Diabetes Prediction Using Various Machine Learning Algorithms

M. Tech Dissertation

by

MOHAMMAD OBAIDA

Department of Applied Sciences & Humanities

Faculty of Engineering & Technology

Jamia Millia Islamia

New Delhi-110025

Diabetes Prediction Using Various Machine Learning Algorithms

M. Tech Dissertation

By

MOHAMMAD OBAIDA

Department of Applied Sciences & Humanities

Faculty of Engineering & Technology

Jamia Millia Islamia

New Delhi-110025

Diabetes Prediction Using Various Machine Learning Algorithms

M. Tech Dissertation

Submitted to

Jamia Millia Islamia

New Delhi-110025

In the partial fulfilment of the requirements of the award of the degree of

Masters of Technology (M. Tech.) in

Computational Mathematics

by

MOHAMMAD OBAIDA

Under the supervision of

Prof. Musheer Ahmad

CERTIFICATE

On the basis of declaration submitted by **Mr. Mohammad Obaida, student of M.Tech. Computational Mathematics,** I hereby certify that the M.Tech dissertation entitled **"Diabetes Prediction Using Various Machine Learning Algorithms"** being submitted to the Department of Applied Sciences & Humanities, Faculty of Engineering & Technology, Jamia Millia Islamia, New Delhi in the partial fulfillment of the requirement for the award of the degree of Masters of Technology "M. Tech". The work carried out by **Mr. Mohammad Obaida** in this dissertation includes the experimental/ theoretical/ data analysis/computational work and comprehensive review of literature under my supervision. The work presented in this dissertation is good enough for the successful completion of major project work, which is an integral part of M. Tech. program.

Supervisor: **Prof. Musheer Ahmed**

Prof. Musheer Ahmad
Co-ordinator,
M. Tech. (Computational Mathematics),
D/o Applied Sciences & Humanities,
F/o Engineering and Technology,
Jamia Millia Islamia, New Delhi-110025

Prof. Zishan Hussain Khan
Head,
D/o Applied Sciences & Humanities
F/o Engineering and Technology
Jamia Millia Islamia, New Delhi-110025

Curiginal

Document Information

Analyzed document	Obaida-Mtech 20MCT007.pdf (D1413017854)
Submitted	6/27/2022 10:38:00 AM
Submitted by	Musheer Ahmad
Submitter email	mahmad@jmi.ac.in
Similarity	8%
Analysis address	mahmad.jmi@analysis.urkund.com

Declaration

I, Mr. Mohammad Obaida, student of M. Tech. (Computational Mathematics) Sem- IV hereby declare that the M. Tech. Dissertation entitled **" Diabetes Prediction Using Various Machine Learning Algorithms**", being submitted to the Department of Applied Sciences & Humanities, Faculty of Engineering & Technology, Jamia Millia Islamia, New Delhi, in the partial fulfillment of the requirement for the award of the degree of Masters of Technology ""M. Tech."" has not previously formed the basis for the award of any degree, diploma or similar title or recognition. This is to declare further that I also fulfilled the requirement.

Date: Signature

Place: Jamia Millia Islamia, New Delhi Mohammad Obaida

ACKNOWLEDGEMENTS

Working in this project has been a highly rewarding and enriching experience for me. I would like to take this opportunity to express my sincere gratitude towards my guide **Dr./ Prof Musheer Ahmed** and for his invaluable guidance, support and encouragement throughout the tenure of this project. I would like to thank my course coordinator **Prof. Musheer Ahmad** and faculty of Jamia Millia Islamia for teaching me and supporting me throughout my masters degree. I would also like to thank all the Ph. D. research Scholar of my department for helping and encouraging me to keep moving forward and being there whenever I needed. Last but not the least, I would like to thank my family members, friends, and all those who have directly or indirectly involved and supported me throughout this project.

Mohammad Obaida
Jamia Millia Islamia
Date

Table of Contents

ABSTRACT

Diabetes is a disease caused by an elevated glucose level in the body. Diabetes should not be overlooked; if untreated, it can lead t0 serious complications such as heart disease, kidney disease, high bl00d pressure, eye damage, and other organ damage. Diabetes is a chronic disease that can cause a global health crisis. Diabetes can be cured if it is detected early. According to the International Diabetes Federation, 382 million people worldwide suffer from diabetes. By 2035, this figure will have more than doubled to 592 million. Over the years, several researchers have attempted to build an accurate diabetes prediction model. Due to a lack of appropriate data sets and prediction approaches, this subject still faces significant open research issues, forcing researchers to use big data analytics and Machine Learning MACHINE LEARNING-based methods. Big Data Analytics is important in the healthcare industry. Databases in the healthcare industry are massive. Through the use of big data analytics, one is able to investigate very large datasets in search of concealed information and patterns in order to derive knowledge from the data and accurately predict outcomes. To achieve this goal, we will employ a variety of Machine Learning techniques to more accurately predict diabetes in a human body or a patient. Machine Learning methods and better prediction results can be obtained by building models from patient datasets. In this paper, we will predict diabetes using Machine Learning Classification and ensemble techniques on a dataset. K-Nearest Neighbor, Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boosting, and Random Forest are the algorithms used. When compared to other models, the accuracy of each model varies. The project work provides an accurate or higher accuracy model, demonstrating that the model is capable of accurately predicting diabetes. Our results show that Random Forest outperformed other Machine Learning techniques in terms of accuracy.

INTRODUCTION

Diabetes is one of the common and rapid growing diseases in the world, even among children. Diabetes is a metabolic disorder also known as diabetes mellitus. It significantly raises blood sugar levels. Before we can get a handle on diabetes and how it develops, we have to get a grasp on what goes on inside the body when there is no diabetes present. Sugar glucose is derived from the foods we consume, specifically carbohydrate foods. Carbohydrate foods are the primary source of energy for our bodies; everyone, including diabetics, requires carbohydrates. Rice, cereal, fruit, dairy products, and vegetables are examples of carbohydrates “especially starchy vegetables). When we consume these foods, our bodies convert them into glucose. In the bloodstream, glucose circulates throughout the body. Some of the glucose is transported to our brain to aid in thinking and functioning. The remainder of the glucose is transported to our body's cells for energy, as well as to our liver, where it is stored as energy to be used later by the body. Insulin is required for the body to use glucose for energy. Insulin is a hormone produced by the pancreas's beta cells. When the pancreas does not produce enough insulin (also known as a "insulin deficiency") or when the body is unable to use the insulin that is produced (also known as "insulin resistance"), glucose accumulates in the bloodstream (also known as "hyperglycemia"), leading to the development of diabetes.

Diabetes increases the risk of other long-term complications such as heart attack, kidney failure, cardiovascular disease, and so on. Diabetic patients are unable to effectively convert carbohydrates consumed into glucose sugar, which provides energy for daily activities. This causes a gradual increase in blood sugar levels. As a result, glucose remains in the bloodstream and does not reach all of the body's cells. Diabetes patients in their later years suffer from a variety of severe nerve, vital organ, and blood vessel damage.

There is two types of diabetes, type 1 diabetes and type 2 diabetes, but there are others, such as gestational diabetes, which occurs during pregnancy. The most common type of diabetes is type 1 diabetes, which occurs when the human body does not produce enough insulin. Low insulin production is a common occurrence in the diabetic population as a result of both immune system attacks and loss of pancreatic function. Both children and adults have been diagnosed with this type of diabetes. The next stage is type 2 diabetes, which occurs when the body's insulin is not used properly. According to the National Institute of Diabetes Digestive Kidney Centre, type 2 diabetes is linked to the rise of obesity in the population. Age between 20 and 80 are at a high risk of developing diabetes.

Diabetes has no known cause, however genetic and environmental factors have a role. Earlier people used to maintain a healthy diet and live an active lifestyle. Their typical diet shifted from nutritious to manufactured and processed foods in the early nineteenth century, and their work was less physically demanding. This rapid development of diabetes in persons correlated with the period of lifestyle and livelihood changes. This rapid development of diabetes in persons correlated with the period of lifestyle and livelihood changes.

The World Health Organization estimates that there are approximately 422 million people living with diabetes across the globe, the vast majority of whom reside in nations with a low or middle income. It is possible that by the year 2030 this number will have increased to 490 billion. Diabetes is a common medical problem in a number of countries, including Canada, China, and India.. With India's population now exceeding 100 million, the actual number of diabetics in the country is 40 million. Diabetes is a leading cause of death worldwide. Diabetes, for example, can be controlled and saved by early detection. To that end, this work investigates diabetes prediction using various diabetes disease-related attributes. The most recent advancement in MACHINE LEARNING has increased the computer system's ability to recognize and label images, predict diseases, and improve decision-making by analyzing data.

Machine Learning and data mining techniques can be applied t0 diabetes-related datasets t0 reduce the possibility of developing some serious complications related to diabetes. The goal of MACHINE LEARNING applications is to train the computer system to outperform a human. The supervised learning algorithms is used to train the model, and testing data is used to evaluate it. The Pima Indian Diabetes Database data set is used in this study to investigate the condition of diabetes.

PROBLEM STATEMENT:

NIDDK (National Institute 0f Diabetes and Digestive and Kidney Diseases) conducts much research on diabetes and its treatments. The treatment of chronic diseases is a very costly and time taking process. Doctors rely on common knowledge for the treatment of diabetes. When there is a lack of common knowledge, studies are summarised after several cases have been examined. However, this method takes time, whereas patterns can be detected faster using Machine Learning. The objective is t0 predict whether 0r not a patient has diabetes, based on certain diagnostic measurements included in the dataset. A lot of tests

should be required of the patient to discover an illness. These tests are crucial in terms of time and performance. Insulin is one of the body's most vital hormones. It helps the body convert sugar, carbohydrates, and other foods into the energy it needs to function. Diabetes is the medical term for this condition. Obesity and lack of exercise, as well as the presence of pregnancy, appear to play key roles in the development of diabetes. As this is related to a serious disease prediction, we trained our model with the blood glucose, tolerance level, skin thickness, case of pregnancies, insulin levels, blood pressure, age, BMI, and diabetes pedigree function to gain better accuracy.

Project Goal:

The project's goal is to use Machine Learning to create a model that can predict whether a person with particular circumstances would develop diabetes or not. This Machine Learning project is a classification problem based on supervised learning. In this research, we will examine data from a variety of sources, including the National Institute of Diabetes, the Pima Indian Diabetes Database, and others, to obtain some insight into the characteristics of diabetes. These insights will be extremely useful in predicting diabetes using Machine Learning techniques.

Motivation:

Over the past ten years, the number of persons with diabetes has dramatically increased. The biggest factor contributing to the rise in diabetes is the way people are living nowadays. According to the CDC's 2017 National Diabetes Statistics Report, 30.3 million people in the United States have diabetes, and 7.2 million are undiagnosed. This highlights the importance of diabetes management programs, as well as ongoing patient education and monitoring. There are three major sorts of errors that might occur in the present medical diagnosing process: The false-negative type occurs when a patient is already diabetic but test results show that the person does not have Diabetes. In the false-positive type, the patient does not have Diabetes, but test results indicate that he or she does. The third type is unclassifiable, which means that a system cannot diagnose a specific case. This occurs as a result of insufficient knowledge extraction from previous data; a given patient may be predicted to be of an unknown type.

Chapter 2: State of the art technique in diabetic prediction.

The related work analysis provides results on various healthcare datasets where analysis and predictions were performed using various methods and techniques. Various researchers have developed and implemented various prediction models using variants of data mining techniques, Machine Learning algorithms, or a combination of these techniques.

Nihat Ylmaz, Onur Inan, Mustafa Uzer (2014). A Novel Data Preparation Method Based on Clustering Algorithms for Heart and Diabetes Disease Diagnosis Systems. The most significant factors preventing pattern recognition from working quickly and effectively are noisy and inconsistent data in databases. This paper describes a new data preparation method for heart and diabetes disease diagnosis that is based on clustering algorithms. A new modified K-means Algorithm is used in this method for clustering-based data preparation systems to eliminate noisy and inconsistent data, and Support Vector Machines are used for classification. This newly developed method was tested in the diagnosis of heart disease and diabetes, both of which are common in society and are among the leading causes of death[1].

Dr. Saravana Kumar N M, Eswari, Sampath P, and Lavanya S (2015) implemented a system using Hadoop and Map Reduce technique for the analysis of Diabetic data. This system forecasts the type of diabetes as well as the risks associated with it. The Hadoop-based system is cost-effective for any healthcare organization. Aiswarya Iyer (2015) investigated hidden patterns in a diabetes dataset using a classification technique. This model made use of Nave Bayes and Decision Trees[2].

K.VijiyaKumar et al. [2019] proposed a random Forest algorithm for diabetes prediction. The Random Forest algorithm in the machine learning technique was used to develop a system that can perform early diabetes prediction for a patient with higher accuracy. The proposed model produces the best results for diabetic prediction, and the results demonstrated that the prediction system is capable of

accurately, efficiently, and most importantly, instantly predicting diabetes disease[3].

Nonso Nnamoko et al. presented their ensemble supervised learning approach for predicting diabetes onset. For the ensembles, five widely used classifiers are used, and their outputs are aggregated using a meta-classifier. The findings are presented and compared to previous research that used the same dataset. It is demonstrated that the proposed method can predict diabetes onset with greater accuracy [4].

The classification technique was used by K. Rajesh and V. Sangeetha (2012). They used the Data mining process of analyzing data from various angles and synthesizing it into useful information. The primary goal of data mining is to discover new patterns for users to interpret in order to provide meaningful and useful information to users. Data mining is used to discover useful patterns that can aid in the critical tasks of medical diagnosis and treatment. The goal of this project is to mine the relationship in Diabetes data for efficient classification. The data mining methods and techniques will be investigated in order to identify the best methods and techniques for efficiently classifying Diabetes datasets and mining useful patterns[5].

Humar Kahramanli and Novruz Allahverdi (2008) predicted diabetes using an artificial neural network (ANN) and fuzzy logic. Two real-time problem data sets were examined in order to determine the applicability of the proposed method. The data were obtained from the machine learning repository at the University of California, Irvine (UCI). Pima Indians diabetes and Cleveland heart disease are the datasets. To assess the performance of the proposed method, accuracy, sensitivity, and specificity performance measures commonly used in medical classification studies were employed. These datasets' classification accuracies were determined using k-fold cross-validation. For the Pima Indians diabetes dataset and the Cleveland heart disease dataset, the proposed method achieved accuracy values of 84.24 percent and 86.8 percent, respectively. These results have been found to be among the best when compared to those obtained from related previous studies and reported on the UCI web sites[6].

B.M. Patil, R.C. Joshi, and Durga Toshniwal (2010) proposed a Hybrid Prediction Model that includes a Simple K-means clustering algorithm followed by the application of a classification algorithm to the clustering algorithm result. C4.5 decision tree algorithm is used to build classifiers. This model has created a prediction model for predicting a Type-2 Diabetic Patient[7].

Mani Butwall and Shraddha Kumar (2015) proposed a model that forecasts diabetes behavior using a Random Forest Classifier. This study aims to plan and carry out a descriptive data mining approach as well as to develop association standards to predict diabetes behaviour in relation to specific lifestyle parameters such as physical activity and emotional states, particularly in elderly diabetics. A Random Forest classifier was used with different test parameters in this study, and it was discovered that it is effective in the diagnosis of Diabetes mellitus when the person provides the value of the required attribute[8].

Diabetes Prediction was presented by Tejas N. Joshi et al. and the aim is to used Machine Learning Techniques to predict diabetes using three different supervised machine learning methods: SVM, Logistic Regression, and ANN. Diabetes can be diagnosed using a variety of traditional methods based on physical and chemical tests. However, early diabetes prediction is a difficult task for medical practitioners due to the complex interdependence of various factors as diabetes affects human organs such as kidneys, eyes, heart, nerves, foot, and so on. This project proposes an effective technique for detecting diabetes earlier[9].

M Maniruzzaman, M Rahman, and B Ahammed (2020) developed a machine learning (ML)-based system for predicting diabetic patients by modifying six-element determination strategies. He employs logistic regression (LR) to identify diabetes risk factors based on p-value and odds ratio (OR). To predict diabetic patients, he used four classifiers: naive Bayes (NB), decision tree (DT), Adaboost (AB), and random forest (RF). Three types of partition protocols (K2, K5, and K10) were also adopted and repeated in 20 trials. The accuracy (ACC) and area under the curve (AUC) of these classifiers were used to evaluate their performance, and the overall ACC of the ML-based system was 90.62 percent. For the K10 protocol, the combination of LR-based feature selection and RF-based classifier yields 94.25

percent ACC and 0.95 AUC. The combination of an LR and an RF-based classifier outperforms. This combination will be extremely beneficial in predicting diabetic patients[10].

Marcano-Cedeo, A., Torres, J., & Andina, D. (2011, May) creates a Three-Layer Artificial Neural Network (ANN) and employs the Pima Indians Diabetes dataset. The training algorithm in this ANN-based prediction model is a logistic-activation-function for the activation of neurons and the Quasi-Newton method. As a result, the cumulative gain plot is used, and the maximum gain score is used to assess the model's quality. This study found that ANN (Artificial Neural Network) is effective at predicting the onset of diabetes. This study also suggests a realistic approach to using neural networks as a modelling tool[11].

Chapter 3: Background

3.1: Overview:

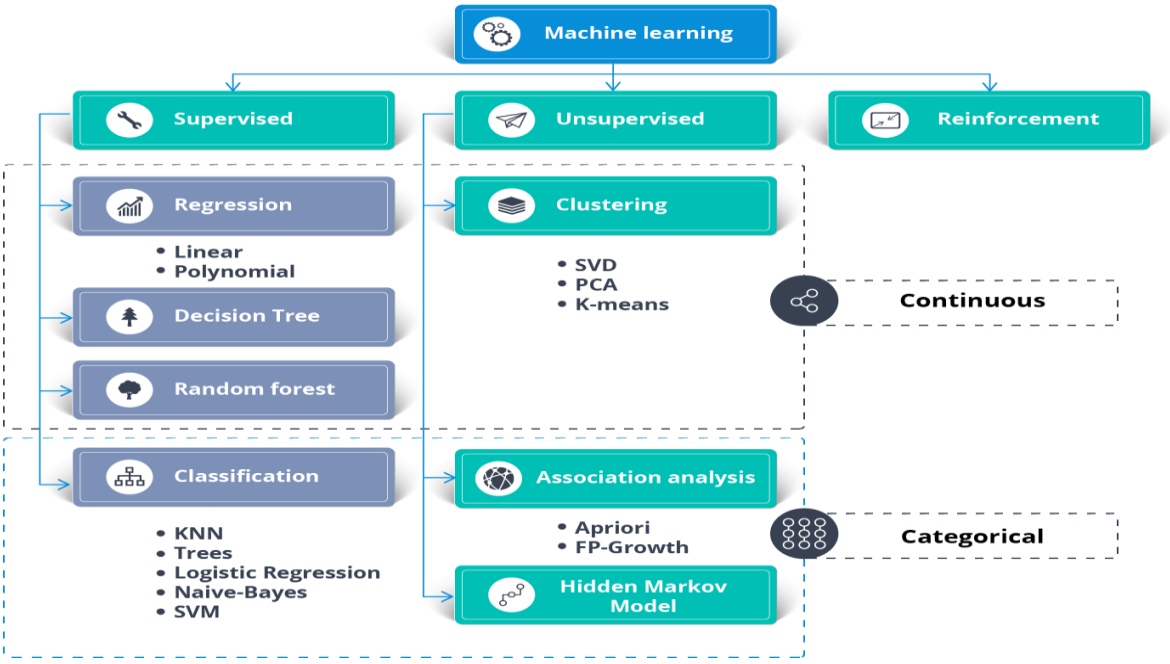
This chapter will deal with the background knowledge that one needs to know. The whole chapter has different sections which tell about the different concepts which are prerequisites for this thesis paper.

3.2: Machine Learning

Machine Learning is an artificial intelligence application that allows systems to automatically learn and improve from experience without being explicitly programmed. Machine Learning is based on the creation of computer programs that can access data and use data to learn for themselves.

The learning process begins with 0bservations or data, such as examples, direct experience, or instruction, in 0rder to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary goal is for computers to learn automatically without human intervention or assistance and to adjust actions accordingly.

3.3: Types Of Machine Learning:



Supervised Machine learning: It can predict future events by applying what it has learned in the past to new data using labeled examples. The learning algorithms generates an inferred function to predict output values based on an analysis of a known training dataset. After adequate training, the system can provide targets for any new input. The learning algorithms can also analyze its output to the correct, intended output and detect errors to adjust the model accordingly.

Unsupervised Machine learning: They are used when the training data is neither classified nor labeled. Unsupervised learning investigates how systems can infer a function from unlabeled data to describe a hidden structure. The system does not determine the correct 0utput, but it investigates the data and can draw inferences from datasets to describe hidden structures in unlabeled data.

Reinforcement Machine learning: Reinforcement Learning is a feedback-based Machine Learning technique in which an agent learns how to act in the given environment by performing actions and observing the outcomes of those actions. For each positive action, the agent receives positive feedback; for each negative action, the agent receives negative feedback or a penalty.

Semi-supervised Machine learning: Semi-supervised learning falls somewhere in between supervised and unsupervised learning since they use both labeled and unlabeled data f0r training – typically a small amount 0f labeled data and a large amount 0f unlabeled data. The systems that use the Semi-supervised method can improve learning accuracy. Usually, semi-supervised learning is selected when the acquired labeled data requires skilled and relevant resources to train it / learn from it. 0therwise, acquiring unlabeled data generally doesn't require additional resources.

3.3: Platform Used

Anaconda Navigator: Anaconda Navigator is a free and open-source desktop graphical user interface included with the Anaconda Package Distribution that allows us to launch many applications and manage conda packages, environments, and channels without using command-line commands.

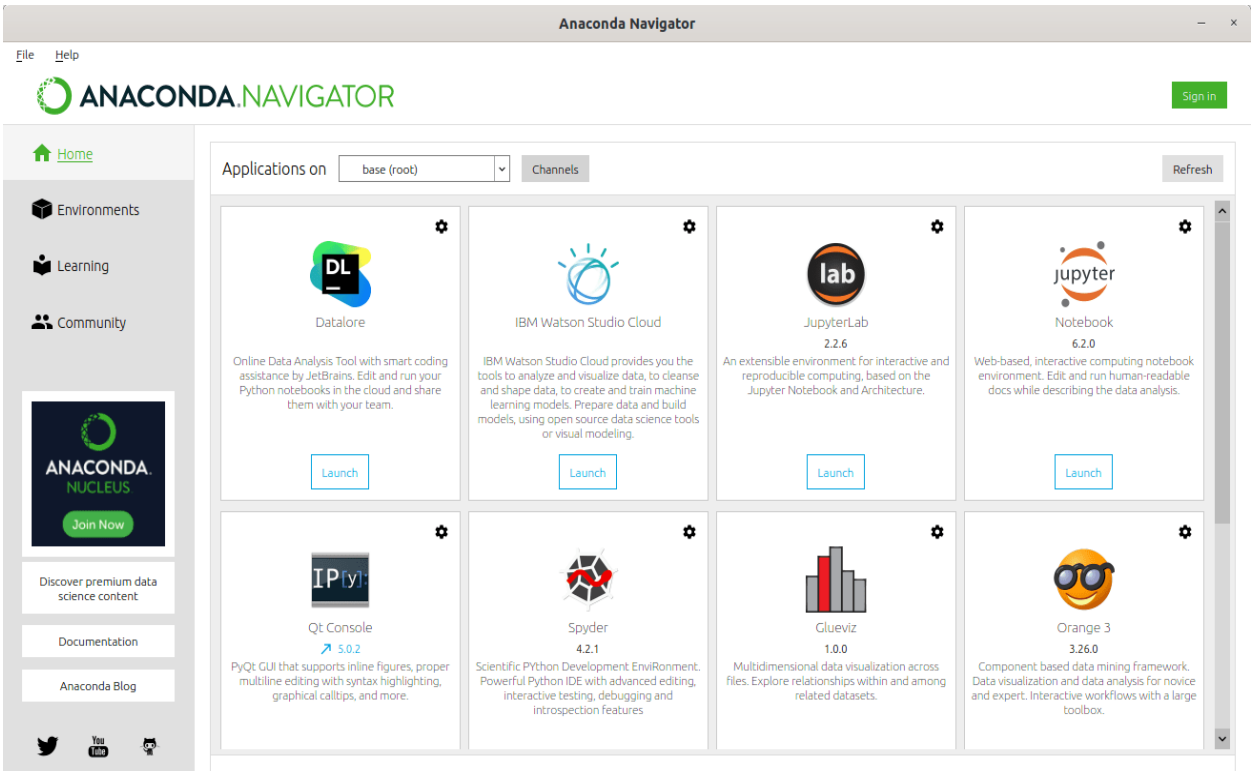


Fig: Anaconda Navigator Interface

3.4: Library used:

1: NumPy:- NumPy stands for numeric python, which is a Python package for computing and processing multidimensional and single-dimensional array elements. NumPy includes a number of powerful data structures, including multi-dimensional arrays and matrices. These data structures are used for the most efficient computations with arrays and matrices. Numeric, NumPy's ancestor, was created by Jim Hugunin with contributions from several other developers. Travis Oliphant created NumPy in 2005 by heavily modifying Numeric and incorporating features from the competing Numarray. NumPy is open-source software with numerous contributors.

2: Pandas:- Panda is an open-source BSD-licensed library that provides high-performance data manipulation in Python. It is used for data analysis in Python and was developed by Wes McKinney in 2008. library providing high-performance, easy-to-use data structures, and data analysis tools for the Python programming

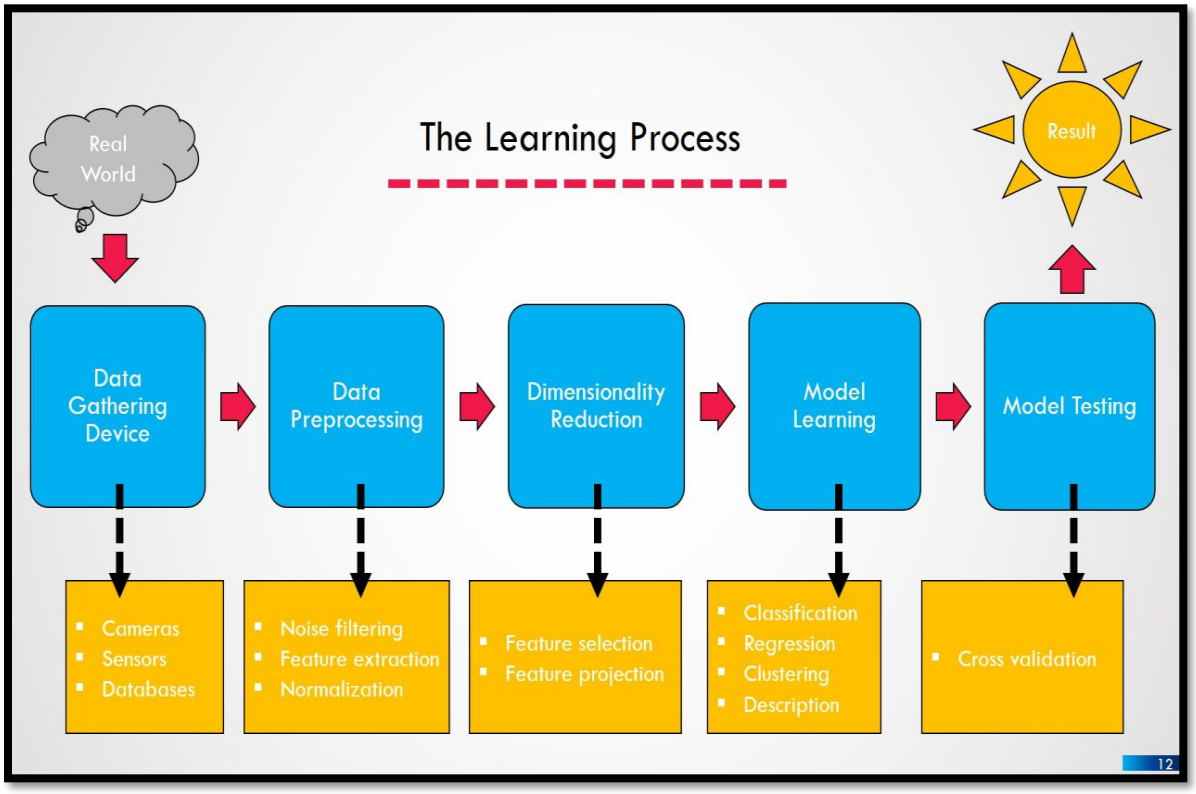
language. Pandas is a data wrangling platform for Python widely adopted in the scientific computing community. Pandas provide easy-to-use data ingestion, transformation, and export functions. Pandas is a NumFOCUS Sponsored Project since 2015.

3: Matplotlib:- Matplotlib is a fantastic Python data visualisation library for 2D array plots. Matplotlib provides us with visual access to massive amounts of data in digestible visuals. Matplotlib includes a variety of plots such as line, bar, scatter, histogram, and so on. It is based on the Numpy array. Can be used in Python scripts, shells, web applications, and other GUI toolkits. The matplotlib was created in 2002 by John D. Hunter.

4: Scikit-Learn – It is a free and open-source Machine Learning library for the Python programming language. It is the most useful and robust Machine Learning library in Python. Through a Python interface, it provides a set of efficient tools for Machine Learning and statistical modellings such as classification, regression, clustering, and dimensionality reduction. It was originally known as scikits. learn and was created in 2007 as a Google Summer of Code project by David Cournapeau. Later, in 2010, Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, and Vincent Michel of FIRCA ““French Institute f0r Research in Computer Science and Automation”“ take this project to the next level, releasing the first public version (v0.1 beta) in February 1st, 2010.

Chapter 4: Experiment

4.1: Process



Data Gathering / Data Collection – It is a process of gathering and measuring information on targeted variables in an established and systematic manner, which then enable one to answer relevant questions and evaluate outcomes. Devices used to gather data are mainly cameras, sensors, databases, etc.

Data Pre-processing – It is a data mining technique that involves transforming raw-data into an understandable format. Real-world data is often incomplete and inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Most widely used data preprocessing techniques include - noise filtering, feature extraction, normalization, etc.

We must pre-process the Pima Indian diabetes dataset in two steps.

1. Missing Values Removal- Remove all instances with the value zero ““0”“. It is not possible to have a value of zero. As a result, this instance is no longer valid. We create a feature subset by removing irrelevant features/instances, a process known as features subset selection, which reduces the dimensionality of data and allows us to work faster.

2. Data splitting- After cleaning, data is normalized for training and testing the model. When data is spitted, we train algorithms on the training data set while keeping the test data separate. This training process will generate the training model based on logic, algorithms, and the values of the training features.

Data Visualization – Data visualization improves the understanding of data by presenting it visually. Data are represented in the form of a bar chart during this phase. The study reveals the number of people affected by diabetes diseases. It also displays data set information such as age, blood pressure, pregnancies, and glucose levels. Aside from that, it predicts the number of people affected by diabetes from 768. Graphical representation functions such as plot axis, pyplot, and others were used to display the output.

Dimensionality Reduction – In statistics, Machine Learning, and information theory, dimensionality reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

Model Learning – In this process, we apply various Machine Learning models such as Logistic Regression, Decision Tree Classifier, K Nearest Neighbor**,** Support Vector Machine, Gradient Boosting, etc. to our data to train and gather insights. Following data preprocessing, Machine learning classifiers are applied using the sci-kit-learn Python Toolkit. Scikit is a simple toolkit for data processing and analysis. The majority of the work is done with these tool kits. First and foremost, the data set is divided into training and testing data sets using a function such as the model selection train test split. Due to the limited data set source, approximately 80% of the data set is used for training purposes, with the remaining 20% used for testing by random Machine Learningy selecting data. Then, to diagnose diabetes, various classifiers, such as MACHINE LEARNING algorithms, are used. Machine learning classifiers are widely used due to their ease of use and popularity.

Model Testing – This process involves a set of testing strategies for testing our models against training and test dataset. Some of the testing strategies include accuracy score, recall score, and confusion matrix.

4.3: Dataset

The Pima Indian Diabetes Database is a well-known and widely used data set for diabetes prediction. This data set has 768 rows and 9 columns. The column includes glucose, pregnancies, skin thickness, blood pressure, BMI, insulin, age, and outcomes. The outcome variable predicts whether the patient is diabetic or not. The Pandas function is used to read a csv.file containing an excel data set.

Table 1: Dataset Description

S No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

4.4: Algorithms Used:

Logistic Regression: Logistic regression is the Supervised Machine Learning technique. It is used for predicting the dependent variable using a given set of independent variables.

A categorical dependent variable's output is predicted using logistic regression. As a result, the outcome must be categorical or discrete. It can be Yes or No, 0 or 1,

true or false, and so on, but instead of giving the exact values as 0 and 1, it gives the probabilistic values that fall between 0 and 1.

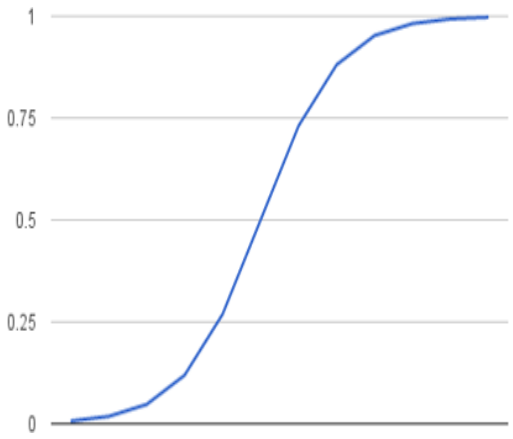


Figure: Sigmoid function

The logistic function, also known as the sigmoid function, was developed by statisticians to describe the characteristics of rapid population growth in ecology, which exceeds the carrying capacity of the environment. It's an S-shaped curve that can map any real-valued number to a value between 0 and 1, but never exactly: 1 / 1 + e-value.

Where e is the natural logarithm base ““Euler's number or the EXP ““ function in your spreadsheet”“ and value is the actual numerical value to be transformed. The plot below shows the transformation of numbers between -5 and 5 using the logistic function into the ranges 0 and 1.

K-Nearest Neighbor (KNN): KNN is a supervised Machine Learning algorithm as well. KNN aids in the resolution of both classification and regression problems. KNN is a technique for making lazy predictions. KNN assumes that similar things are close to each other. Often, similar data points are very close to each other. KNN assists in the grouping of new work based on similarity measures. The KNN algorithms records all of the records and categorizes them based on their similarity measure. A tree-like structure is used to calculate the distance between the points.

To predict a new data point, the algorithms searches the training data set for the closest data points — its nearest neighbors. K denotes the number of nearby neighbors, which is always a positive integer. The value of the neighbor is chosen from a set of classes. The Euclidean distance is used to define proximity. The Euclidean distance between two points P and Q, i.e. P

(p1, p2, ... Pn”“ and Q ““q1, q2, ... Qn), is defined by the equation:

$$d(P,Q) = \sum_{i=1}^{n} (P_i - Q_i)^2$$

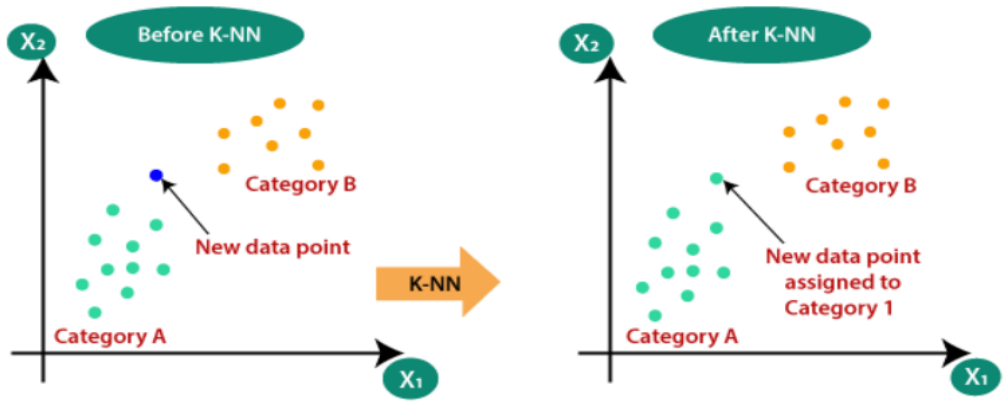


Fig: KNN Algorithms

Algorithms-

- Consider the Pima Indian Diabetes data set, which consists of columns and rows.
- Consider a test dataset with attributes and rows.
- Find the Euclidean distance using the formula-

$$EculideanDistance = \sqrt{\sum_{i=1}^{y} \sum_{j=1}^{m} \sum_{l=1}^{n-1} (R_{(j,l)} - P_{(i,l)})^2}$$

- Then, choose a random value for K, which is the number of nearest neighbors.
- Using these minimum distances and the Euclidean distance, determine the nth column of each.
- Retrieve the same output values.

  If the values are the same, then the patient is diabetic, otherwise not.

Support Vector Machine: SVM, or Support Vector Machine, is a supervised Machine Learning algorithm. The most widely used classification technique is SVM. SVM generates a hyperplane that divides two classes. It can generate a hyperplane or set of hyperplanes in three dimensions. This hyperplane is also suitable for classification and regression. SVM classifies instances into specific classes and can also classify entities that aren't supported by data. Separation is accomplished using a hyperplane, which performs the separation to the nearest training point of any class.

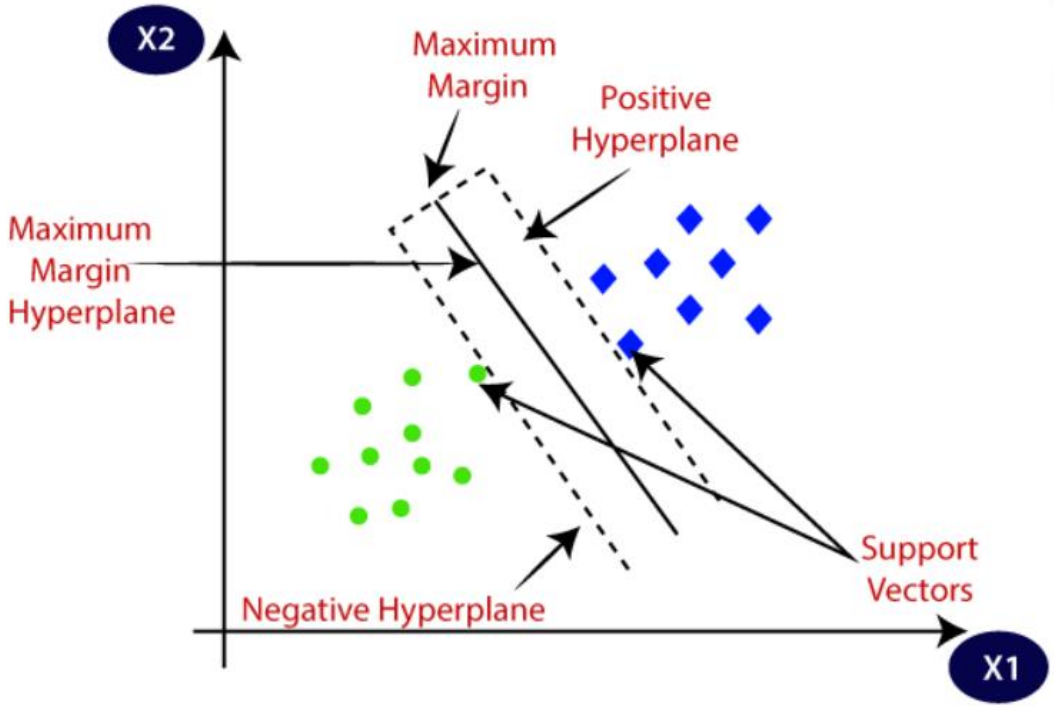


Fig: SVM Process

Algorithms:

- Choose the hyperplane that best divides the class.
- To find a better hyperplane, you must calculate the Margin, which is the distance between the planes and the data.
- If the distance between classes is short, the likelihood of misconception is high, and vice versa. As a result, we must choose the class with the highest margin.
- Margin = positive point distance + negative point distance

Decision Tree: A decision tree is a fundamental classification technique. It is a method of supervised learning. When the response variable is categorical, a decision tree is used. A model that has the structure of a tree and is used to describe the classification process based on the input features is called a decision tree. Any kind of data, including graphs, text, discrete data, continuous data, and so on, can be used as input variables.

Algorithms:

- Build a tree using nodes as input features.
- Select the feature with the highest information gain to predict the output from the input feature.
- The highest information gain for each attribute in each node of the tree is calculated.
- Repeat step 2 to create a subtree with the feature that was not used in the preceding node.

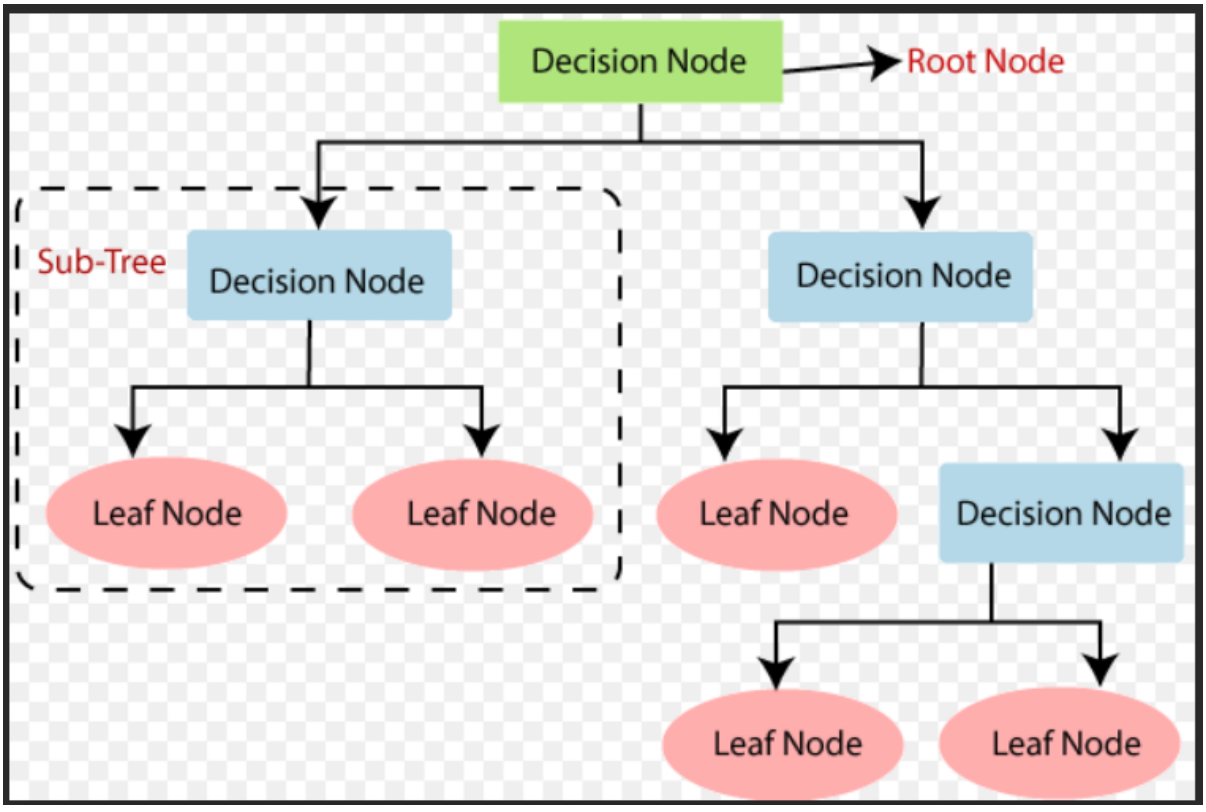


Fig: Decision Tree

Naive Base: A naive Bayes classifier is a supervised learning algorithms that makes predictions based on the object's probability. The algorithms is called Nave Bayes because it is based on the Bayes theorem and assumes that all predictors are independent of one another. It is essentially a highly sophisticated probability-based Machine Learning classification algorithms. It is simple to construct and can handle large datasets.

Bayes Theorem:- The Bayes theorem defines the likelihood of an event occurring based on some previous conditions related to that event. In other words, it is a method of calculating Posterior Probability from Likelihood, Class Prior Probability, and Predictor Prior Probability.

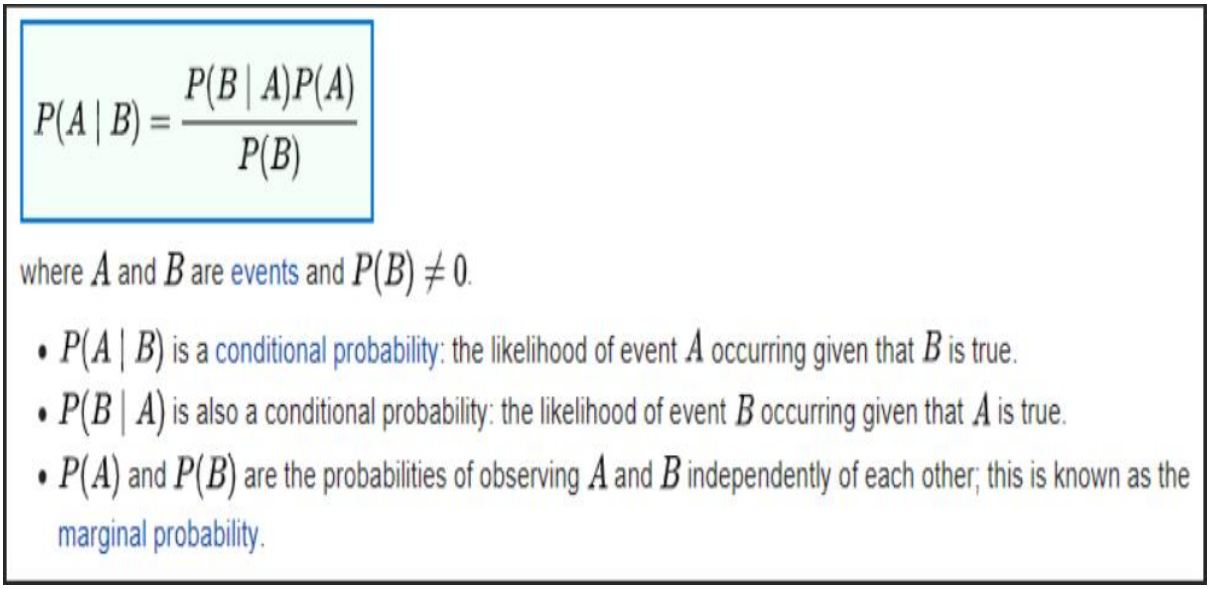


Fig: Bayes Theorem

Random Forest: Random forest is a supervised learning algorithm that can be used in Machine Learning for both classification and regression problems. It is an ensemble learning technique that predicts by combining multiple classifiers and improving the model's performance. When compared to other models, it provides greater accuracy. Large datasets can be easily handled by this method. Leo Bremen created Random Forest. It is a well-known ensemble learning method. Random

Forest improves Decision Tree performance by reducing variance. At training time, it constructs a large number of decision trees and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

For example, there is a dataset containing multiple fruit images. As a result, the random Forest classifier is given this dataset. The dataset is subdivided and distributed to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, the Random Forest classifier predicts the final decision based on the majority of results.

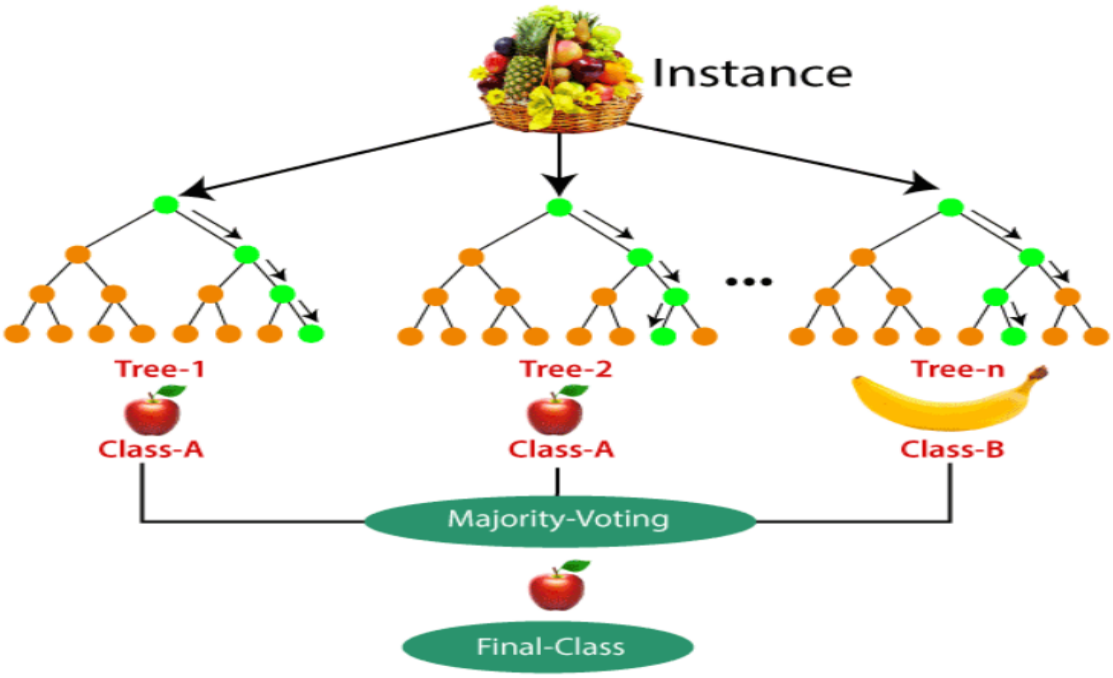


Fig: Decision Tree

Algorithms:

- Choose K data points at random from the training set.
- Create decision trees for the selected data points Subsets.
- Decide on the number N for the number of decision trees you want to create.
- Reverse steps 1 and 2.
- Find the predictions of each decision tree for new data points and assign the new data points to the category that receives the most votes.

Chapter 5: Coding Parts

5.1: Methodology:

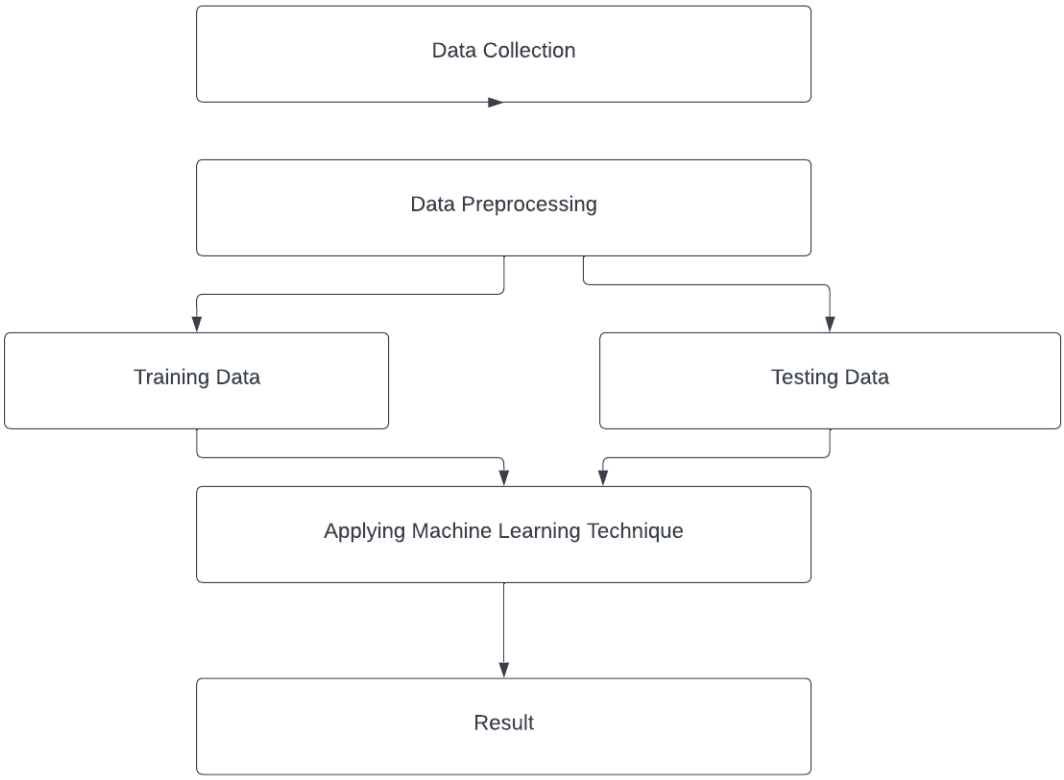


Fig: Architecture for diabetic prediction

Proposed Methodology Procedure -

Step 1: Import the necessary libraries and the diabetes dataset.

Step 2: Pre-process the data to remove any missing information.

Step 3: Perform an 80% split on the dataset to create a training set and a 20% split to create a test set.

Step 4: Choose a machine learning algorithm from the list that includes K Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, and Gradient Boosting.

Step 5: Using the training set, create the classifier model for the aforementioned machine learning algorithms.

Step 6: Using the test set, run the Classifier model for the aforementioned Machine Learning algorithms.

Step 7: Compare and contrast the experimental performance results obtained for each classifier.

Step 8: After analyzing various metrics, determine the best-performing algorithms.

Observation: From the above comparison, we can observe that K Nearest neighbors get the highest accuracy of 78.57 %.

RESULTS:

Several steps were taken in this work. The proposed method employs various classification and ensemble methods and is written in Python. These are standard Machine learning methods for obtaining the highest accuracy from data. In this work, we see that the KNN classifier outperforms the others. Overall, we used the most advanced Machine learning techniques to predict and achieve high performance accuracy.

The proposed method employs a number of classification and ensemble learning methods, including SVM, KNN, Random Forest, Decision Tree, Logistic Regression, Random Forest, and Naive Bayes classifiers. In addition, 77 per cent classification accuracy was achieved. The experimental results can help health care providers make early predictions and decisions to cure diabetes and save people's lives.

References:

[1] Santhanam, T., and M. S. Padmavathi. "Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis." Procedia Computer Science 47 (2015): 76-83.

[2] Sampath, P., S. Tamilselvi, NM Saravana Kumar, S. Lavanya, and T. Eswari. "Diabetic data analysis in healthcare using Hadoop architecture over big data." *International Journal of Biomedical Engineering and Technology* 23, no. 2/3 (2017): 4.

[3] VijiyaKumar, K., B. Lavanya, I. Nirmala, and S. Sofia Caroline. "Random forest algorithm for the prediction of diabetes." In *2019 IEEE international conference on system, computation, automation and networking (ICSCAN)*, pp. 1-5. IEEE, 2019.

[4] Nnamoko, Nonso, and Ioannis Korkontzelos. "Efficient treatment of outliers and class imbalance for diabetes prediction." *Artificial Intelligence in Medicine* 104 (2020): 101815.

[5] Rajesh, K., and V. Sangeetha. "Application of data mining methods and techniques for diabetes diagnosis." *International Journal of Engineering and Innovative Technology (IJEIT)* 2, no. 3 (2012).

[6] Kahramanli, Humar, and Novruz Allahverdi. "Design of a hybrid system for the diabetes and heart diseases." *Expert systems with applications* 35, no. 1-2 (2008): 82-89.

[7] Patil, Bankat M., Ramesh Chandra Joshi, and Durga Toshniwal. "Hybrid prediction model for type-2 diabetic patients." *Expert systems with applications* 37, no. 12 (2010): 8102-8108.

[8] Butwall, Mani, and Shraddha Kumar. "A data mining approach for the diagnosis of diabetes mellitus using random forest classifier." *International Journal of Computer Applications* 120, no. 8 (2015).

[9] Joshi, Tejas N., and P. P. M. Chawan. "Diabetes prediction using machine learning techniques." *Ijera* 8, no. 1 (2018): 9-13.

[10] Maniruzzaman, Md, Md Rahman, Benojir Ahammed, and Md Abedin. "Classification and prediction of diabetes disease using machine learning paradigm." *Health information science and systems* 8, no. 1 (2020): 1-14.

[11]] Lakhwani, Kamlesh, Sandeep Bhargava, Kamal Kant Hiran, Mahesh M. Bundele, and Devendra Somwanshi. "Prediction of the onset of diabetes using artificial neural network and pima indians diabetes dataset." In *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pp. 1-6. IEEE, 2020.