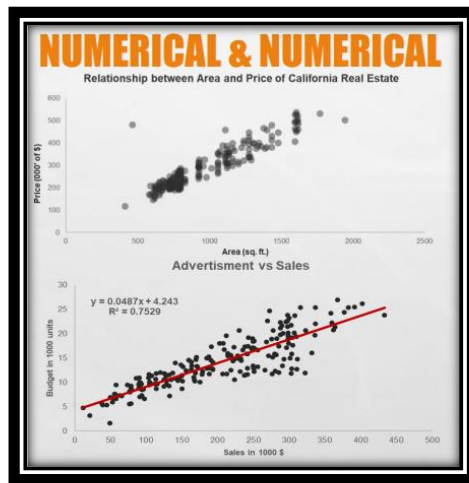


Fig 7: Numerical & Numerical Variables

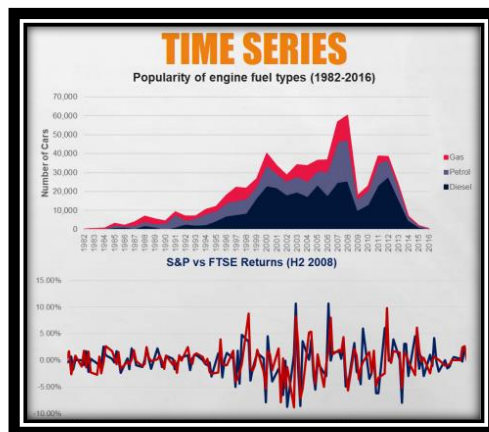


Fig 8: Time Series

For pictorial representation and for it we have colors as it can be taken from predetermined as company colors and client predefined or desirable colors thus, we can use any templates to design our charts and graphs. Secondly, we have online tools by using this we can design our own color palettes according to our desires with aid of online tools, and color online palettes.

BAR CHART

Bar chart is hence intuitive, appropriate for non-technical audiences, one of the most commonly used charts.

Thus, bar chart is used to communicate our intentions clearly, make sure our chart isn't misleading.

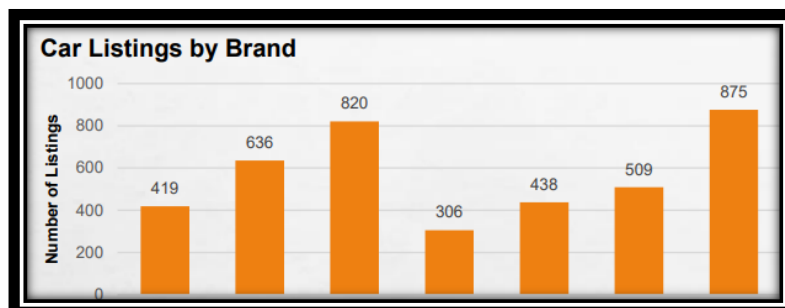


Fig 9: Bar Chart

PIE CHART

- Appropriate for non-technical audiences
- Widely used, despite criticism
- A few categories
- Data sums up to 100%
- Don't use when data \neq 100%
- Don't use when there are too many categories
- No 3d or doughnut

STACKED AREA CHART

Compare volume among features • at least three features • ordering for at least two of them • time series data • avoid when you have too many categories – a line chart works better • avoid

with categories of similar size – difficult to determine size of non-rectangular shapes • order categories by size – to improve readability • y-axis must start at 0 – we’re measuring volume.

LINE CHART

Up to several categories • time series data • y-axis doesn’t have to start at 0 • when you have a large period of time, narrow it down to gain more insight • be careful not to include too many categories, to avoid a spaghetti chart.

HISTOGRAM

Distribution of a numeric variable • the variable’s range of values is split into intervals or bins • y-axis –number of observations within each interval(or density) • similar to a bar chart, no gap between bins • to create a histogram • determine the interval size • choose the number of bins.

SCATTER PLOT

• Displays each point from the data, instead of showing aggregated form • shows relationship between variables • use transparency to avoid overplotting • a third variable could be used with a color parameter.

REGRESSION PLOT

Used to determine relationships between predictor(s) and outcome • regression line & equation help us quantify the relationship • there exist many types of relationships between variables • sometimes there is no apparent relationship between features.

BAR AND LINE CHART

Combination chart of bar chart & line chart • dual y-axis • it’s essential for the two charts to be

well labeled • pareto chart is a specific case of bar & line chart • the secondary y-axis on a pareto shows the cumulative frequency and sums up to 100%.

Graphics are used a great deal in many different fields, and one might expect more progress to have been made along theoretical lines. Sometimes in science the theoretical literature for a subject is considerable while there is little applied literature to be found. The literature on data visualization is very much the opposite. Examples abound in almost every issue of every scientific journal concerned with quantitative analysis. There are occasionally articles published in a more theoretical vein about specific graphical forms, but little else.

Machine Learning

By comparing well-established machine learning models trained on datasets against each other. Our study provides two main contributions. First, we contribute to the literature by systematically comparing the predictive capability of different prediction models (e.g., K nearest neighbor, support vector machine, principal component analysis, random forest, decision trees, and most importantly neural networks, feature sets (e.g., technical, blockchain-based), and prediction horizons. Thereby, our study establishes a thorough benchmark for the predictive accuracy by comparing different machine learning algorithms. The general picture emerging from the analysis is that decision trees appear well-suited for this prediction problem and technical features remain most-relevant.

We hope to use Machine Learning Algorithms which also are widely utilized by many organizations in. This report will walk through a simple implementation of analyzing and forecasting the datasets by using various Machine Learning Algorithms.

Basic components of learning process are the machine learning process, whether by a human

or a machine, can be divided into four components, namely, data storage, abstraction, generalization and evaluation.

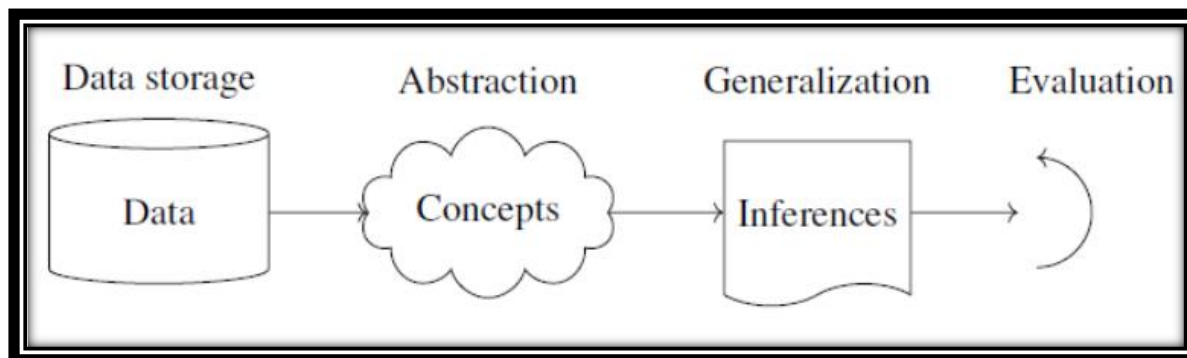


Fig 10: **Components of Machine learning process**

Application of Machine Learning

Application of machine learning methods to large databases is called data mining. In data mining, a large volume of data is processed to construct a simple model with valuable use, for example, having high predictive accuracy. The following is a list of some of the typical applications of machine learning.

1. In retail business, machine learning is used to study consumer behavior.
2. In finance, banks analyze their past data to build models to use in credit applications, fraud detection, and the stock market.
3. In manufacturing, learning models are used for optimization, control, and troubleshooting.

Machine learning algorithms can be classified into three types.

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised learning is a training set of examples with the correct responses (targets) is provided and, based on this training set, the algorithm tends to respond correctly to all possible inputs. This is also called learning from exemplars. Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pair.

Unsupervised learning Correct responses are not provided, but instead the algorithm tries to identify similarities between the inputs so that inputs that have something in common are categorized together. The statistical approach to unsupervised learning is known as density estimation.

Reinforcement learning This is somewhere between supervised and unsupervised learning. The algorithm gets told when the answer is wrong, but does not get told how to correct it. It has to explore and try out different possibilities until it works out how to get the answer right.

METHODOLOGY

Application of machine learning methods to large databases is called data mining. In data mining, a large volume of data is processed to construct a simple model with valuable use, for example, having high predictive accuracy. The following is a list of some of the typical applications of machine learning.

1. In retail business, machine learning is used to study consumer behavior.
2. In finance, banks analyze their past data to build models to use in credit applications, fraud detection, and the stock market.
3. In manufacturing, learning models are used for optimization, control, and troubleshooting.

As a result, the paper aims to achieve the following by using deep learning algorithms, which can discover hidden patterns from data, integrate them, and create far more efficient predictions.

Simple Linear Regression in Machine Learning

Simple Linear Regression is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable. The relationship shown by a Simple Linear Regression model is linear or a sloped straight line, hence it is called Simple Linear Regression. The key point in Simple Linear Regression is that the dependent variable must be a continuous/real value. However, the independent variable can be measured on continuous or categorical values.

Multiple Linear Regression

But there may be various cases in which the response variable is affected by more than one predictor variable; for such cases, the Multiple Linear Regression algorithm is used. Moreover, Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable.

We can define it as: “Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.”

Logistic Regression

It is used to find casual affect relationship between independent variables and dependent variables. In which dependent variables can either be binary categorical or multinomial or multiclass categorical.

Type of dependent variable is binary which is categorical or multinomial. And type of independent variable is categorical and continuous.

Discriminant Analysis

It is used to find most discriminant independent variables. And the type of dependent variable is categorical in nature and type of independent variable is continuous in nature.

Decision Tree

We can actually construct a flow chart to help us decide a classification for something with machine learning. And it is called decision tree. Another form of supervised learning and gives it some sample data and the resulting classifications. That comes out as a tree. Its example as we want to build a system to filter out resumes based on historical hiring data. And we have a database of some important attributes of job candidates, and we know which ones were hired and what ones weren't. We can also train a decision tree on this data and arrive at a system for predicting whether a candidate will get hired based on it.

Decision tree works at each step, we find the attribute we can use to partition the data set to minimize the entropy of the data at the next step. Fancy term of this sample algorithm is ID3. It is greedy algorithm as it goes down the tree, it just picks the decision that reduce the entropy the most at that stage. That might not result in an optimal tree but it works. It is used to classify the records in a pictorial format example hierarchical etc. And the type of dependent variable is categorical in nature and type of independent variable is categorical and continuous or numerical in nature.

Random Forest

In random forest, decision trees are very susceptible to overfitting and to fight this thing, we can construct several alternate decision trees and let them “vote” or the final classification.

- Randomly re-sample the input data for each tree or fancy term for this is bootstrap aggregating or bagging
- Randomize the subset of the attributes each step is allowed to choose from.

Naïve Bayes

Naïve bayes is a classification technique used to classify records in the posterior property and type of dependent variable is categorical in nature and type of independent variable is categorical in nature. Thus, talking about Bayes theorem which we studied in statistics. As how would we express the probability of an email being spam if it contains the word “free”. The numerator is the probability of a message being spam and containing the word “free”. The denominator is the overall probability of an email containing the word “free”.

Now we construct conditional probability of spam and word for every meaningful word we encountered during training. Then multiply these together when analyzing a new email to get the probability of it being spam. Assumes the presence of different words are independent of each other – one reason this is called “Naïve Bayes”. The CountVectorizer let us operate on lots of words at once. MultinomialNB does all the heavy lifting on Naïve Bayes.

K Nearest Neighbor

Type of dependent variable is categorical in nature and type of independent variable. This is used to classify records help of Euclidean distance.

Used to classify new data points based on “distance” to known data. Find the k nearest neighbors, based on our distance metric. And its better to go with the classification. Although it is one of the simplest machine learning models there is – it still qualifies as “supervised learning”.

K – Means Clustering

- Attempts to split data into k groups that are closest to k centroids.
- Unsupervised learning- uses only the positions of each data point.
- Can uncover interesting groupings of people, things or behavior.
- Randomly pick k centroids(k-means)
- Assign each data point to the centroid its closest to.
- Recompute the centroids based on the average position of each centroid’s points.
- Iterate until points stop changing assignment to centroids.
- If we want to predict the cluster for new points, just find the centroid they are closest to.

K – Means Clustering Gotchas

- Choosing K- Try increasing K values until we stop getting large reductions in squared error (distance from each point to their centroids).
- Avoiding local minima – The random choice of initial centroids can yield different results. Also run it a few times just to make our initial results aren’t wacky
- Labeling the clusters – k-means does not attempt to assign any meaning to the clusters we find, and it’s up to us to dig into the data and try to determine that. Unsupervised learning- uses only the positions of each data point.

The obtained dataset was then averaged into one dataset for consistency and in order to fill in the gaps created by missing data in dataset. Building Neural Network Model Machine Learning is the most suitable technique. The model to be built had to achieve several goals in order to produce a near to accurate prediction. This included selecting the framework which could produce a good prediction accuracy, take in consideration of other parameters in its prediction algorithm and be trainable.

Support Vector Machine

- It works well for classifying higher- dimensional data (lots of features)
- Finds higher- dimensional support vectors across which to divide the data (mathematically, these support vectors define hyperplanes. Needless to say that.
- Uses something called the kernel trick to represent data in higher- dimensional spaces to find hyperplanes that might not be apparent in lower dimensions.
- The important point is that SVM's employ some advanced mathematical trickery to cluster data, and it can handle data sets with lots of features.
- It's also fairly expensive – “the kernel trick” is the only thing that makes it possible.

Support Vector Classification

In practice we will use something called SVC to classify data using SVM. We can use different “kernels” with SVC. Some will work better than others for a given data set.

Below is the organization of our information.

UNSUPERVISED LEARNING ALGORITHMS

Factor Analysis

Factor analysis purpose is used for data reduction technique, reduce number of variables into number of factors with the help of eigen values which is input and output range.

Cluster Analysis

Cluster analysis purpose is a grouping technique used to group homogeneous characteristics of responses with the help of Euclidean distance.

Entropy

- A measure of a data set's disorder – how same or different it is.
- If we classify a data set into N different classes (example: a data set of animal attributes and their species)
- The entropy is 0 if all of the classes in the data are the same (everyone is an iguana)
- The entropy is high if they are all different
- Again, a fancy word for a simple concept.

Ensemble Learning

- Random forests are an example of ensemble learning.
 - It just means we use multiple models to try and solve the same problem, and let them vote on the results.
 - Random forest uses bagging (bootstrap aggregating) to implement ensemble learning.
 - Many models are built by training on randomly – drawn subsets of the data.
 - Boosting is an alternate technique where each subsequent model in the ensemble boosts attributes that address data mis-classified by the previous model.
-
- A bucket of models trains several different models using training data, and picks the one that works best with the test data.
 - Stacking runs multiple models at once on the data, and combines the results together.
 - This is how the Netflix prize was won.

Advanced Ensemble Learning

- Bayes Optimal Classifier
Theoretically the best- but almost always impractical
- Bayesian Parameter Averaging
Attempts to make BOC practical- but it's still misunderstood, susceptible to overfitting, and often outperformed by the simpler bagging approach.
- Bayesian Model Combination
Tries to address all of those problems
But in the end, it's about the same as using cross-validation to find the best combination of models.

Principal Component Analysis

Curse of dimensionality

Many problems can be thought of as having a huge number of “dimensions”. For example, in the recommending movies, the ratings vector for each movie may represent a dimension – every movie is its own dimension. That make us head hurt. Its tough to visualize. Dimensionality reduction attempts to distill higher- dimensional data down to a smaller number of dimensions, while preserving as much as the variance in the data as possible.

This is an example of a dimensionality reduction algorithm and it reduces data down to K dimensions. Thus, principal component analysis involves a fancy math- but at a high level. Finds the eigen vectors in the higher dimensional data and these define hyperplanes that splits the data while preserving the most variance in it. The data gets projected onto these hyperplanes, which represents the lower dimensions we want to represent an also a popular implementation of this is called singular value decomposition (SVD). Also, really useful for things like image compression and facial recognition.

Principal Analysis Component let us visualize this in 2 dimensions instead of 4, while still preserving the variance.

REINFORCEMENT LEARNING

- We have some sort of agent that “explores” some space.
- As it goes, it learns the values of different state changes in different conditions.
- Those values inform subsequent behavior of the agent. Example: Pac man and cat mouse game.
- Yields fast on-line performance once the space has been explored.

CHAPTER 02

LITERATURE SURVEY

FEASIBILITY STUDY ON MAJOR PROJECT

Machine learning (ML) is a type of artificial intelligence that can predict the future based on past data. And as we seen so far and discussed each algorithm in detail. We will further in this project see the complete exploratory data analysis on various data sets. And also see the implementations of various algorithms i.e. SVM (support vector machine, random forest, KNN (k nearest neighbor), and principal component analysis. ML-based models have various advantages over other forecasting models as prior research has shown that it not only delivers a result that is nearly or exactly the same as the actual result, but it also improves the accuracy of the result. Examples of machine learning also include neural networks (NN), support vector machines (SVM), and deep learning also. The authors demonstrate that incorporating and seeing various algorithms on different datasets into a portfolio improves its effectiveness in many ways. The authors of focus on timeseries data forecasting in particular and apply two machine learning algorithms, random forests (RF) and stochastic gradient boosting machine (SGBM). The results show that the MLensemble technique can be used to anticipate different values.

REQUIREMENTS

TOOLS AND TECHNOLOGIES TO BE USED:

- ❖ **Google Colab:** An open-source platform for solving problems. Includes variety of a comprehensive, flexible tools and multi-purpose libraries along with community resources that help researchers to develop ML models and easily build and deploy Machine Learning based powered applications.

- ❖ **Jupyter Notebook:** Jupyter Notebook provides the necessary visualization tools which are needed for machine learning model building. Also helps us in:
 - a) Keep track and visualize metrics like loss and accuracy
 - b) Model the graph and its various layers.
 - c) Develop histograms, boxplots, line graphs, bar charts, and other tensors which change over (i.e., include new and additional features) from time to time.

- ❖ **Pytorch:** PyTorch is an open-source platform for solving problems. Used for high level tensor computing which is generally done using python libraries like NumPy with strong acceleration via graphics processing units.

DESCRIPTION OF PROPOSED APPROACH

TABLE 1: DESCRIPTION OF PROPOSED APPROACH

Parameter Name	Description
Processor	Intel Core i7 processor, up to 3.8 GHz
Operating System	WINDOWS
RAM	RAM 16 GB
Graphics Processor	NVIDIA GeForce 930M

CHAPTER 03

SYSTEM DEVELOPMENT

IMPLEMENTATION OF VARIOUS ALGORITHMS ON DIFFERENT DATA SETS:

BENCHMARK DATA SET

Train Data set and Test Dataset. The dataset was freely available for use on the Internet.

Table 2: Train and Test Dataset

Parameter	Value/Description
Dataset Details	(Huge in number stated as Lone)
Memory usage	919.8 MB
Range Index	614 entries, 0 to 613
Data Columns in Total	13
Loan Id	integer
Gender	string
Married	bool
Dependent	Int64
Education	string
Self employed	bool
Applicant Id	int64
Co applicant Id	int64
Loan Amount	int64
Credit History	int64
Property Area	string
Loan Status	string

TRAIN DATASET

Loan_ID	Gender	Married	Dependen	Education	Self_Empl	ApplicantI	Coapplicar	LoanAmou	Loan_Amc	Credit_His	Property_#	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Gradu	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Gradu	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y

TEST DATASET

Loan_ID	Gender	Married	Dependen	Education	Self_Empl	ApplicantI	Coapplicar	LoanAmou	Loan_Amc	Credit_His	Property_Area
LP001015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban
LP001022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban
LP001031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban
LP001035	Male	Yes	2	Graduate	No	2340	2546	100	360		Urban
LP001051	Male	No	0	Not Gradu	No	3276	0	78	360	1	Urban
LP001054	Male	Yes	0	Not Gradu	Yes	2165	3422	152	360	1	Urban

Fig11: Train & Test Dataset

FEATURES AND THEIR DEFINITIONS

- **Loan Id:**
It is the loan Id of a particular employee.
- **Gender:**
The gender of a particular employee.
- **Married:**
Whether the particular employee is married or not.
- **Education:**
Whether the particular employee is graduated or not.
- **Self employed:**
Whether the particular employee is graduated or undergraduate

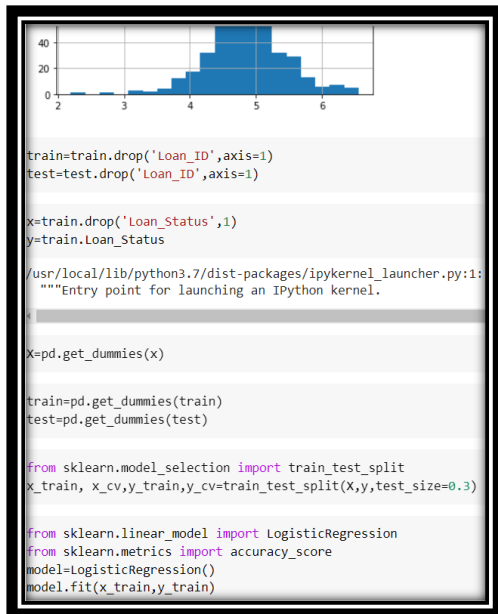
- Loan amount:
The amount of loan being taken by the particular employee.
- Property area:
Whether the property is in rural or urban area of that particular employee.
- Loan status:
Whether that particular employee has taken loan or not.

IMPLEMENTATION

Exploratory Data Analysis

Performing a detailed Exploratory Data Analysis on this Train and Test dataset. When we deal with machine learning we segregate it into dependent or output variable and input or predictors variables. In case of continuous variables, we need to understand central tendency and dispersion of variable. And in case of scatter plot it will not show correlation, for that we need to go to regression. Thus, in regression range is from 0 to 1. When data is normally distributed then mean is equal to median is equal to mode is equal to skewness which is 0.





RESULT:

Thus, after performing detailed and complete exploratory data analysis we see that accuracy score and the model is 76% accurate.



Fig 12: EDA (Exploratory data analysis)

THE PROPOSED APPROACH

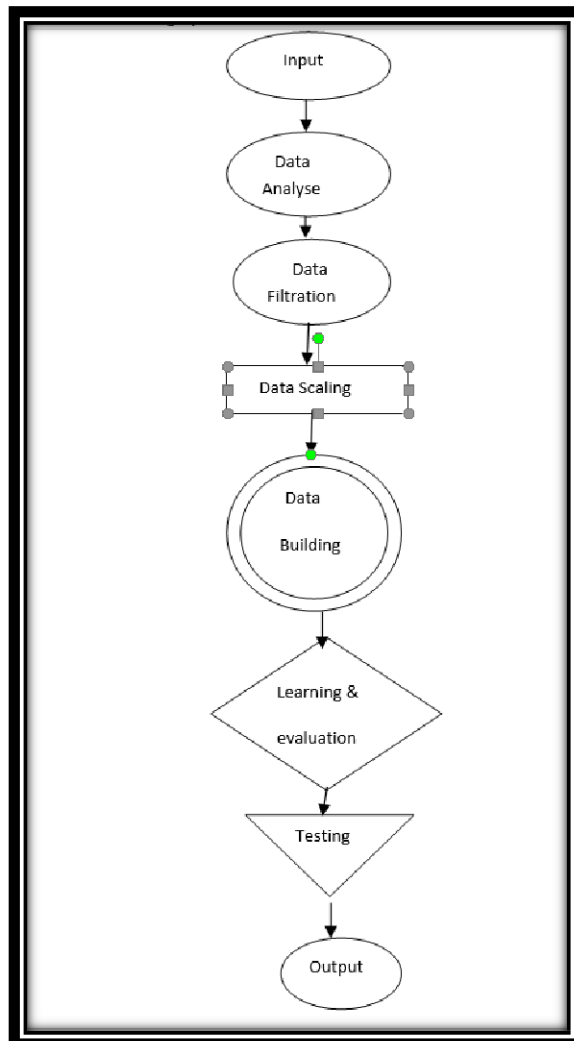


Fig 13: Proposed Approach

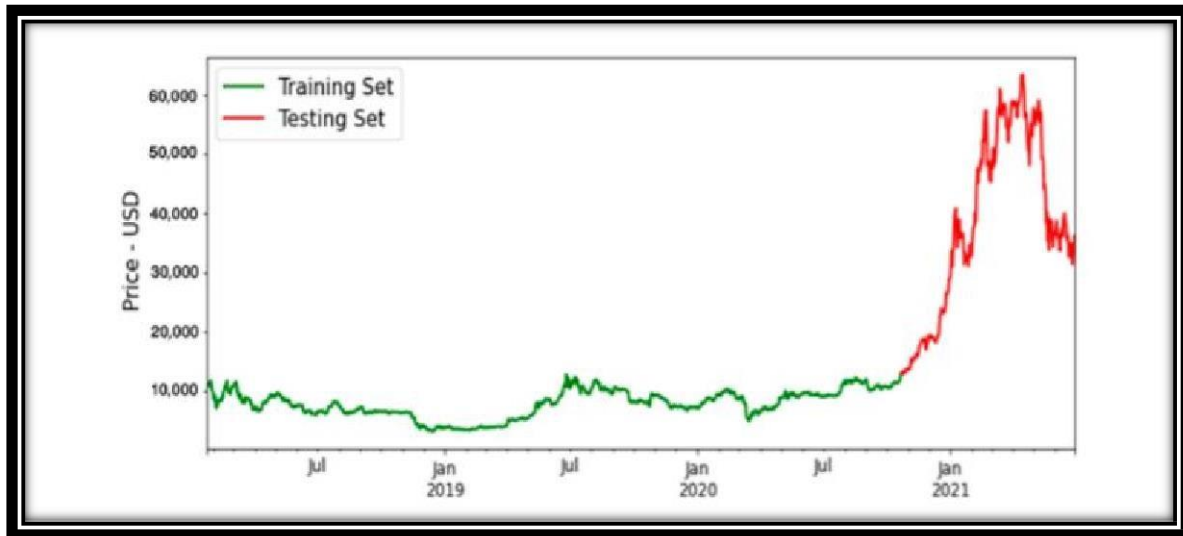


Fig14: Training and Testing Datasets

Support Vector Machine

Voice Data set

meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm
0.059781	0.064241	0.032027	0.015071	0.090193	0.075122	12.86346	274.4029	0.893369	0.491918
0.066009	0.06731	0.040229	0.019414	0.092666	0.073252	22.42329	634.6139	0.892193	0.513724
0.077316	0.083829	0.036718	0.008701	0.131908	0.123207	30.75715	1024.928	0.846389	0.478905
0.151228	0.072111	0.158011	0.096582	0.207955	0.111374	1.232831	4.177296	0.963322	0.727232
0.13512	0.079146	0.124656	0.07872	0.206045	0.127325	1.101174	4.333713	0.971955	0.783568
0.132786	0.079557	0.11909	0.067958	0.209592	0.141634	1.932562	8.308895	0.963181	0.738307

Implementation

```
from sklearn.preprocessing import LabelEncoder
y=df.iloc[:,1]
# Encode label category
# male -> 1
# female -> 0
gender_encoder = LabelEncoder()
y = gender_encoder.fit_transform(y)
y
array([1, 1, 1, ..., 0, 0, 0])

# Scale the data to be between -1 and 1
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(X)
X = scaler.transform(X)

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)

from sklearn.svm import SVC
from sklearn import metrics
svc=SVC() #default hyperparameters
svc.fit(X_train,y_train)
y_pred=svc.predict(X_test)
print('Accuracy Score:')
print(metrics.accuracy_score(y_test,y_pred))

Accuracy Score:
0.9763406940063092

svc=SVC(kernel='linear')
svc.fit(X_train,y_train)
y_pred=svc.predict(X_test)
print('Accuracy Score:')
print(metrics.accuracy_score(y_test,y_pred))
```

```
svc=SVC(kernel='linear',gamma=0.01)
scores = cross_val_score(svc, X, y, cv=10, scoring='accuracy')
print(scores)
print(scores.mean())

[0.91167192 0.97160883 0.97160883 0.97791798 0.95899054 0.9873817
 0.99369085 0.97791798 0.95253165 0.99367089]
0.9696991175178692

from sklearn.svm import SVC
svc= SVC(kernel='poly',degree=3)
svc.fit(X_train,y_train)
y_predict=svc.predict(X_test)
accuracy_score= metrics.accuracy_score(y_test,y_predict)
print(accuracy_score)

0.9589905362776026

svc=SVC(kernel='poly',degree=3)
scores = cross_val_score(svc, X, y, cv=10, scoring='accuracy')
print(scores)
print(scores.mean())

[0.89274448 0.94952681 0.93059937 0.92744479 0.94952681 0.99369085
 0.98422713 0.96529968 0.87974684 0.9778481 ]
0.9450654873617378

from sklearn.svm import SVC
svm_model= SVC()
```

RESULT

Thus after performing Support Vector Machine on voice dataset we see that accuracy score or percentage is 95.899%.

Which means this model is a good model.

```
y_pred= model_svm.predict(X_test)
print(metrics.accuracy_score(y_pred,y_test))

0.9589905362776026
```

Fig 15: SVM Accuracy Score

KNN (K Nearest Neighbor)

Performing KNN (K nearest neighbor) on diabetes data set

Diabetes Data set

Pregnancies	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesPe	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1

Implementation

```
train_score= []
test_score= []
k_vals = []

for k in range(1, 21):
    k_vals.append(k)
    knn = KNeighborsClassifier(n_neighbors = k)
    knn.fit(X_train, y_train)

    tr_score= knn.score(X_train, y_train)
    train_score.append(tr_score)

    te_score= knn.score(X_test, y_test)
    test_score.append(te_score)

plt.figure(figsize=(10,5))
plt.ylabel('Model score')
plt.xlabel('Different Values of K')
plt.plot(k_vals, train_score, color = 'r', label="training score")
plt.plot(k_vals, test_score, color = 'b', label = 'test score')
plt.legend (bbox_to_anchor= (1, 1),
bbox_transform=plt.gcf().transFigure)
plt.show()
```

```
#trying with various values for n_neighbours
knn=KNeighborsClassifier(n_neighbors=14)

#fit the model
knn.fit(X_train,y_train)

#get the score
knn.score(X_test,y_test)

0.7489177489177489
```

RESULT:

```
confusion_matrix(y_test,y_pred)

array([[123, 15],
       [ 45, 48]])

from sklearn.metrics import accuracy_score
accuracy_score(y_test,y_pred,normalize=True) # other way of finding with knn.score(x_test,y_test)

0.7402597402597403
```

Fig 16: KNN Accuracy score

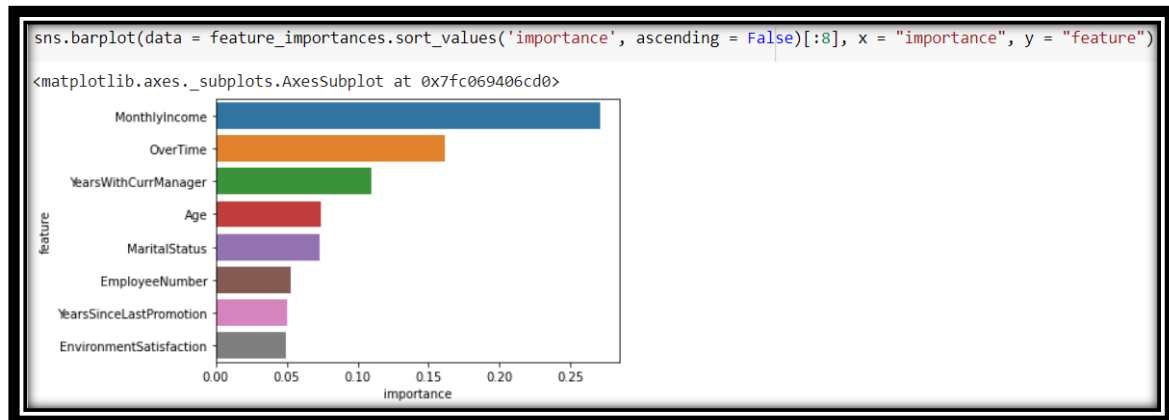
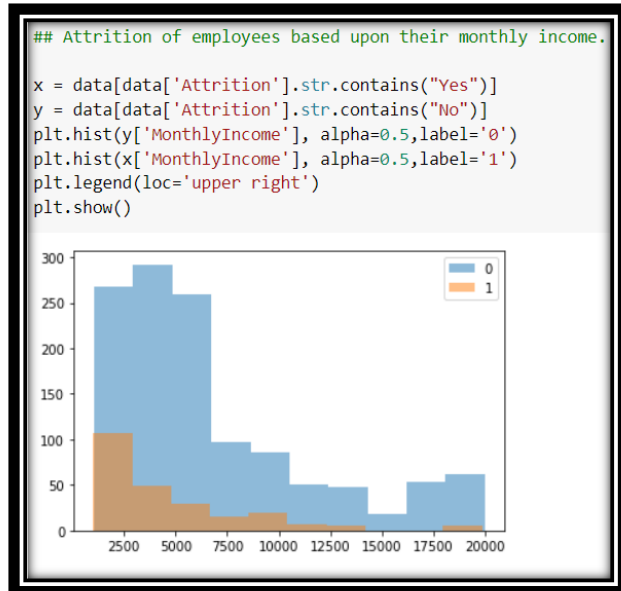
The accuracy score is 74% when we perform KNN (k nearest neighbor) algorithm on diabetes data set.

Decision Tree and Random Forest

HR- Attrition Dataset

Age	Attrition	BusinessTr	DailyRate	Departmen	DistanceFr	Education	EducationI	EmployeeC	EmployeeI	Environme
41	Yes	Travel_Rai	1102	Sales	1	2	Life Scienc	1	1	2
49	No	Travel_Fre	279	Research &	8	1	Life Scienc	1	2	3
37	Yes	Travel_Rai	1373	Research &	2	2	Other	1	4	4
33	No	Travel_Fre	1392	Research &	3	4	Life Scienc	1	5	4
27	No	Travel_Rai	591	Research &	2	1	Medical	1	7	1
32	No	Travel_Fre	1005	Research &	2	2	Life Scienc	1	8	4
59	No	Travel_Rai	1324	Research &	3	3	Medical	1	10	3
30	No	Travel_Rai	1358	Research &	24	1	Life Scienc	1	11	4
38	No	Travel_Fre	216	Research &	23	3	Life Scienc	1	12	4

Implementation



RESULT:

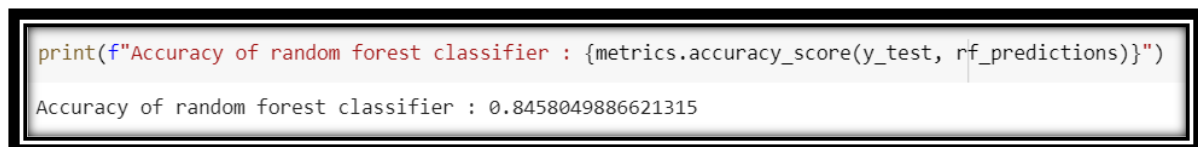


Fig 17: Random Forest and decision accuracy score

Accuracy score of random forest classifier is 84.58%.

SCREEN SHOTS OF THE VARIOUS STAGES OF THE PROJECT

Discovering/Analyzing the data:

Used to understand trends and patterns in the data.

```
Discovering/Analyzing the data
```

```
[ ] df.shape      #shows number of rows and columns in the dataset
```

```
(1258, 14)
```

```
[ ] df.columns
```

```
Index(['symbol', 'date', 'close', 'high', 'low', 'open', 'volume', 'adjClose',  
      'adjHigh', 'adjLow', 'adjOpen', 'adjVolume', 'divCash', 'splitFactor'],  
      dtype='object')
```

```
[ ] df.describe()
```

	close	high	low	open	volume	adjClose	adjHigh	adjLow	adjOpen	adjVolume	divCash	splitFactor
count	1258.000000	1258.000000	1258.000000	1258.000000	1.258000e+03	1258.000000	1258.000000	1258.000000	1258.000000	1.258000e+03	1258.000000	1258.000000
mean	188.950545	190.749544	186.896469	188.723430	4.927098e+07	71.238460	71.953692	70.444818	71.179304	1.205522e+08	0.009380	1.002385
std	70.181322	70.961984	69.061847	69.900707	3.873169e+07	39.502144	39.950835	39.005628	39.483081	5.591478e+07	0.079172	0.084583
min	106.840000	110.030000	103.100000	104.540000	1.136204e+07	25.672732	25.889200	25.470380	25.708025	4.099995e+07	0.000000	1.000000
25%	143.187500	144.300000	142.202500	143.433750	2.371339e+07	41.041037	41.458644	40.611750	41.047695	8.360832e+07	0.000000	1.000000
50%	170.415000	171.860000	169.120000	170.415000	3.336960e+07	51.330761	52.059583	50.904935	51.303798	1.053918e+08	0.000000	1.000000
75%	209.557500	211.997500	207.605000	209.870000	6.390696e+07	112.518578	114.510343	111.335425	113.101473	1.406886e+08	0.000000	1.000000
max	506.090000	515.140000	500.330000	514.790000	3.326072e+08	165.300000	170.300000	164.530000	167.480000	4.479402e+08	0.820000	4.000000

Data validation and Preprocessing:

Cleaning up the data (Cleansing the data). Converting raw data to clean dataset.


```
Cleaning up the data (Cleansing the data).
Converting raw data to clean dataset.

df.duplicated() #Checking for duplicate values and eliminating them one by one
0      False
1      False
2      False
3      False
4      False
...
1253   False
1254   False
1255   False
1256   False
1257   False
Length: 1258, dtype: bool

[ ] df.isnull().sum() #Checking for missing values
symbol      0
date        0
close       0
high        0
low         0
open        0
volume      0
adjClose    0
adjHigh     0
adjLow      0
adjOpen     0
adjVolume   0
divCash     0
splitFactor 0
dtype: int64

[ ] df1=df.reset_index()['close']
```

Exploration of data and it's visual analysis:

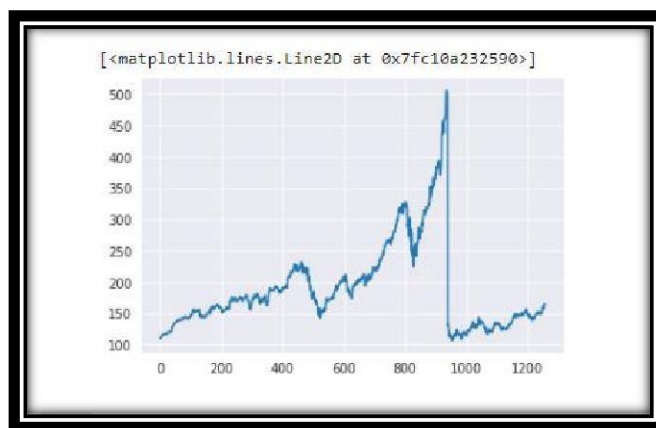


Fig 18: XY Function Plot

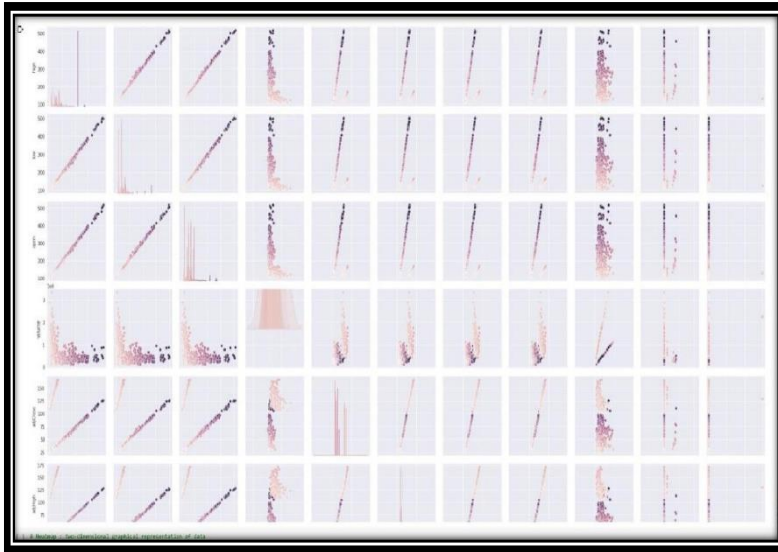


Fig 19: Pair Plot

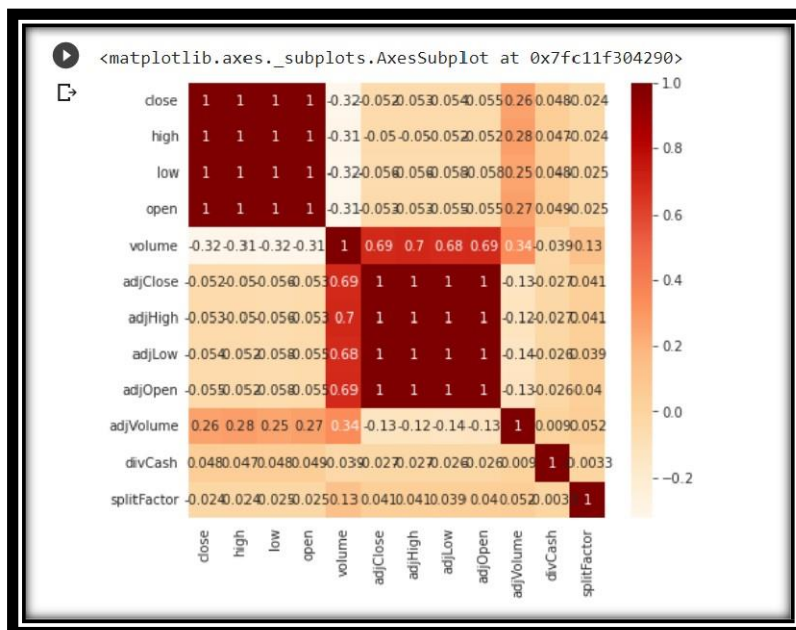


Fig 20: Heat Map

Data Scaling:

Before data are passed to the model, the data are scaled according to model requirements. In this way, this phase reshapes data to make them more suitable for the model.

```
[ ] df1=df.reset_index()['close'] # Storing our close values in df1.

LSTM are sensitive to the scale of the data. so we apply MinMax scaler

from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler(feature_range=(0,1)) #range of values for which we want to scale down
df1=scaler.fit_transform(np.array(df1).reshape(-1,1))

[ ] print(df1) #new df1 and it has now been transformed to values b/w 0 to 1

[[0.00766437]
 [0.00568566]
 [0.00778961]
 ...
 [0.12515967]
 [0.13375078]
 [0.14642455]]
```

Fig 21: Scaling Data

Train-Test Split:

```
Splitting dataset into train and test split

training_size=int(len(df1)*0.65) # 65% of the total length
test_size=len(df1)-training_size # 35 %
train_data,test_data=df1[0:training_size:],df1[training_size:len(df1),:]

training_size,test_size
(817, 440)

[ ] train_data

array([[0.00766437],
 [0.00568566],
 [0.00778961],
 [0.01049468],
 [0.0132248 ],
 [0.01780839],
 [0.01618034],
 [0.02091421],
 [0.02091421],
 [0.02249217],
 [0.02286788],
```

Fig 22: Splitting

The final Model evaluated will be discussed ahead.

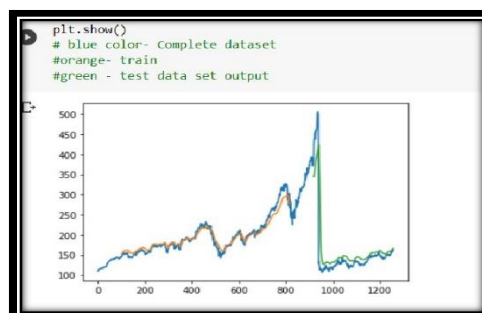


Fig 23: Train-Test Graph

The given blue line represents the whole dataset whereas orange is our training data used for Model construction and green is the testing data.

CHAPTER 05

CONCLUSION

RESULTS AND DISCUSSIONS

Since we have seen that Support vector machine brings out the most and highest accuracy score thus, we will be regarding support vector machine to be the best algorithm for accuracy.

Support Vector Machine

```
y_pred= model_svm.predict(X_test)
print(metrics.accuracy_score(y_pred,y_test))

0.9589905362776026
```

FUTURE SCOPE

- I will try out new machine learning algorithms and try to improve factors like accuracy and minimize errors.
 - I will develop a website with full-containing all the information regarding our project.
 - I will also deploy this project on cloud platforms to make it accessible to everyone.
-