DIABETES PREDICTION MODEL

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering/Information Technology

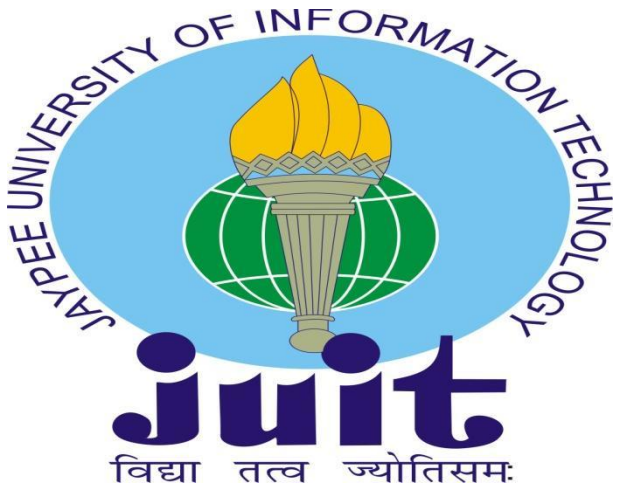
By

Divyansh Srivastava 191309
Rajan Madhav Sharma 191444

Under the supervision of

Prof. Dr. Vivek Kumar Sehgal

to



Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology
Waknaghat, Solan-173234, Himachal Pradesh

Candidate's Declaration

We hereby declare that the work presented in this report entitled "**Diabetes Prediction Model"** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of our own work carried out over a period from July 2022 to May 2023 under the supervision of (**Prof. Dr. Vivek Kumar Sehgal**) (Professor and Head, Computer science & engineering and Information Technology).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)
Divyansh Srivastava, 191309.
Rajan Madhav Sharma, 191444.

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature)
Supervisor Name: Prof. Dr. Vivek Kumar Sehgal.
Designation: Professor And Head of the Department.
Department name: CSE & IT.
Dated:

Plagiarism Certificate

As Provided by the LRC of JUIT

Acknowledgement

We are very grateful to our project supervisor, Prof. Dr. Vivek Kumar Sehgal, for his competent guidance. We would want to express our gratitude to him for his support, counsel, and direction. We would also like to thank Mr. Mohan Sharma (Lab In-charge), who offered facilities as the research progressed. Last but not least, we would like to thank everyone who contributed directly or indirectly to the design and fulfilment of this project.

Thanks and Regards

(Student Signature)

Project Group No. :148
Divyansh Srivastava
191309
Rajan Madhav Sharma
191444

Table of Contents

List of Figures

Abstract

Diabetes is considered as one of the deadliest and chronic diseases which is caused by increased levels of blood sugar. Many complications occur if diabetes remains untreated and unidentified. According to the International Diabetes Federation, 382 million people worldwide have diabetes. By 2035, this figure will have more than doubled to 592 million. A variety of traditional approaches based on physical and chemical investigations are available for diagnosing diabetes. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc.

The primary goal of this research is to is to design a model which can predict the likelihood of diabetes in patients with maximum accuracy by merging the findings of several machine learning algorithms.

This study attempts to predict diabetes using four distinct machine learning algorithms: SVM, Logistic Regression, Decision Tree and Random Forest. Furthermore, using machine learning techniques, this study aims to propose a best solution for early diabetes detection.

Chapter-1 INTRODUCTION

1.1 Introduction

Diabetes is a chronic condition that could lead to a global health care disaster. 382 million people worldwide have diabetes, according to the International Diabetes Federation. This will double to 592 million by 2035.

Sugar (glucose) is derived from the meals we eat, notably those high in carbohydrates. Everyone, including those with diabetes, requires carbohydrates since they are the body's primary source of energy. Bread, cereal, pasta, rice, fruit, dairy products, and vegetables are examples of foods high in carbohydrates (especially starchy vegetables). These foods are converted into glucose by the body when we eat them. The bloodstream carries glucose around the body. Our brain receives some of the glucose to aid in our ability to think properly and perform. The remaining glucose is sent to our body's cells for usage as fuel, and it is also stored as energy in our liver for later use by the body.

Insulin is necessary for the body to utilize glucose as fuel. The beta cells in the pancreas create the hormone known as insulin. Insulin functions as a door's key. In order to allow glucose to enter the cell from the blood stream, insulin binds to the cell's doors, opening them. Glucose builds up in the circulation (hyperglycemia) and diabetes occurs if the pancreas is unable to generate enough insulin (shortage) or if the body is unable to utilise the insulin it produces.

Obesity, a high blood glucose level, and other factors can cause diabetes. It alters the function of the hormone insulin, which causes crabs to have an irregular metabolism and raises blood sugar levels.

Diabetes causes large number of deaths all over the world, This

can be managed and controlled early on. This study investigates

diabetes prediction using a variety of diabetes disease-related factors in order to achieve this.

1.1.1 Types of Diabetes:

1) When an individual having type 1 diabetes, their immune system is not strong enough and the white blood cells cannot to make enough insulin. There are no convincing studies that demonstrate the
2) causes of type 1 diabetes, and there are also no effective preventative measures till now.

3) Type 2 diabetes is characterised by either insufficient insulin production by the cells or improper insulin use by the body. 90% of people with diabetes have this kind of diabetes, making it the most prevalent type. Both genetic and lifestyle factors contribute to its occurrence.

4) Gestational diabetes manifests as in pregnant women who have high blood sugar levels unexpectedly. It will return in two-thirds of patients during consecutive pregnancies. There is a high likelihood that type 1 or type 2 diabetes will develop during a gestational diabetes-affected pregnancy.

   Diabetes is also caused by genetic conditions, It is caused by at least two defective genes on chromosome 6, the chromosome that controls the body's response to numerous antigens. The incidence of type 1 and type 2 diabetes may also be influenced by viral infection. Infection with viruses such as rubella, mumps, hepatitis B virus, and cytomegalovirus increase the risk of having diabetes.

   The goal of this study is to create a system that, by fusing the findings of several machine learning approaches, can more accurately conduct early diabetes prediction for a patient.

To predict diabetes, we use a variety of Machine Learning classification and ensemble techniques. Machine learning is a technique used to intentionally train computers or other machines. By creating various categorization and ensemble models from the obtained dataset, various machine learning techniques efficiently capture knowledge.

Many machine learning techniques are capable of making predictions, but selecting the right method can be challenging. Therefore, we use well-known classification and ensemble algorithms on the dataset for this aim to make predictions.

1.2 Problem Statement

Diabetes is recognized as one of the most dangerous and long-lasting conditions that cause blood sugar to rise. Diabetes will cause a number of dangers if it is not detected and managed. The time-consuming identification method leads to the diabetic contacting medical care and seeing a doctor. However, growing the usage of machine learning algorithms aids in the resolution of this significant difficulty. The goal of this study is to create a model that can accurately predict the likelihood of diabetes in patients. As a result, machine learning algorithms like Naive Bayes, KNN and SVM are utilized for early diabetes prediction.

Diabetes patients have to go through a variety of tests in order to be correctly diagnosed. These tests also include unnecessary or repetitive medical procedures that cause difficulty and waste of time and resources. Diabetes lowers people's quality of life and affects work productivity, therefore the financial effect of the disease far above the direct medical bills in the care sector. The major causes of these bad impacts include a lack of a proper diagnosis scheme, a lack of financial resources, and a general lack of knowledge. As a result, avoiding the illness entirely by early identification will almost certainly reduce the economic burden and benefit the patient in diabetes care.

In some leading nations, the average lifetime cost of direct medical treatment for a diabetes patient is anticipated to be $85,000 to $1,00,000. Diabetes has a substantially greater financial consequences than direct medical expenses in the healthcare industry since it lowers the quality of life and reduces work efficiency. The primary causes of these bad impacts include a lack of a suitable diagnosis scheme, a shortage of financial resources, and a general lack of knowledge. As a result, limiting the diabetes completely by early diagnosis may substantially decrease the financial cost and support patients in diabetes prevention.

The major aim is to construct a predictive engine that will enable user to determine if they have diabetes or not. If the user has diabetes, the user may not need to consult any doctor for additional treatment.

To forecast the existence of the illness, the prediction engine needs a huge dataset and powerful machine learning techniques.

For treatment, doctors rely on common knowledge. When there is a lack of common knowledge, investigations are summarised after a certain number of instances have been analyzed. However, this procedure takes time, whereas machine learning can identify patterns faster. A massive quantity of data is necessary to use machine learning. Depending on the ailment, there is very little data accessible. Furthermore, the number of samples with no illnesses is far higher than the number of samples with the condition.

This project involves conducting two case studies to examine the efficacy of several machine learning algorithms in assisting in the identification of such patterns in I and developing a platform for easier data sharing and cooperation.

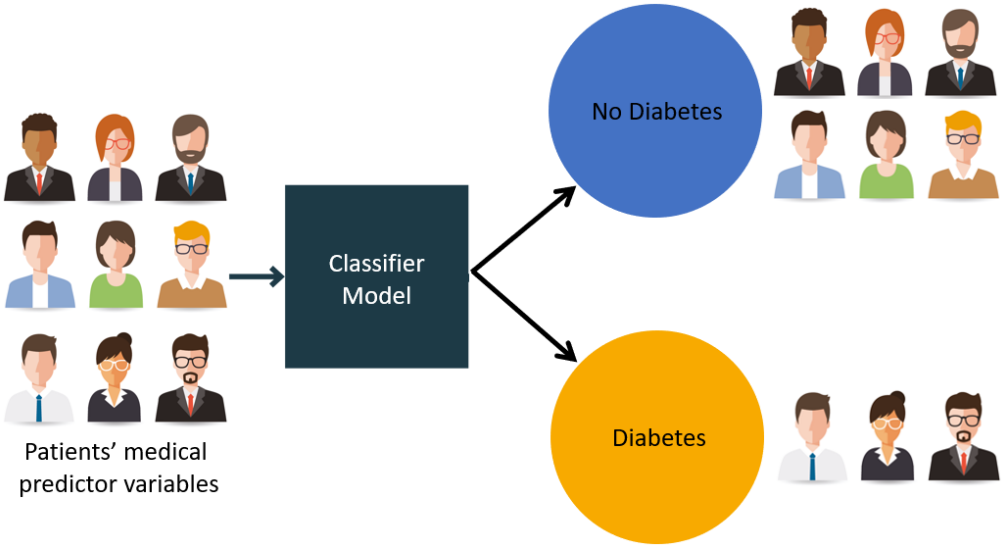


Fig 1.1: Diabetes Classifier

1.3 Objectives

The purpose of this study is to evaluate the Diabetes dataset, develop, and implement a Diabetes prediction and recommendation system built on machine learning classification algorithms.

Leukemia, anaemia, diabetes, haemophilia, blood cholesterol, cancer, HIV/AIDS, and other blood problem illnesses exist. Diabetes Mellitus affects around 400 million individuals worldwide. Hundreds of thousands of people are affected by this chronic illness. These technologies are intended to detect their medical issues.

1.3.1 Goals:

- The goal is to raise awareness about the significance of diabetes as a worldwide public health concern.
- To examine the literature on diabetes diagnosis and prediction.
- Create a model using machine learning techniques.
- Diabetes prevention and management are being promoted in underserved populations.
- Diagnosis of diabetes at an early stage using food intake.
- The importance of lifestyle in identifying individuals with diabetes and avoiding complications, especially health and food.

Serious actions must be undertaken to reduce the impacts of diabetes at an initial stage, which also helps to reduce the number of diabetic patients. Aside from that, if someone believes they have diabetes, they should focus on preventing complications such as blindness, common illness that involves dialysis, amputation, or perhaps death. Therefore, a balanced diet is necessary to prevent the progression of diabetes.

1.4 Methodology

The proposed technique begins by obtaining the dataset, followed by visualizing and displaying the dataset's original values. The dataset is subjected to some machine learning algorithms.

Here we will use many machine learning methods:

1: Import the necessary libraries and the diabetes dataset.

2: Preprocess the data to eliminate any missing details.

3: Scale the set data by 80% to create a training set and a test set.

4: Choose a machine learning method, such as Support Vector Machine, Decision Tree, logistic regression, or Random Forest.

5: Using the training data, build a model classifier using the stated machine learning technique.

6: Using the test set, run the classifier model for the stated machine learning technique.

7: Conduct a comparative analysis of the test performance results for each classifier.

8: Determine the best performing algorithm after reviewing it by using various factors.

1.4.1 The steps in methodology are:-

1) Data selection is the process of choosing the highest relatable data from a certain area in order to obtain values that are useful and aid learning. The diabetes dataset contains eight variables that help us to predict and avoid diabetes at early stage.

2) Data pre-processing is a Machine Learning approach which involves converting raw input into a suitable configuration. It consists of data cleansing, integration, transformation, and discretization.
3) Feature Extraction Using Principle Component Analysis: Feature Extraction of the dataset to obtain the best collection of characteristics for improved categorization. The feature vector is the collection of properties supplied by the dataset. Feature or dimensionality reduction will aid us by lowering compute and space complexity.

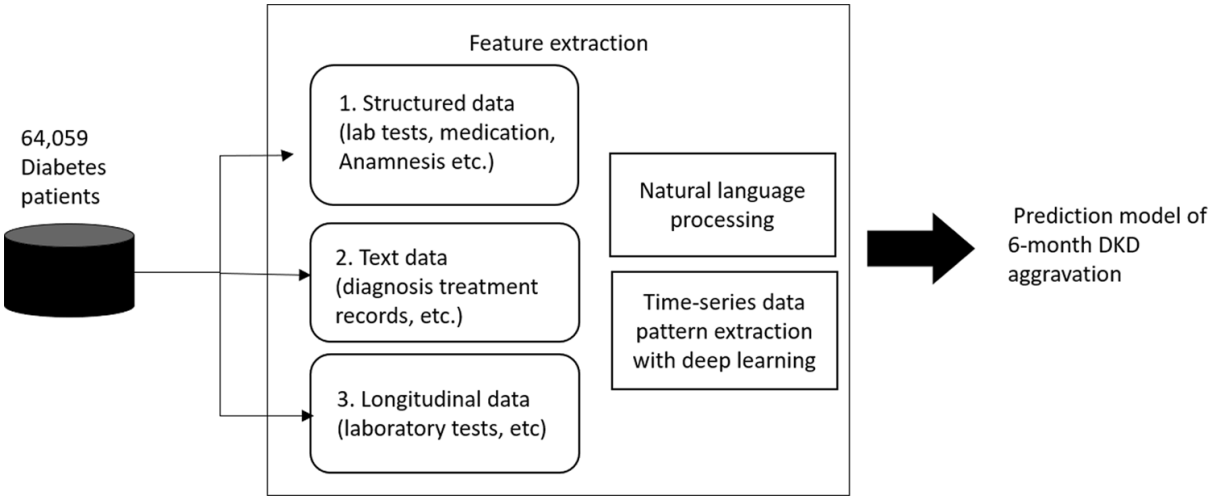


Fig 1.2: Feature Extraction

4) Validation: Many research employed two validation approaches, namely the hold-out method and the k-fold cross validation method, to assess the model's capabilities. Here we can use alternative techniques to tackle the problem depending on the purpose of each challenge and the magnitude of the data. The dataset is separated into two sections in the hold-out method: training set and test set. The training set is used to train the machine learning algorithm, whereas the test set is used to assess the model's performance. The training set is distinct from the test set, The entire dataset is utilized to train and test the classifier in the k-fold cross validation approach.

- First of all, the dataset is split into k parts called as folds. The approach employs k-1 folds to train the model and onefold to test it during the training phase. This procedure will be repeated k times, with each fold having the opportunity to be the test set. The final result is the average of all test results for all folds.
- The advantage of this technique is that each and every sample of the dataset are trained and tested, minimizing larger variance. We applied the five-fold cross validation approach for this research.

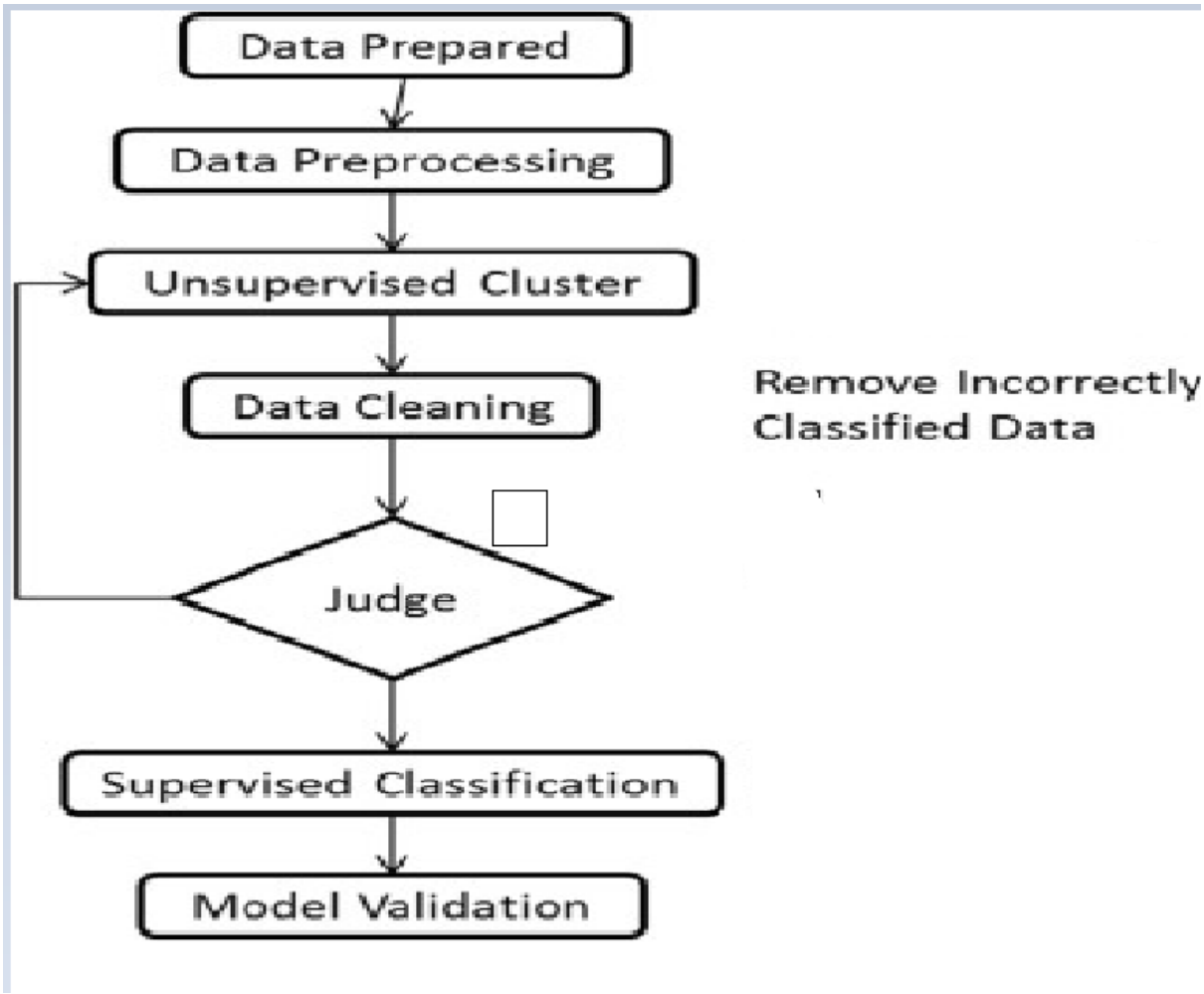


Fig 1.3 : Methodology Flowchart 1

To predict the diabetic patient, we offer a classification model with improved accuracy. For the model, we used several machine learning techniques such as classification, regression. The primary goal is to improve accuracy by applying various machine learning algorithms and selecting the best method for the task.

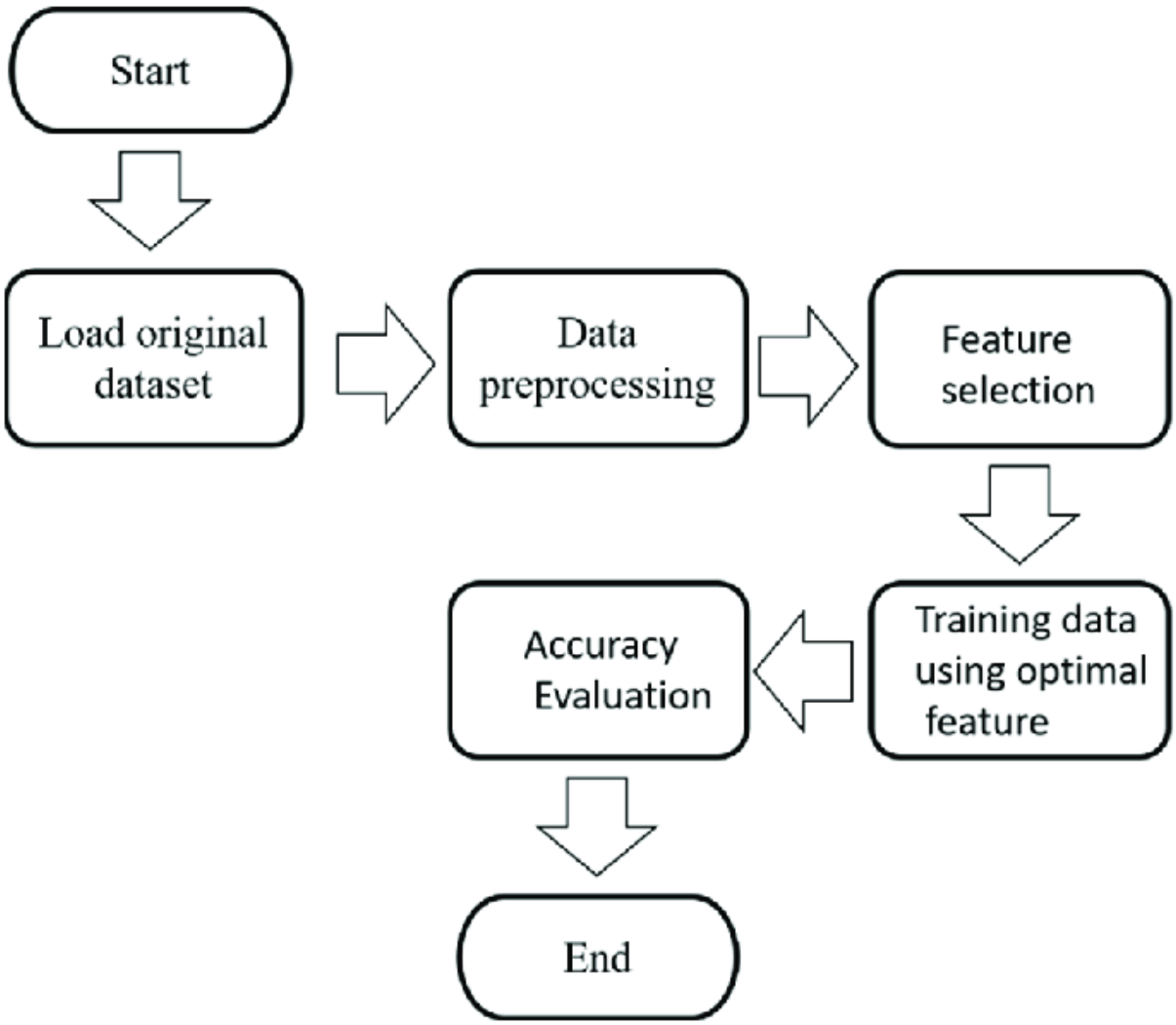


Fig 1.4 : Methodology Flowchart 2

1.5 Organization

The project Report has been organized into five distinct chapters, which are as follows:

Chapter 1: It includes the basic introduction of the project, problem statement includes the issues that our project addresses including the main aim of the project. The Methodology have methods that have been used to achieve it's goal, i.e in our case to predict if a person is diabetic or not.

Chapter 2: In this chapter, the knowledge behind Diabetes Prediction Model which includes Logistic Regression, SVM, Decision Tree and Random Forest Classifiers. This chapter goes into deep research on how Diabetes Prediction is done on various peoples by researching different Research Papers that have different types of classifiers and techniques that we may use in our project.

Chapter 3: In this chapter, the main purpose was to explain step by step build the Model using the research done before. The first step was to research on Diabetes Prediction using Machine Learning, then visualizing the data from the data set, then preprocessing the dataset to train the model. It also incudes different types of accuracy measurements that we will acquire in this model by applying various algorithms.

Chapter 4: In this chapter, the report contains the analysis of Model created and the corresponding results obtained at several stages of the project. Therefore we can get a clear idea of the accuracy of this Model, several times it ranges from 70 to 80% . Next we compare the different accuracies from the different Machine Learning Algorithms, so we can get the best accuracy of them all.

Chapter 5: In this chapter, the report contains the conclusion obtained from the research and work done. It also contains the outputs obtained from our final project model and also the accuracies of the different Machine Learning Algorithms. It also contains the future work than can be added to the Model.

Chapter-2 LITERATURE SURVEY

Priyanka Rajendra and Shahram Latifi, 2021, in their Research Paper focused on building a predictive model for diabetes to identify if a certain patient has diabetes or not by using Logistic Regression to build the model, Where they used PIMA Indian Dataset. In the Dataset, there are all female patients who are at least 21 years old. This Paper explains the step by step process of the model, - from its design to its implementation. They also used a second dataset from Vanderbilt, which is based on a study of rural African Americans in Virginia. In this dataset they have used 16 features, also the dataset consists of both male and female patients.

Ramesh, 2015, in his research paper proposed a framework for prediction of diabetes using Support Vector Machine. The dataset used in this research paper contains the amount of oxygen in the blood, pulse rate, diastolic blood pressure, medication status, systolic blood pressure, number of calories consumed in a day, count of the steps, and which type of activity is performed by the user. These Vital Information was extracted by the mobile application system, Two cloud services named Google fit and iHealth were used by the author to extract the vital information obtained from smart wearable devices. Vitals extracted from smart devices are then visualized by the server to predict diabetes, then accuracy of SVM was found to be 79%.

In the Research Paper of Aishwarya Mujumdar and V Vaidehi, 2019 used several machine learning algorithms such as Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, etc. Here they used two different datasets- the PIMA Indian and another Diabetes dataset for testing the various models, Logistic Regression gave them an accuracy of 96% .

Tejas Joshi and Pramila chawan, 2018, in their research paper used two algorithms- Logistic Regression and SVM to design a diabetes prediction model. The model was built and executed and was found that SVM gave them an accuracy of 79% .

Chapter-3 SYSTEM DEVELOPMENT

We will learn about the different classifiers used in machine learning to predict diabetes in this part. We will also describe the technique we have suggested in order to increase accuracy.

We have used the following algorithms in our model.

3.1 Logistic Regression

3.2 SVM

3.3 Random Forest

3.4 Decision Tree

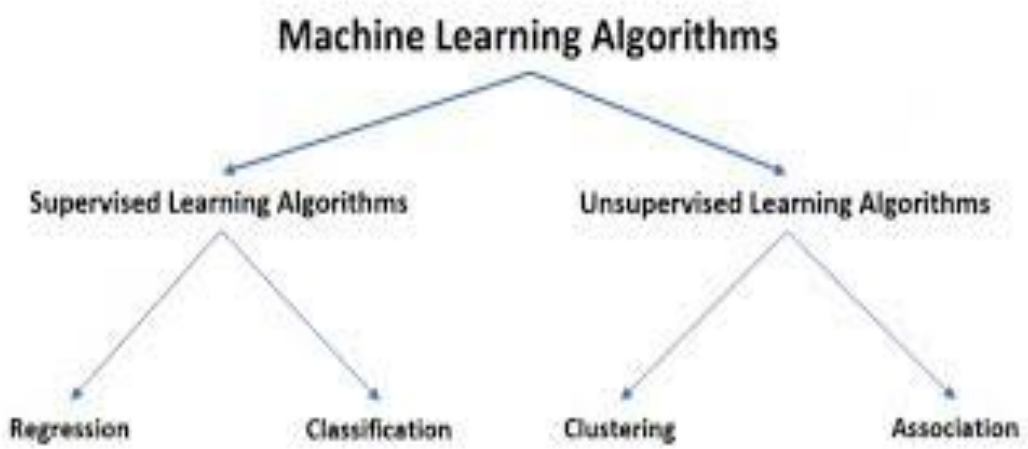


Fig 3.1 : Machine Learning Algorithms

3.1 Logistic Regression:-

Using prior observations from a data set, a statistical analysis method known as logistic regression predicts a binary outcome, such as yes or no. A logistic regression model predicts a dependent data variable by looking at the association between one or more already existing independent variables.

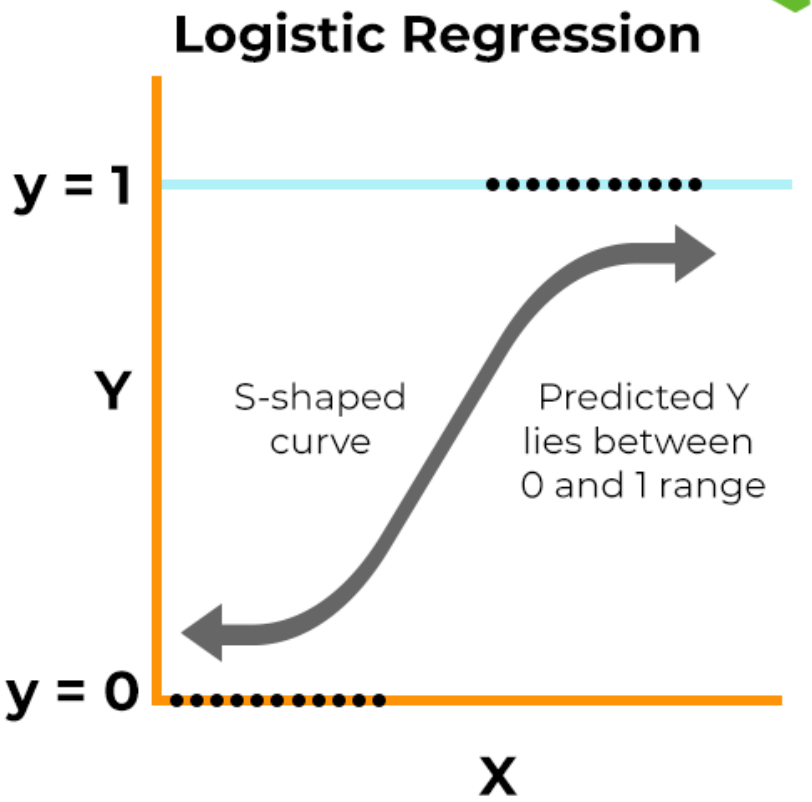


Fig 3.2 : Logistic Regression Graph 1

3.1.1 Working:-

Predicting a numeric outcome or a qualitative class is a common task for machine learning. A regression problem is a popular term for the first situation. In a linear regression scenario, a continuous variable serves as the input, and a numerical value is predicted. A classification challenge is what the task is when anticipating a qualitative result (class). Predicting what a user will purchase or whether a target user will click on an internet advertising are two examples of categorization challenges.

The logistic regression technique is used in practice to examine correlations between variables. Following the Sigmoid function, that changes numerical values into an formula of probability between 0 and 1.0, it assigns probabilities to discrete outcomes. Depending on whether or not the event occurs, probability ranges from 0 to 1. With a cut-off of 0.5, you may divide the population into two groups for binary predictions. Group A includes everything that is greater than 0.5, while group B includes everything that is less than 0.5.

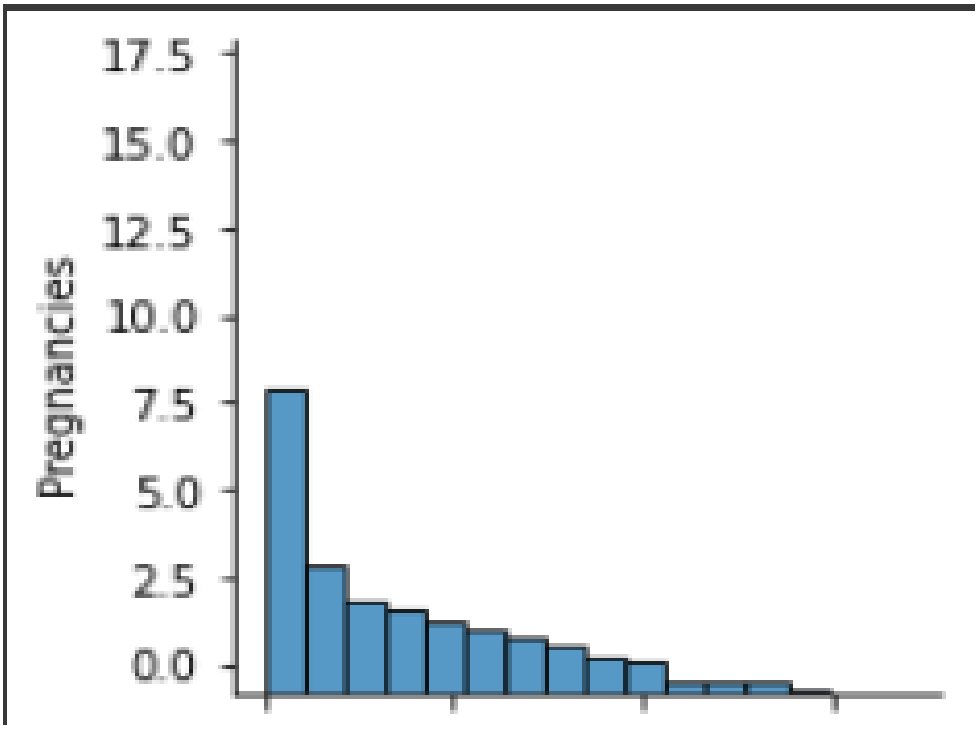


Fig 3.3 : pregnancies Regression graph

After data points are classified using the Sigmoid function, a hyperplane is employed as a decision line to divide two groups (as much as feasible). The decision boundary can then be used to forecast the kind of upcoming data points.

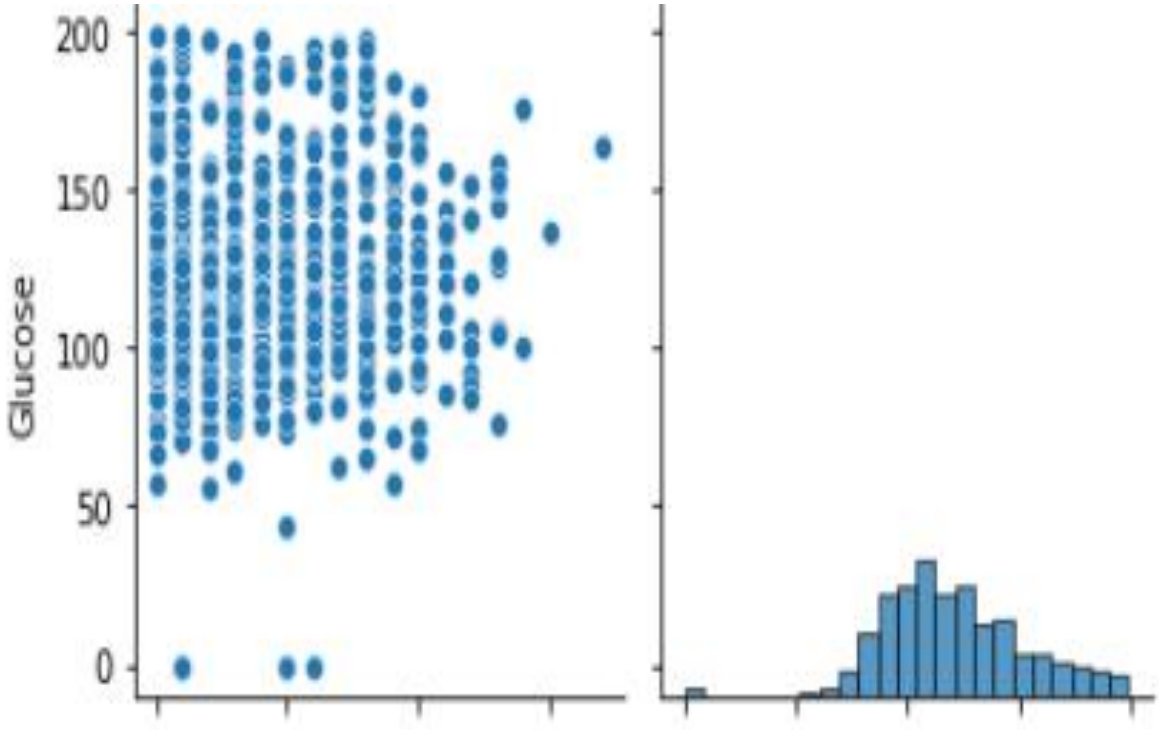


Fig 3.4 : Glucose Graph

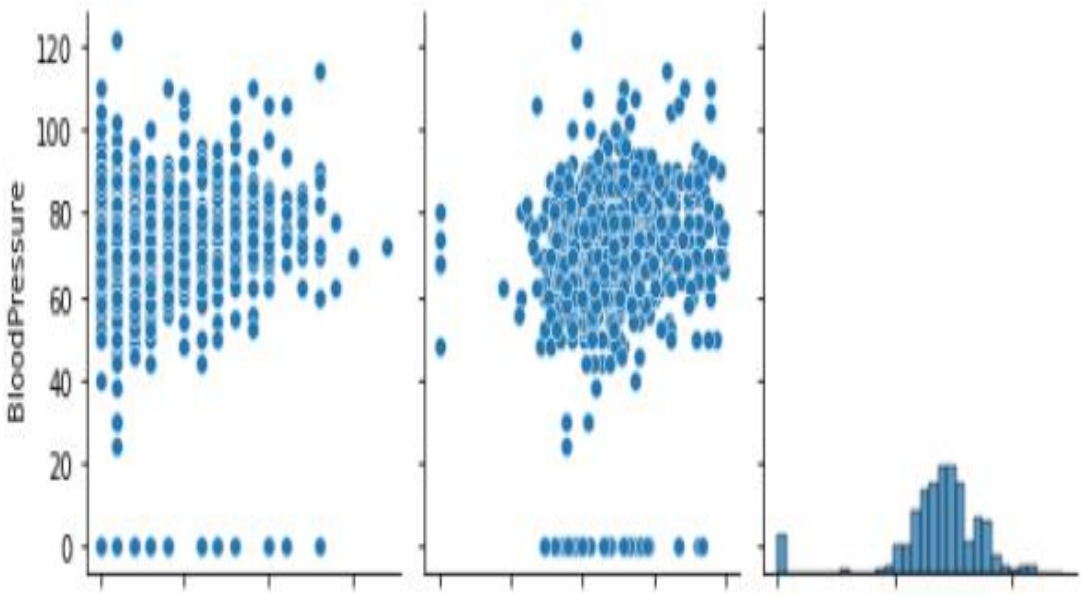


Fig 3.5 : Bloodpressure Graph

3.1.2 Uses:-

3.1.2.1 Fraud Detection:

Teams may use to identify data anomalies that are suggestive of fraud using logistic regression models. Banking and other financial institutions may discover that particular behaviours or characteristics are more frequently linked to fraudulent activities in order to better protect their consumers. SaaS-based businesses have also begun utilising these techniques to eliminate fake user accounts from their datasets while conducting data analysis on business performance.

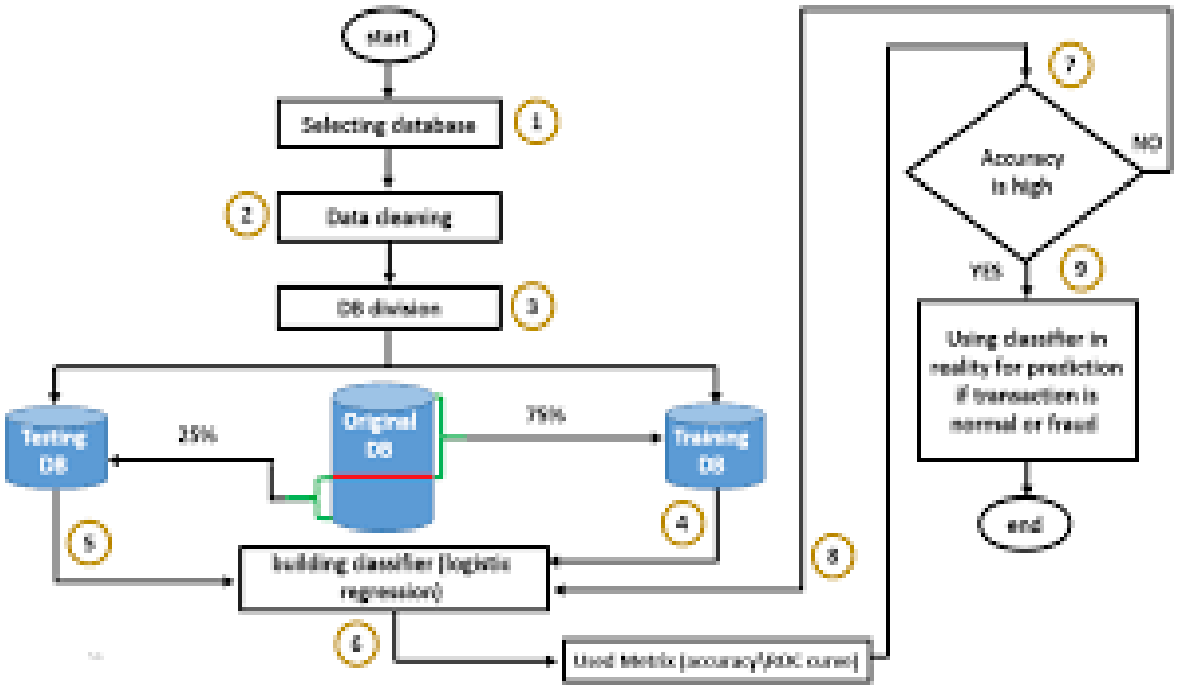


Fig 3.6 : Flow Chart of the proposed approach(fraud prediction)

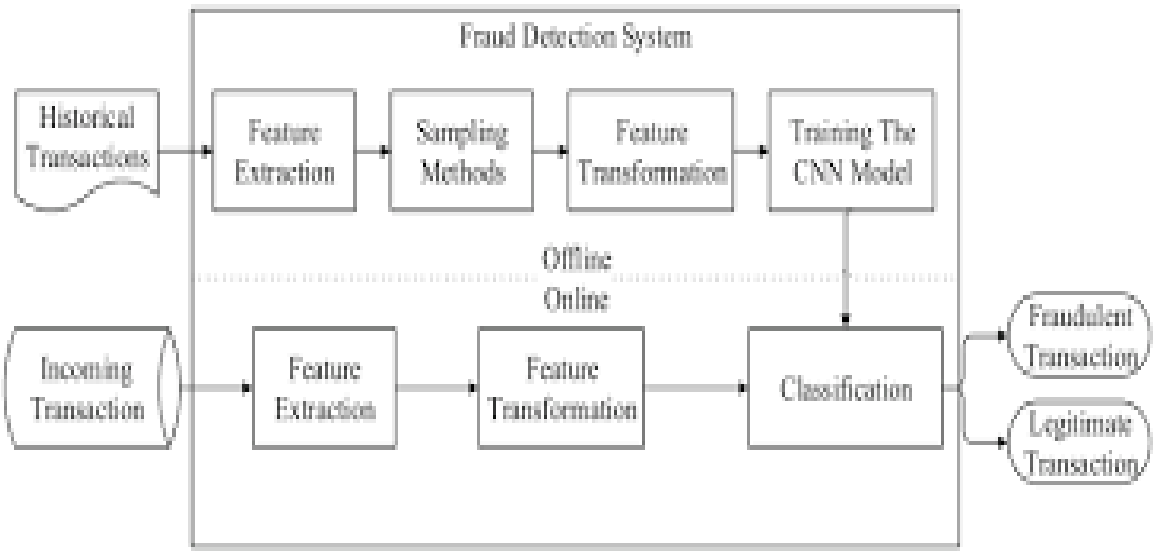


Fig 3.7 : General Scenario of the fraud detection system

3.1.2.2 Disease Prediction:

This analytics strategy may be applied to medicine to forecast the likelihood of sickness or illness in a certain group. Healthcare institutions can set up preventative treatment for those who have a higher risk of developing a certain ailment.

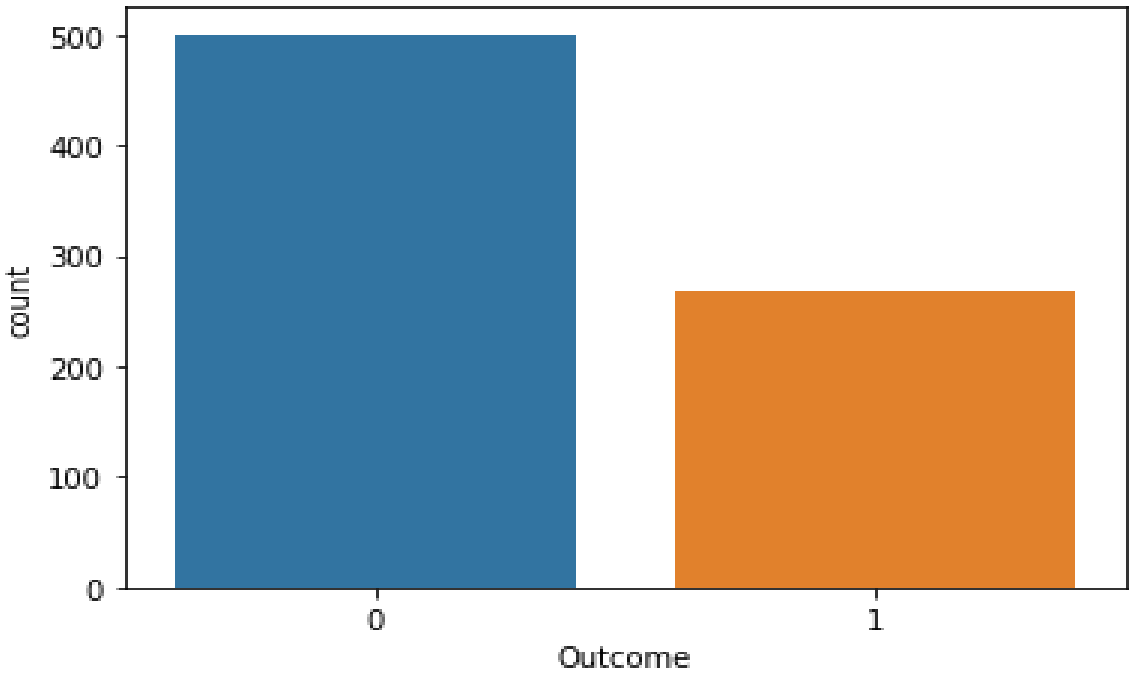


Fig 3.8 : Disease/not Disease graph

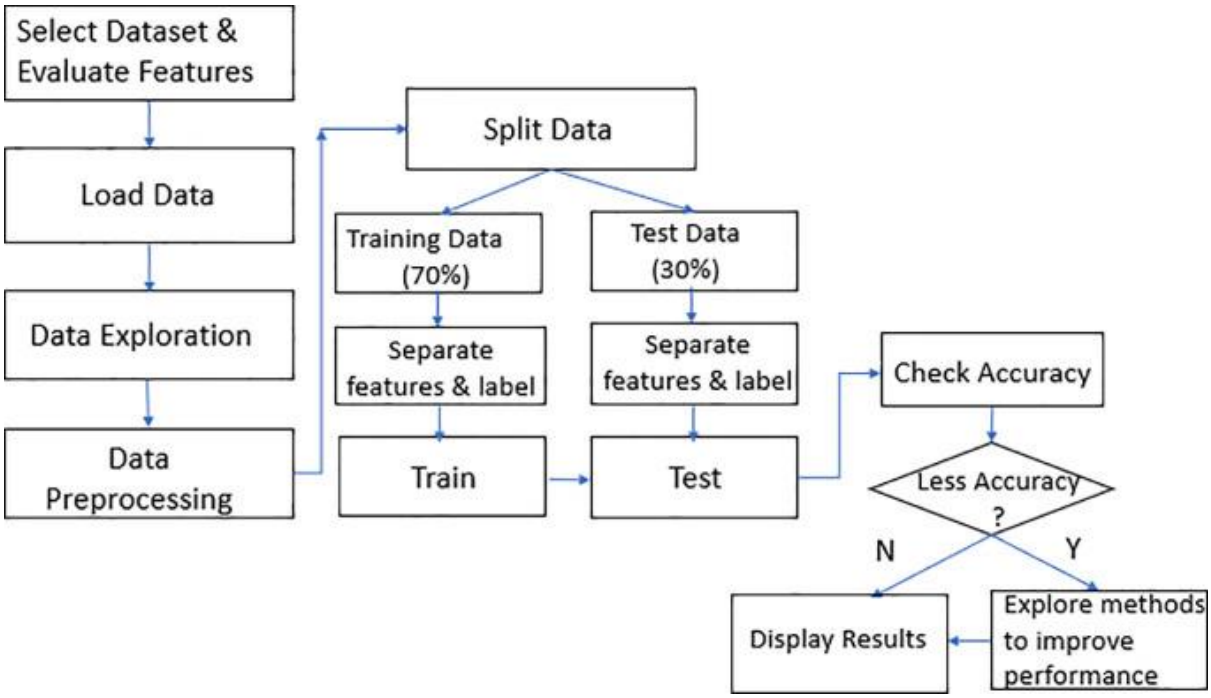


Fig 3.9 : Split Flowchart (data splitting)

3.1.2.3 Churn Prediction:

Churn in various organisational tasks may be indicated by certain actions. If strong performers are at risk of leaving the firm, for instance, human resources and management teams may be interested in finding out. This sort of

information might spark discussions about the company's culture or pay practices. As an alternative, the sales team would try to find out which of its customers might decide to do business elsewhere. In order to prevent income loss, this may inspire teams to develop a retention plan.

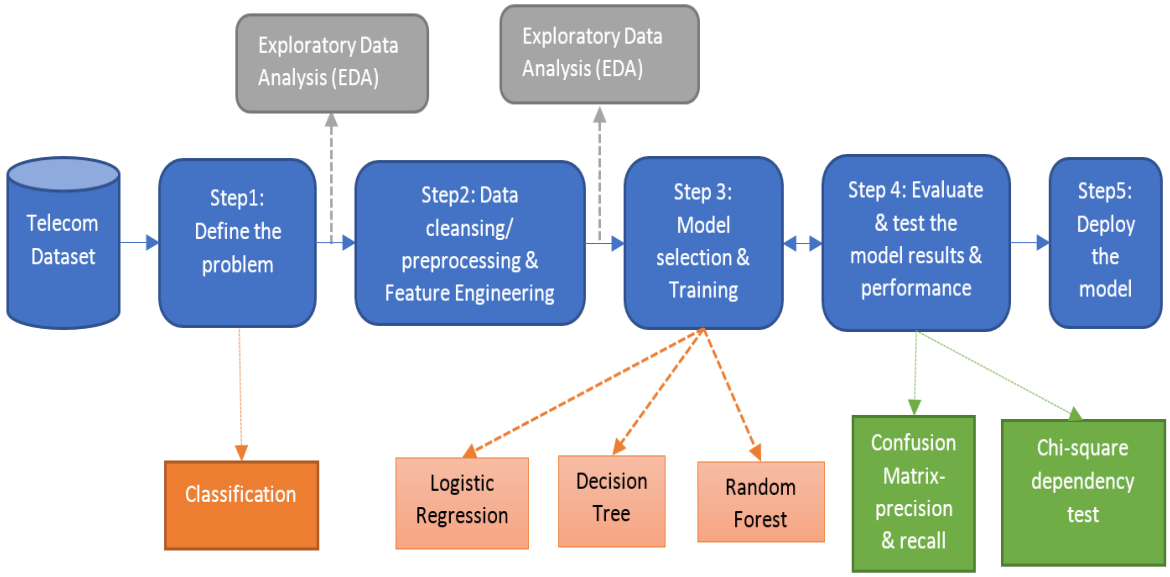


Fig 3.10 : churn prediction flowchart (telecom)

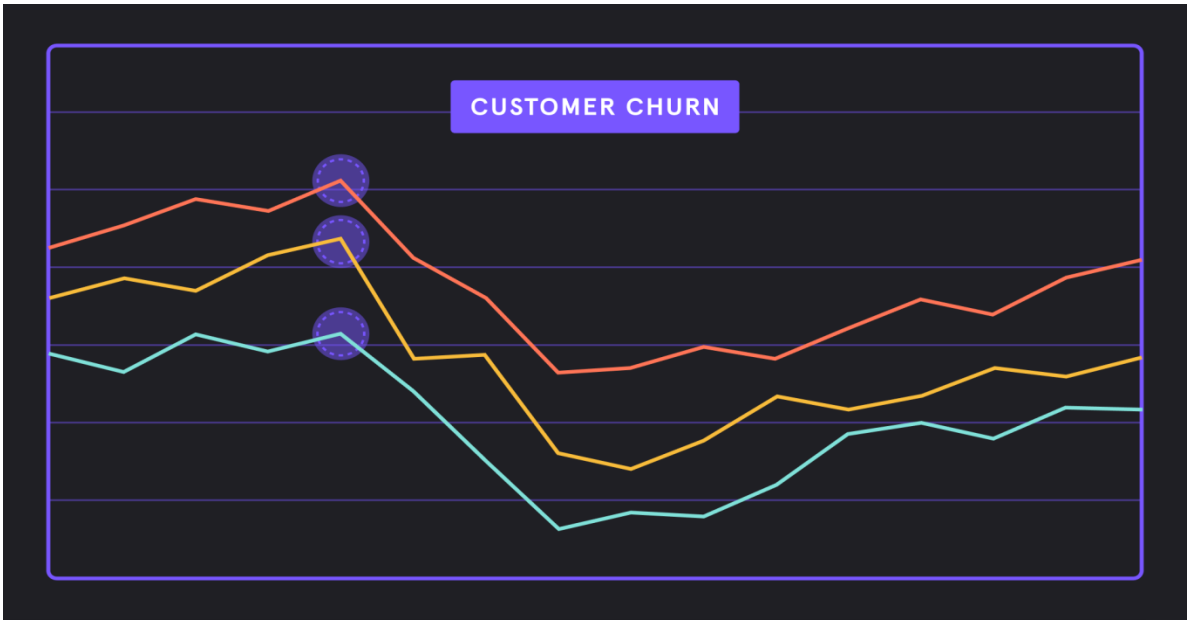


Fig 3.11 : Customer Churn Graph(Internet Provider)

3.1.2.4 Credit Scoring:

A financial organisation called ID Finance creates credit score prediction algorithms. They require simple to understand models for their systems. A regulator may at any time inquire of them about a specific judgement.

One such stage in the data preparation for credit score modelling is the reduction of correlated variables. If your model has more than 15 variables, it becomes challenging. It is easy to conclude what factors afftects a higher and lesser effect on the predictions' result when applying logistic regression. Moreover, applying methods such as recursive feature elimination, it is easier to calculate the best amount of characteristic and get rid of not to the point variables.

The last stage allows them to export the forecast findings to an Excel file, which allows analysts—even those without technical knowledge—to draw conclusions from the information.

Fig 3.12 : Credit Scoring Figure

At some time, ID Finance rejected the usage of external statistical software and switched to Python for their model-building procedures. The speed of model creation has significantly increased as a result. They didn't, however, give up on logistic regression in favour of more sophisticated algorithms. The findings of logistic regression, which is frequently employed in credit scoring, are impressive.

3.1.2.5 Hotel Booking:

On Booking.com, machine learning methods are essentially applied everywhere. They seek to recognise objects and foresee user intentions. What are you doing, where are you heading, and where do you prefer to stop? Certain predictions are made even if the user hasn't yet written anything in the

search line. But how did they start doing this? Nobody has ever been able to build a complex, large-scale system from scratch using a variety of machine learning algorithms. They created some simple models and initially gathered some fundamental data.

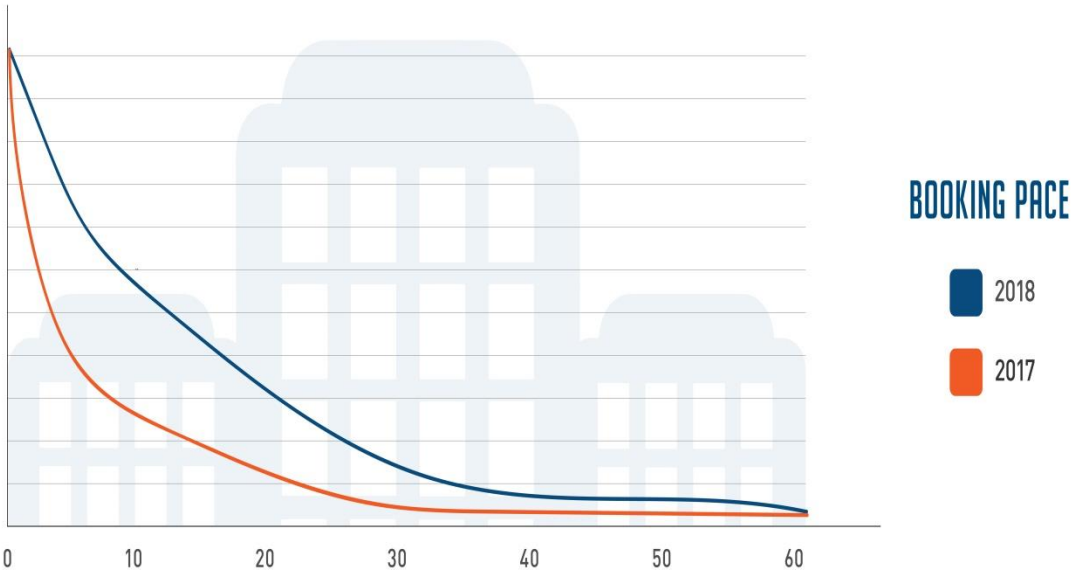


Fig 3.13 : Hotel Booking Graph

At websites like Booking.com, the majority of the features are more category than numerical. Prediction of an occurrence without particular user information is sometimes essential. For instance, the only information they know is the user's home address and intended destination. For these purposes, logistic regression is excellent.

This histogram represents logistic regression's attempt to forecast whether a user would alter a departure date or not. 2018 saw its presentation at the HighLoad++ Siberia conference.

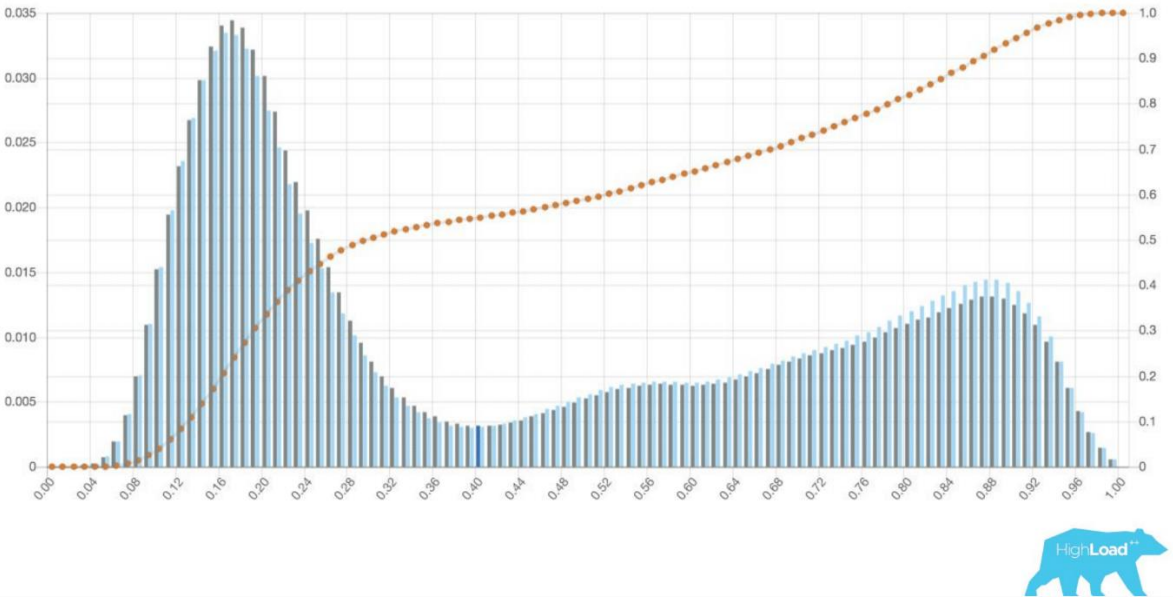


Fig 3.14 : Logistic Regression Application Graph (booking)

Two kinds of users could be separated through logistic regression. The business may then determine whether to update an interface for a certain type of users based on this data.

If you have used Booking, it's likely that you have seen this capability. You now know that this application uses logistic regression in some way.

3.1.2.6 Gaming:

The benefits of logistic regression includes is speed, which helps us in the gaming world. Speed comes in handy when playing in games. Games that involve in-game purchases to enhance your character's gaming capabilities, or for a different appearance and player interaction, are quite famous nowadays.

Fig 3.15 : ML in Gaming Industry

The largest video game giant in the world is Tencent. It offers equipment recommendations for players utilizing several platforms. Their programme calculates a reachable quantity of gamer activity data and makes suggestions on the equipment a certain gamer would like to pick up fastly. Logistic regression is the algorithm used here.

Three different kinds of recommendation systems exist. Based on reviews from users with similar tastes who have made prior purchases as well as other activities, the collaborative system anticipates what the user would want to purchase. A content-based algorithm bases its choice on the characteristics listed in the item description and the interests the user listed in her profile. The third form is a hybrid, which combines the first two categories.

Fig 3.16 : Game Image

3.2 SVM (Support Vector Machine):-

A deep learning system known as a support vector machine (SVM) uses supervised learning to classify or predict the behaviour of groupings of data. Supervised learning systems in AI and machine learning give input and intended output data that are labelled for categorization.

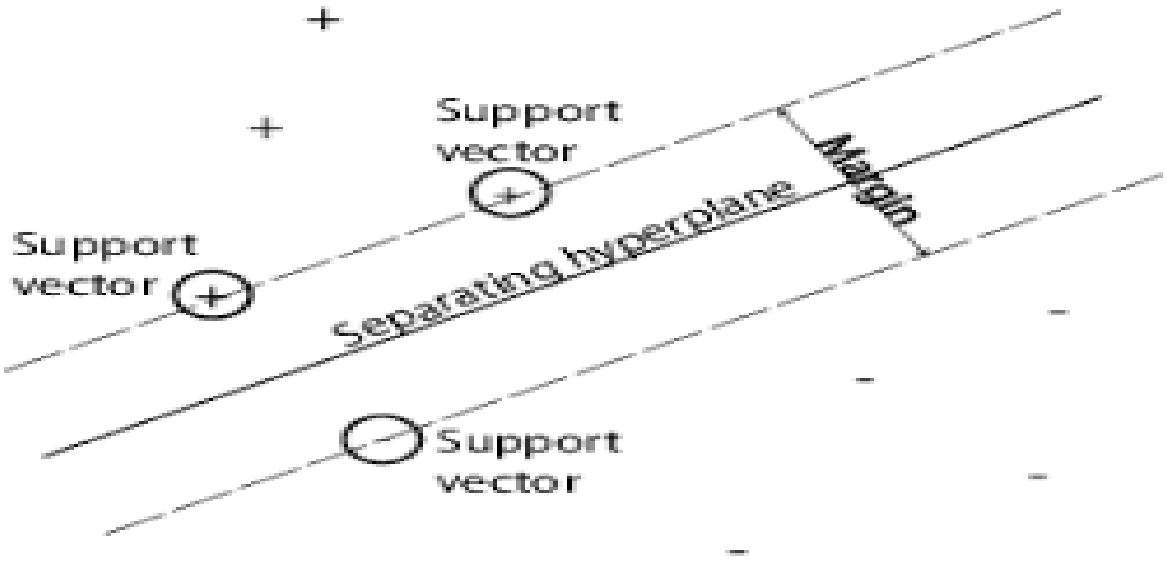


Fig 3.17 : Support Vector Figure

3.2.1 Working:-

The best way to comprehend Support Vector Machines' fundamentals and how they operate is with a straightforward example. Imagine that our data contains two features—x and y—and that we have two tags—red and blue. Given a pair of (x,y) coordinates, we want a classifier that can determine if an object is red or blue. We use a plane to plot our already-labeled training data:

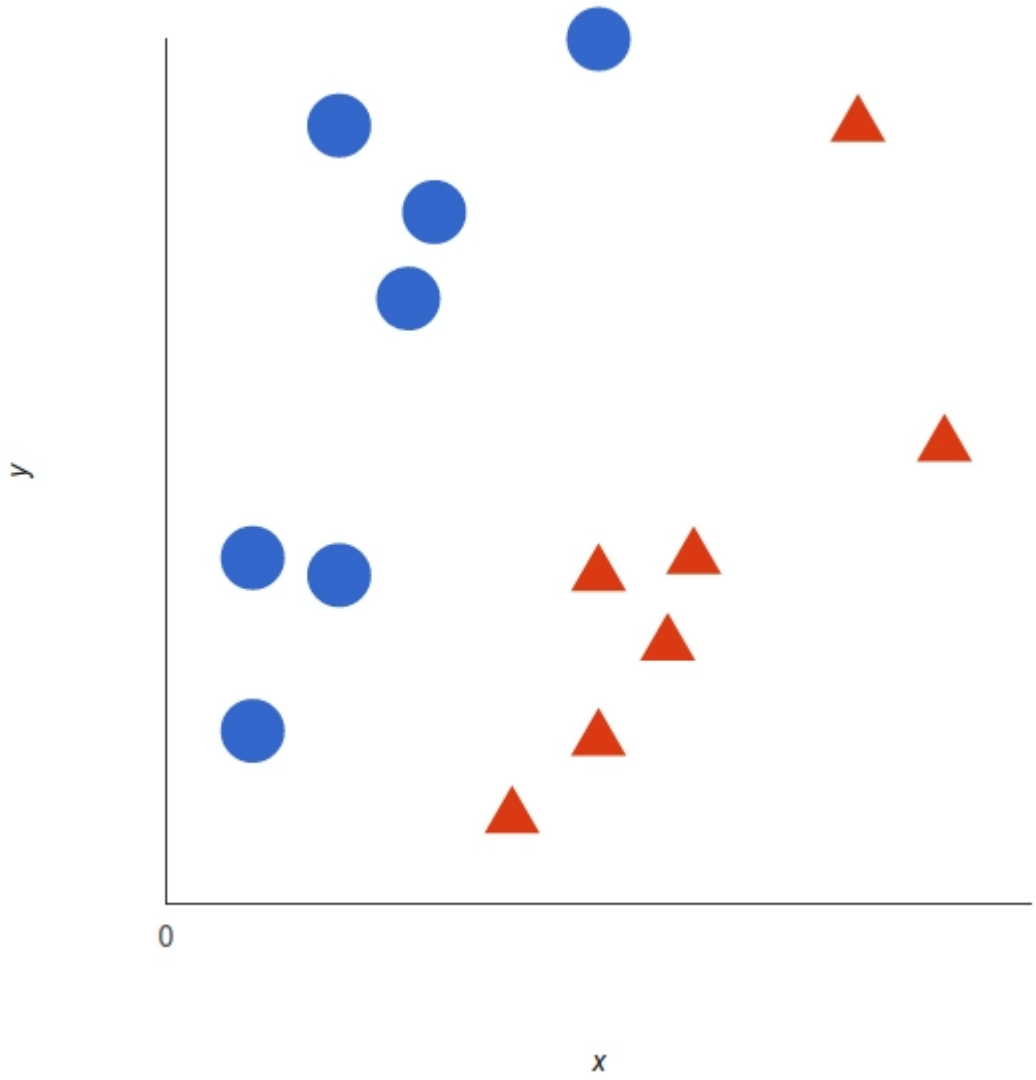


Fig 3.18 : Text Classification Graph using SVM

These data points are sent into a support vector machine, which produces the hyperplane—which is just a line in two dimensions—that best separates the

tags. The decision boundary is represented by this line; everything falling on one side of it will be classified as blue, and anything falling on the other as red.

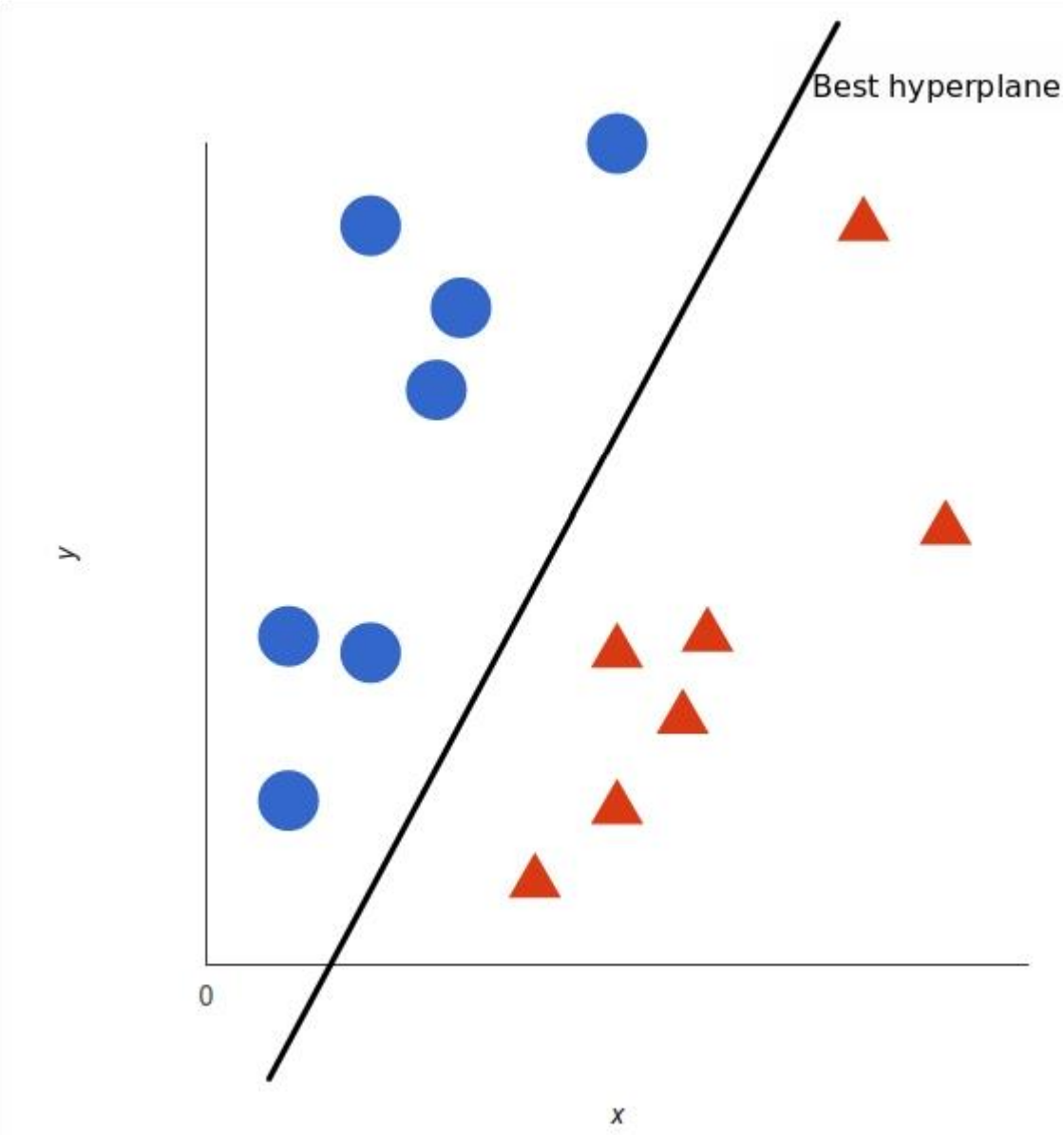


Fig 3.19 : Decision Boundry Graph

Tell us the finest hyperplane, though? It is the only which optimises the edges from both sides according to SVM. Or, to assemble it in different way, the hyperplane with the largest distance from the highest element of each tag.

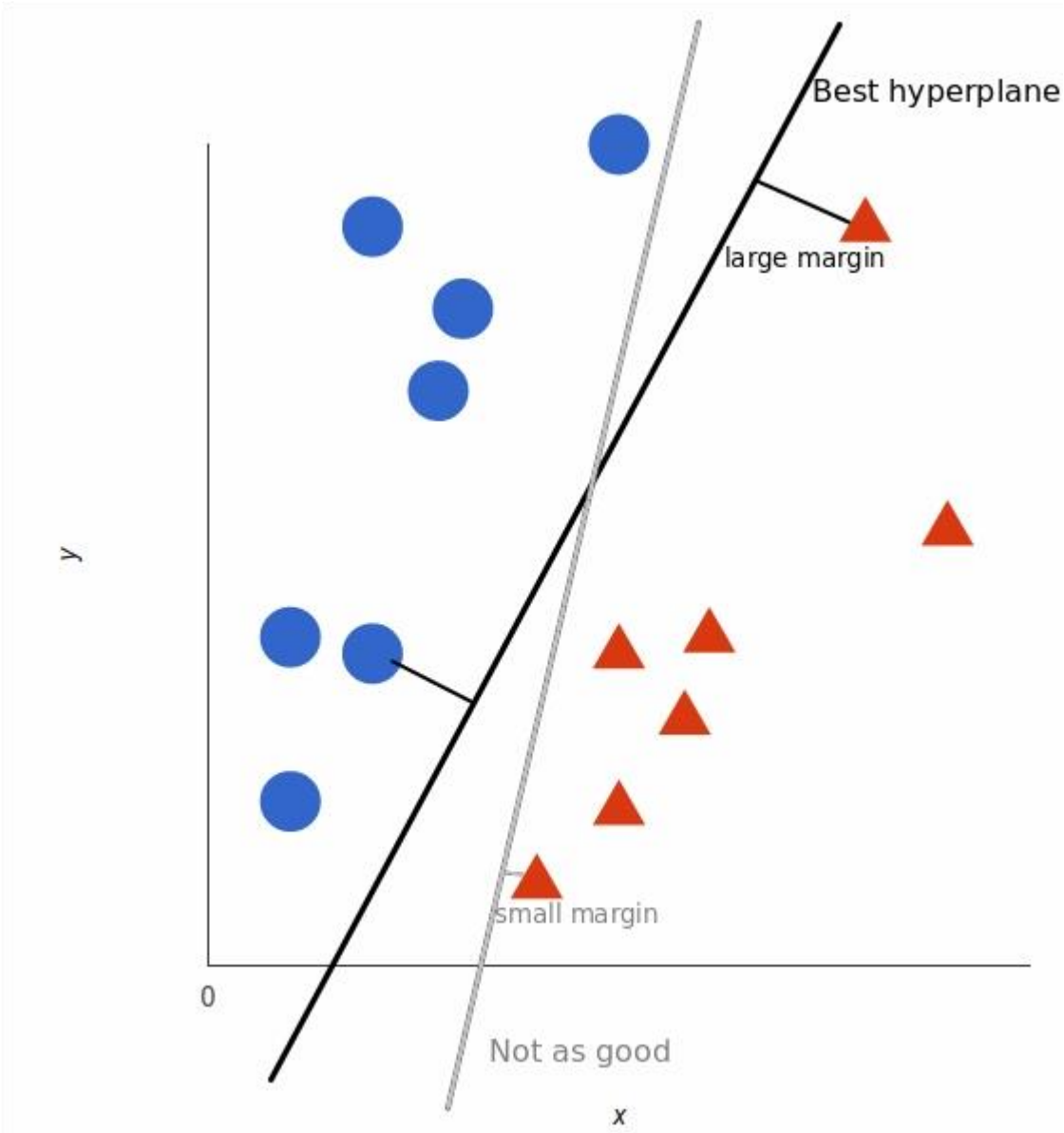


Fig 3.20 : Linear Data Graph

This example was simple since red and blue could easily be separated using a straight line because the data was obviously linearly separable. Unfortunately, things are rarely that easy. Look at this scenario:

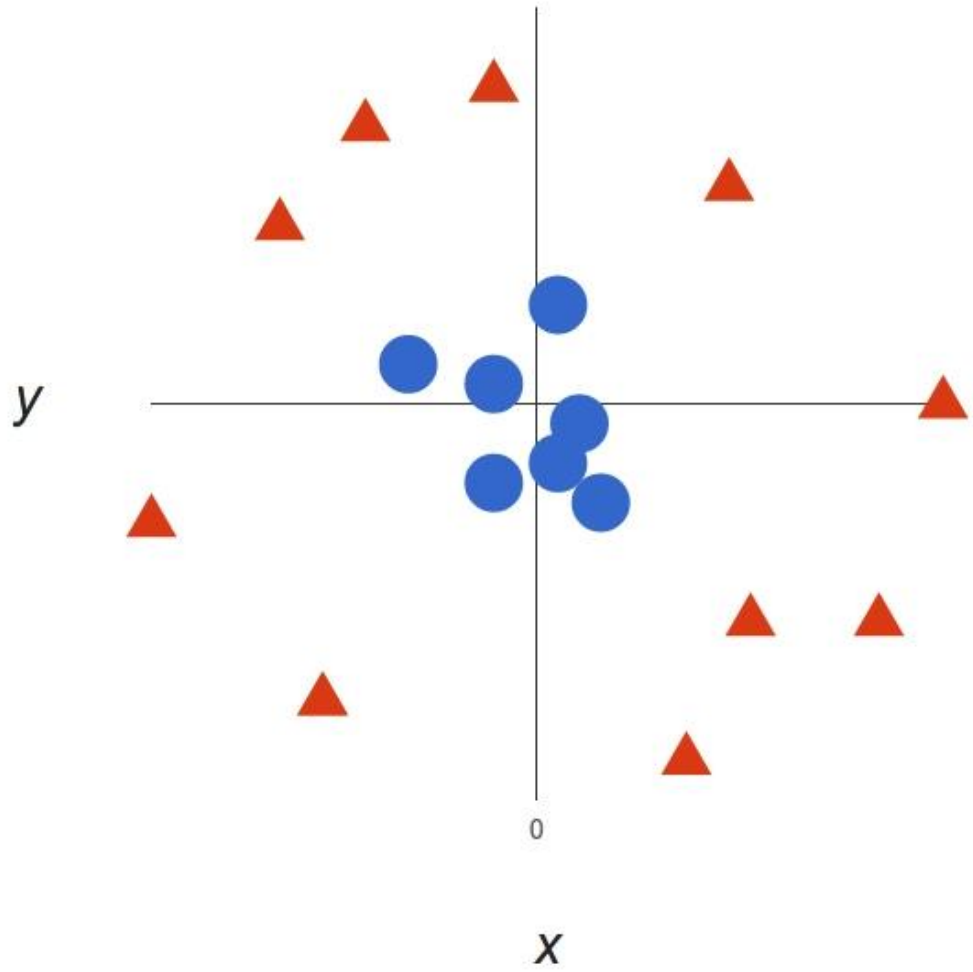


Fig 3.21 : Non Linear Data Graph 1

It is quite obvious that the decision boundary is not linear (a single straight line that separates both tags). However, the vectors appear to be easily separable because of how clearly they are divided.

Therefore, this is what we'll do: we'll introduce a third dimension. We only had x and y up to this point. We add a new dimension called z, and we decide how it will be computed so that it is easy for us to use: z = x2 + y2 (you'll note it's the equation for a circle).

We shall have a three-dimensional space as a result. A portion of that area appears as follows:

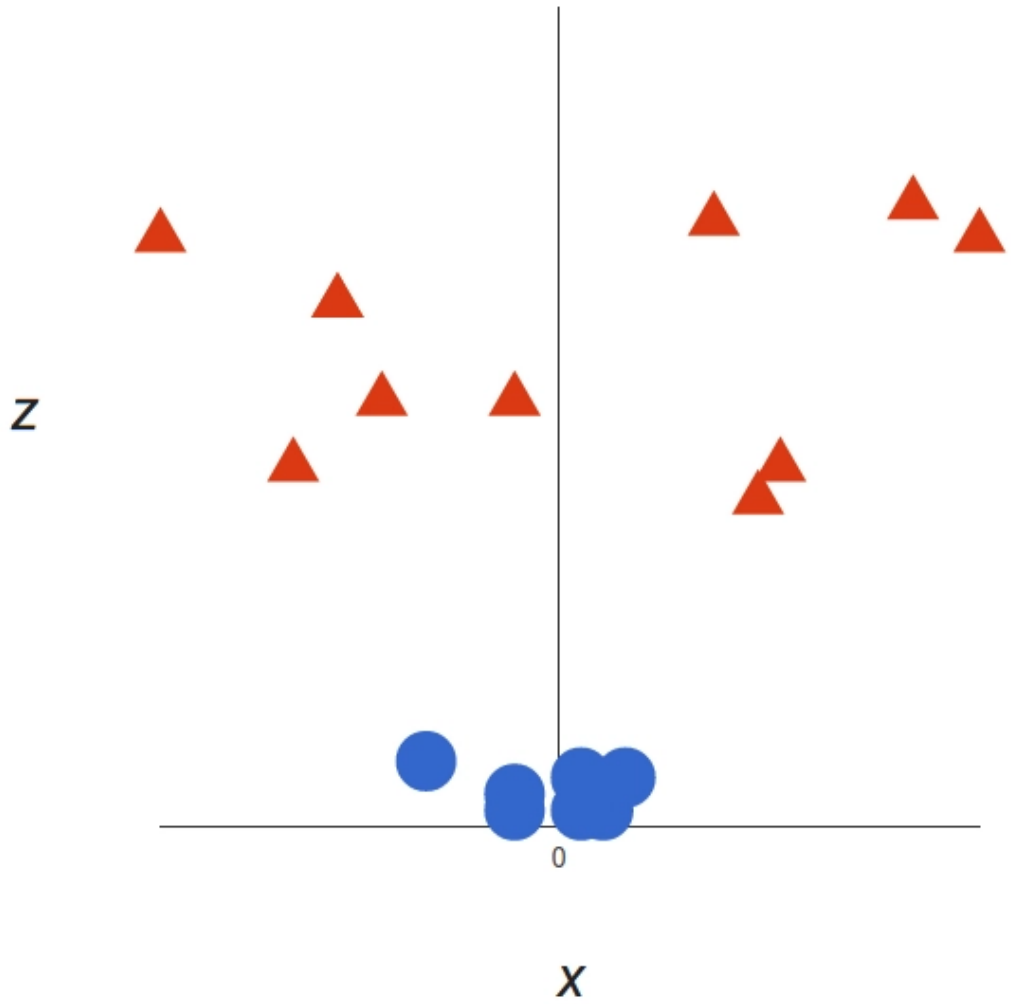


Fig 3.22 : Non Linear Data Graph 2

What can SVM do with this? Let’s see:

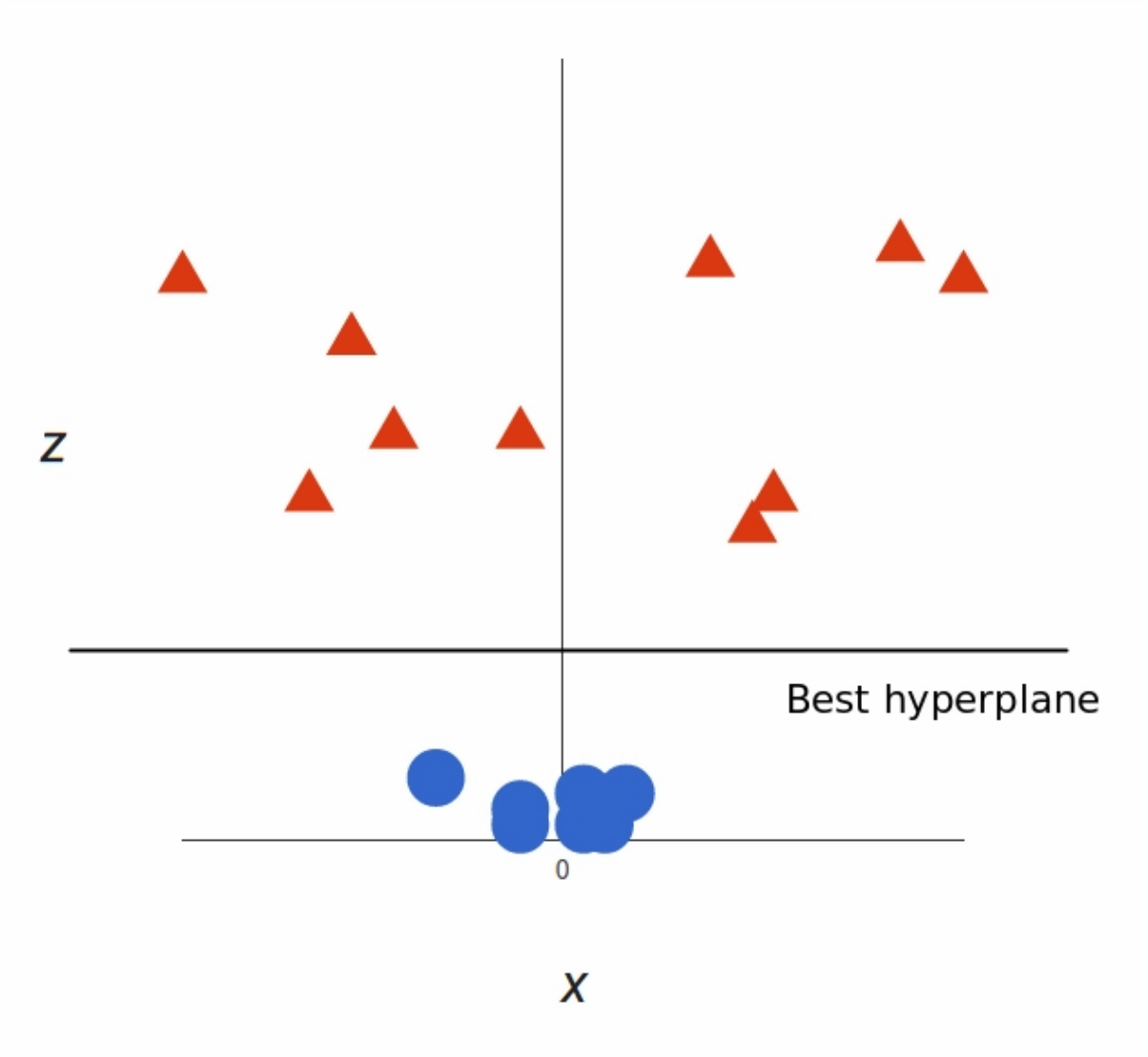


Fig 3.23 : Non Linear Data Sepration 1

I love that! The hyperplane is a plane that is perpendicular to the x axis at a particular z (let's assume z = 1) because we are now in three dimensions.

The final step is to map it back to two dimensions.

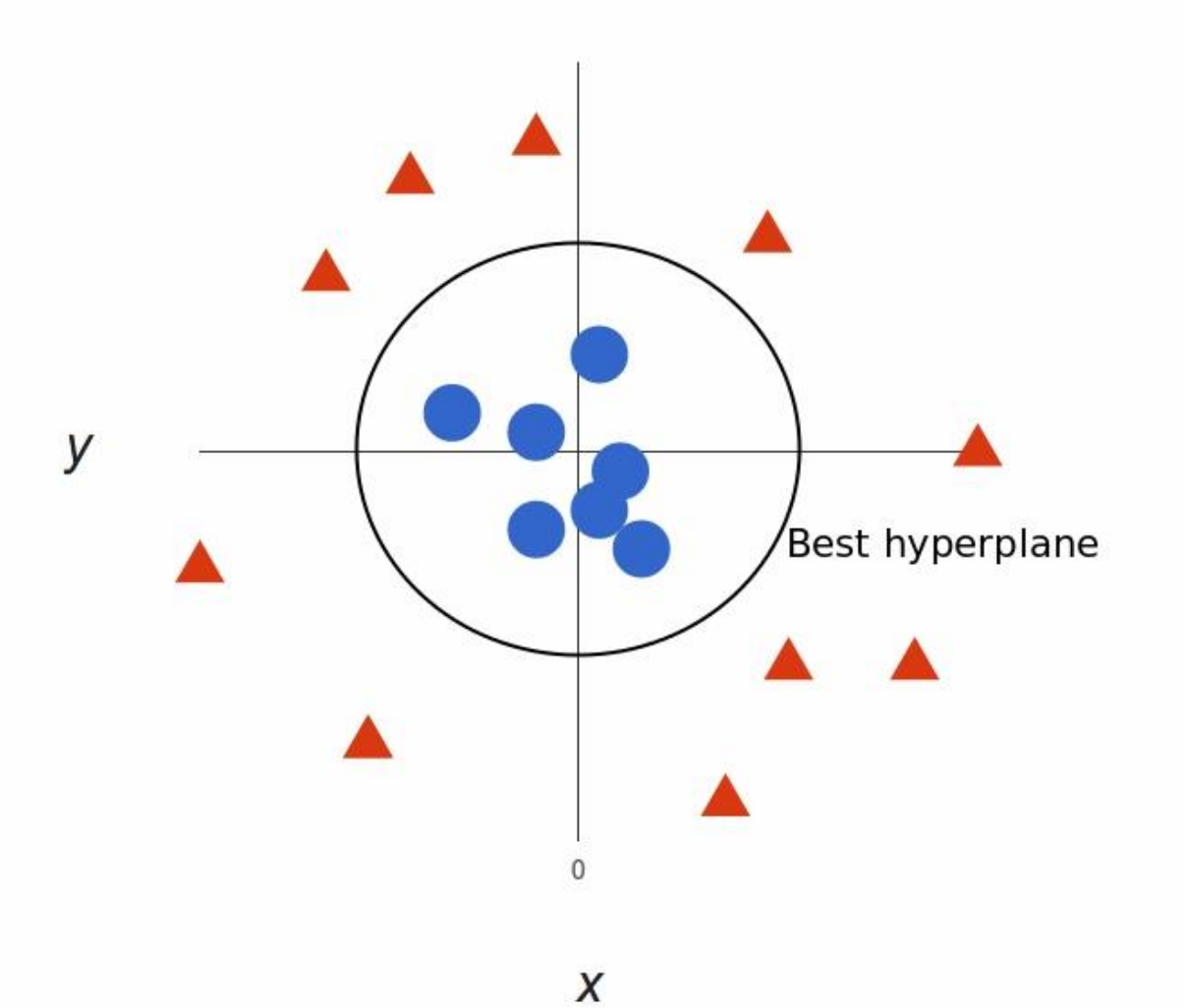


Fig 3.24 : Non Linear Data Sepration 2

And we're off! Our decision boundary is a circle with radius 1, which uses SVM to divide the two tags.

3.2.2 Uses:-

As we've seen, supervised learning techniques are necessary for SVMs. SVM is used to accurately categorize unknown data. SVMs have several uses in a variety of industries.

Some common applications of SVM are-

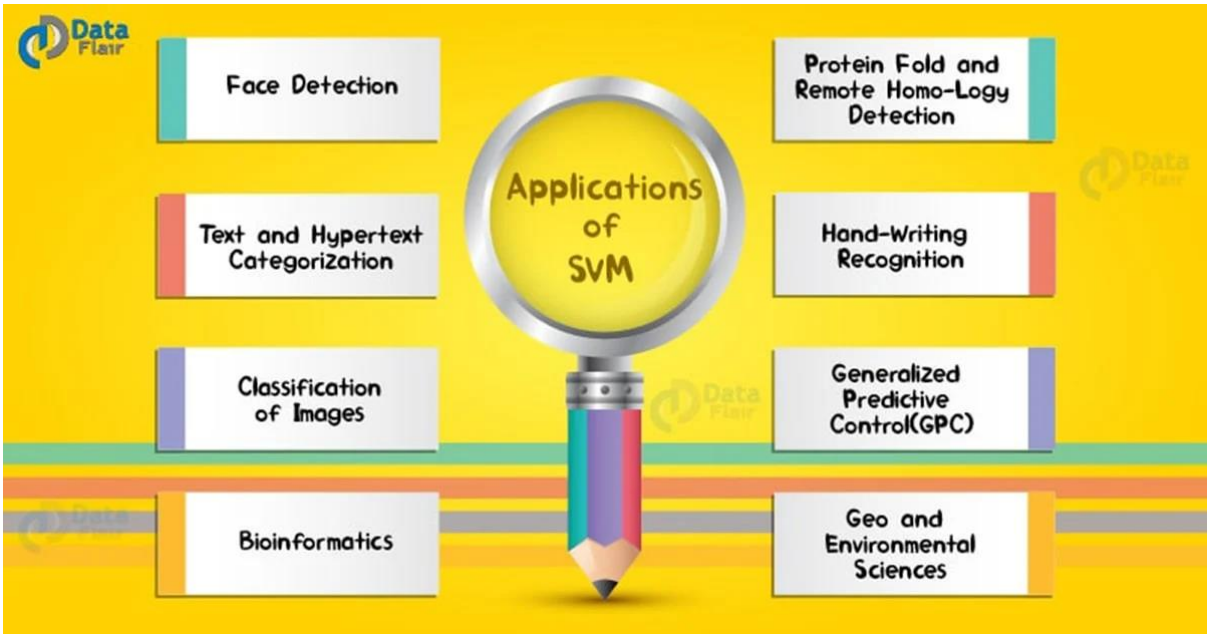


Fig 3.25 : Applications

3.2.2.1 Face Detection:

It divides the elements of the picture into face and non-face categories. It has training information for n × n pixels with the face (+1) and non-face classes (-1). Then, it separates each pixel's characteristics into face- or non-face categories. based on pixel brightness, draws a square border around faces and classifies each image using the same method.

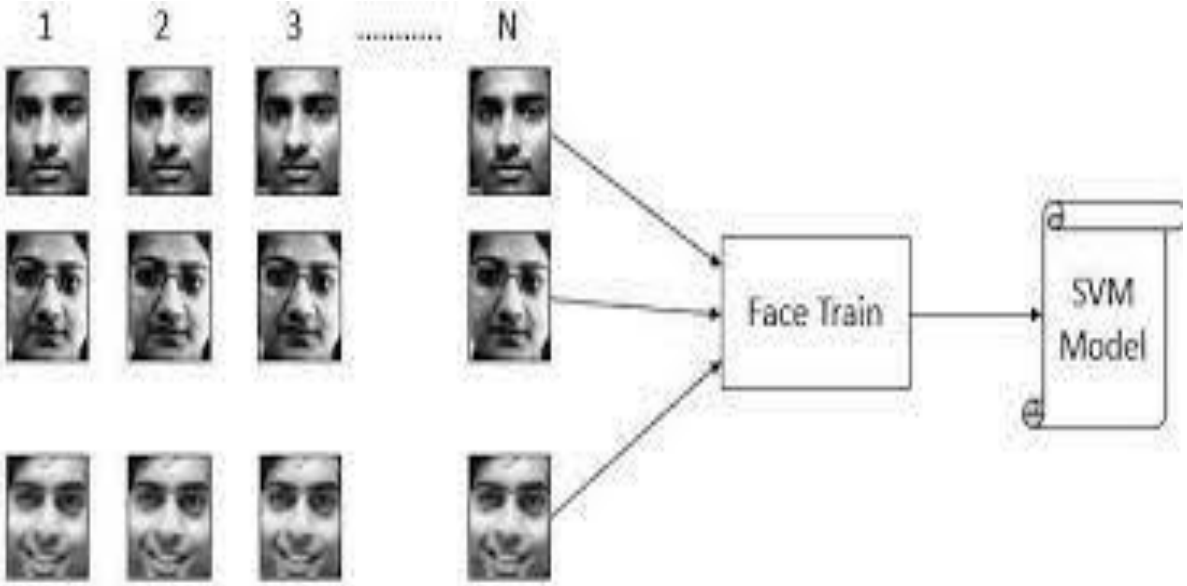


Fig 3.26 : Face Detection

3.2.2.2 Text and Hypertext Categorization:

allows both inductive and transductive models to be categorized using text and hypertext. It classifies documents into many categories, including news stories, emails, and web pages, using training data.

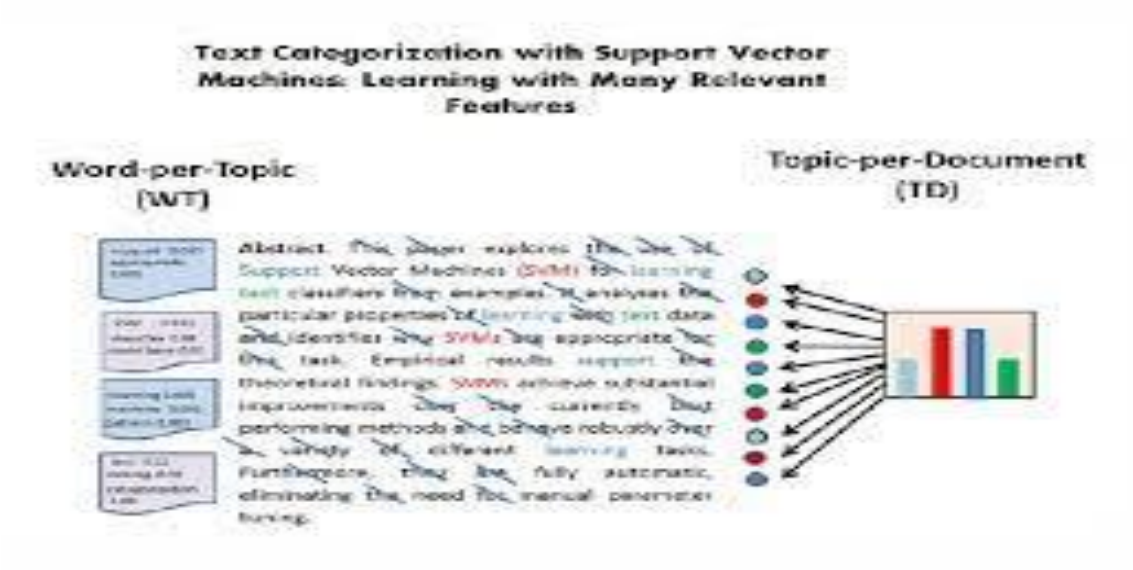


Fig 3.27 : Word Prediction

3.2.2.3 Classification of Images:

SVMs are able to categorise pictures more accurately. Compared to conventional query-based refining systems, it has a greater accuracy.

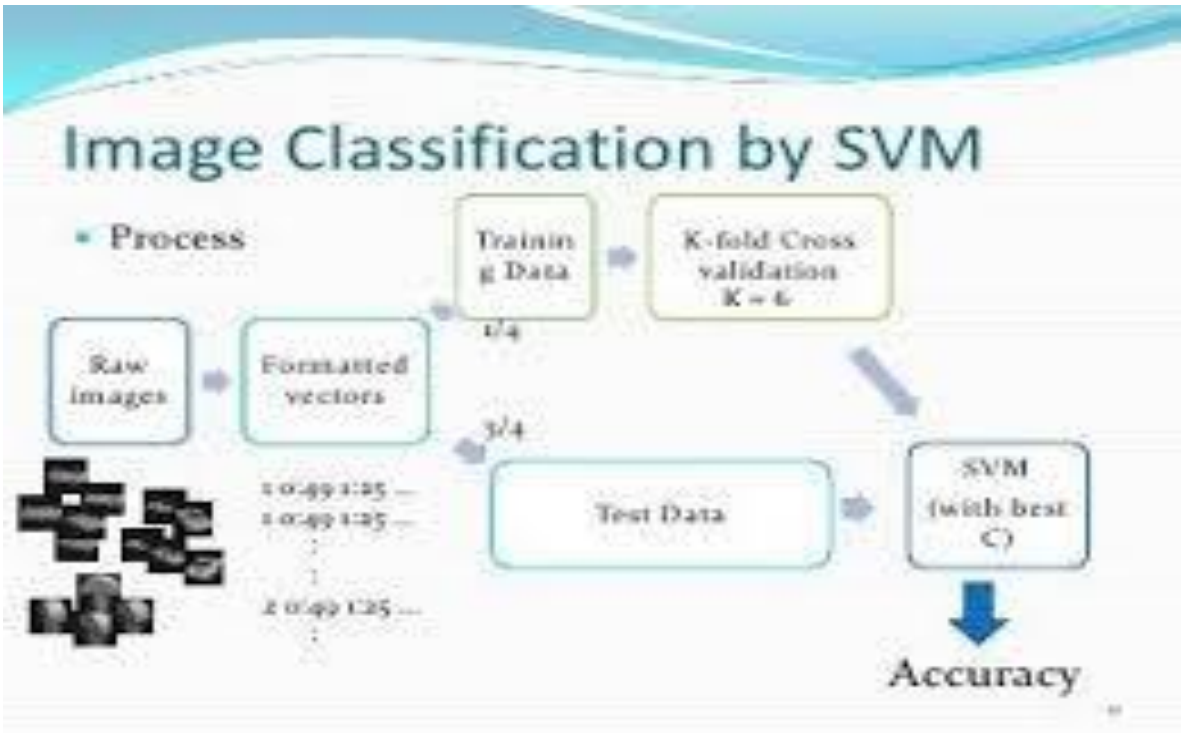


Fig 3.28 : Image classification

3.2.2.4 Bioinformatics:

The discovery of protein distant homologies is a frequent issue in the field of computational biology. SVM is the most efficient approach for resolving this issue. SVM algorithms have been widely used in recent years to detect protein distant homology.

These methods have been frequently employed for biological sequence identification. Examples include classifying genes, treating patients based on their genes, and many other biological issues.

Fig 3.29 : Bio Informatics Application

3.2.2.5 Protein Fold and Remote Homology Detection:

One of the main issues in computational biology is the identification of protein distant homology. The best techniques for distant homology identification use supervised learning algorithms using SVMs. How the protein sequences are modelled affects how well these strategies work. how the kernel function was calculated in between them.

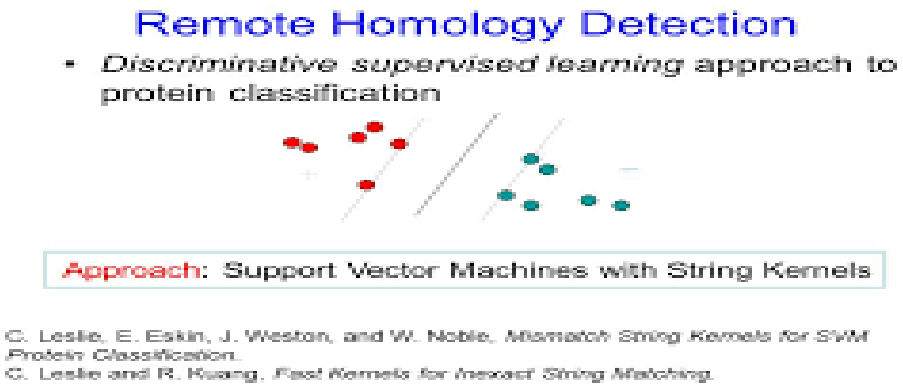


Fig 3.30 : Remote Homology detection

3.2.2.6 Handwriting Recognition:

SVMs can also be used to read handwritten characters for data entry and document signature verification.

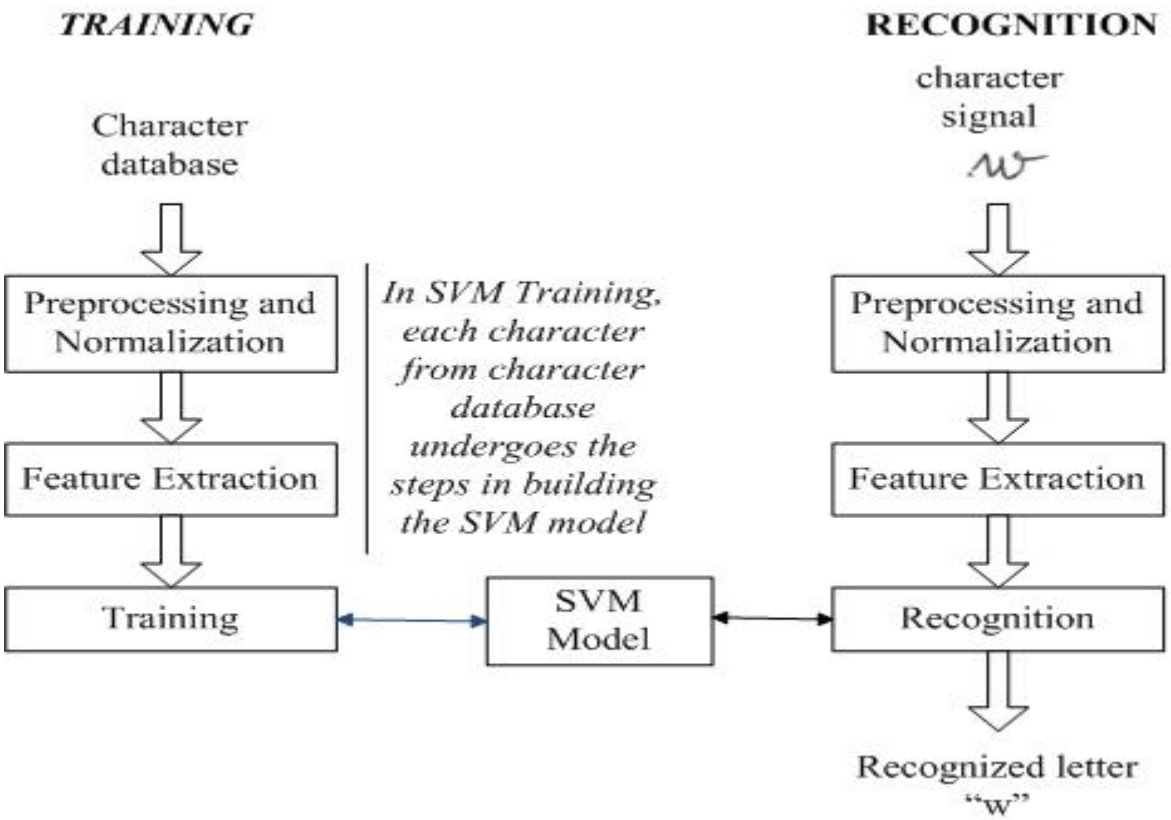


Fig 3.31 : SVM Working Flowchart

3.2.2.7 Geo and Environmental Sciences:

SVMs are employed in the analysis and modelling of geospatial and spatiotemporal environmental data.

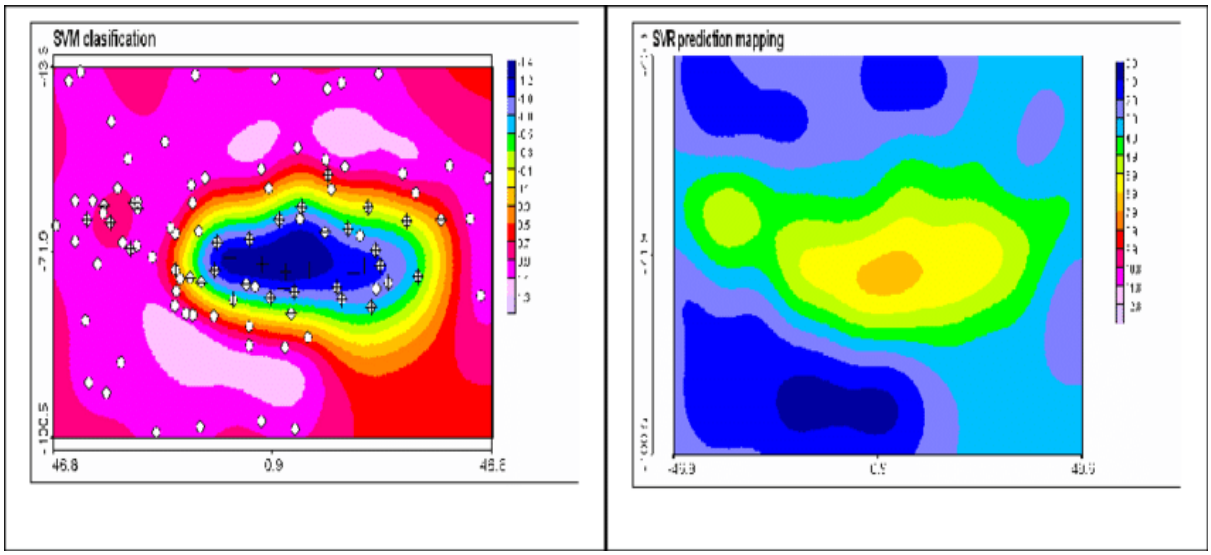


Fig 3.32 : Geospatical Application of SVM

3.2.2.8 Generalized Predictive Control:

To regulate chaotic dynamics using usable parameters, we employ SVM-based GPC. It performs superbly while controlling the systems. When it comes to the local stabilisation of the target, the system complies with chaotic dynamics.

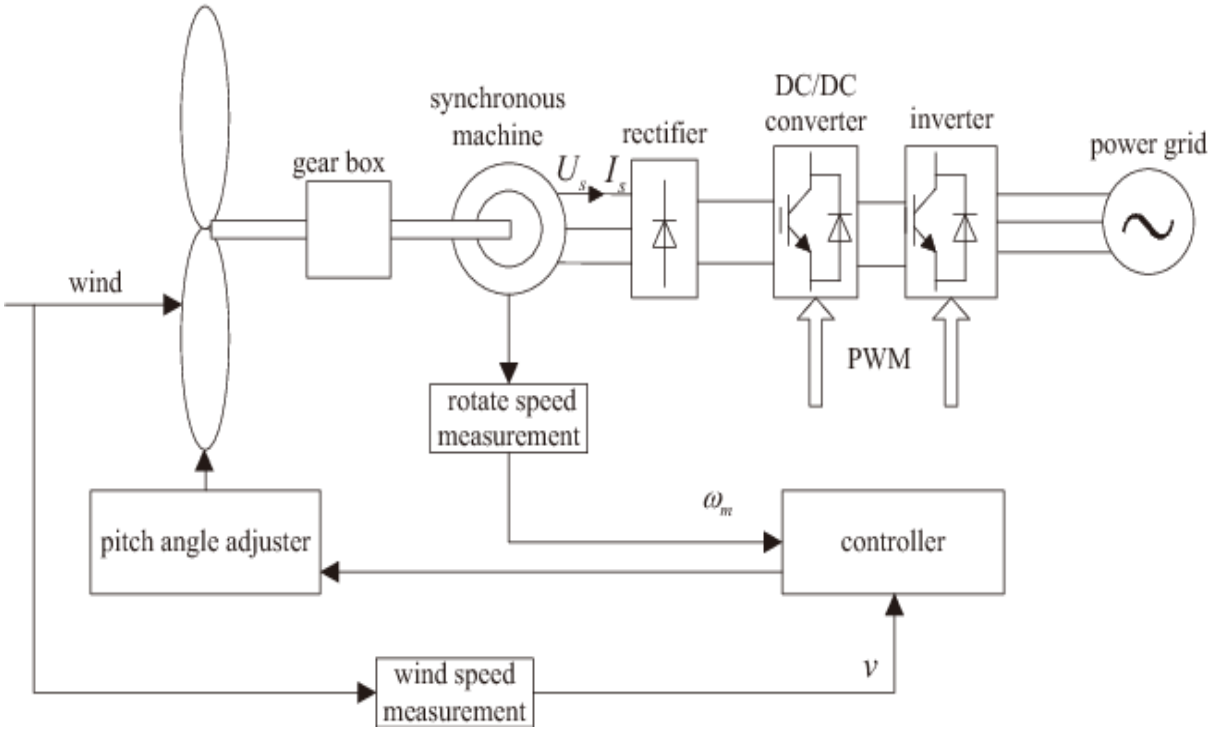


Fig 3.33 : Wind Turbine Application of SVM

3.3 Random Forest:-

A large number of decision trees are built during the training phase of the random forests or random decision forests ensemble learning approach, which is used for classification, regression, and other tasks. The class that the majority of the trees choose is the output of the random forest for classification problems.

Random decision forests adjust for decision trees' propensity to overfit to their training set. Random forests often outperform decision trees, although their accuracy is lower than gradient boosted trees. For regression tasks, the mean or average prediction of the individual trees is returned. However, their effectiveness may be impacted by data peculiarities.

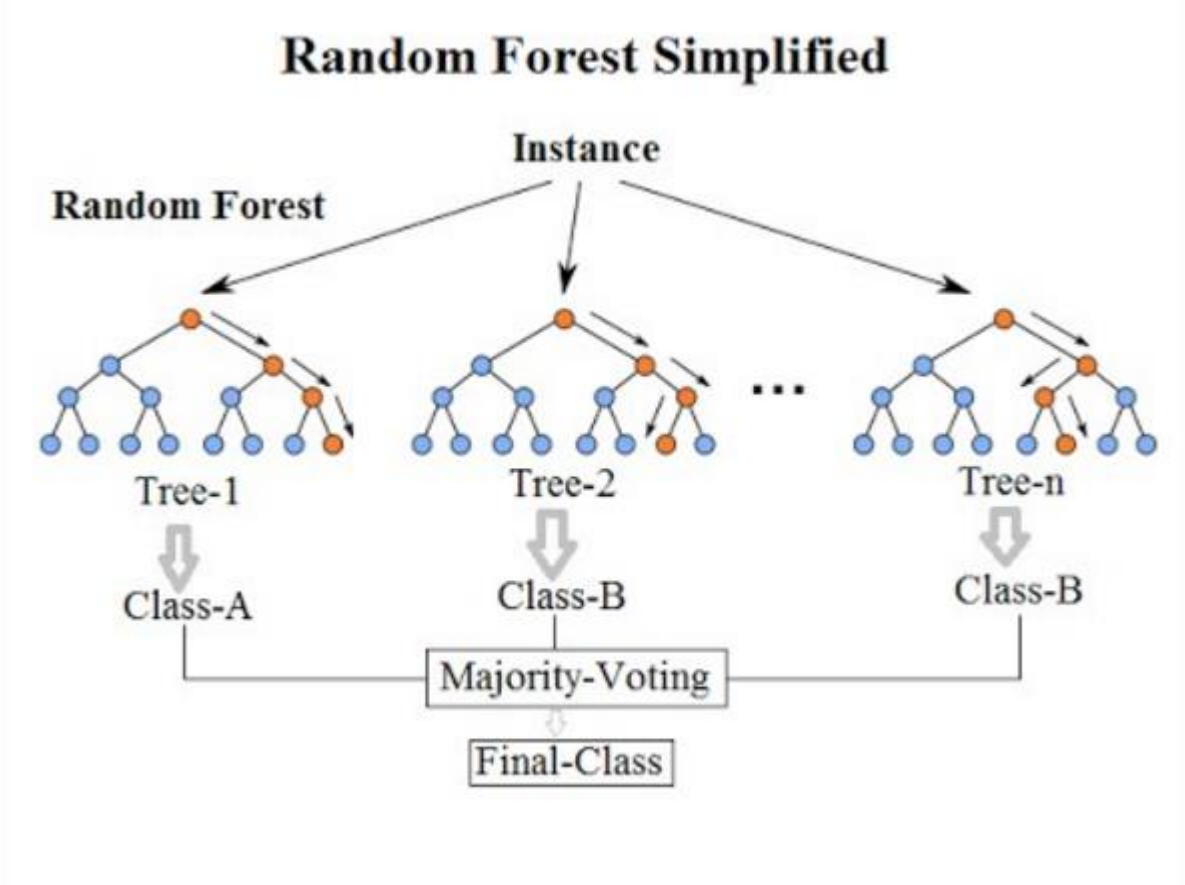


Fig 3.34 : Random Forest Simplified Figure

3.3.1 Working:-

The following steps explain the working Random Forest Algorithm

Choose random samples from a specified data collection or training set in step

The second step of this technique is to create a decision tree for each training set of data.

The decision tree will be averaged to determine the results of the vote.

Finally, choose the prediction result that received the most votes as the final prediction result.

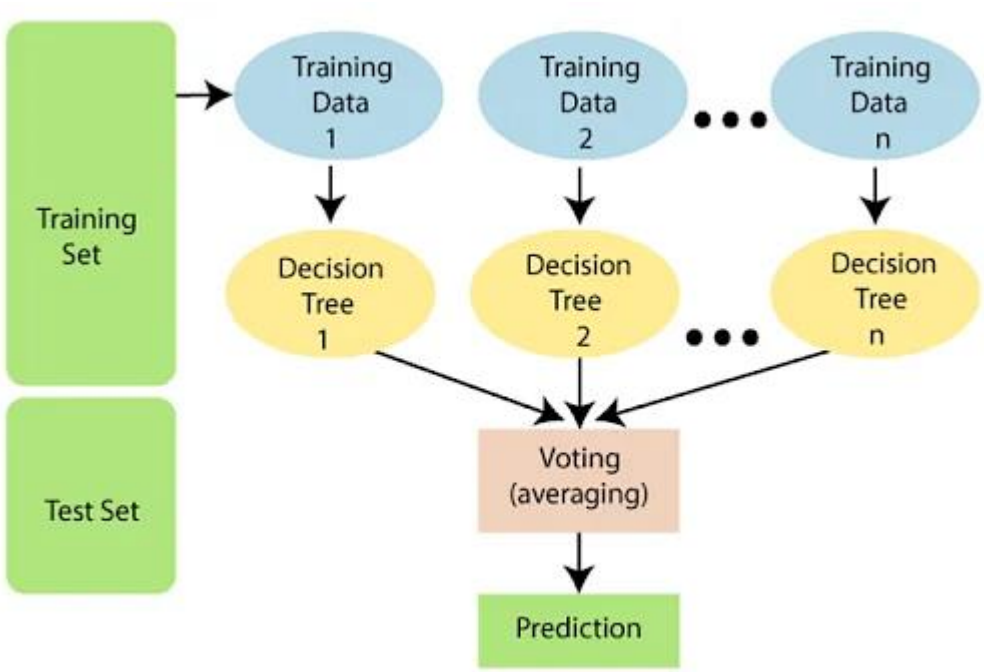


Fig 3.35 : Decision Tree Figure

This combination of multiple models is called Ensemble. Ensemble uses two methods:

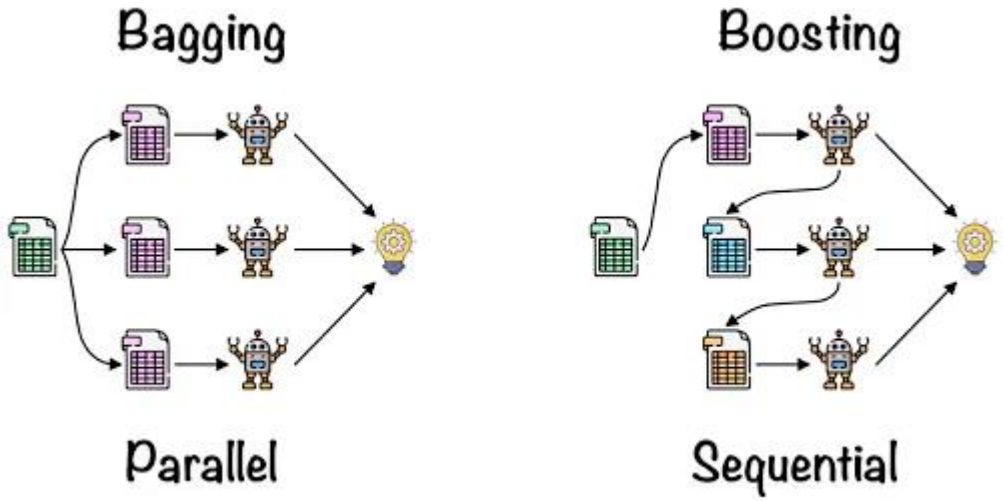


Fig 3.36 : Bagging and Boosting Figure

Bagging: The aforem entioned notion explains how Random forest use the Bagging code. Let's now examine this idea in further depth. In random forest, bagging is sometimes referred to as Bootstrap Aggregation. Starting with any initial random data, the procedure begins. After organising, it is divided into Bootstrap Sample samples. Bootstrapping is the name for this procedure. Additionally, each model is trained separately, producing distinct outcomes known as Aggregation. The last stage combines all the findings, and the output that is produced is based on majority vote. The Bagging phase of the process makes use of an Ensemble Classifier.

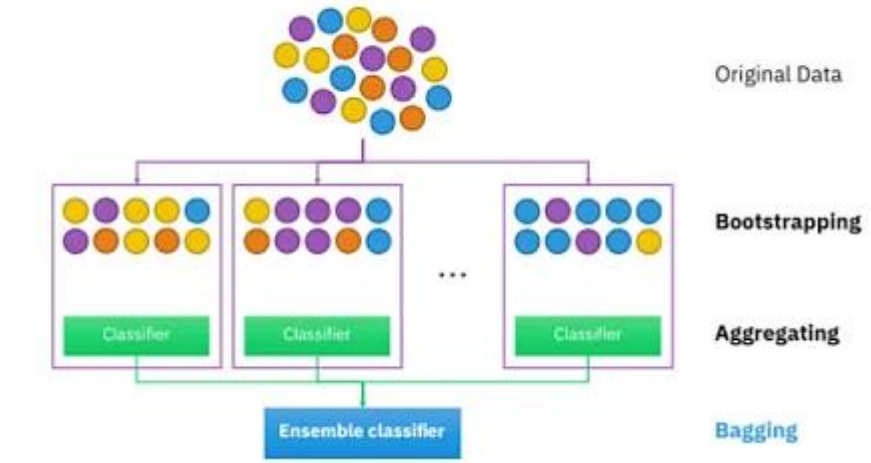


Fig 3.37 : Ensemble Classifier Flowchart

3.3.2 Advantages and Disadvantages:-

3.3.2.1 Advantages:

Popularity is warranted for Random Forest. It has several benefits, ranging from relative simplicity to precision and efficiency. Scikit-learn provides a straightforward and effective random forest classifier framework for data scientists looking to utilise Random Forests in Python.

Fig 3.38 : Random Forest Figure 2

The main practical upper hand of using random forest is that it undoubtedly corrects for decision trees' propensity to overfit their training set. The problem of overfitting is closely completely eliminated when applying the bagging method with random feature selection, which is above par because overfitting results in wrong results. Furthermore, Random Forest often keeps its accuracy even in the leave of certain data.

When analyzing a big database, random forest is far more effective than a single decision tree. However, compared to a neural network, Random Forest is less effective. A neural network, sometimes known as a "neural net," is a collection of algorithms that, by modelling how the human brain processes information, expose the underlying relationships within a dataset.

3.3.2.2 Disadvantages:

Although Random Forest has few drawbacks, every tool has some drawbacks. On bigger projects, random forest may aquire a high memory because it develops use of several decision trees. It could be slimmer as a result than some other effective algorithms.

Fig 3.39 : Random Forest Figure 3

Because this strategy is based on decision trees, which frequently experience overfitting, this issue can occasionally have an impact on the entire forest. Because Random Forest generates smaller trees using random selections of the characteristics, this issue is typically avoided by default. This may result in a slower processing rate but more accuracy.

3.4 Decision Tree:-

A decision tree is a graph that use the branching approach to show each potential result for a certain input. Drawing decision trees by hand, using a graphics tool, or using specialist software are all options. When a group has to make a decision, decision trees can help concentrate the conversation.

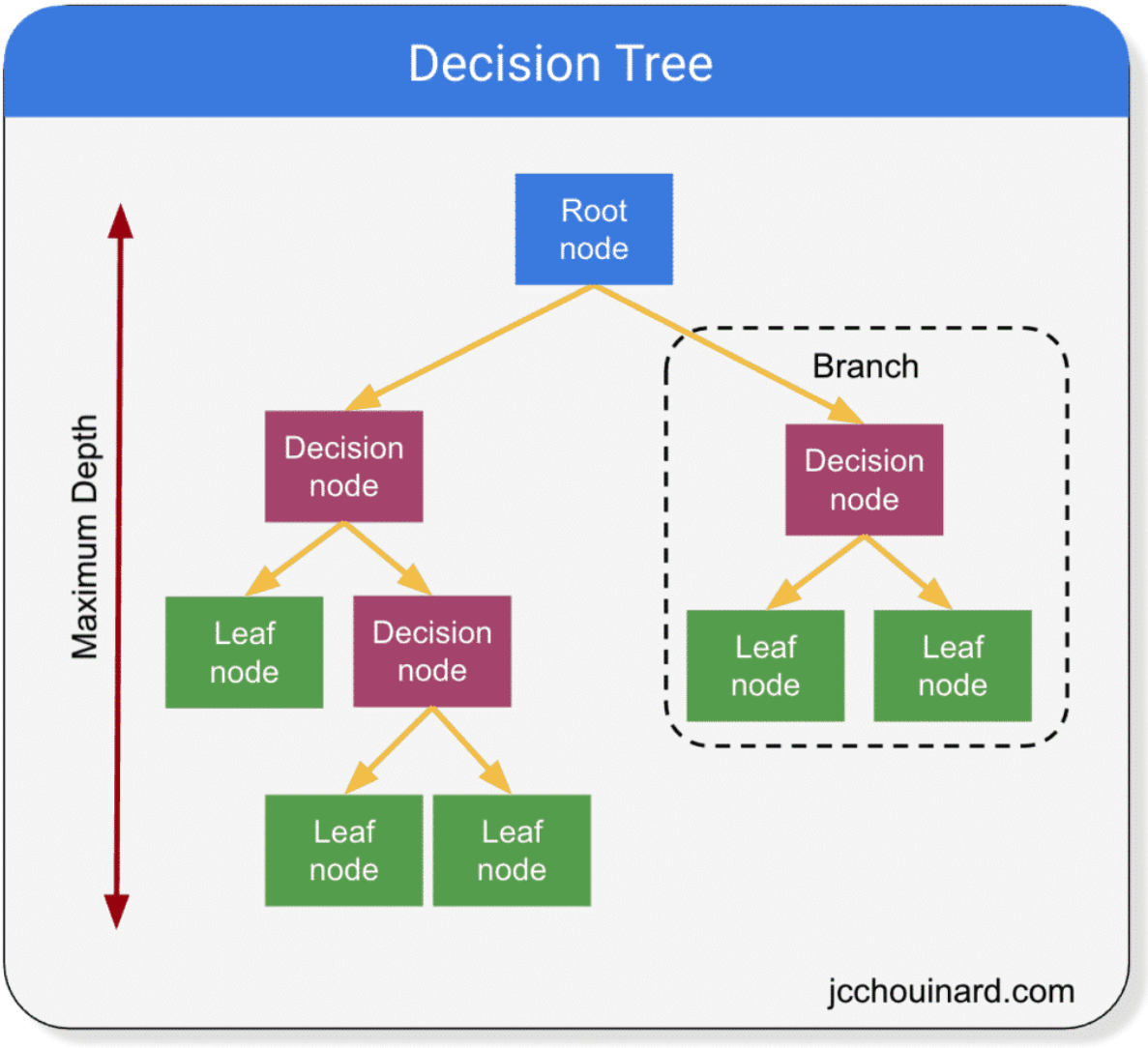


Fig 3.40 : Decision Tree Figure

3.4.1 Working:-

A decision tree is a graphical depiction of every option for making a choice depending on certain circumstances. We attempt to create a condition on the features for each step or node of a classification decision tree in order to fully separate all of the labels or classes present in the dataset.

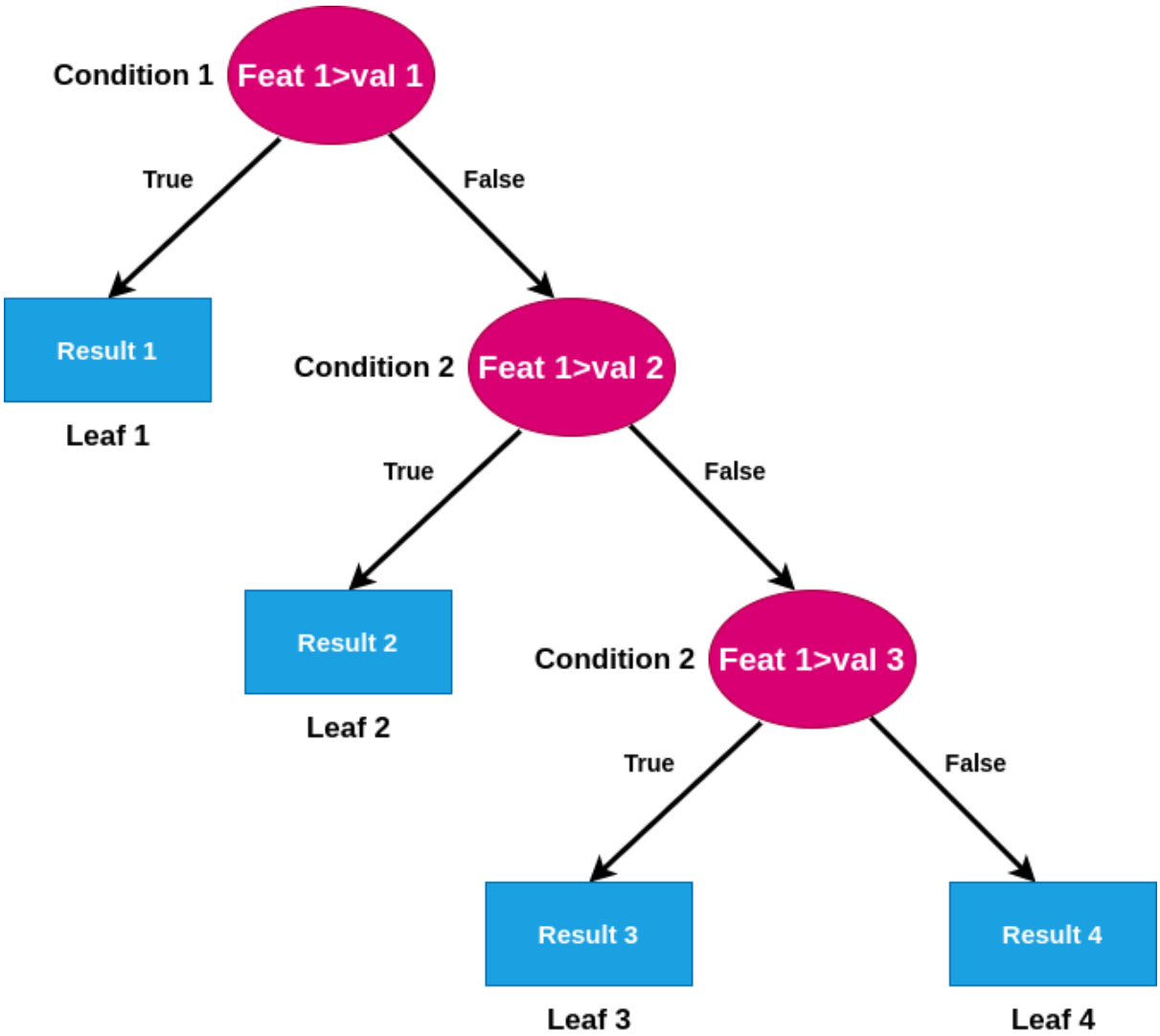


Fig 3.41 : Decision Tree Figure 2

3.4.2 Advantages and Disadvantages:-

Given below are the advantages and disadvantages mentioned:

3.4.2.1 Advantages:

Both classification and regression issues may be solved with it: Decision trees are effective in both classification and regression applications because they may be used to predict both continuous and discrete values.

Decision trees need minimal work to comprehend an algorithm since they're basic.

It may be avail to categorise data that is not easily separable along linear lines.

Because decision trees does not concurrently apply into consideration numerous heavy calculations, they have the privilage of not needing any specific modification when working with non-linear data.

When analyzed to KNN and different classification algorithms, they are amazingly fast and effective.

One of the fastest methods to conclude the most salient factors and bonds between two or more differences is to apply a decision tree. We may add on additional variables or characteristics to the final variable more effeciently using decision trees.

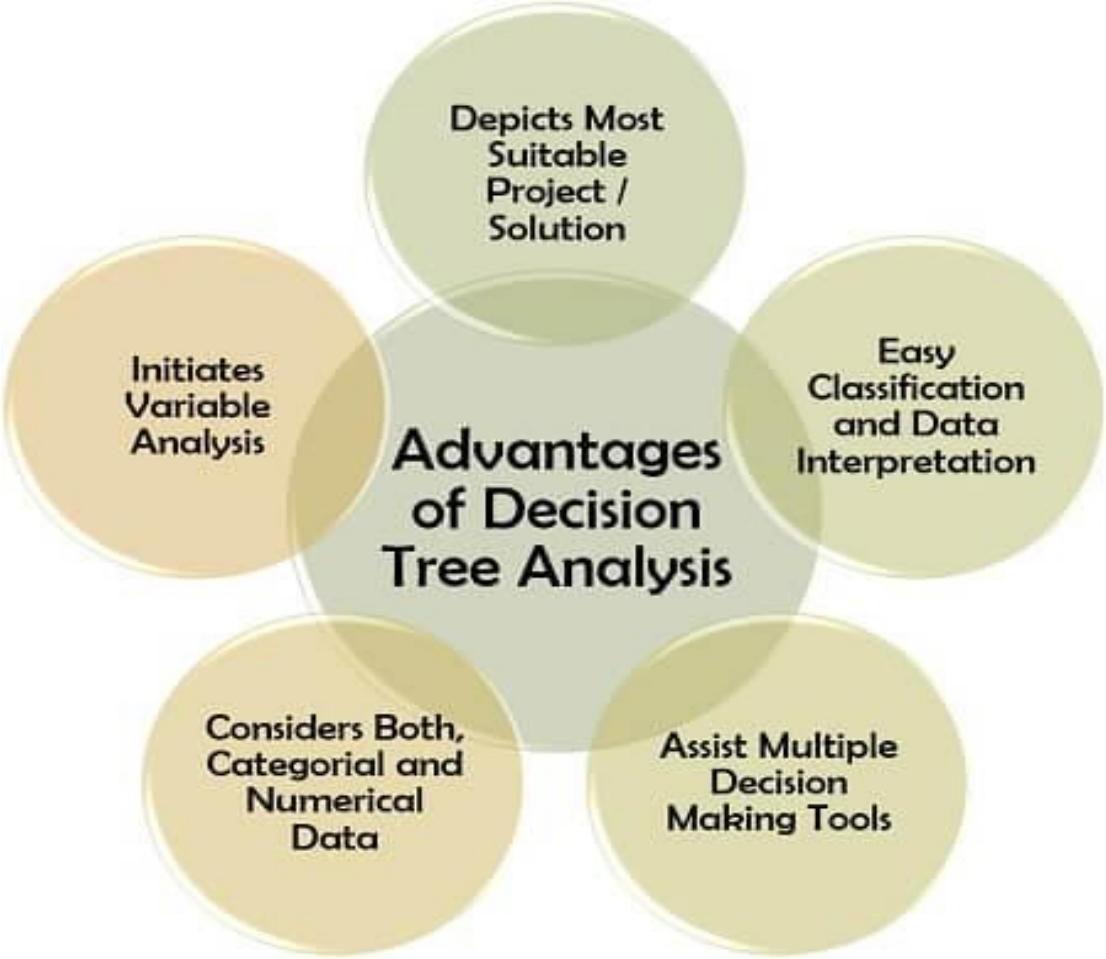


Fig 3.42 : Decision Tree Advantages

3.4.2.2 Disadvantages:

The time complexity of conducting this operation is quite high and keeps increasing as the number of records increases for the decision tree split for millions of records with numerical variables. It could take a while to train a decision tree using numerical variables.

Decision-making tree for many features: Give the training-time complexity more time to rise as the input does.

If we talk about overfitting at all, it's one of the toughest methods for decision tree models. The overfitting problem may be remedied by employing the pruning technique and putting constraints on the parameters model.

Reusability of decision trees: A decision tree may produce a complicated new tree as a result of tiny alterations in the input. In the decision tree, this is referred to as variance, and it may be reduced using techniques like bagging and boosting.

Fig 3.43 : Decision Tree Disadvantages

Chapter-4 PERFORMANCE ANALYSIS

4.1 Performance table of the ML Algorithms:-

Name of the Algorithm	**Accuracy(%) of the Algorithm**
Logistic Regression	79.6%
Support Vector Machine	80.9%
Random Forest	75.3%
Decision Tree	70.5%

Table 4.1 : Accuracy Table

From the above table we can see that Support Vector Machine(SVM) has the highest accuracy(we can see from the above) that means SVM applied over Diabetes Prediction Model has 80.9% more chances to detect wether a person is Diabetic or not.

So the Model has 20% less chances of a person getting falsely Diagnosed with Diabetes, Through this Peoples can go to Medical Centre and start getting proper treatment to minimize the risks that may occur from Diabetes as a Drawback.

A support vector machine is a strong and adaptive machine learning algorithm which can support linear and nonlinear classification, regression, and outlier identification. It is favoured over other classification methods since it requires less processing while provides greater accuracy. It is beneficial since it produces credible findings even with limited data.

Let us examine the operation of SVM using an example of two classes, class A: Circle and class B: Triangle. Now we'll use the SVM technique to discover the optimal hyperplane that separates the two classes.

SVM considers all of the data points and produces a line called a 'Hyperplane,' which separates both groups. This line is known as the 'decision border.' Anything in the circular class belongs to class A, and vice versa. SVM takes into account all of the data points and generates a line known as a 'Hyperplane,' which divides both groups. This is referred to as the 'decision boundary.' Everything in the circular class is a member of class A, and vice versa. Different dimensions are possible depending on the features we have. When there are more than three aspects, it is difficult to visualize, Consider two cases, A and B, which are red and yellow, respectively. We must choose the optimum hyperplane between them that differentiates the two cases. A number of the aforementioned data points may be miscategorized due to the flexible margin. It may seeks to meet an agreement between locating an hyperplane which mis-categorizations and maximizing classification accuracy.

Chapter-5 CONCLUSIONS

5.1 Conclusions

Diabetes diagnosis and prediction are two of the most common medical issues in the real world. Its long-term presence in the human body causes microvascular problems of diabetes.

Doctors have traditionally utilized diagnostic tests to determine whether a person is diabetic. They started with monitoring the serum and plasma glucose levels each hour. Diabetes was typically been diagnosed by fasting blood glucose levels that are greater than the authorized rate. Another key component in the diagnosing of a diabetic pregnant woman is body mass index. Compared to women with a pre pregnancy BMI of 29 kg/m2, women with a BMI > 29 kg/m2 have a higher chance of developing type 2 diabetes. As a consiquence, both of these parameters, such as BMI and Plasma glucose, were shown to be strongly co-related throughout our research.

The primary goal of this project was to design and implement Diabetes Prediction Model Using Machine Learning Techniques and to evaluate the results of such methods, which was been conducted efficiently. In this Project we have used eight variables, like Age, Diabetes Pedigree function, BMI, insulin, skin thickness etc. This Project employs a number of classification algorithms, such as SVM, KNN, Random Forest, Decision Tree, and Logistic Regression classifiers. The strategies may also assist researchers in developing an accurate and useful tool that will reach physicians' tables to assist them in making better decisions about illness state. The combined approach makes use of the total of two or more methods' values. We thought that our approach would may offer us with greater than 98% accuracy. In this data pre-processing is a vital stage in any study in order to construct a better and more reliable model for the prediction process. furthermore, a good classification accuracy was obtained. The experimental results will help health care professionals make

early predictions and actions to treat diabetes and save people's lives.

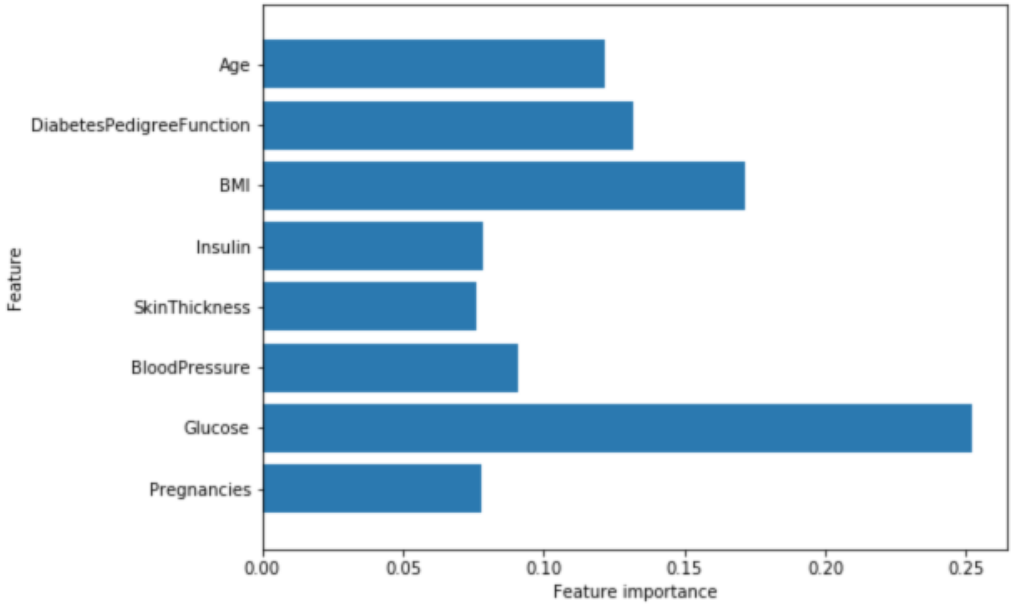


Fig 5.1 : Diabetes Features Comparison

In this Project we use four Machine learning algorithms:

5.1.1 Logistic regression:

Logistic Regression has shown to be one of the most effective algorithms for developing prediction models. This study also reveals that the accuracy of the model is affected by aspects such as data preparation, elimination of redundant and null values, normalisation, cross-validation, feature selection, and ensemble approaches. When the features differ on a broad scale, feature selection plays a vital role in improving the accuracy and decreasing the runtime. In contrast to other methods, the study showed that the Logistic Regression classifier performed best, with the maximum accuracy of 79.6% and the lowest misclassification rate of 23.8. Using ensemble machine learning techniques, this study may be extended to enhance prediction accuracy.

Combining numerous algorithms, as shown in ensemble approaches, helps to improve the model's performance. Cross-validation is also important for increasing accuracy.

Logistic Regression gives an accuracy of 79.6%

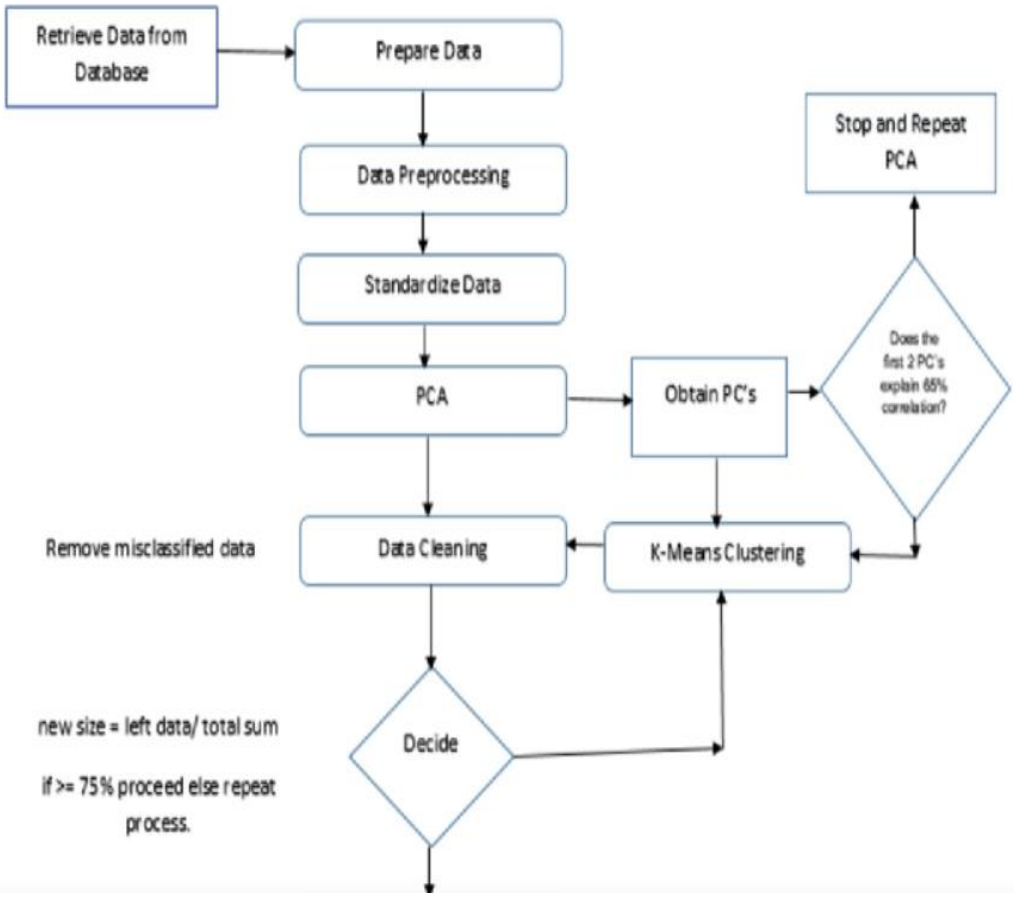


Fig 5.2 : Proposed Algorithm Flowchart

5.1.2 Support Vector Machine(SVM):

SVM is a supervised machine learning model that is suitable for classification and regression and it's a model-free approach for solving classification problems that makes no assumptions about the distribution or interdependence of the data. In epidemiologic research and population health surveys, the SVM methodology has the potential to outperform classic statistical methods like as logistic regression, particularly when multivariate risk variables are present, as in Diabetes.

Support vector machine modelling is an effective classification method for diagnosing a complicated disease such as diabetes utilising common, basic data. Validation shows that SVM models had discriminative qualities similar

to frequently used logistic regression techniques. Our Diabetes Classifier, a web-based tool designed only for demonstration reasons, exemplifies one possible use of the SVM technique: the recognition of persons with undiagnosed common illnesses like diabetes and pre-diabetes.

The SVM classifier gives an accuracy of 80.9%

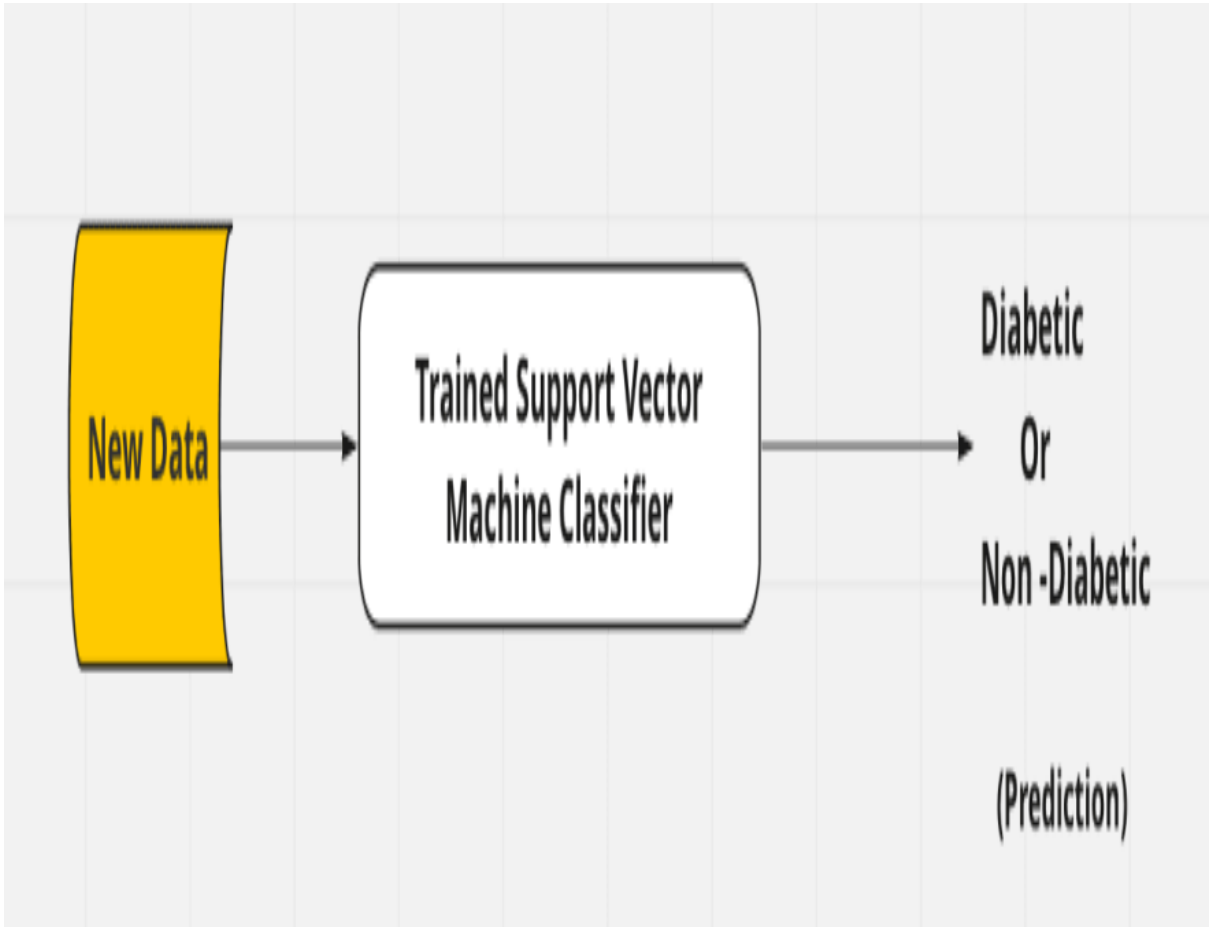


Fig 5.3 : SVM Conclusion Figure

5.1.3 Random Forest:

Diabetes is a condition that can lead to a variety of consequences. How can machine learning be used to precisely predict and diagnose this disease? Based on the results of the earlier trials, we should use this approach this algorithm to improve accuracy.

Furthermore, when we examine the outcomes of three classifications, we can see that there is difference between random forest, SVM, and Logistic Regression, but random forests are just as reliable as logistic regression but

lower than SVM and Logistic Regression.

The Random Forest approach is a machine learning ensemble model that is used for regression and classification tasks. It is also used for tasks that involve the generation of a large number of decision trees while training and the output of the class that is the mode of the classes or the mean estimate of the individual trees. It boosts the overall result by applying variability to the model when developing the trees.

The Random Forest Classifier along with SVM and Logistic Regression classifiers approach outperforms individuals in detecting high-risk diabetics. Furthermore, the combination strategy presented might be a useful tool for early detection of Diabetes.

Random Forest gives an accuracy of 75.3% .

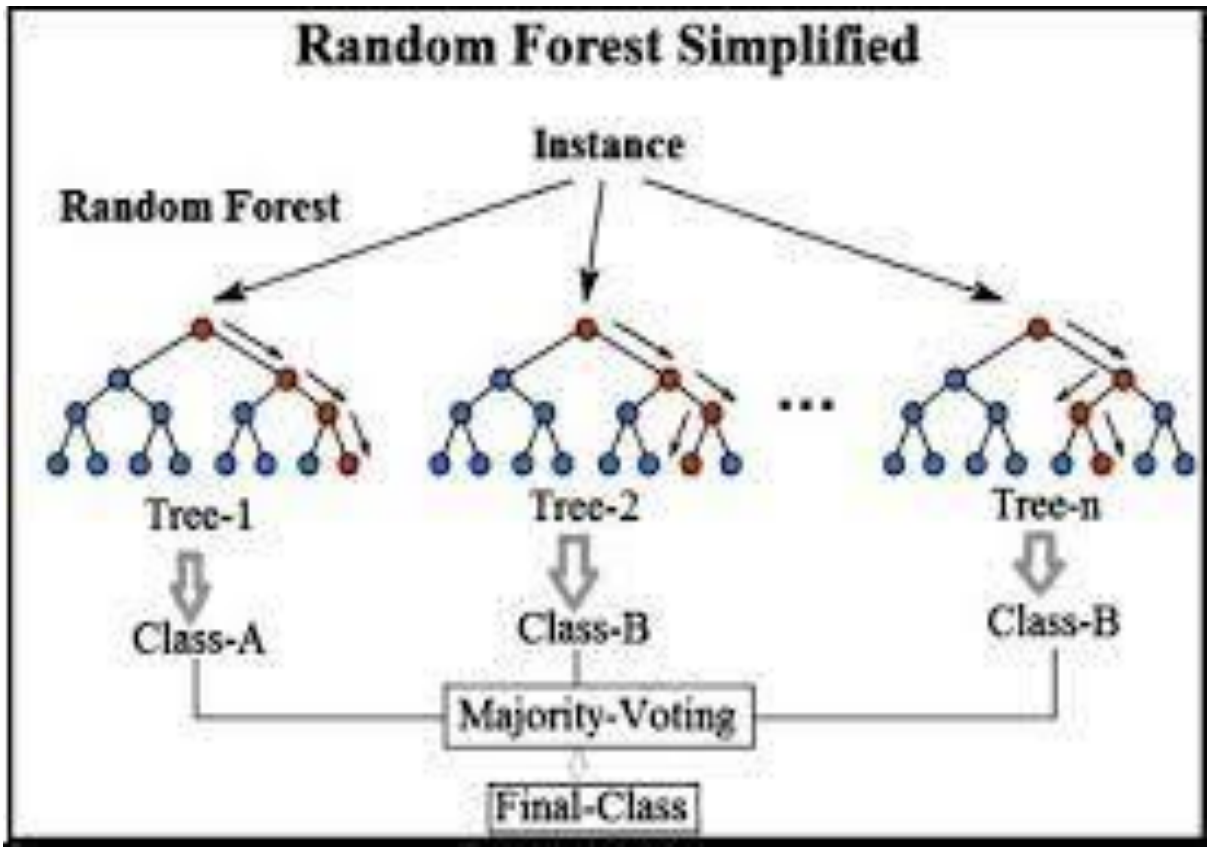


Fig 5.4 : Random Forest sample flowchart

5.1.4 Decision Tree:

Overall, decision tree classifier is a predictive modelling approach that may be used in a variety of situations. An algorithmic technique that could be used to create decision trees that divide the dataset in multiple ways based on the circumstances. We may deduce from our study that the more data we dedicate for training the model, the greater the accuracy estimate we acquire. In the instance, the ideal strategy is to split the data in half for training and half for testing.

The following benefits can be observed after studying the constructed diagnostic model: Quick learning process; creation of rules in domains where an expert's knowledge is difficult to formalize.
However, the Decision Tree is vulnerable to overfitting, and its prediction accuracy is worse when compared to other classifiers/algorithms.

Decision Tree offers a accuracy of 70.5% .

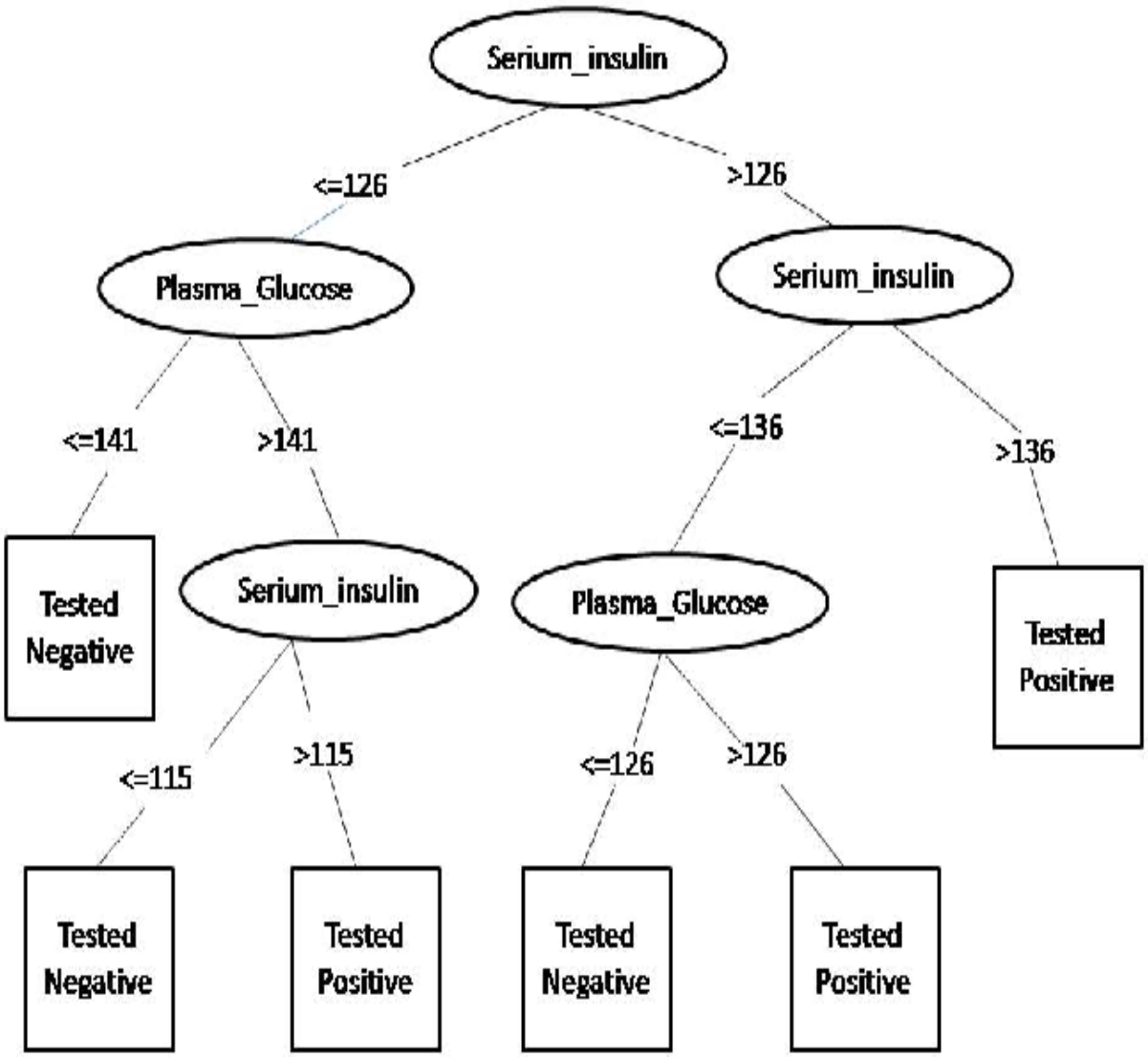


Fig 5.5 : Diabetes Conclusion Flowchart

This Model can assist doctors in identifying and treating diabetes patients. We can propose that boosting classification accuracy will help the Diabetes Prediction Models perform better.

We also notice that the given model's accuracy ranges between 70 and 80%, indicating that we can adopt a mix of classifiers. The combined approach makes use of the total of two or more methods. Thus Increasing it's accuracy up to 98%

5.2 Future Scope:-

In the future development of this Model we shall use XGBoost Algorithm to Obtain more higher accuracy than the previously used Algorithms.

XgBoost is the short form of Extreme Gradient Boosting,

XGBoost is a gradient boosting ensemble Machine Learning technique based on decision trees. Artificial neural networks excel all other algorithms or frameworks in prediction issues involving unstructured data (pictures, text, etc.) However, decision tree-based algorithms are now regarded best-in-class for small-to-medium structured/tabular data.

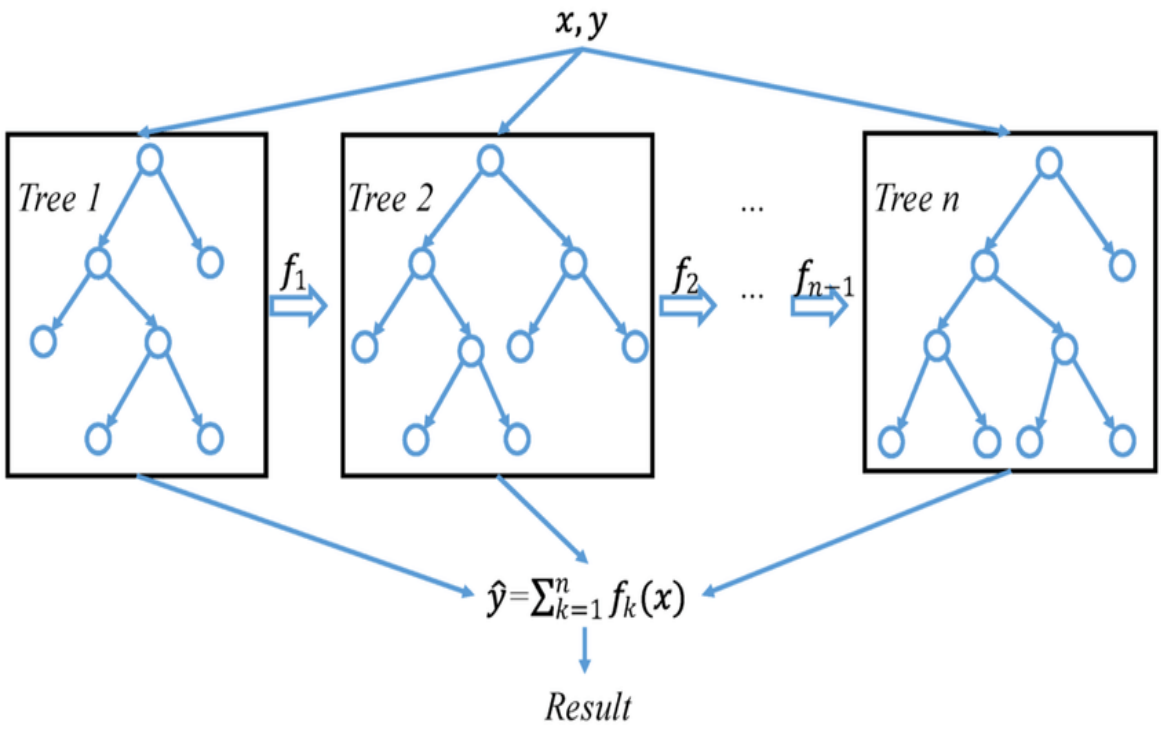


Fig 5.6 : General Structure of XGBoost

The XGBoost algorithm was created as part of a university research effort. Tianqi Chen and Carlos Guestrin created this algorithm in 2014. Since its creation, this algorithm has been recognized as the driving force behind various cutting-edge industry applications. As a response, the XGBoost Algorithm is highly recommended by a large community of data scientists.

5.2.1 Advantages and Disadvantages:

- A diverse set of applications: may be applied to solve confusions combining of regression, classification, ranking, and user-defined prediction.
- Portability: Works well on Windows, Linux, and Mac OS X.
- Languages: C++, Python, R, Java, Scala, and Julia are some of the major programming languages supported by XGBoost.

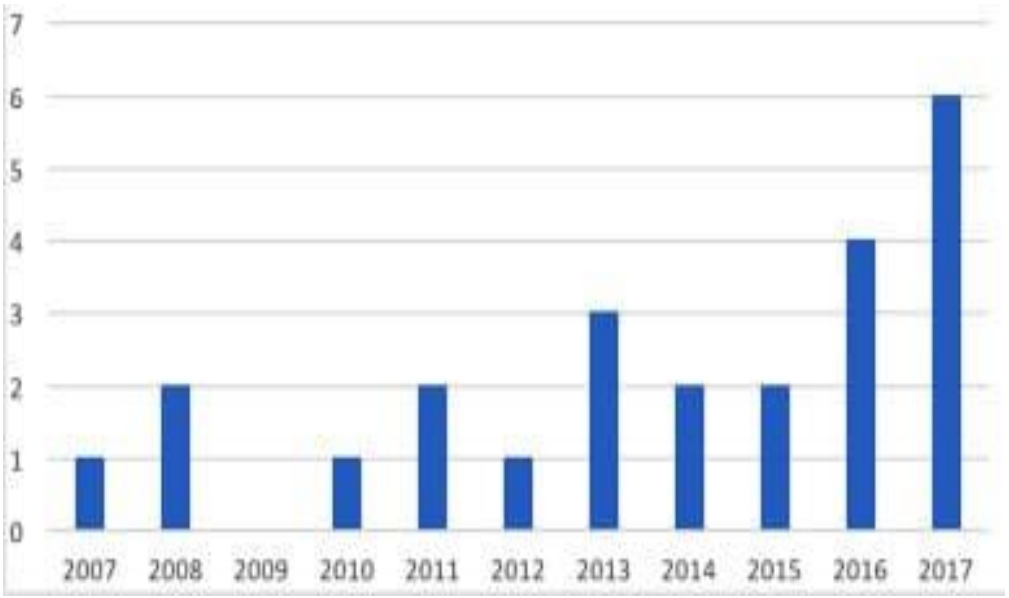


Fig 5.7 : Ml Algorithms Comparison Graph

5.2.2 XGBoost is made up of five algorithms that are:

Decision Tree: A Decision tree is a tree structure that looks like a flowchart, with each internal node representing a test on an attribute, each branch representing a test outcome, and each leaf node holding a class label.

Bagging: A Bagging classifier is an aggregate meta-estimator that fits base classifiers on random sections of the entire dataset, then combines their individual predictions to produce a final prediction.

Random Forest: Every decision tree has a significant variance, but when we aggregate all of them in parallel, the resultant variance is minimal since each decision tree is completely trained on that specific sample data, and therefore the 2outcome does not depend on one decision tree but on several decision trees.

Boosting: Boosting is a strategy in ensemble modelling that seeks to construct a strong classifier from a large number of weak classifiers. The next classifier corrects the mistakes of the previous one. This approach is repeated, and models are added until either the whole training data set is properly predicted.

Gradient Boosting: In gradient boosting, each prediction corrects the inaccuracy of its previous. Unlike Adaboost, the values of the training instances are not changed; instead, each predictor is trained using the predecessor's residual mistakes.

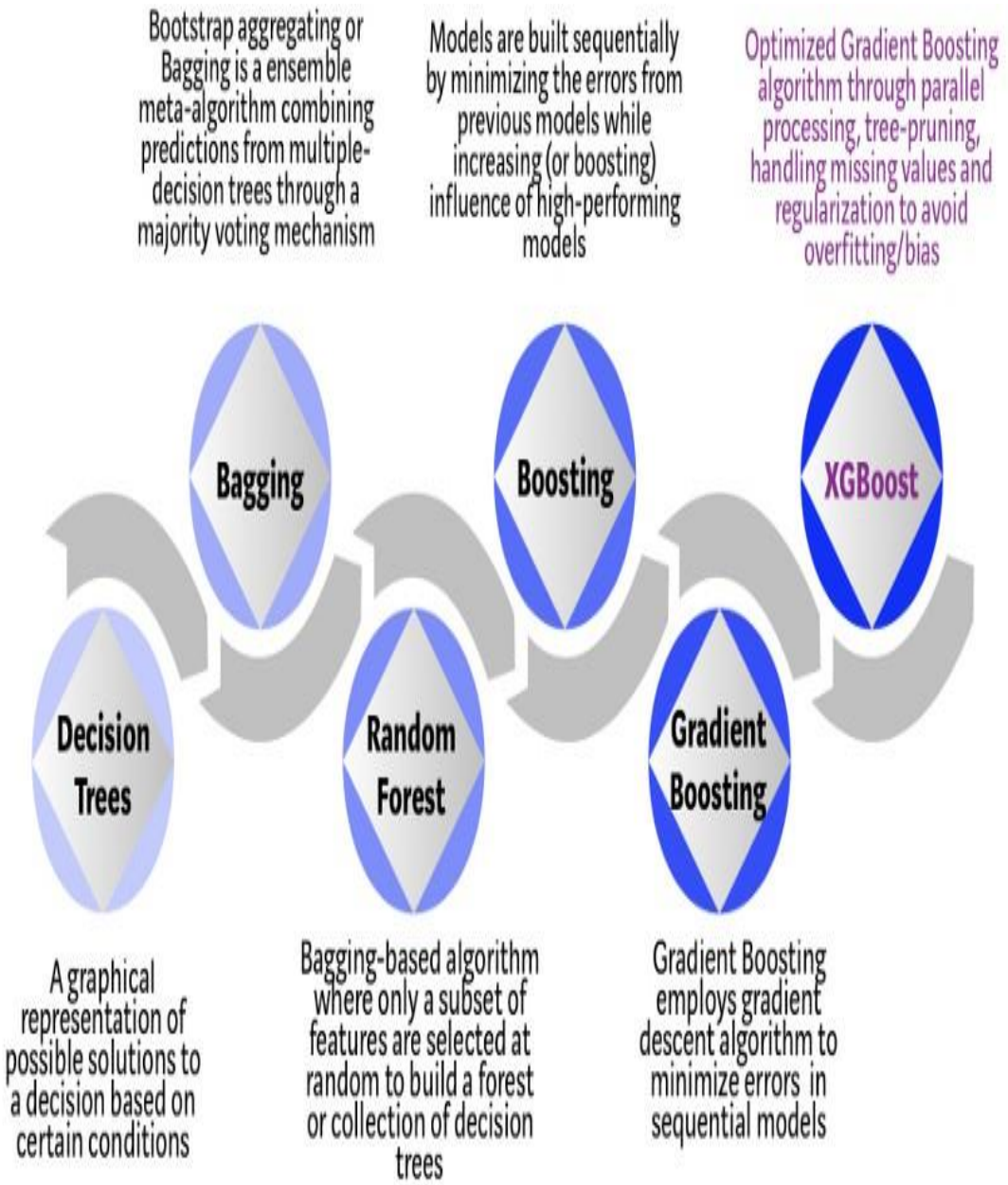


Fig 5.8 : XGBoost Structure Figure

REFERENCES

[1] Hwapyeong Song;Sanghoon Lee 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) Year: 2021 | Conference Paper | Publisher: IEEE

[2] Nachiket Dunbray;Rashmi Rane;Sparsh Nimje;Jayesh Katade;Shreyas Mavale 2021 2nd Global Conference for Advancement in Technology (GCAT) Year: 2021 | Conference Paper | Publisher: IEEE

[3] Jianchao Han;Juan C. Rodriguez;Mohsen Beheshti 2008 Second International Conference on Future Generation Communication and Networking Year: 2008 | Volume: 3 | Conference Paper | Publisher: IEEE

[4] Adiwinata Gani;Andrei V. Gribok;Yinghui Lu;W. Kenneth Ward;Robert A. Vigersky;Jaques Reifman IEEE Transactions on Information Technology in Biomedicine Year: 2010 | Volume: 14, Issue: 1 | Journal Article | Publisher: IEEE

[5] Krittika Kantawong;Supan Tongphet; Panu Bhrommalee;Napa Rachata;Sakkayaphop Pravesjit 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON) Year: 2020 | Conference Paper | Publisher: IEEE

[6] F. Ståhl;R. Johansson;Eric Renard 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Year: 2010 | Conference Paper | Publisher: IEEE

[7] Eslam Montaser;José-Luis Díez;Paolo Rossetti;Mudassir Rashid;Ali Cinar;Jorge Bondia IEEE Journal of Biomedical and Health Informatics Year: 2020 | Volume: 24, Issue: 7 | Journal Article | Publisher: IEEE

[8] Liu Lei 2020 International Conference on Virtual Reality and Intelligent Systems (ICVRIS) Year: 2020 | Conference Paper | Publisher: IEEE

[9] Ramya Akula;Ni Nguyen;Ivan Garibay 2019 SoutheastConYear: 2019 | Conference Paper | Publisher: IEEE

[10] Pınar Cıhan;Hakan Coşkun 2021 29th Signal Processing and Communications Applications Conference (SIU) Year: 2021 | Conference Paper | Publisher: IEEE

[11] Ayush Anand;Divya Shakti 2015 1st International Conference on Next Generation Computing Technologies (NGCT) Year: 2015 | Conference Paper | Publisher: IEEE

[12] Xiaoyu Sun;Xia Yu;Jianchang Liu;Honghai Wang 2017 36th Chinese Control Conference (CCC) Year: 2017 | Conference Paper | Publisher: IEEE

[13] Prakhar Saxena;Subhadeep Saha;S. Kiruthika Devi 2022 International Mobile and Embedded Technology Conference (MECON) Year: 2022 | Conference Paper | Publisher: IEEE

APPENDICES

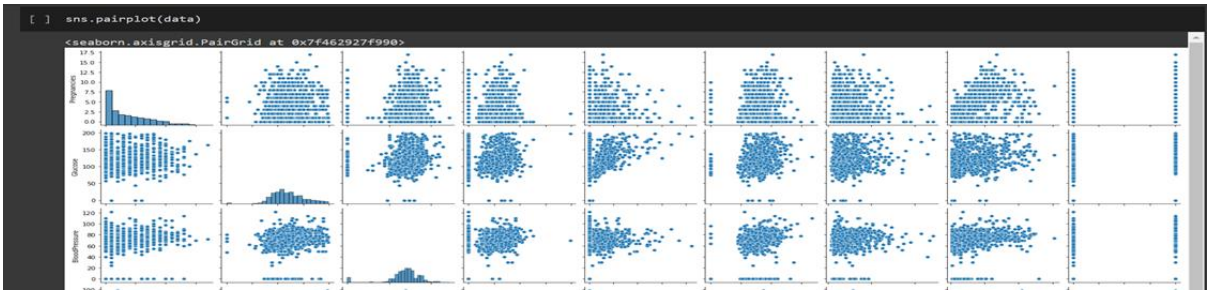
Here are few screenshots of the code:-

```
[1] import numpy as np
    import pandas as pd
    import tensorflow as tf
    import matplotlib.pyplot as plt
    import seaborn as sns
    import xgboost as xgb

[2] data = pd.read_csv('diabetes.csv')
```

The above libraries have been used to run the dataset

```
[ ] sns.pairplot(data)
```



Here we use the seaborn library to plot the dataset given.

```
[4] from sklearn.linear_model import LogisticRegression

[5] logreg = LogisticRegression()

[13] logreg.fit(X_train, Y_train)

[14] X_train.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
429	1	95	82	25	180	35.0	0.233	43
524	3	125	58	0	0	31.6	0.151	24
212	7	179	95	31	0	34.2	0.164	60
558	11	103	68	40	0	46.2	0.126	42
66	0	109	88	30	0	32.5	0.855	38

```
[15] Y_test.head()
```

implementation(1)

```
[ ] data.corr()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

implementation(2)