Project Report on

"Intelligent Crop Recommendation System Using Machine Learning"

Submitted by:

Miss. TEJAL D. SURVASE (PRN:2019032500207344)
Miss. SIDDHI S. VELAPURE (PRN:2019032500207197)
Miss. MAYURI S. SWAMI (PRN:2019032500209173)
Miss. RATAN G. KORE (PRN:2019032500210192)
Miss. PRAPTI P. ATKALE (PRN:2019032500207441)

UNDER THE GUIDANCE OF

Mrs. S. S. Kadam

In partial fulfillment for the award of the

degree of

BACHELOR

OF ENGINEERING IN
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
at

SHRI VITHAL EDUCATION and RESEARCH INSTITUTES's, COLLEGE OF ENGINEERING, PANDHARPUR

AFFILIATED TO PUNYASHLOK AHILYADEVI HOLKAR SOLAPUR UNIVERSITY, SOLAPUR

2022-2023

SVERI's COLLEGE OF ENGINEERING,PANDHARPUR

CERTIFICATE

This is to certify that the project report entitled **"Intelligent Crop Recommendation System Using Machine Learning"** is submitted for partial fulfillment of Bachelor Of Technology as per requirement of Punyashlok Ahilyadevi Holkar Solapur University, Solapur for the academic year 2022-2023.

Mrs. S. S. kadam
(Project Guide)

Mr. P. D. Mane
(Project Coordinator)

Dr. S. P. Pawar
(HOD, CSE)

Dr. B. P. RONGE
(PRINCIPAL)

EXTERNAL EXAMINAR

Acknowledgement

We are pleased to acknowledge **Dr. B. P. Ronge** (Principle) And **Dr. S. P. Pawar** (HOD CSE) for his valuable guidance during the course of this project work. We extend our sincere thanks **Mrs. S. S. Kadam** to who continuously helped usthroughout the project and without his guidance this project would have been an uphilltask.

We are also grateful to other members of the CSE faculty members and technical staff who cooperated with us regarding some issues.

Signature

Miss. Survase T. D.	**Sign. .....**
Miss. Velapure S. S.	**Sign. .....**
Miss. Swami M. S.	**Sign .....**
Miss. Kore R. G.	**Sign. .....**
Miss. Atkale P. P.	**Sign .....**

ABSTRACT

Agriculture and its allied sectors are undoubtedly the largest providers of livelihoods in rural India. The agriculture sector is also a significant contributor factor to the country's Gross Domestic Product (GDP). Blessing to the country is the overwhelming size of the agricultural sector. However, regrettable is the yield per hectare of crops in comparison to international standards. This is one of the possible causes for a higher suicide rate among marginal farmers in India. This Mini - Project proposes a viable and user-friendly yield prediction system for the farmers. The proposed system provides connectivity to farmers. The user provides the area & soil type as input. Machine learning algorithms allow choosing the most profitable crop list or predicting the crop yield for a user-selected crop. To predict the crop yield, selected Machine Learning algorithms such as Decision Tree, Support Vector Machine (SVM), Gaussian Naïve Bayes, Random Forest (RF), Logistic Regression, and Xgboost are used. Among them, the Random Forest showed the best results with 99% accuracy. Agriculture is a major contributor to the Indian economy. The common problem existing among the Indian farmers are they don't choose the right crop based on their soil requirements. Due to this they face a serious setback in productivity. This problem of the farmers has been addressed through precision agriculture. Precision agriculture isa modern farming technique that uses research data of soil characteristics, soil types, crop yield data collection and suggests the farmers the right crop based on their site- specific parameters. This reduces the wrong choice on a crop and increases the productivity. In this project, we are building an intelligent system, which intends to assist the Indian farmers in making an informed decision about which crop to grow depending on the sowing season, his farm's geographical location and soil characteristics. Further the system will also provide the farmer, the yield prediction if he plants the recommended crop.

CONTENTS

List Of Figures

Chapter 1 INTRODUCTION

1.1 Introduction

Agriculture is a significant area for the Indian economy and human survival. It is one of the primary occupations which are essential for human life. It likewise contributes a huge part to our day-to-day life [1]. In most cases,Farmers commit suicide due to production loss because they are not able to pay the bank loans taking for farming purposes[12]. We have noticed in present times that the climate is changing persistently which is harmful to the crops and leading farmers towards debt and suicide [18]. These risks can be minimized when various mathematical or statistical methods are applied to data and by using these methods, we can recommend the best crop to the farmer for his Agricultural land so that it helps him to get maximum profit [12]." Nowadays agriculture has developed a lot in India. "site-specific" farming is the key to Precision agriculture. Although precision agriculture has achieved better enhancements it is still facing certain issues. Precision agriculture plays an important role in the recommendation of crops. The recommendation of crops is dependent on various parameters." Precision agriculture focuses on identifying these parameters in a site-specific way to identify issues. Not all the results given by precision agriculture are accurate to result but in agriculture, it is significant to have accurate and precise recommendations because, in case of errors, it may lead to heavy material and capital loss. Many research works are being carried out, to attain an accurate and more efficient model for crop prediction [11]."

Machine Learning focuses on the algorithm like supervised, unsupervised, and Reinforcement learning and each of them has its advantages and disadvantages. Supervised learning the algorithm assembles a mathematical model from a set of data that contains both the inputs and the desired outputs. An unsupervised learning-the algorithm constructs a mathematical model from a set of data that contains only inputs and no desired output labels. Semi-supervised learning- algorithms expand mathematical models from incomplete. This project aims to recommend the most suitable crop based on input parameters like Nitrogen (N), Phosphorous (P), Potassium (K), PH value of soil, Humidity, Temperature, and Rainfall. This paper predicts the accuracy of the future production of eleven different crops such as rice, maize, chickpea, kidney beans, pigeon peas, moth beans, Mungbam, black gram, lentils, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee crops using various supervised machine learning approaches in India and recommends the most suitable crop.

The dataset contains various parameters like Nitrogen (N), Phosphorous (P), Potassium (K), PH value of soil, Humidity,Temperature, and Rainfall. This proposed system applied different kinds of Machine Learning algorithms like Decision Trees, Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression, Random Forest (RF), and XGBoost [12]

1.2 Need Of Work

Agriculture in India plays a major role in economy and employment. The common difficulty present among the Indian farmers are they don't opt for the proper crop based on their soil necessities. Because of this productivity is affected. This problem of the farmers has been solved through precision agriculture. This method is characterized by a soil database collected from the farm, crop provided by agricultural experts, achievement of parameters such as soil through soil testing lab dataset. The data from soil testing lab given to recommendation system it will use the collect data and do ensemble model with majority voting technique using support vector machine (SVM) and ANN as learners to recommend a crop for site specific parameter with high accuracy and efficiency.

Chapter 2 OBJECTIVE

2.1 Objectives

- To build a robust model to give correct and accurate prediction of crop sustainsability in each state for the particular soil type and climatic conditions.
- Provide recommendation of the best suitable crops in the area so that the farmer does notincur any losses.
- Provide profit analysis of various crops based on previous year's data.
- To recommend optimum crops to be cultivated by farmers based on several parameters and help them make an informed decision before cultivation

2.2 Problem Statement

Failure of farmers to decide on the best suited crop for his land using traditional and nonscientific methods is a serious issue for a country where approximately 50 percent of the population is involved in farming. Both availability and accessibility of correct and up to date information hinders potential researchers from working on developing country case studies. With resources within our reach we have proposed a system which can address this problem by providing predictive insights on crop sustainability and recommendations based on machine learning models trained considering essential environmental and economic parameters.

Chapter 3 LITERATURE REVIEW

- Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique by Kumar et al. This paper proposed a method named Crop Selection Method (CSM) to solve crop selection problem, and maximize net yield rate of crop over season and subsequently achieves maximum economic growth of the country. The proposed method may improve net yield rate of crops.

- Agro Consultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms by Zeel et al, This paper proposed and implemented an intelligent crop recommendation sys- tem, which can be easily used by farmers all over India. This system would assist the farmers in making an informed decision about which crop to grow depending on a variety of environmental and geographical factors. We have also implemented a secondary system, called Rainfall Predictor, which predicts the rainfall of the next 12 months.

- Development of Yield Prediction System Based on Real-time Agricultural meteorological Information Haedong et al. This paper contains about the research and the building of an effective agricultural yield forecasting system based on real-time monthly weather. It is difficult to predict the agricultural crop production because of the abnormal weather that happens every year and rapid regional climate change due to global warming. The development of agricultural yield forecasting system that leverages real-time weather information is urgently required. In this research, we cover how to process the number of weather Intelligent data (monthly, daily) and how to configure the prediction system. We establish a non-parametric statistical model on the basis of 33 years of agricultural weather in- formation. According to the implemented model, we predict final production using the monthly weather information. This paper contains the results of the simulation.

- Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach Monali et al Computer science and Engineering GGITS, Jabalpur. This work presents a system, which uses data mining techniques in order to predict the category of the analyzed soil datasets. The category, thus predicted will indicate the yielding of crops. The problem of predicting the crop yield is formalized as a classification rule, where Naive Bayes and KNearest Neighbor methods are used.

- Crop Recommendation System for Precision Agriculture S.Pudumalar* et al, This paper, proposes a recommendation system through an ensemble model with majority voting technique using Random tree, CHAID, K-Nearest Neighbor and Naive Bayes as learners to recommend a crop for the site specific parameters with high accuracy and efficiency.

Chapter 4 PROJECT PLANNING

The goal of this study is to demonstrate the impact of meteorological variables on agricultural production in order to improve crop yields, which will help farmers. The linear regression system model is created in Python. Other design steps will be explored in detail in the following sections.

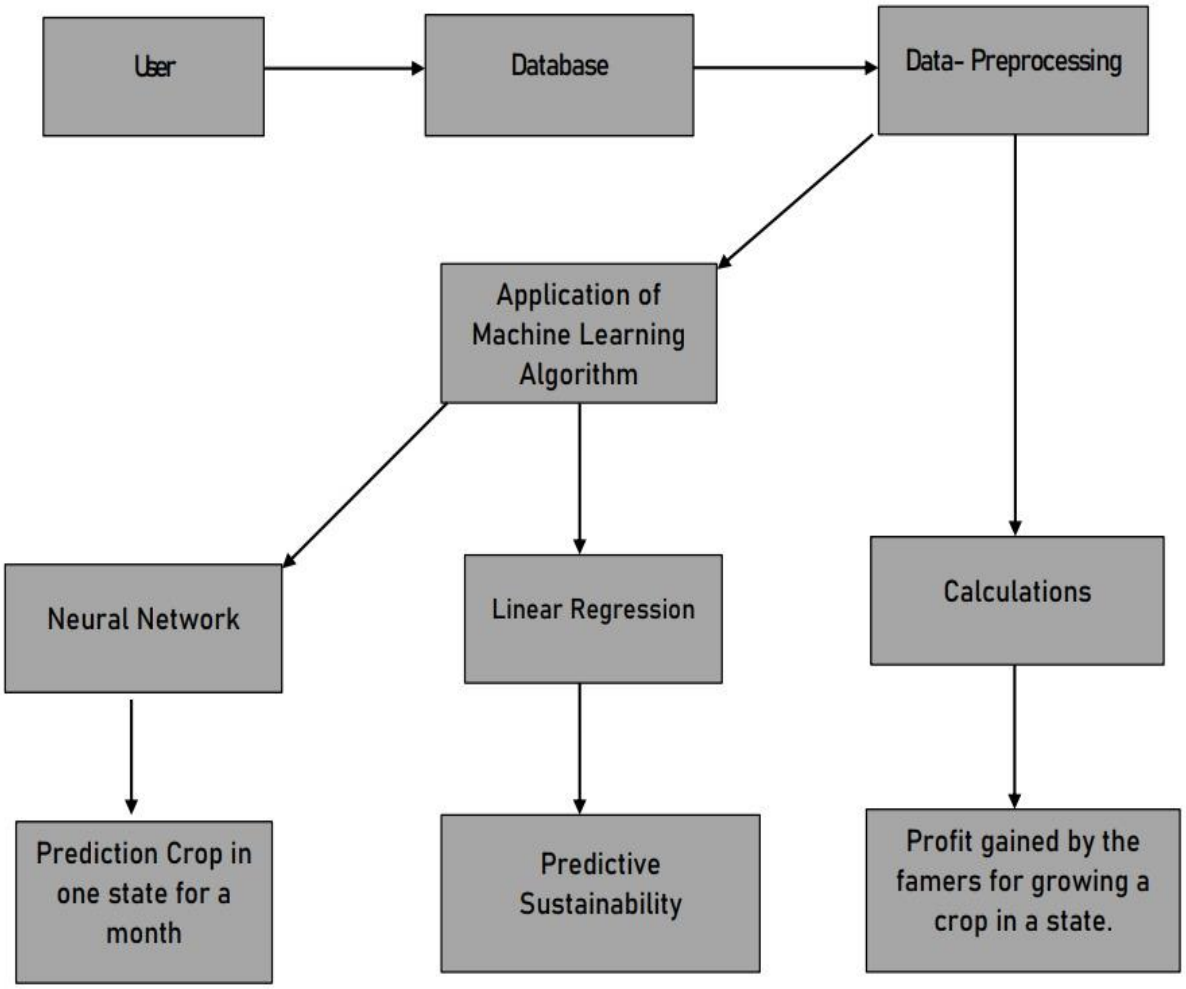

Fig 4.1: System Architecture

- **Data Collection**

  Our information includes the temperature, N, P, K, humidity, ph, and rainfall as an attribute, as well as the outcomes of crops that may be cultivated in that soil type. The dataset consists of a few major crops which are mostly cultivated as wheat, sugarcane, rice, maize, chickpea, kidney-beans, pigeon-peas, moth-beans, mung-bean, blackgram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, coffee.

- **Data Preprocessing**

  Data preprocessing is a data mining approach that entails converting raw data into a format that can be understood. Because the original dataset may have a large number of missing values, all of them should be eliminated at first. Missing values are represented by a dot in the dataset, and their existence can degrade the overall value of the data as well as impair performance. As a result, we replace these numbers with the mean values to fix this problem. The second step is to create the class labels. Because we're utilizing supervised learning, there should be a class label for each entry in the dataset, which is produced during the preprocessing phase

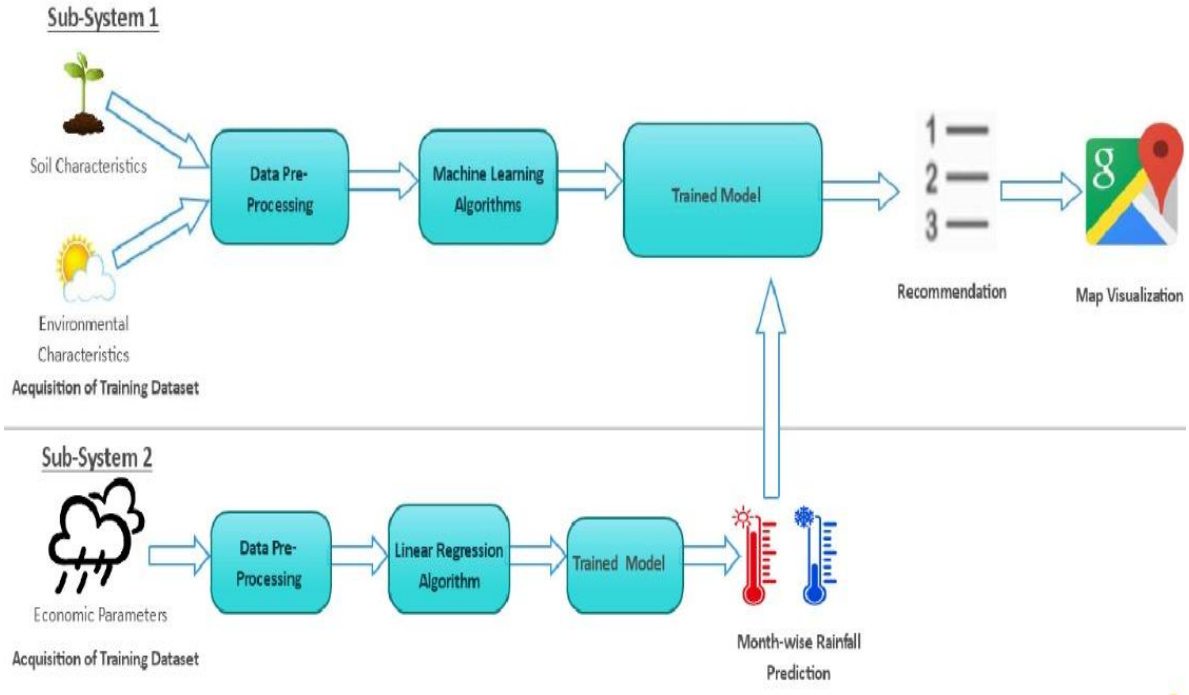

Fig 4.2 : Data Preprocessing

- **Regression Analysis**

  Regression Analysis is a predictive modelling technique that examines the relationship between a dependent or target variable and an independent or predictor variable. It covers linear, multiple linear, and non-linear regression models, among others. Simple linear regression is the most used model.Polynomial regression is a type of regression method in which the link between the independent variable x and the dependent variable y is described as an nth degree polynomial in x.

  polynomial regression fits a nonlinear relationship between the value of x and the associated conditional mean of y, denoted by E(y |x). Despite the fact that polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear in the sense that in the unknown parameters inferred from the data, the regression function E(y |x) is linear. As a result, polynomial

regression is regarded as a subset of multiple linear regressions. The predicted value of y may be modelled as an nth degree

- **Choosing Machine Learning Model**

  When choosing a machine learning algorithm, Random Forest is one of the most popular and widely accepted supervised learning techniques. It consists of a number of decision trees for different subsets of data, rather than using the whole data as a single unit.This helps maximize the accuracy of the prediction of each variable. It makes sense to use Random Forest on a large dataset as it can provide results with maximum accuracy in a short span of time. We chose to use Google Colab for executing python code that trained our machine learning model. A lot of crucial libraries such as pandas, Sklearn were imported to perform a proper analysis of the dataset. Using pandas, a data frame was created that helped read the CSV file. The dataset was split into a specific ratio to train and test respectively. We performed training of the model using random forest and tested its output with testing data. We found the accuracy to be 99%.

- **Serving Machine Learning Prediction as a Services through Web App**

  We leveraged modern web infrastructures such as Python Flask API, Node.JS, Next.JS, MDX, PostgreSQL, Prisma 2 ORM, and Google Firebase PaaS to serve an online web app that directly enables users to submit soil information along with required location data. We made use of the OpenWeather API to accurately forecast and send rainfall data and pH information. The web app is written in JavaScript and Python Language. It is end-to-end protected through OTP-based authentication.

Chapter 5 Design

5.1 Libraries

- **Numpy :**

  NumPy (**Numerical Python**) is an open source Python library that's used in almost every field of science and engineering. It's the universal standard for working with numerical data in Python, and it's at the core of the scientific Python and PyData ecosystems. NumPy users include everyone from beginning coders to experienced researchers doing state-of-the-art scientific and industrial research and development. The NumPy API is used extensively in Pandas, SciPy, Matplotlib, scikit-learn, scikit-image and most other data science and scientific Python packages.

  The NumPy library contains multidimensional array and matrix data structures (you'll find more information about this in later sections). It provides **ndarray**, a homogeneous n-dimensional array object, with methods to efficiently operate on it. NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrice.

- **Pandas :**

  Pandas is an add-on software library created by Wes McKinney for the Python programming language.The main scope of the pandas library is the manipulation of data sets, i.e. to edit, change, and replace particular elements of a DataFrame class object. However, pandas provides a broad range of functions and can also be used for other tasks such as the calculation of descriptive statistics and the visualization of the columns and rows in a data set.Similar to other Python libraries, packages, and modules, pandas is open source, i.e. freely available for usage, modification, and redistribution

- **Scikit-Learn:**

  Scikit-learn is an ML library for python. It has various algorithms for classification and regression like logistic regression, random forest classifiers and support vector machines. We can operate it along with other python libraries like NumPy and pandas. Scikit-learn is one of the best libraries especially for supervised learning which involves training the model by loading a sample dataset which it can observe and structure its learning accordingly. It also gives us the provision to use train_test_split for making training and testing datasets.

- **Matplotlib :**

  Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc. It is used along with NumPy to provide an environment that is an effective open source alternative for MatLab. It can also be used with graphics toolkits like PyQt and wxPython.

- **SkyLearn :**

  We will learn about the sklearn library and how to use it to implement machine learning algorithms. In the real world, we don't want to construct a challenging algorithm each time we need to utilise it. Although creating an algorithm from the beginning is a terrific approach to grasping the underlying concepts behind how it operates, we might not achieve the efficiency or dependability we require.

  A Python module called Scikit-learn offers a variety of supervised and unsupervised learning techniques. It is based on several technologies you may already be acquainted with, including NumPy, pandas, and Matplotlib.

5.2 Algorithms

- **Supervised Learning**

Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized. Supervised learning can be separated into two types of problems when data mining—classification and regression:

Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, which are described in more detail below.

Regression is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given business. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

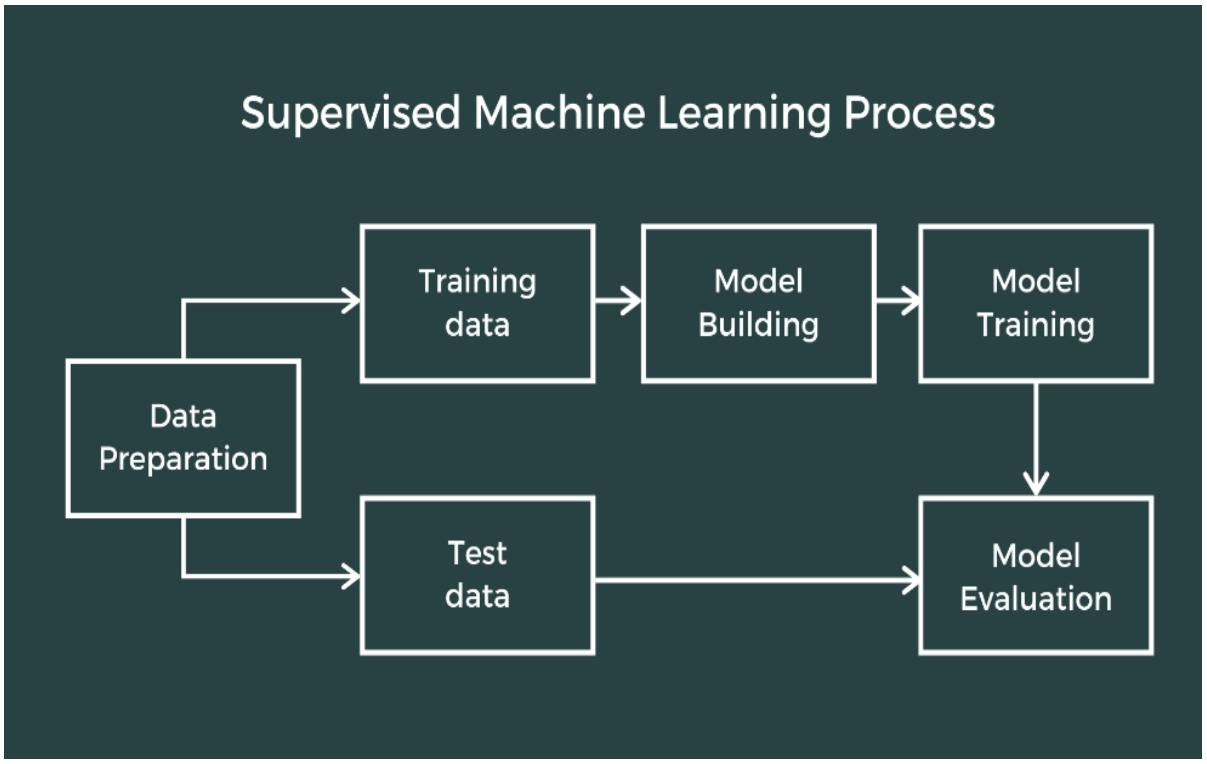

Fig 5.1 : Supervised Learning Process

- **SVM**

  Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane

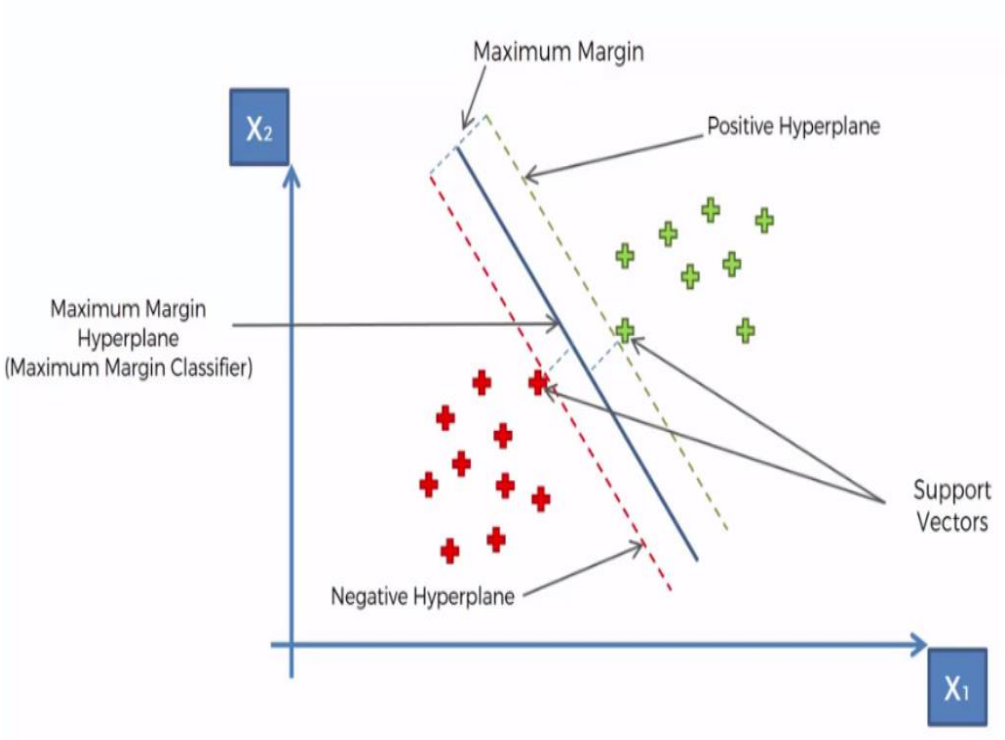

Fig 5.2 : SVM

- **Logistic Learning**

  Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). The logistic model (or logit model) is used to model the probability of a certain classor event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

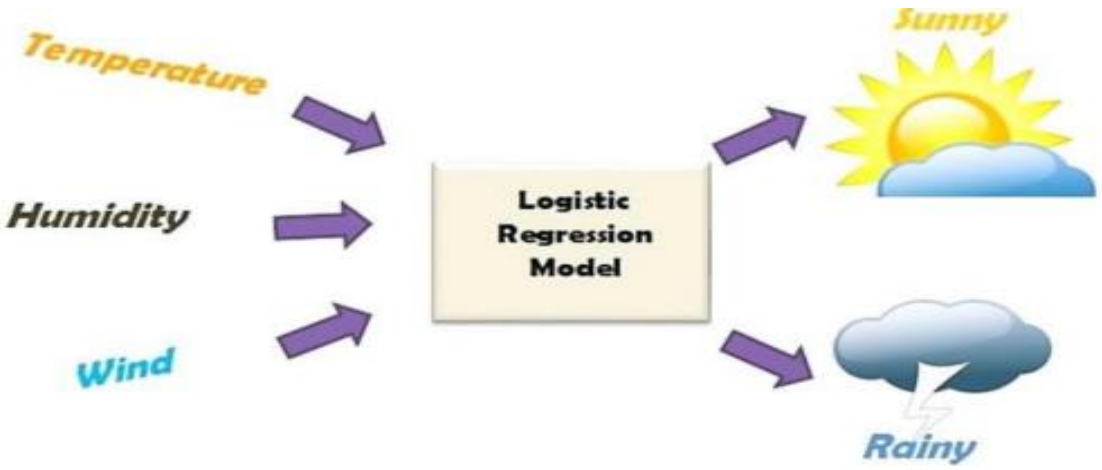

Fig 5.3 : Logistic Regression Model

- **Navie Bayes**

  Naive Bayes is classification approach that adopts the principle of class conditional independence from the Bayes Theorem. This means that the presence of one feature does not impact the presence of another in the probability of a given outcome, and each predictor has an equal effect on that result. There are three types of Naïve Bayes classifiers: Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Gaussian Naïve Bayes. This technique is primarily used in text classification, spam identification, and recommendation systems.

  Formula :

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

- **Heatmap:**

  Heat maps in Python is a type of a graph which represents different shades of a colour to distinguish the values in the graph. The higher values are represented in the darker shades and the lesser values are represented in lighter shades. There can also be a different colour in the graph when the value is more different from the other data values. The graphical representation of the values through color differentiation is known as heat map.

  Heatmap is also called a shading matrix. The heatmap can also include normalizing the matrices, performing and analyzing the clusters, customizing the colors and permuting the rows and columns so that the user can place similar values nearby to each other.

- **Logistic Regression :**

  Linear regression is a linear approach to modeling the relationship between a scalar response (and dependent variable) and one or more explanatory variables (or independent variables). Linear regression is used for finding linear relationship between target and one or more predictors. It fits a linear model with coefficients w = (w1, . . . , wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. Linear regression is used for finding linear relationship between target and one or more predictors

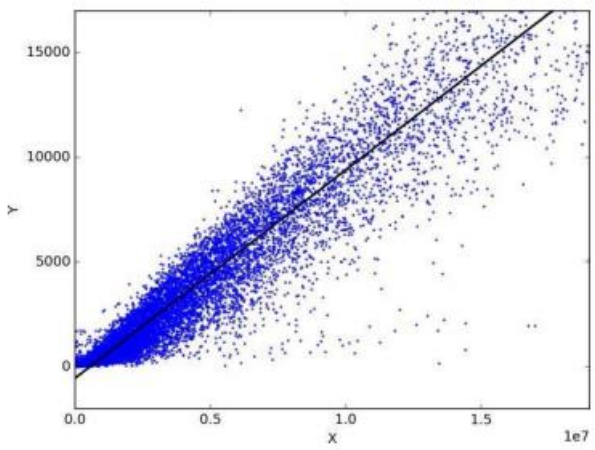

Fig 5.4 : Linear Regression

- **Decision Tree :**

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record‟s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

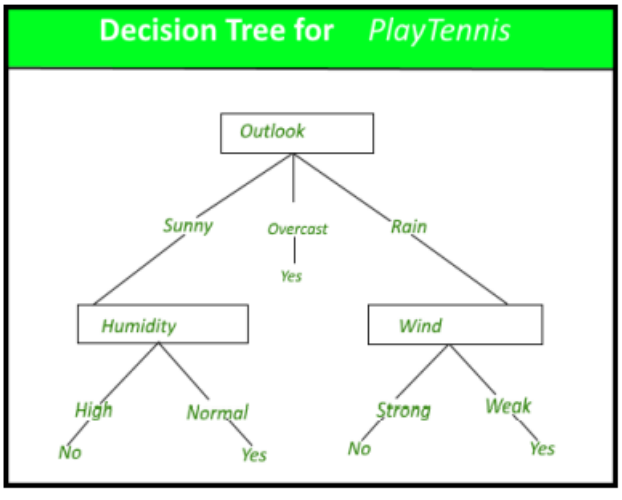

Fig 5.5 : Decision Tree

- **Confusion Matrix :**

A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm. The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as an **error matrix**. Some features of Confusion matrix are given below:

- For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, and so on.
- The matrix is divided into two dimensions, that are **predicted values** and **actual values** along with the total number of predictions.
- Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.

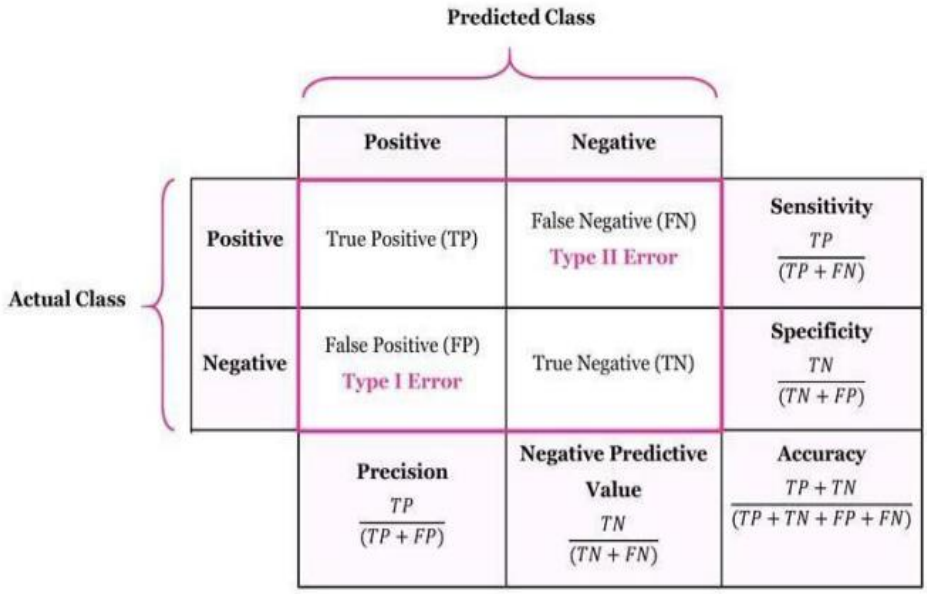

Fig 5.6 Confusion Matrix