INTRODUCTION

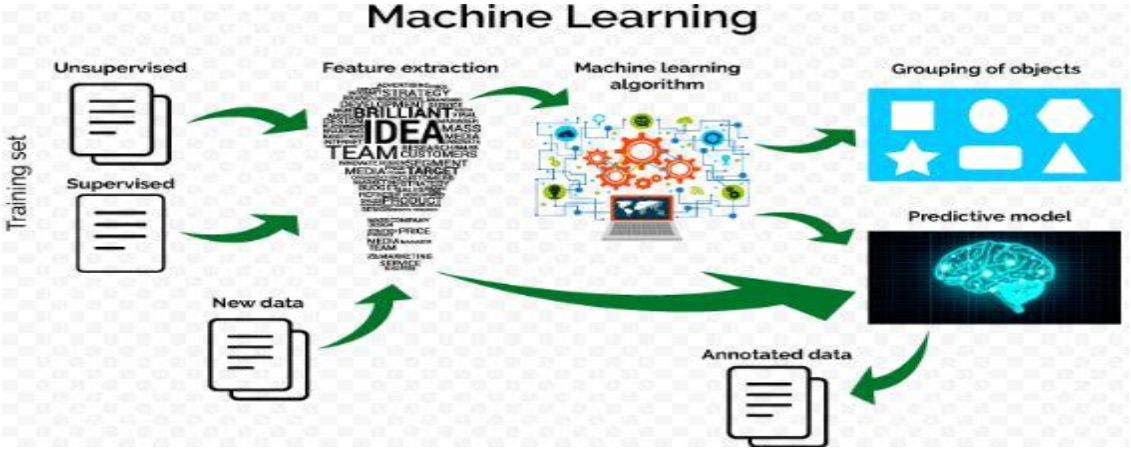

Fig.1.1

Machine Learning is 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 which 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.Fig.1.1 describes about the Machine Learning.

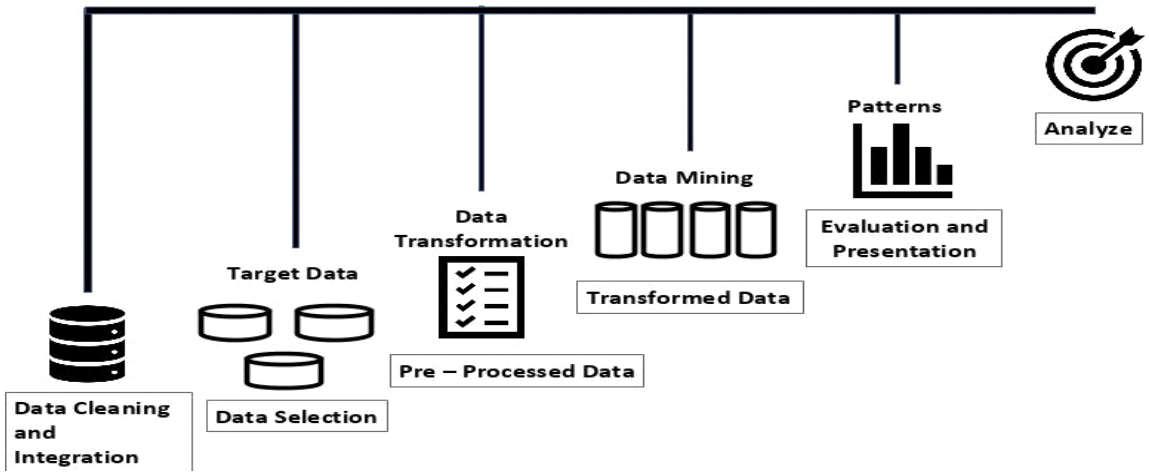

Fig.1.2

Fig.1.2 describes about the steps involved in Machine Learning. There are mainly 6 steps in Machine Learning which are Data Cleaning and Integration, Data Selection, Data Transformation, Data Mining, Evaluation and Patterns, Analyze.

- **Data Cleaning and Integration**

  Quality of data is critical in getting to final analysis. Any data which tend to be incomplete, noisy and inconsistent can affect your result.

  Some data Cleaning methods are

  1. Removal of Unwanted Data.
  2. Managing Unwanted outliers.

3. Handling Missing data.

- **Data Selection**
  Data Selection is the process where data relevant to the analysis task are retrieved from the database. Sometimes data transformation and consolidation are performed before the data selection process.
- **Data Transformation**
  Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system.
- **Data Mining**
  Data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analyzing data patterns in large batches of data using one or more software. Data mining is also known as Knowledge Discovery in Data (KDD).
- **Evaluation and Patterns**
  The pattern evaluation identifies the truly interesting patterns representing knowledge based on different types of interestingness measures. A pattern is considered to be interesting if it is potentially useful, easily understandable by humans, validates some hypothesis that someone wants to confirm or valid on new data with some degree of certainty.
- **Analyze**
  The information mined from the data needs to be presented to the user in an appealing way. Different knowledge representation and visualization techniques are applied to provide the output of data mining to the users.

1.1.Advantages

- Machine Learning(ML) is used in so many industries of applications such as banking and financial sector, healthcare, retail, publishing and social media, etc.
- ML is used by Google and Facebook to push relevant advertisements based on users search history. ML allows time cycle reduction and efficient utilization of resources.
- Due to Machine Learning there are tools available to provide continuous quality improvement in large and complex process environments.
- From our project (i.e., Data Analysis on loan approval) we can easily predict whether a person can get Loan or not.

1.2. Real time Applications

- Machine Learning is an application of Artificial Intelligence(AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- Machine Learning focuses on the development of computer programs that can access data and use it learn from themselves.
- Netflix, Amazon, Google Search and Google Maps employ Machine Learning too. When you start typing in the search box it automatically anticipates what you might be

looking for and provides suggested search terms. The suggestions could be based on past searches, what is popular now, or where you are at the time.

- Google Assistant is a new example of Machine Learning on Android, helping you with everyday tasks. Assistant makes it easy to buy movie tickets while on the go, to find a perfect restaurant for your family to grab a quick bite before the movie starts, and then help you navigate to the theater. Today, Google Assistant is available on the new Google Pixel phone. Older phones have an earlier version called Google Now.
- Visual Search technology helps customers use visual information for search, discovery and shopping. We create Augmented Reality solutions on mobile devices, overlaying relevant information over camera-phone views of the world around us powers solutions that lets customers search for products based on their visual attributes such as color, shape or even texture. Such solutions appear on Amazon and Zappos, allowing customers to quickly find the shoes or watches they like based on the appearance of the product.

1.3. Challenges

- Getting relevant data is the major challenge. Based on different algorithms data need to be processed before providing as input to respective algorithms. This has significant impact on results which should be achieved.
- Understanding of results is also a major challenge to determine effectiveness of machine learning algorithms.
- Based on which action to be taken and when to be taken, various machine learning techniques are need to be tried.

1.4.Methodology.

- First we will clean the dataset by replacing the null values. We can replace the null values by mean or median or most frequent values which will be available in Imputer class also we can replace the null values using bfill or ffill method.
- Then we need to plot different graphs against different columns to analyze the dataset. We have different plots like bar, pie, histogram, box, area, line which will be available in matplotlib class.
- Finally, we will apply different algorithms to obtained dataset and we will choose which model fits to the dataset. Now this model is used to find whether a person is applicable for Loan.

CHAPTER-2

SOFTWARE REQUIREMENT ANALYSIS

2.1. Problem Statement

1. Removing the null values in the dataset by actualizing suitable method in a specific column to increase the efficiency of the result.
2. To convert the dataset into graphs for easy understanding to the User.
3. Select the specific columns that would be significant for the result, to decrease the dataset into graphs.
4. Analyzing the every column with the required data column and presenting the data into graph, because we acquire large amount of data.

2.2. Modules

2.2.1. Numpy

NumPy is a Python package which stands for 'Numerical Python'.

Syntax: import numpy

2.2.2. Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive.

Syntax: import pandas

2.2.3. Matplotlib.pyplot

Matplotlib is used for drawing graphs ,creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.It is also possible to create a plot using categorical variables.

Syntax: import matplotlib.pyplot

2.2.4. SeaBorn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Syntax: import seaborn

2.2.5. sklearn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.

Syntax: import sklearn

2.3 Environmental Setup

2.3.1. Install Anaconda Prompt

- Download the Anaconda installer.
- Double click the installer to launch.
- Click Next.
- Read the license terms and click "I Agree".
- Select an install for "Just Me" unless you're installing for all users (which requires Windows Administrator privileges) and click Next.
- Select a destination folder to install Anaconda and click the Next button. See FAQ.
- Choose whether to add Anaconda to your PATH environment variable. We recommend not adding Anaconda to the PATH environment variable, since this can interfere with

other software. Instead, use Anaconda software by opening Anaconda Navigator or the Anaconda Prompt from the Start Menu.

- Choose whether to register Anaconda as your default Python. Unless you plan on installing and running multiple versions of Anaconda, or multiple versions of Python, accept the default and leave this box checked.
- Click the Install button. If you want to watch the packages Anaconda is installing, click Show Details.
- Click the Next button.
- After a successful installation you will see the "Thanks for installing Anaconda" dialog box:
- If you wish to read more about Anaconda Cloud and how to get started with Anaconda, check the boxes "Learn more about Anaconda Cloud" and "Learn how to get started with Anaconda". Click the Finish button.
- After your install is complete, verify it by opening Anaconda Navigator, a program that is included with Anaconda: from your Windows Start menu, select the shortcut Anaconda Navigator from the Recently added or by typing "Anaconda Navigator". If Navigator opens, you have successfully installed Anaconda. If not, check that you completed each step above, then see our Help page.

2.3.1. Accessing the Jupyter Notebook

- On Windows, a Jupyter notebook can be started from the Anaconda Prompt, the Windows start menu and Anaconda Navigator.
- 2 ways to open a Jupyter notebook

2.3.1.1 Anaconda Prompt

Go to Windows Start Menu and Select Anaconda Prompt under Anaconda3.

2.3.1.2 Anaconda Navigator

One additional way to open a Jupyter notebook is to use Anaconda Navigator. Anaconda Navigator comes with the Anaconda distribution of Python. Open Anaconda Navigator using the Windows start menu and select [Anaconda3(64-bit)] --> [Anaconda Navigator]**.**

An Anaconda Navigator window will open. In the middle of the page, in the Jupyter notebook tile, click [Launch

ALGORITHMS

Fig.4.1

Machine Learning contains different types of algorithms.Fig.4.1 shows the different algorithms present in Machine Learning. Some of them are Linear Regression, Logistic Regression, Decision Tree and Support Vector Machine etc.

4.1. Linear Regression

The objective of a linear regression model is to find a relationship between one or more features(independent variables) and a continuous target variable(dependent variable). When there is only feature it is called Uni-variate Linear Regression and if there are multiple features, it is called Multiple Linear Regression.Fig.4.2 shows how a linear regression line is plotted against data points.

The coefficient of determination, denoted as R^2, tells you which amount of variation in y can be explained by the dependence on **x** using the particular regression model. Larger R^2 indicates a better fit and means that the model can better explain the variation of the output with different inputs.

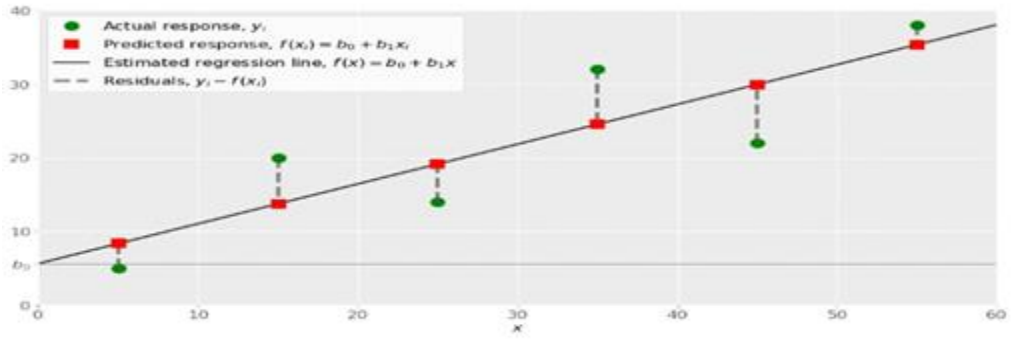

Fig.4.2

4.2. Logistic Regression

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X. Fig.4.3 shows difference between Linear and Logistic Regression.

Contrary to popular belief, logistic regression is a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function,

Logistic regression models the data using the sigmoid function.

$$g(z) = \frac{1}{1+e^{-z}}$$

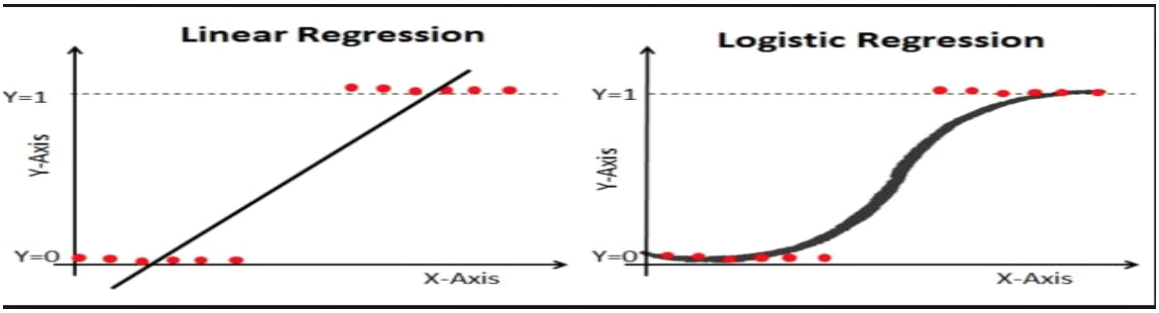

Fig.4.3

4.3. Decision Tree

Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.Fig.4.4 shows how to create a decision tree.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery.

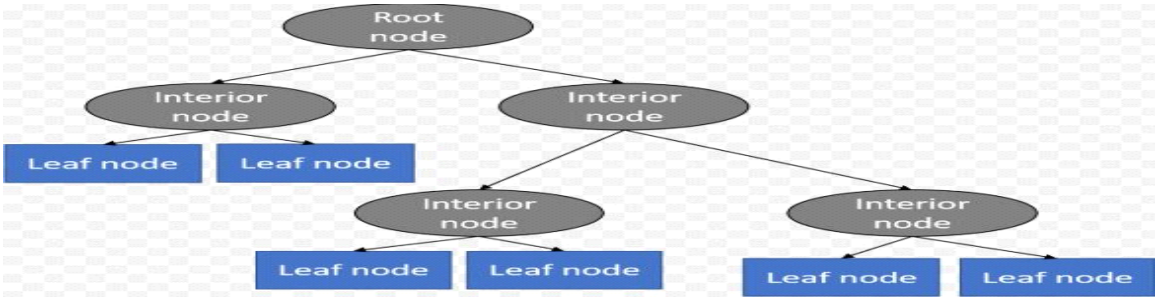

Fig.4.4

4.4. Support Vector Machine

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces.Fiig.4.5 is Support Vector Machine Diagram.

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the

information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

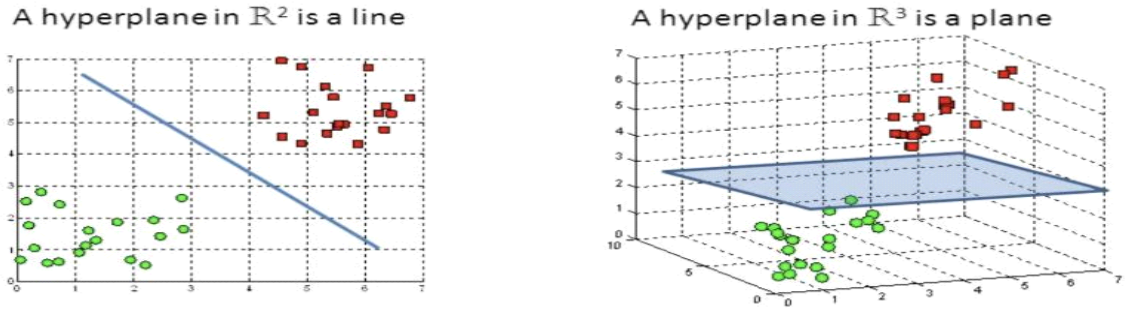

Fig.4.5

4.5. Naïve Bayes

Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.

Naive Bayes is among one of the most simple and powerful algorithms for classification based on Bayes' Theorem with an assumption of independence among predictors. Naive Bayes model is easy to build and particularly useful for very large data sets. There are two parts to this algorithm:

1. Navie
2. Bayes

The Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that a particular fruit is an apple or an orange or a banana and that is why it is known as "Naive". Fig.4.6 shows how Naïve Bayes Tables are created.

n Statistics and probability theory, Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It serves as a way to figure out conditional probability.

Given a Hypothesis H and evidence E, Bayes' Theorem states that the relationship between the probability of Hypothesis before getting the evidence P(H) and the probability of the hypothesis after getting the evidence P(H|E) is :

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Fig.4.6