PREDICTION OF DIABETES USING MACHINE LEARNING ALGORITHMS

Submitted By

Sk. Salman Rafsun
Id no: 0242220005153007

This Report Presented in Partial Fulfilment of the Requirements for the Degree of Masters of Science in Electronics and Telecommunication Engineering.

Supervised By

Md. Taslim Arefin
Associate Professor and Head
Department of ETE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2025

APPROVAL

This research titled **"Prediction of Diabetes Using Machine Learning Algorithms "** submitted by **Sk Salman Rafsun Id: 0242220005153007** to the department of Information and Communication Engineering (ICE), Daffodil International University, has been accepted as satisfactory for the partial fulfilments of the requirements for M.Sc. in Electronics and Telecommunication Engineering (ETE) and approved as to its style and contents. The presentation was held on 25th January 2025.

Board of Examiners

Md. Taslim Arefin — **Chairman**
Associate Professor and Head
Department of ICE
Faculty of Engineering
Daffodil International University

Professor Dr. A. K. M. Fazlul Haque — **Internal Examiner**
Professor
Department of ICE
Faculty of Engineering
Daffodil International University

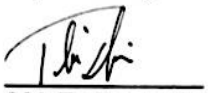
Engr. Md. Zahirul Islam — **Internal Examiner**
Assistant Professor
Department of ICE
Faculty of Engineering
Daffodil International University

Md. Ashraful Alam Bhuiyan — **External Examiner**
Chief Technology Officer (CTO)
Bank Alfalah, Bangladesh

DECLARATION

I hereby declare that this research is my own work and effort under the supervision of **Md. Taslim Arefin, Assistant Professor and Head, Department of Information and Communication Engineering, Daffodil International University, Dhaka.** It has not been submitted anywhere for any award. Where other sources of information have been used, they have been acknowledged.

Supervised By



Md. Taslim Arefin
Associate Professor and Head
Department of ICE
Faculty of Engineering
Daffodil International University

Submitted By

Sk Salman Rafsun
Student ID: 0242220005153007
Department of ETE
Faculty of Engineering
Daffodil International University

ACKNOWLEDGEMENT

Firstly, I would want to thank God Almighty for His wonderful grace that enabled me to finish this job successfully.

I am sincerely grateful to and deeply indebted to Md. Taslim Arefin, Associate Professor and Head of the ETE Department at Daffodil International University, Dhaka. He has been very supportive in providing useful advice, guidance, instructions, and general supervision during the inquiry of this project which I appreciate very much. I also thank him for reviewing the first drafts of my project and giving me valuable comments and suggestions for improvement.

I'd want to sincerely regards the instructors and employees of Daffodil International University's ETE department.

I want to express my gratitude to every student at Daffodil International University who participated in this discussion while finishing the assignment.

Lastly, I must respectfully thank my parents for their unwavering patience and support.

ABSTRACT

Diabetes is a frequent condition in humans that is brought on by a collection of metabolic diseases in which the body's sugar levels remain abnormally high for an extended length of time. Because it damages many of the body's systems by affecting various organs, we are all trying to prevent diabetes at an early stage by anticipating its symptoms using a variety of techniques.

Human life can be saved by controlling such diseases early on. In order to accomplish the goal, this research project primarily uses machine learning approaches to investigate different risk variables associated with this disease. Effective knowledge extraction is accomplished by machine learning methods that build prediction models using diagnostic medical datasets from diabetic patients. It may be possible to forecast diabetic people by gleaning information from such data. K-Means Cluster, Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Linear Regression, Decision Tree (DT), Logistic Regression, Random Forest (RF) and Hierarchical Cluster are nine well-known machine learning algorithms that I use in this work to predict diabetic disease using data from the adult population. In comparison to other machine learning methods, the findings from my experiments indicate that when compared to alternative approaches, the C4.5 decision tree attained a greater level of accuracy.

Table of Contents

Topic	Page
Approval	ii
Declaration	iii
Acknowledgement	iv
Abstract	v
Table of Contents	vi
List of figures	vii
List of tables	vii
Chapter 1: Introduction	1-4
1.1 Objective	1
1.2 Statement of Problems	2
1.3 Motivation	2
1.4 Methodology	2
Chapter 2: Background study	5-8
2.1 Introduction	5
2.2 Pregnancy	6
2.3 Glucose	6
2.4 Blood Pressure	6
2.5 Skin Thickness	6
2.6 Insulin	7
2.7 BMI	7
2.8 Age	7
Chapter 3: Algorithm Learning	9-28
3.1 Introduction	9
3.2 Supervised Machine Learning Algorithm	9-8
3.2.1 Linear Regression	11
3.2.2 Logistic Regression	13
3.2.3 Decision Tree Algorithm	16
3.2.4 Random Forest Algorithm	17
3.2.5 K-Nearest Neighbour (KNN) Algorithm	19
3.2.6 Support Vector Machine Algorithm	20
3.2.7 Naïve Bayes Algorithm	23
3.3 Unsupervised Machine Learning Algorithm	25
3.3.1 K-Means Clustering Algorithm	25
3.3.2 Hierarchical Clustering Algorithm	27
Chapter 4: Implementation and Performance Analysis	29-37
4.1 Introduction	29
4.2 Implementation of Dataset	29
4.2.1 Overview of Dataset and Attributes	30
4.2.2 Data Pre-processing	30
4.3 Implementation of Machine Learning Algorithm on Data set	31
4.3.1 Implementation of Supervised Machine Learning	31
4.3.2 Implementation of Unsupervised Machine Learning	34
4.4 Performance Analysis of Algorithms on diabetes data set	35

Chapter 5: Result and Discussion 38-40
5.1 Results 38
5.2 Discussion
Chapter 6: Conclusion and Future work 39
6.1 Conclusion 40
6.2 Future work and scope 40
References 41

List of Figure

Figure No	Title	Page No
Fig 1.4	An overview of the overall process	4
Fig 3.2.	Machine Learning Supervise Process.	10
Fig 3.2.1.1	Gradient Descent	12
Fig 3.2.1.2	Convex vs non-convex function	13
Fig 3.2.2.1	Linear Regression VS Logistic Regression Graph Image	13
Fig 3.2.2.2	Sigmoid Function Graph	14
Fig 3.2.3	Decision Tree Algorithm	17
Fig 3.2.6.1	Possible hyper planes	21
Fig 3.2.6.2	Hyper planes in 2D and 3D feature space	21
Fig 3.3.1	K-means iteration	26
Fig 4.2.1	Data set head	30
Fig 4.2.2	Data cleaning heat map	31
Fig 4.3	Confusion Matrix	35
Fig 4.3.2	Accuracy comparison Plotting	37

List of Table

Table No	Table Name	Page No
Tab 4.1	Accuracy of All Proposed Algorithms	36
Tab 5.1	Supervised and Unsupervised Accuracy Table	38

Chapter 1

Introduction

1.1 Objective

When a person has diabetes mellitus (DM), their blood glucose levels are unusually high., a condition where the body either fails to respond appropriately to it or does not make enough affront. One of the most prevalent chronic diseases worldwide is diabetes. Those who are not attempting to lose weight may experience increased thirst and urination. A characteristic of diabetes mellitus is persistently elevated blood sugar levels. In diabetes, blood vessel damage occurs, neurons, and feeling, increasing risks for eyesight, chronic renal disease, heart attacks, and strokes. Although there is no cure for diabetes, it is certainly manageable. In the past, the illness only affected adults. Some people these days are afflicted with this illness. This is primarily due to their inadequate diet. A balanced diet is the key to managing diabetes. Hyperthyroidism, pancreatitis diabetes during pregnancy, Acanthosis Nigricans, Polycystic Ovarian Syndrome, genetic disorders, excess body weight, weight gain, and a lack of regular exercise are some of the numerous causes of diabetes. Eating too much junk food also raises the body's calorie and fat content. As a result, the body produces more insulin with sugar.

As stated by WHO, diabetes specifically affects globally 422 million individuals, most of whom are low- and middle-income nations, and results in 1.6 million deaths annually. Both the prevalence and diabetes cases have been gradually increased during the last several decades. The diabetes epidemic is predicted to spread at an alarming rate in the absence of appropriate prevention and treatment measures [1].

It is possible to manage diseases and save lives by predicting them early. This study primarily investigates the early prediction of diabetes by accounting for a number of risk variables associated with the prerequisite for achieving this objective. I gather diagnostic datasets for the study that include 16 characteristics of 200 diabetes patients. Age, nutrition, high blood pressure, eyesight issues, genetics, etc. are some of these characteristics. These characteristics and their associated values are covered in more detail in the next section. In order to forecast diabetic disease, I build an expectation demonstration based on these characteristics using many machine learning techniques. When machine learning approaches are used to build

prediction models from diagnostic medical datasets obtained from diabetes patients, they yield effective results for knowledge extraction. Predicting diabetic people may be possible by gleaning information from such data. Numerous machine learning methods are capable of forecasting diabetes. But selecting the appropriate prediction method based on these characteristics is really challenging. For this reason, To predict diabetic disease, I use a few popular machine learning algorithms on data from the adult population: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and decision tree (DT), Naive Bayes (NB).

1.2 Statement of problems

In order to accomplish this, the research depicted in this thesis examines a number of medical predictor variables unique to each patient. The patient's age, BMI, insulin and glucose levels, number of pregnancies, and other factors are predictor variables. Based on specific diagnostic parameters included in the dataset, we develop a machine learning model that reliably predict whether or not the individuals in the dataset have diabetes. Additionally, supervised and unsupervised learning methods were used to compare prediction accuracy.

1.3 Motivation

To find the best methods for completing any complicated assignment with ease, research is essential. Consequently, I attempted to investigate the best machine learning method for diabetes disease prediction in this study. Additionally, I demonstrated the accuracy of nearly nine widely used methodologies and made comparisons using a confusion matrix. This helps me discover answers I didn't know existed. to addressing information shortages and altering the methods used by medical practitioners.

1.4 Methodology

Without a question, one of the most significant and powerful developments in the modern world is machine learning. The various types of actions that we would normally find difficult to uncover inside complex data are thus found through the use of machine learning

techniques. Better learning from data and its analysis without human intervention is made possible by the rapidly expanding discipline of machine learning. In the medical field, it is often utilized for assessing and identifying complex and critical illnesses.

The strategy consists of several steps to achieve the goal: collecting diabetes data sets with important persistent characteristics, pre-processing the numerical value attributes, using many machine learning classification methods, and comparing predictive research using the data.

In the bellow, we quickly discuss these phases;

i. Data set and it's attributes

For this study, I gathered the data set from Kaggle, which includes 768 patients' varied characteristics or risk factors for diabetes.

ii. Data Pre-processing

A small amount of data processing has been performed on gathered data set in order to meet the research goal. I replaced the null value with the mean value using a function to exclude null values from the data collection.

iii. Apply Machine learning techniques

I used both supervised and unsupervised machine learning methods to accomplish the objective after the data was processed.

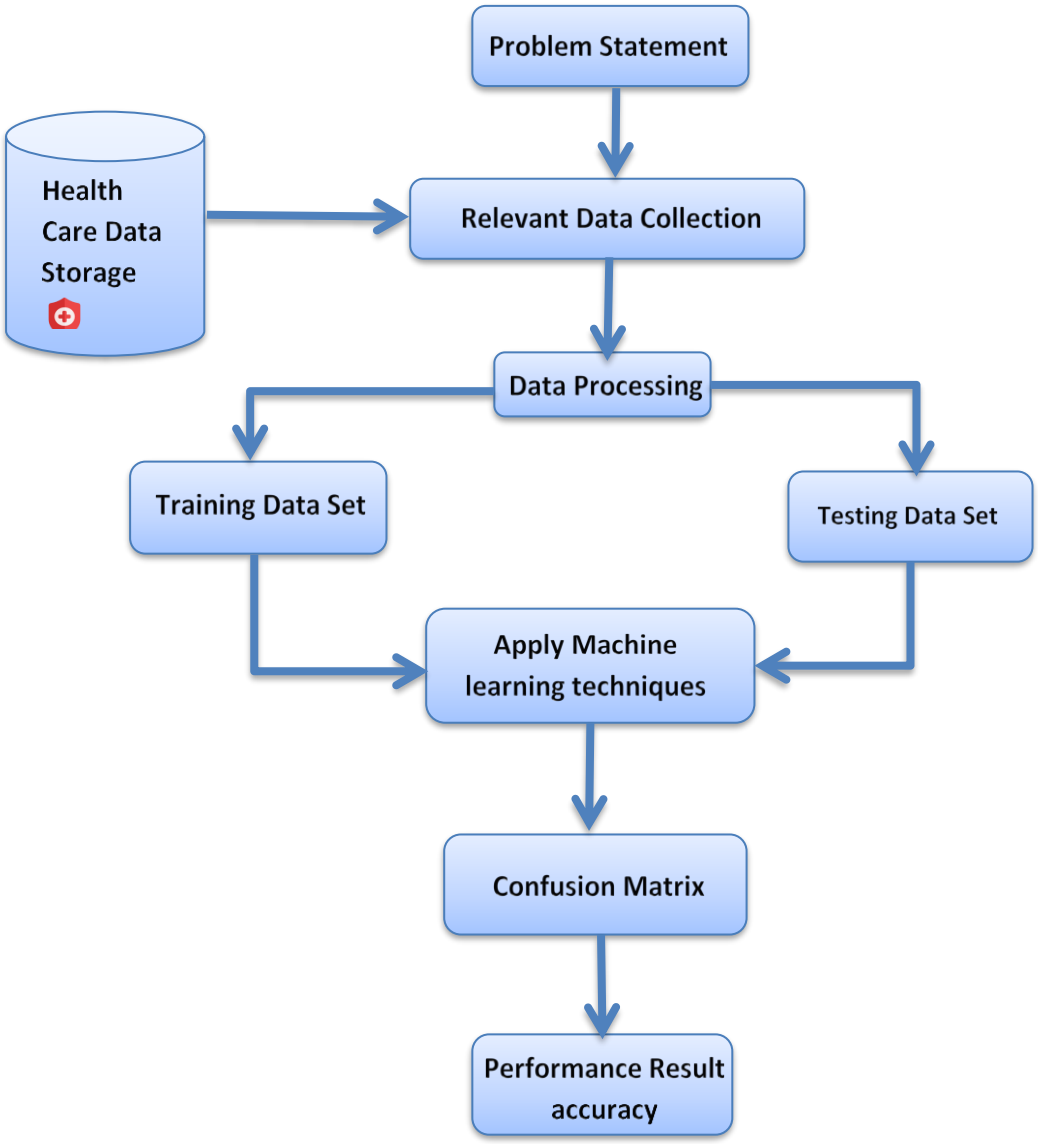


Fig 1.4 An overview of the overall process

Chapter 2

Background study

2.1 Introduction

In recent years, medical science has made numerous advancements. Many diseases can now be identified in their early stages because to scientific advancements, diligent study, and cutting-edge technologies. Medical science has seen amazing advances that have allowed for the curing of numerous fatal illnesses. In recent years, doctors have discovered a number of new facets in the realm of diabetes. Numerous studies have been written about the elements that contribute to the development of diabetes. Moreover, a great deal of study has been done to determine the effects of diabetes on health and the kinds of issues it might lead to. After testing a predetermined group of patients, doctors discovered that diabetes can eventually cause cardiovascular, retinopathy, and nephropathy. In artificial intelligence, the field referred to as machine learning has produced prediction models for a variety of applications, including stock market trading, weather forecasting, traffic patterns, and the identification of suitable habitats. These days, the beginning of many diseases is predicted using an embedded pipeline of machine learning-based algorithms. Numerous articles explain how scientists are using predictive algorithms with adequate accuracy in the cases of diseases like cancer, mental health issues, and cardiovascular disorders. There are numerous machine learning algorithms that can determine whether a person has diabetes or is at risk of developing the disease. Nevertheless, there are relatively few models that can forecast when diabetes-related health issues would manifest. One such model used data from electronic health records and a data mining pipeline to predict complications associated to type 2 diabetes [2]. In order to simplify the complexity of the model, the researchers employed a whole dataset and only a few features. Another noteworthy study by Cho (2008) describes a model that predicted diabetic nephropathy using feature selection and visualization. Almost no other noteworthy research has directly addressed how machine learning can help forecast health issues caused by type 2 diabetes, other from the work listed above. Even so, a model that can precisely forecast when issues related to diabetes would arise can be created. A useful tool that can assist physicians in overcoming their limits is machine learning, which enhances disease detection and prediction. Pregnancy, skin thickness, blood pressure, insulin, age, BMI, glucose, and blood pressure data can all be used to predict diabetes [3].

2.2 Pregnancy

Diabetes Prediction is associated with pregnancy. The likelihood of developing diabetes rises with the length of pregnancy. Pregnant women frequently have either type 1 or type 2 diabetes, and between 6% and 9% of them have gestational diabetes.

2.3 Glucose

Blood sugar in the body is called glucose. Type 1 diabetes is the result of this. You will have more sugar-attached hemoglobin the higher your blood sugar levels. You have diabetes if two different tests show an A1C result of 6.5% or above. The A1C range of 5.7 to 6.4% is indicative of prediabetes. 5.7 is seen as typical. Blood sugar is considered normal if it is less than 140 mg/dl. After two hours, if it is greater than 200 mg/dl, diabetes is suspected. A result that falls between 140 and 199 mg/dl is indicative of prediabetes.

2.4 Blood Pressure

Blood pressure should be kept within a healthy range to prevent the development of conditions including diabetes, hypertension, and hypotension. People with diabetes should have that is less than 130/80 mmHg or less than 140/80 mmHg blood pressure.

2.5 Skin Thickness

Collagen content in skin thickness contributes to an increased risk of insulin-dependent diabetic mellitus, or IDDM. We assessed the skin thickness of 66 individuals with IDDM, ages 24 to 38, and looked into any potential correlations between the occurrence of certain diabetes problems and long-term glycemic management [4].

2.6 Insulin

Insulin is connected with blood glucose. A chart is given below

Blood Glucose (mg/dl)	Insulin units
61-150	0
151-200	3
201-250	5
251-300	8
301-350	10
351-400	12
>400	15[a]

2.7 BMI

BMI is calculated by dividing a person's weight in kilograms by their height in meters squared. Type 2 diabetes is considerably more likely to strike those who are overweight (BMI of 25–29.9), obese (BMI of 30-39.9), or extremely obese (BMI of 40 or above). The more weight you gain, the more resistant your muscle and tissue cells are to the insulin hormone.

2.8 Age

Diabetes comes in two varieties: type 1 and type 2. Middle-aged people continue to have the highest risk of type 2 diabetes and older persons. In 2021, there were approximately 2 million

new diagnoses of diabetes in adults, according to the CDC's National Diabetes Statistics Report Trusted Source.

A chart is given below:

AGE	NEW CASES
18-44	455000
45-64	1009000
65 to older	466000

Chapter 3

Algorithm Learning

3.1 Introduction of Machine Learning Algorithm:

The development of algorithms that facilitate computer Machine learning's goal is learning. Learning is the process of discovering data patterns or other statistical regularities; awareness is not necessarily involved. For learning tasks, Numerous algorithms for machine learning will therefore hardly resemble their human counterparts. However, learning algorithms can reveal how difficult learning is in different contexts.

I used two types of machine learning algorithm in this work.

- Supervised learning algorithm
    - Logistic Regression
    - Linear Regression
    - KNN
    - SVM
    - Decision Tree
    - Random Forest
    - Naive Bayes
- Unsupervised learning algorithm
    - Hierarchical Clustering
    - K-Means Clustering

3.2 Supervised Machine Learning Algorithm

The concept of a teacher or supervisor whose main duty is to provide the agent with an accurate assessment of its inaccuracy (one that is directly compared with output values) defines a supervised scenario. In actual algorithms, this function is provided by a training set that consists of two parts: input and anticipated outputs.

In supervised learning, the probability of inputs is frequently left unspecified. frequently fails to define the likelihood given inputs. Although this is not necessary provided that the inputs are accessible, if any of the input values are missing, it is difficult to make any inferences about the outputs. The learning algorithm's objective in the classification issue is to reduce the error with regard to the inputs provided. This is a crucial point to note. Inputs that are frequently referred to as the samples that the agent looks at in order to learn are known as the "training set". However, it's not always the greatest course of action to fully understand the training set. If I were to instruct you exclusively-or, for instance, and only allowed you to see blends of one true and one false, never both true and false, you may discover that the answer is true. The same is true for machine learning algorithms, which frequently overfit the data and effectively learn the training set by heart rather than a more general categorization technique [5].

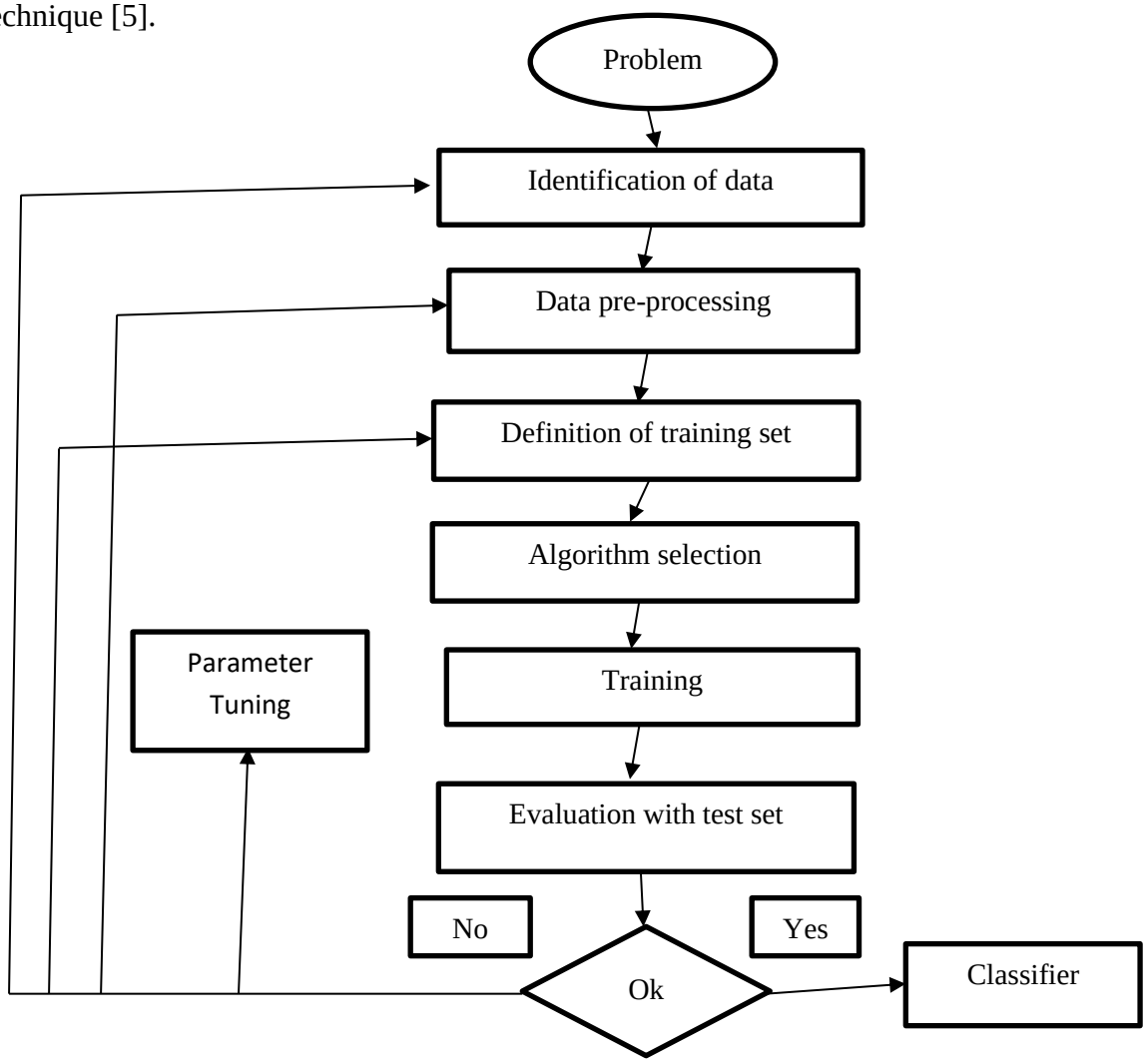


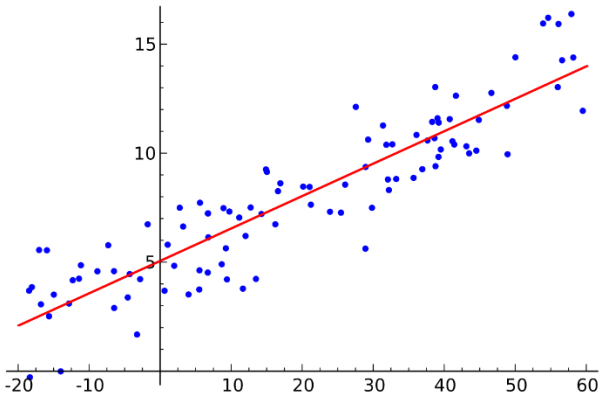
Fig 3.2. Machine Learning Supervise Process

3.2.1 Linear Regression

Regression is the process of using independent predictors to simulate a desired value. Predicting and identifying the cause-and-effect correlations between variables are the primary uses of this methodology. Regression process changes are often determined by the amount of independent variables and the nature of the relationship between them and the dependent variables [6].

Three different application types make use of regression analysis:

1. Determining how input factors affect target variables.
2. Determining how the target variable has changed in relation to another variable or variables.
3. To learn about emerging trends.



One independent variable and a linear relationship between the independent (x) and dependent (y) variables are features of a regression analysis type called simple linear regression. In the graph above, the red line represents x and y, the independent and dependent variables. The best-matched straight line is the red line above the graph. We try to create a line that most accurately depicts the given data points. The line contains the dependent (y) and independent (x) variables. Based on the linear equation below, the red line in the graph above is celled.

$$y = a\_0 + a\_1 * x$$

Finding the ideal values for a_0 and a_1 is the aim of the linear regression procedure.

Cost Function

It is made easy by the cost function to ascertain the ideal values for a_0 and a_1 in order to produce the data points' best fit line. Finding the ideal values for a_0 and a_1 and minimizing the error between the actual and anticipated values turns this search problem into a minimization problem.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Minimization and Cost Function

The discrepancy between expected result and the base truth is known as the discrepancy in error. The sum of every piece of information and the square of the error difference are divided by the total amount of information. Thus, every data point's average squared error is displayed. This cost function is hence alternatively referred to as the Mean Squared Error (MSE) function. Now, we'll use this MSE function to modify the a_0 and a_1 variables so that MSE value reaches the minimum.

Gradient Descent

Gradient descent is a method for reducing the cost function that entails updating a_0 and a_1. The idea is to start with a set of values for both a_0 and a_1, followed by systematically modify these values in order to minimize the cost. Gradient descent can help us modify the value.

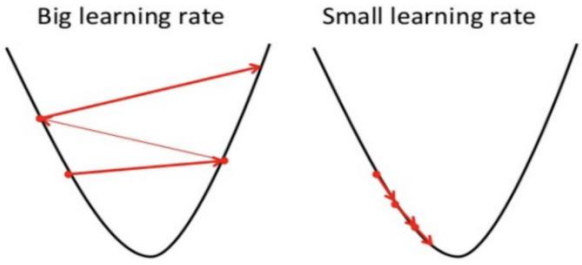


Fig 3.2.1.1 Gradient Descent

The number of actions you take determines the gradient descent algorithm's learning rate. This establishes the speed at which the algorithm comes together to the minima.

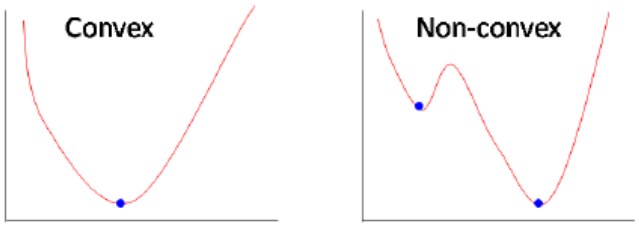


Fig 3.2.1.2 convex vs non-convex function

For linear regression, the cost function is always convex, but occasionally it may be non-convex, allowing you to settle at local minima.

3.2.2 Logistic Regression

Observations are categorized into many groups using a process known as logistic regression. It is a forecasting analytical method it is grounded in the concept of likelihood.

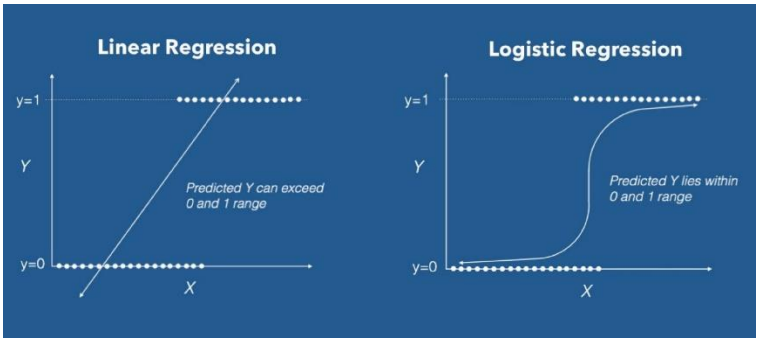


Fig 3.2.2.1 Graph Image Comparing Linear and Logistic Regression

Although it use a more intricate cost function—sometimes known as the "logistic function" or "sigmoid function"—rather than a linear function, a logistic regression model is also called a linear regression model.

The logistic regression hypothesis states that it tends to restrict The cost function ranges from 0 to 1. Linear functions cannot accurately describe it since it might possess a value that is either more than 1 or less than 0, which is impossible. according to the logistic regression hypothesis [7].

$$0 \leq h_\theta(x) \leq 1$$

Logistic regression hypothesis expectation

Sigmoid Function

To convert expected principles into likelihoods and the sigmoid function is employed. The function transforms every real integer from 0 to 1 into a different value. Sigmoid converts predictions into probabilities in machine learning.

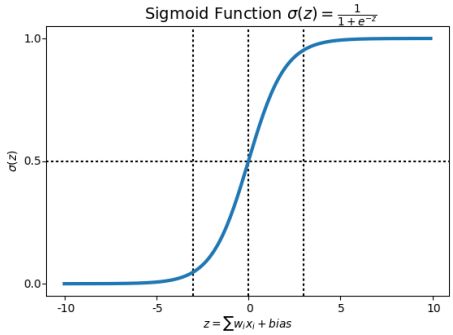


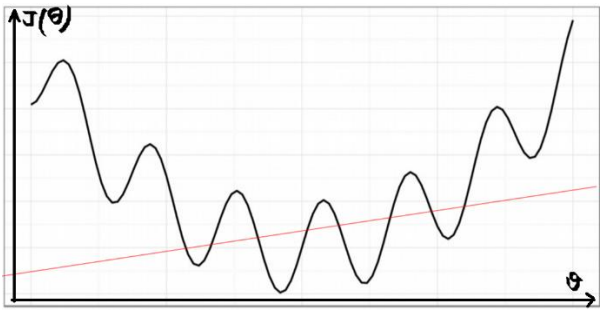
Fig 3.2.2.2 Sigmoid Function Graph

Cost Function

A cost function J (θ) was introduced to us in linear regression. As an optimization objective, the cost function is determined and minimized in order to provide an accurate model with the least amount of inaccuracy.

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2.$$

Finding the global minimum and minimizing the cost value in "Logistic Regression" would be quite challenging," since the cost function of linear regression would be Multiple local minimums in a non-convex function, rendering it meaningless.

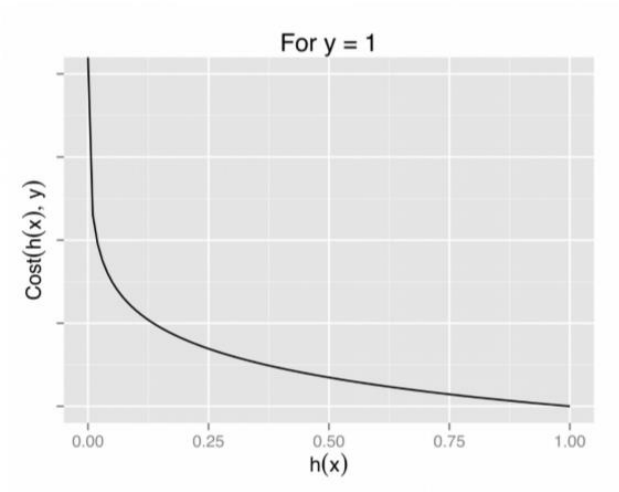


The cost function for logistic regression is described as:

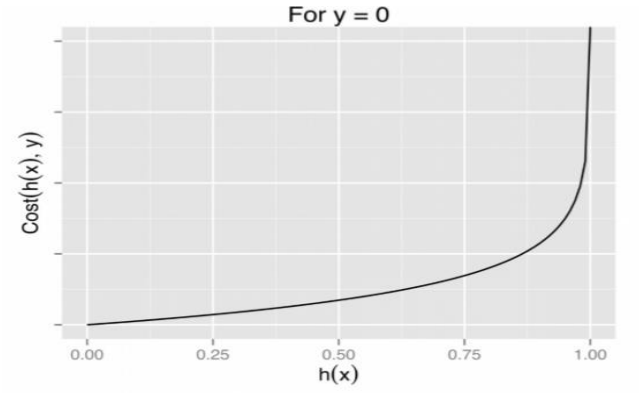
$-log\ (h\theta(x))\ if\ y = 1$

$-log\ (1-h\theta(x))\ if\ y = 0$

$$Cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & \text{if } y = 1 \\ -log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



Graph of logistic regression



Graph of logistic regression

One function may be created by compressing the two functions mentioned above, i.e.

$$J(\theta) = -\frac{1}{m}\sum\left[y^{(i)}\log(h\theta(x(i))) + \left(1 - y^{(i)}\right)\log(1 - h\theta(x(i)))\right]$$

3.2.3 Decision Tree Algorithm

The objective of a decision tree, in contrast to other supervised learning algorithms, is to learn basic choice rules using training data in order to generate a training model to predict the class or value of the target variable. Issues with regression and classification may also be addressed using this approach.

Using decision trees, the class label of a record is predicted from the tree's root. The root attribute's values and the record's attribute values are compared. To go on to the next node, we compare the two and follow the branch that matches that value.

Terminology related to Decision Trees

1. Root Node: It further splits it into two or more groups and reflects the full population or sample homogenous groupings.
2. Splitting: Splitting is the process of dividing a node into two or more smaller nodes.
3. Decision Node: A sub-node that splits into more sub-nodes is known as the Decision node.
4. Leaf/Terminal Node: Nodes that do not split are known as terminal or leaf nodes.
5. Pruning: Pruning is getting rid of a decision node's sub-nodes. One may argue that dividing is the reverse procedure.
6. Branches and Sub-Trees: A branch or sub-tree is a segment of the main tree.
7. Parent and Child Node: Sub-nodes are a parent node's offspring, and a node that is divided into sub-nodes is referred to as a parent node of sub-nodes.

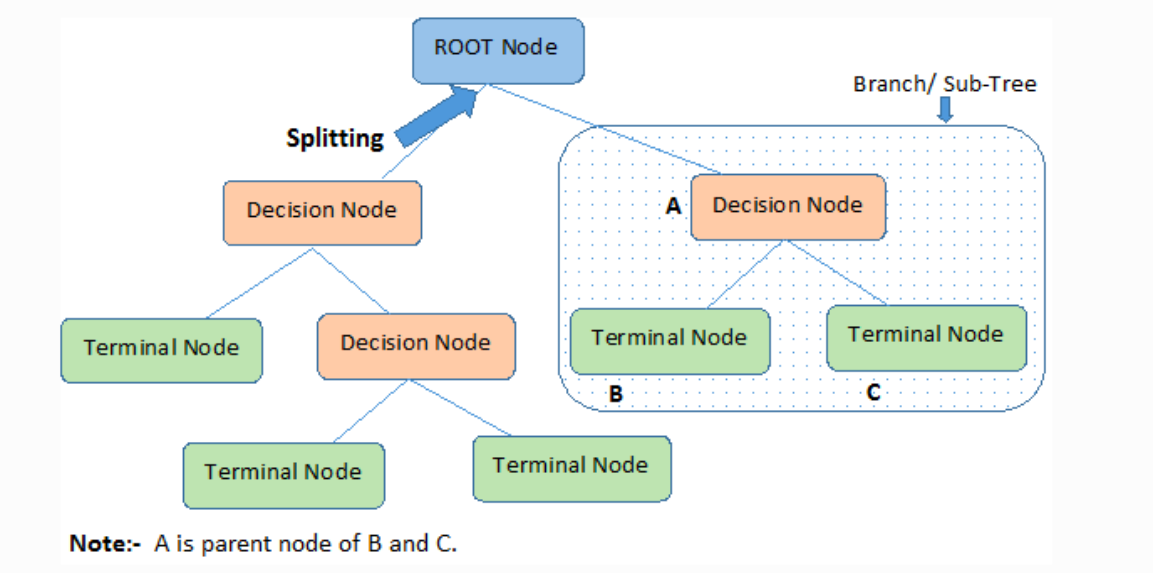


Fig 3.2.3 Decision Tree Algorithm

The classification of the cases is determined sorted from the root to a leaf or terminal node in decision trees.

An example of a test case for an attribute is represented by each node in the tree, and a possible response to the test case is represented by each edge that descends from the node. This recursive process is carried out for each subtree rooted at the new node.

3.2.4 Random Forest Algorithm

Random forests are supervised machine learning algorithms constructed from decision tree methods. It is a method of machine learning for fixing problems with regression and classification. It employs ensemble learning, a method for solving complicated problems that involves several classifiers.

Decision trees make up a random forest algorithm. The random forest algorithm's "forest" is trained via bootstrap aggregation or bagging. Machine learning algorithms are made more precise by an ensemble meta-algorithm known as bagging.

The (random forest) technique the result according to the forecasts made by the decision trees. By figuring out the mean or average of the outcomes from many trees, it generates predictions. The outcome will be more accurate the more trees there are.

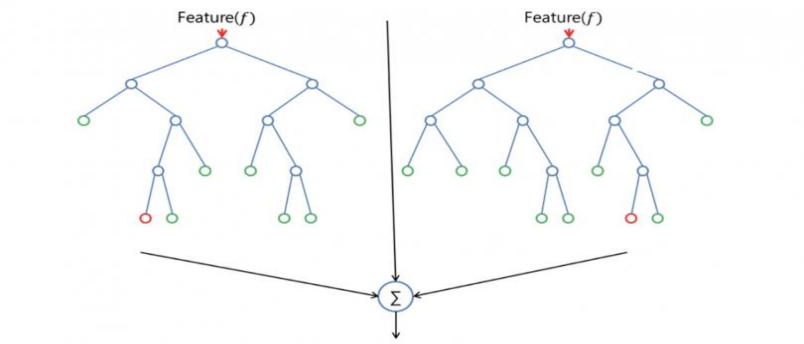
A random forest algorithm eliminates the disadvantages of a decision tree method. It lessens the tendency for the dataset and increases accuracy. Without requiring several package setups, it generates forecasts.

Features of a Random Forest Algorithm

- Compared to the decision tree approach, it is more accurate and give a workable alternative for handling missing data.
- To get a believable forecast, hyper-parameter tweaking is not necessary.
- It fixes the overfitting issue with decision trees.
- At the node's splitting point, a subset of attributes is chosen at random for each random forest tree.

How do Random Forest Works

In essence, random forest builds many decision trees and merges them. One of random forest's primary benefits is that it can use for both classification and regression tasks, which include the majority of machine learning systems nowadays. Given that classification is sometimes seen as the foundation of machine learning, let's examine random forests in classification. Two trees in a random forest hown below; these trees work together to produce a prediction that is more reliable and accurate.



Hyperparameters for a random forest are nearly the same as Those of a bagging classifier or decision tree. Thankfully, the random forest classifier-class may be utilized without integrating a bagging classifier and a decision tree. By employing the algorithm's regressor, random forest may also be used to address regression jobs.

The model's unpredictability is increased by random forest when the trees are being developed.

Instead, then partitioning a node based on the most important characteristic, it seeks for the great feature from a random selection of features.

There is therefore a lot of variety, which frequently produces a superior model. As a result, In a random forest, only a random portion of the characteristics is taken into account when splitting a node. By adding random thresholds for each characteristic, you may further improve the unpredictability of trees rather of searching for the ideal thresholds, as a normal decision tree does.

3.2.5 K-Nearest Neighbor (KNN) Algorithm

For regression and classification issues, K-Nearest Neighbors (KNN) is among the most basic machine learning techniques.

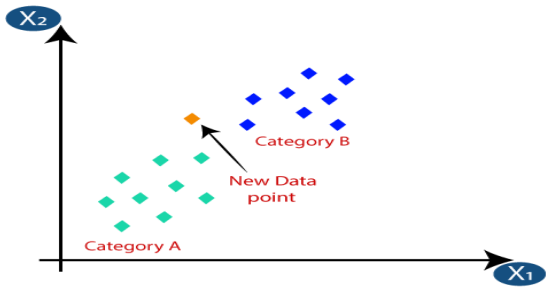
The KNN algorithm utilizes data to categorize new data points using similarity criteria.

How K-NN Works

The following algorithm steps may be used to describe how the K-NN operates:

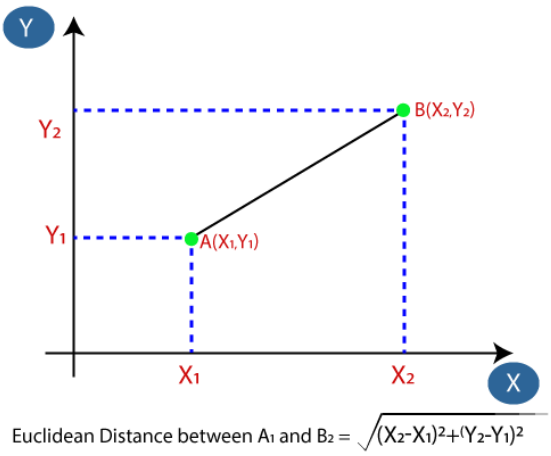
1. Determine the neighbors' K number.
2. Determine the K neighbors' Euclidean distance from one another.
3. Determine the K nearest neighbors using the computed Euclidean distance.
4: Find the proportion of data items from these k neighbors that fall into each category.
5. Assign the newly added data points to the category with the most neighbors.
6: We have our model ready.

Assume that I have to allocate a new data point to the relevant category. Go look at this image:

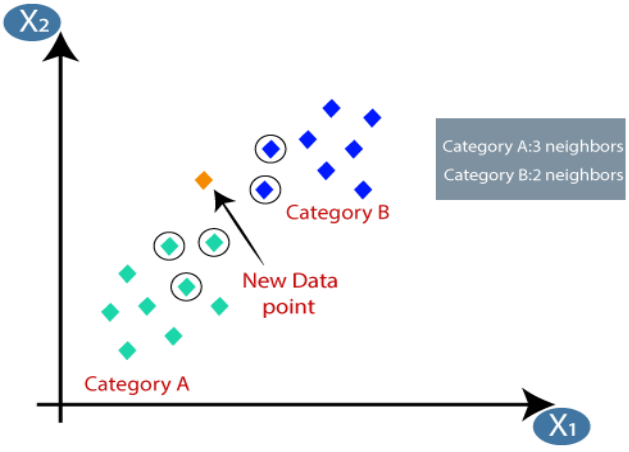


First, I'll decide how many neighbors I have, so I'll go with k=5

The Euclidean distance between each data point will then be determined. The distance between two points, which I have already covered in geometry class, is known as the Euclidean distance. It may be calculated in this way:



By calculating the Euclidean distance, I was able to identify the nearest neighbors, who were two in category B and three in category A. Look at the image below:



from category A, hence this new piece of information the three closest neighbors, as far as I can tell, must fall under group A.

3.2.6 Support Vector Machine Algorithm

The goal of the support vector machine approach is to locate a hyperplane that clearly categorizes the data points in an N-dimensional space, where N is the number of attributes.

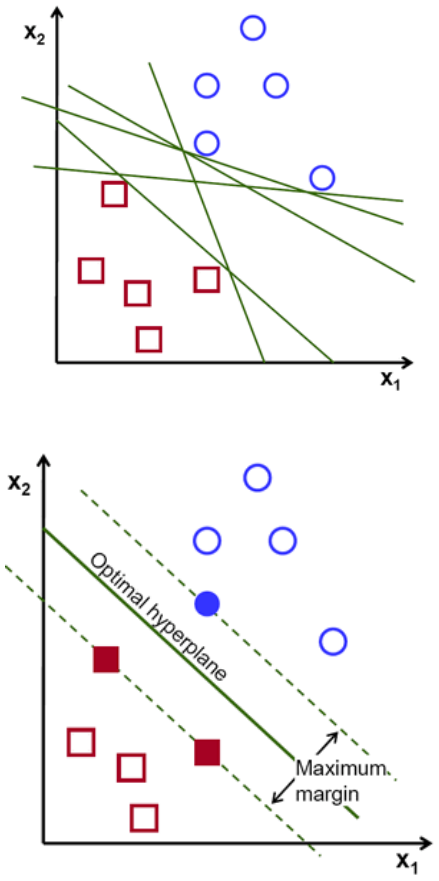


Fig 3.2.6.1 possible hyperplanes

There are several hyperplanes from which to choose in order to divide the two types of data. Finding the plane with the largest margin—that is, the greatest separation—between the two groups' data points is our aim. By providing some reinforcement, boosting the margin distance allows future data points to be categorized more confidently.

Support Vectors and hyperplanes

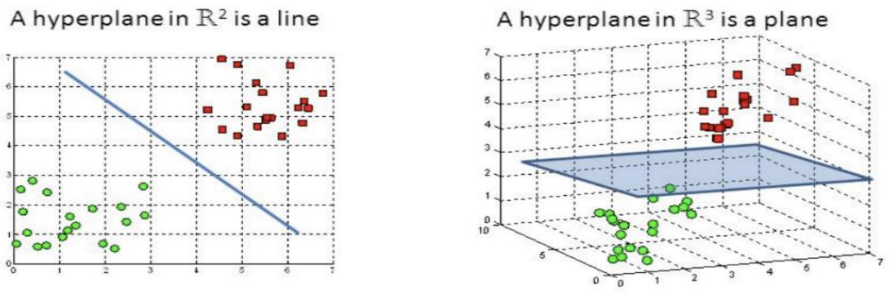
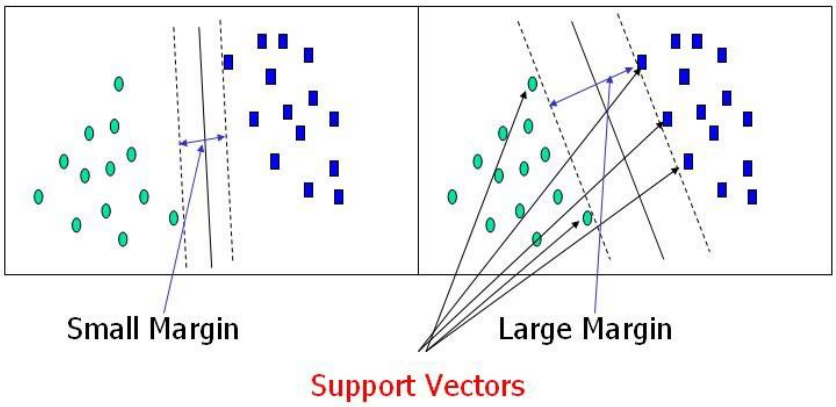


Fig 3.2.6.2 2D and 3D feature space hyperplanes

Decision boundaries, or hyperplanes, can be used to classify data points. It is possible to classify data points that are located on each side of the hyperplane differently. The number of features has an impact on the hyperplane's size as well. A line is all that the hyperplane is if there are two input characteristics. Three input qualities cause the hyperplane to become a two-dimensional plane. More characteristics than three make it more difficult to visualize.



Nearer data points that influence the hyperplane's location and orientation are known as support vectors. By employing these support vectors, we optimize the margin of the classifier. If the support vectors are eliminated, the hyperplane's position will shift. These elements support the development of our SVM [8].

Cost Function and Gradient Updates

Optimizing the distance between the data points and the hyperplane is the aim of the SVM approach. The hinge loss The loss function aids in margin maximization.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

Loss function (function on left can be represented as a function on the right)

There is no expense if the expected and actual numbers belong to the same sign. The loss value is then determined if they are not. Additionally, we include a regularization parameter in the cost function. Finding a balance between loss and margin maximization is the goal of the parameter for regularization. The cost functions look like this once the regularization argument is added.

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

Loss function for SVM

In relation to the weights, we take partial derivatives after getting the loss function in order to locate the gradients. We may use the gradients to update our weights.

$$\frac{\delta}{\delta w_k} \lambda \parallel w \parallel^2 \quad = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} \left(1 - y_i \langle x_i, w \rangle\right)_+ = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

The gradient from the regularization parameter only has to be updated when there is no misclassification, that is, when our model accurately predicts the class of our data point.

$$w = w - \alpha \cdot (2\lambda w)$$

Gradient Update — No misclassification

In the event of a misclassification—that is, we do a gradient update by include the regularization parameter and the loss when our model incorrectly predicts the class of our data point.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Gradient Update — misclassification

3.2.7 Naive Bayes Algorithm

Even when dealing with millions of records of data, it is suggested to utilize the Naive Bayes machine learning model since it performs well with large data sets. It achieves outstanding performance in NLP tasks such as emotional analysis. This technique of classification is quick and easy to use.

Based on Bayes' Theorem, naive classifiers a subset of classification algorithms. Is not a method, but rather a family of algorithms based on the same idea: each pair of qualities being classified is unrelated to the others.

The Bayes theorem must be understood before we can comprehend the naïve Bayes classifier.

Bayes Theorem

Theoretically, conditional probability provides the basis for this. Conditional probability is the likelihood of an event occurring provided that another event has already occurred. Based on prior information, we may use the conditional probability to calculate the likelihood of an event.

Conditional probability:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Where,

P(A): The likelihood that hypothesis H is correct. We call this the prior probability.

P(B): The evidence's likelihood.

P(A|B): The likelihood that the hypothesis is correct in light of the available data.

P(B|A): The likelihood that the hypothesis is correct in light of the available data.

Naive Bayes Classifier

• The Bayes theorem provides the basis for this type of classifier.

• Membership probabilities, like the probability that data points fall into a particular class, are predicted for each class.

• The best appropriate class is determined to be the one with the highest probability.

• Maximum a Posteriori (MAP) is another name for this.

• For a hypothesis, the MAP is:

- $MAP\ (H) = \max P((H|E))$
- $MAP\ (H) = \max P((H|E) * (P(H))\ /P(E))$
- $MAP\ (H) = \max(P(E|H) * P(H))$

The evidence probability, represented by $P\ (E)$, is used to normalize the outcome. Eliminating (E) will not change the outcome.

• The existence or lack of a variable has no bearing on the existence or lack of any other variable. NB classifiers determine that none of the variables are related to one another or features.

3.3 Unsupervised Machine Learning Algorithm

To evaluate and categorize unlabeled data, unsupervised machine learning applies machine learning algorithms. Without the assistance of a human, these algorithms reveal hidden patterns or data groupings. For client segmentation, cross-selling tactics, exploratory data analysis, and picture identification, it is the perfect tool because to its capacity to identify similarities and contrasts in data.

Since the objective of unsupervised learning is to teach the computer to perform an action without our guidance, it seems much more challenging! There are two methods for approaching unsupervised learning. The initial approach involves teaching the agent by employing some kind of reward mechanism to show achievement instead of by giving clear classifications. It should be noted that since the objective of this kind of training is to make decisions that maximize rewards rather than to generate a classification, it will typically fit within the decision problem framework [9].

3.3.1 K-Means Clustering Algorithm

K-means clustering, a well-known illustration of an exclusive clustering algorithm, divides data points into K groups, where K is the quantity of clusters based on each group's separation from the centroid. The facts that are near a certain centroid will be grouped together in the same category. A higher K value will imply smaller groups with more resolution, whereas a lower K number will indicate broader groupings with less granularity. K-means clustering's fundamental phase is straightforward. We start by figuring out how many clusters there are (K) and assuming that these clusters are cantered. Another option is to utilize the first K items in a series as the starting center, or we can choose any random object. The next three phases will then be followed by the K means algorithm until it converges.

Repeat until the situation is stable (i.e., no object moves group):

1. Find the coordinates of the centre.
2. Calculate each object's distance from the centre.
3. Sort the items according to their minimum distance.

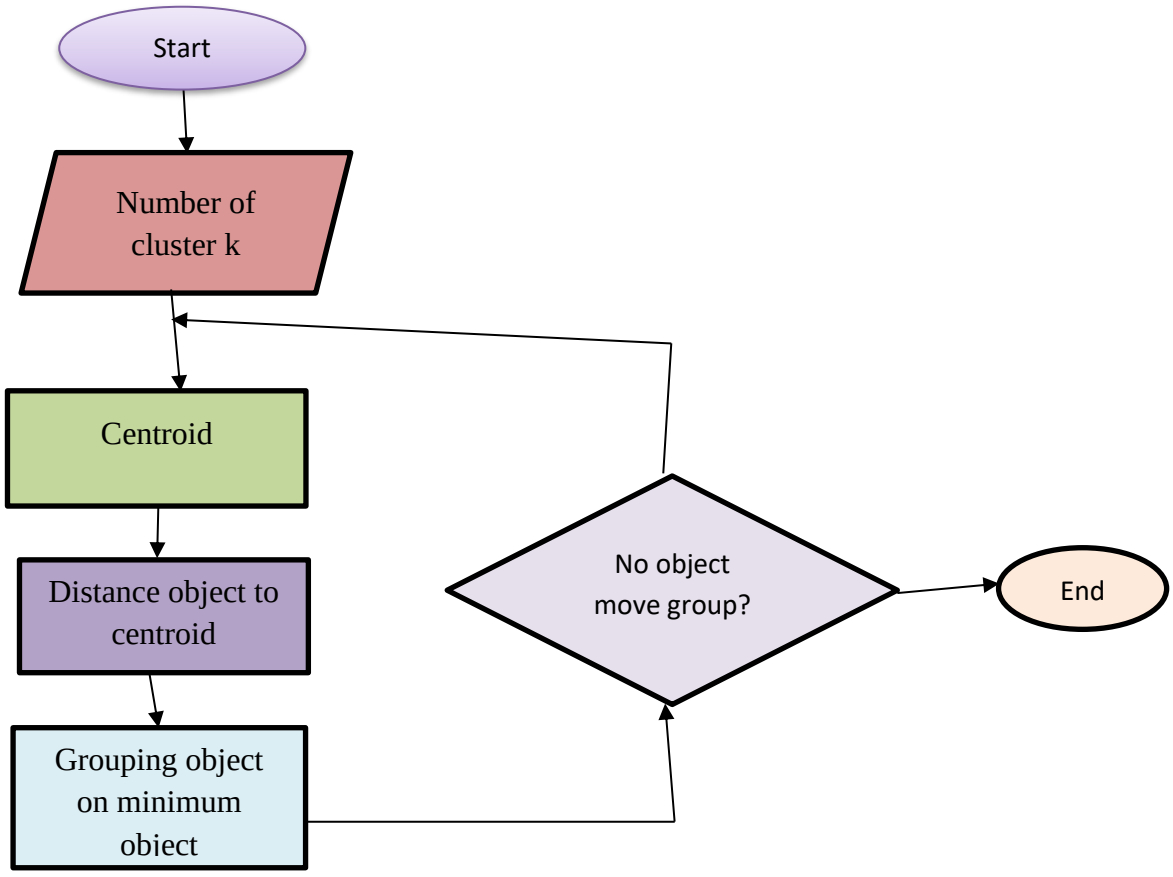


Fig 3.3.1 K-means iteration

The squared error function, an objective function supplied by, is what this method aims to reduce:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

Where,

'$||x_i - v_j||$' is the distance measured in Euclides between x_i and v_j.

'c_i' the quantity of information points in i^{th} cluster.

'c' the quantity of cluster centers.

Algorithmic steps for k-means clustering

Let V = {v1,v2,…….,vc} be the set of centers and X = {x1, x2, x3,….., xn } be the set of data points.

1) Choose "c" cluster centers at random.

2) Calculate the separation between each data point and the cluster centers.

3) Out of all the cluster centers, the one closest to the data point should be allocated to it.

4) Make a new cluster center calculation using the formula below:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, *'c_i'* stands for the number of data points in i^{th} cluster.

5) Recalculate the distances between each data point and the recently identified cluster centers.

6) If no data points were redistributed, stop; if not, return to step 3.

3.3.2 Hierarchical Clustering Algorithm

Alternatively known as hierarchical cluster analysis, hierarchical clustering, is a technique that creates clusters—groups of related objects. A collection of clusters makes up the endpoint; each cluster is unique, but the items that are part of each cluster are generally comparable.

Hierarchical Clustering is of two types.

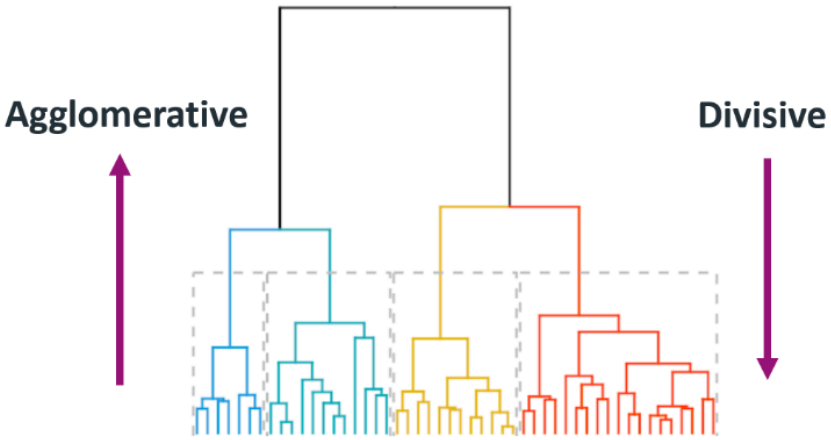
1. Divisive

2. Agglomerative Hierarchical Clustering

Divisive Hierarchical Clustering is another term for top-down clustering. By using this technique, a single cluster is assigned to all of the data or observations. Each collection of data or observation is given its own cluster, which is then further subdivided.

Sometimes called a bottom-up method, agglomerative hierarchical clustering, treats each observation or data item as a separate cluster.

Until every cluster is combined into a single, sizable cluster that contains all of the data, two clusters are connected.



Both algorithms are exactly the opposite of each other

Chapter 4

Implementation and Performance Analysis

4.1 Introduction

In the application area of diabetes prediction, machine learning approaches or algorithms are yielding positive outcomes. Among the world's biggest causes of death is diabetes. The need for efficient learning algorithms to help medical professionals assess diabetes illness stems from the accessibility of a large amount of medical data. The prediction of diabetes is greatly aided by machine learning algorithms, which seem to be capable of doing so. Nevertheless, the process of observing knowledge and helpful patterns is adversely affected by these redundant and noisy data. Machine learning techniques have drawn a lot of attention from analysts who want to transform this data into useful knowledge.

It is feasible to obtain significant data in advance from large records by using highlight determination techniques. Supervised and unsupervised machine learning methods are used in our paper, such as feature extraction and the following: K-nearest neighbor, Random Forest, Decision Tree, Naive Bayes, Support Vector, and Logistic Regression using linear models, K-means cluster, and Hierarchical cluster. The accuracy of each algorithm allows these predictive models to be compared. Here, a comparison is made between a number of different approaches based on diabetes condition, including detection strategies (data collection strategy, handling missing data strategy, data cleaning strategy, data cluster strategy, relationship strategy, classifier strategy). In this case, artificial neural networks were used to estimate the algorithms' performance. According to the discussion, feature extraction techniques improve machine learning algorithms' ability to accurately predict and diagnose diabetes [10].

4.2 Implementation of Datasets

Numerous machine learning methods may be used to predict diabetes. It can be challenging to select the finest prediction method based on these features, though. Therefore, in order to

predict diabetic diseases, we use nine renowned algorithms for machine learning on population data for the study: Linear regression, Decision tree, K-nearest neighbor, Random Forest, Support Vector, Logistic, Naive Bayes, K-means cluster, and Hierarchical cluster.

Consider the following steps in order to accomplish our goal: gathering diabetes data sets with key patient attributes, pre-processing the numerical esteem attributes, applying different machine learning classification strategies, and conducting corresponding predictive research using the data. We look at these stages in brief in the following.

4.2.1 Overview of Data set and Attributes

We gathered the data set for this project from Kaggle. The data collection includes 768 people' different diabetic illness characteristics.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig. 4.2.1 Data set head

4.2.2 Data Pre-processing

The diabetic data set has undergone a number of data processing procedures to enable to meet the objectives of our study. Therefore, we translate the numerical values of the attributes into nominal values. As an illustration, we dealt with the missing values, plotted null value removal, and aimed for input null values for the columns before visualizing a heat map for clean data.

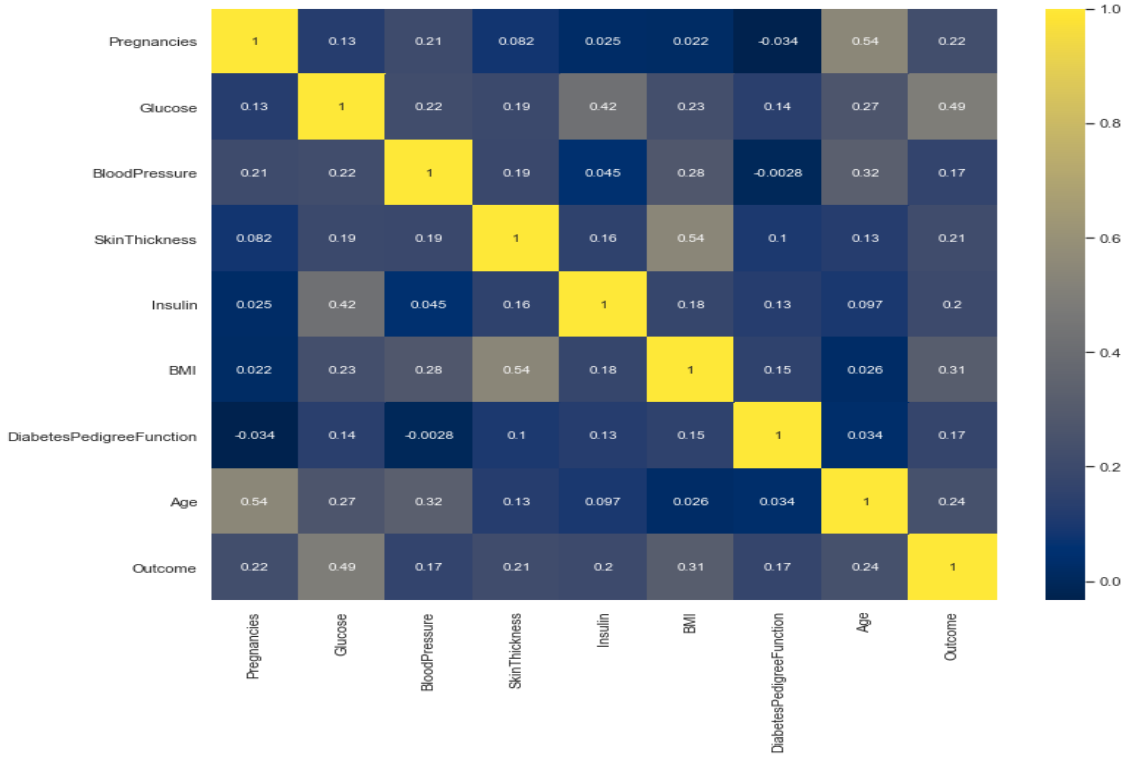


Fig 4.2.2 Data cleaning Heat map

4.3 Implementation of Machine Learning Algorithms on Data Set

When the data set is prepared for modeling, we use nine well-known supervised and unsupervised machine learning methods to forecast diabetes in order to get the highest performance accuracy possible.

4.3.1 Implementation of Supervised learning Algorithm

Supervised learning is the process of creating an algorithm that maps an input to a particular output and is memorized. Using the labeled datasets we recently gathered, this can be achieved. If the mapping is accurate, the algorithm has trained successfully. In any other case, we modify the algorithm in a crucial way to enable accurate learning.

1. Decision Tree Algorithm

This approach is intended for machine learning under supervision. Trees are the foundation of the algorithm. In essence, it uses individual attributes from our data set and has the ability to question its own judgment. In order for the algorithm to run smoothly, The data set is initially separated into training and testing sets, as well as some features. Next, we import GridsearchCV to apply the fit and score approach to the model. The decision tree technique was selected for our study in order to assess the diabetic data's performance analysis.

Best Score ==> (75.10%)

Tuned Parameters ==> {'criterion': 'gini', 'max_depth': 4, 'min_samples_leaf': 1}

Accuracy on Train set ==> (81.71%)

Accuracy on Test set ==> (72.44%)

Decision tree best accuracy: 75.10%

2. KNN algorithm

A straightforward nonparametric categorization is K-nearest neighbor and relapse calculation method. The system categorizes new attributes according to their degree of similarity and logs all significant qualities. It employs a tree-like information structure to identify the point of interest to concentrate on when compiling the data collection. The neighbors are used to categorize the quality. The value of k in a classification process is consistently a positive number of the closest neighbor. From a list of lesson or protest property values, the nearest neighbors are selected.

In our study, we achieved 100% results when K=1 by applying the KNN classification algorithm to the training set. However, when we apply the method to the test dataset, we obtain an accuracy of 73.62% for K=12.Thus, we obtain a comparable prediction accuracy for diabetes. The performance is then obtained using the Confusion matrix.

KNN accuracy: 73.62%

3. Random Forest algorithm

This paper's goal is to develop a framework that uses the machine learning method's Arbitrary Timberland computation to more precisely calculate a patient's early-stage

probability of developing diabetes. In addition to being frequently used for both classification and regression tasks, irregular Woodland computations may also be a type of outfit learning technique. Compared to other calculations, the level of precision is more notable. The best results for diabetic prediction are obtained using the proposed model, and the findings showed that the prediction system can anticipate the beginning of diabetes with high accuracy, speed, and significance.

RFC Accuracy: 74.41%

4. Support Vector Machine Algorithm

The Support Vector Machine is a popular supervised machine-learning approach that is used for issues with categorization and pattern recognition. By building a hyperplane in several dimensions that optimizes the edge between data clusters to split two groups as efficiently as possible, the SVM algorithm does classification.
To adhere to the SVM model's the need of normalizing variable values to values between -1 and +1, the input data set's first column, which represents case status of 0 was assigned.

SVM Accuracy: 74.02%

5. Logistic Regression algorithm

We employed this supervised learning method to forecast a categorical dependent variable. Thus, after preparing the data set, we use the training data set to fit the model. Then, to make sure our model is functioning, we utilize a test data set to forecast diabetes.
Then, in order to accomplish the primary objective, we employed XGBClassifier to improve the liner model and confusion matrix in order to obtain the diabetes prediction accuracy.

Logistic Regression accuracy: 74.80%

6. Naive Bayes algorithm

Naive Bayes may be a well-liked probabilistic classification method. Through frequency counting and combining the values provided in the data set, the method generates probabilistic results. By applying the Bayesian hypothesis, It recognizes that every attribute depends on the values of class-variables and is independent. The conditional autonomy

Assumptions seldom come to pass in practice and results in more complicated and superior classifier output. In our study, we import GaussianNB to obtain live model parameter changes. We employed a confusion matrix to obtain the greatest accuracy result after fitting the model using training and testing data sets.

Naive Bayes accuracy: 73.63%

7. Linear regression algorithm

Linear regression is among the easiest machine learning techniques. This is a static model that uses a linear equation to try and illustrate the relationship between two variables. But this algorithm cannot help us reach our objective. This model's accuracy is appalling.

Linear regression accuracy: 39.60%

4.3.2 Implementation of Unsupervised Learning

The datasets are automatically grouped by clustering according to their shared characteristics. Unusual data points in our databases can be found through anomaly detection. It is helpful in identifying fraudulent transactions. Groups of objects that frequently appear together in our collection are found by association mining. Preprocessing data is a common application for latent variable models. such as breaking down a dataset into several components or lowering the number of features in the dataset.

1. K-means cluster

An iterative clustering process called the K-means cluster algorithm aids in determining the maximum value for each iteration. In order to group the data set in our research, we chose the appropriate number of clustering. A collection of labels is the algorithm's output. After fitting the prediction model to the datasets, we translate the clusters into Boolean. We use a confusion matrix after conversion to obtain the intended accuracy result.

K-means f1-score accuracy: 79%

2. Hierarchical cluster

Another kind of unsupervised machine learning cluster algorithm is this one. We import the model from Sklearn and fit it using prediction data to obtain an exact accuracy. We then run the cluster and obtain the label data. Next, we contrast the dataset and label data. The accuracy is then obtained.

Hierarchical Clustering Accuracy: 32.42%

4.4 Performance analysis of Algorithms on diabetes datasets

Accuracy is the criterion we use to compare algorithms. In our study, we used the confusion matrix to calculate these requirements. Below is a general picture of the confusion matrix.

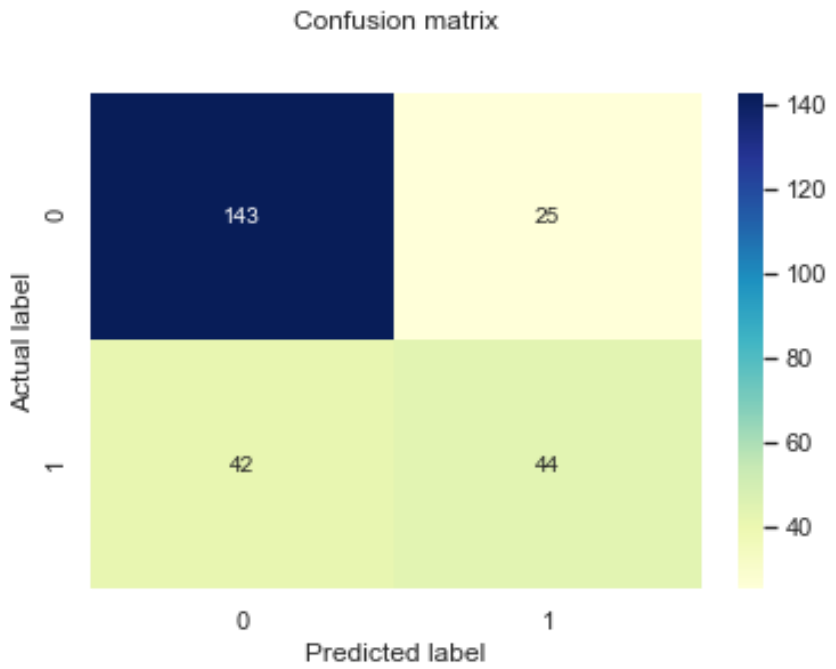


Fig: 4.3 Confusion Matrix

In a confusion matrix, the actual label is the label found in the data set, while the predicted label is the label that the machine learning algorithms predict.

The True Positive (TP) indicator shows how many records were accurately detected.

The True Negative (TN) indicates the quantity of valid entries that are properly categorized.

Inaccurate classification of the records is indicated by a False Negative (FN).
Records that are mistakenly categorized as positive are known as False Positives (FP).

The accuracy is computed using the confusion matrix as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Nine machine learning methods for diabetic illness prediction are suggested in our study. After training and testing the algorithms, the outcomes are documented. Here, the optimum algorithm for prediction is determined by comparison.

Serial No.	Machine-Learning Algorithms	Accuracy
1	**Decision Tree**	**75.10%**
2	**KNN algorithm**	**73.62%**
3	**Random Forest algorithm**	**74.41%**
4	**Support Vector Machine Algorithm**	**74.02%**
5	**Logistic Regression**	**74.80%**
6	**Naive Bayes**	**73.63%**
7	**Linear regression**	**39.60%**
8	**K-means cluster**	**79%**
9	**Hierarchical cluster**	**32.42%**

Table 4.3.1: Accuracy of All Proposed Algorithms

From figure 4.3.1 Compared to the other eight algorithms, we discover that the K-means method produces more accurate results. The table also shows that the Random Forest algorithm predicts with 74.41% accuracy, the Decision Tree algorithm predicts with 75.10% accuracy, the Logistic Regression algorithm predicts with 74.80% accuracy, and KNN achieves better accuracy with 73.62% When the k value is raised, accuracy increases with a k value of 3.

Where, the performance of linear regression and Hierarchical cluster algorithm are very poor.

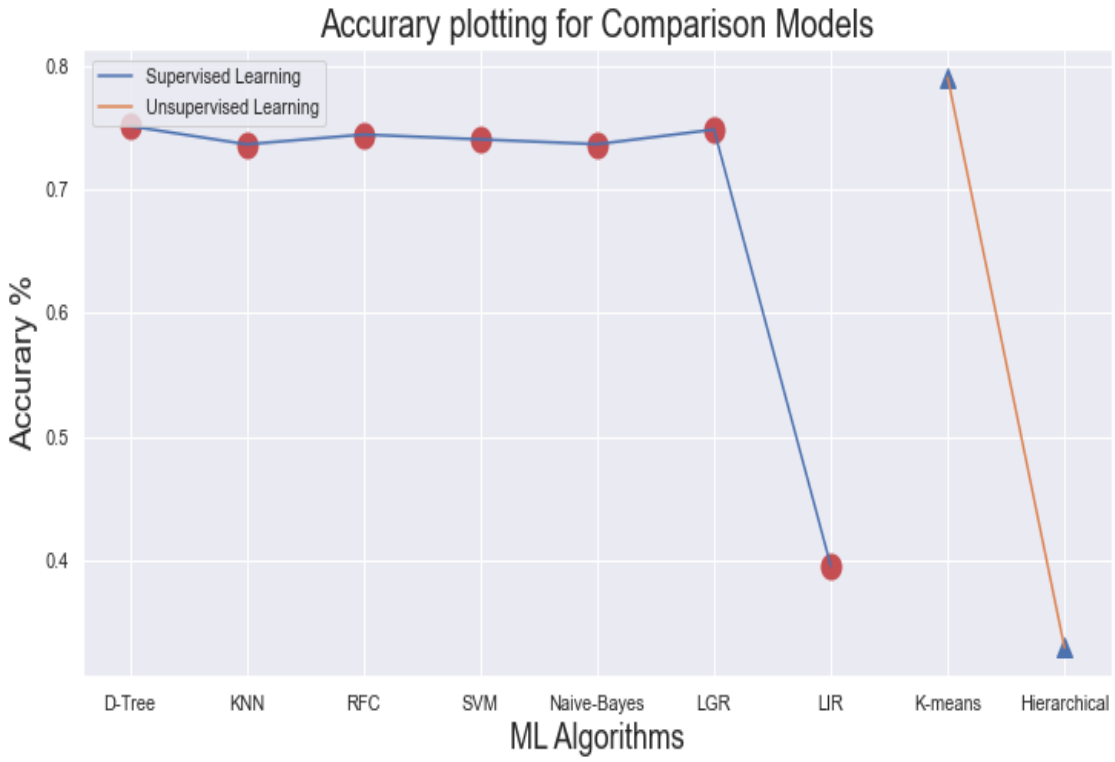


Fig 4.3.2 Accuracy comparison Plotting

This indicates that our datasets will yield positive outcomes for our issues once they have been cleaned. The optimal algorithm for diabetes prediction can be chosen using the results of this study.

Chapter 5

Results and Discussion

5.1 Results

When various categorization algorithms were applied to our data set, the outcomes for each strategy varied marginally. The outcomes were assessed based on accuracy, allowing us to observe how well each algorithm performed in making predictions. The algorithms' performances were compared in order to predict the result. K-means algorithms outperformed decision trees among supervised learning algorithms while decision trees outperformed k-means algorithms among unsupervised algorithms in the provided table.

Algorithms	**Accuracy**
Supervised Machine-Learning	
Decision Tree	75.10%
KNN algorithm	73.62%
Random Forest algorithm	74.41%
Support Vector Machine Algorithm	74.02%
Logistic Regression	74.80%
Naive Bayes	73.63%
Linear regression	39.60%
Unsupervised Machine-Learning	
K-means cluster	79%
Hierarchical cluster	32.42%

Table 5.1: Supervised and Unsupervised Accuracy Table

We can more accurately determine whether a patient has diabetes thanks to the table. The suggested models' accuracy has been compared. The accuracy of the decision tree approach was 75.10%, that of the logistic regression method was 74.80%, and that of the K-means clustering method was 79%.

5.2 Discussion

The study's findings demonstrate how well machine learning algorithms predict diabetes and provide information on their real-world uses. With an accuracy of 75.10%, the Decision Tree algorithm proved to be the most successful supervised learning technique in this investigation. Its effectiveness in managing both numerical and categorical data is responsible for its performance.

With an accuracy of 39.60%, linear regression was the least successful supervised technique. This outcome emphasizes how inappropriate it is for binary classification issues such as diabetes prediction. Regression applications with a continuous target variable are best suited for linear regression. Binary categorization (diabetes vs. no diabetes) is the aim of this investigation. The sensitivity of linear regression to outliers is quite great. Extreme values in the dataset have the potential to significantly impact the model and lower its forecast accuracy.

Conversely, with an accuracy of 32.42%, Hierarchical Clustering performed the least. It's possible that this method's poor performance was caused by its computing complexity and susceptibility to outliers. Furthermore, because of its high memory needs, it is less appropriate for huge datasets.

Chapter 6

Conclusion and Future work

6.1 Conclusion

In order to diagnose diseases, machine learning and data mining approaches are useful. The K-Means approach yields the maximum accuracy of 79 percent when applied to the datasets in this research study. Several machine learning techniques are implemented based on accuracy for medical diagnosis of diabetic patients, and classification is done using various algorithms. Thanks to sophisticated computational techniques and the abundance of genetic and epidemiological data sets for diabetes risk, machine learning holds enormous promise to transform the prediction of diabetes risk. Early detection is essential for effective diabetes treatment. This research described a strategy that uses machine learning to determine if a patient has diabetes or not. This study is limited by the fact that a structured dataset was used.

6.2 Future Work and Scope

• Using deep learning models, such neural networks, might increase accuracy by identifying intricate links in the data.
• Adding more features and using sophisticated feature selection techniques might improve the model's performance.
• Testing the models' robustness and generalizability on datasets from other sources would be beneficial.

The experimental findings may help medical professionals prevent diabetes early and make wiser treatment choices to manage the illness and preserve lives. Algorithms for supervised and unsupervised learning can assist in forecasting fresh, unseen data that we acquire in the future.

Reference

[1] Karthikeyani V**,**Parvin Begum I**,**Shahina Begam I Comparison of Diabetes Disease Prediction Using Data Mining Classification Algorithms (CDMCA)**,** International Journal of Computer Applications, 2012; **60**:26-31.

[2] Olaiya F. Comparative Analysis of the Effectiveness of Various Data Mining Techniques in Medical Database Knowledge Discovery, *International Journal of Advanced Research in Computer Science and Software Engineering*, 2013; **3**:11-15.

[3] Elsevier B.V. Performance Analysis of Classifier Models to Predict Diabetes Mellitus.J. Pradeep Kandhasamy*, S. Balamurali. Department of Computer Applications, Kalasalingam University, Krishnankoil -626126, Tamilnadu, India.Procedia Computer Science 47 (2015) 45 – 51.

[4] Md.Faisal Faruque,Asaduzzaman,Iqbal H. sarkaer. 2019. Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus.ResearchGate.(February. 2019)

[5] Edureka!(2020). Supervised machine learning [Online]. Available: https://www.edureka.co/blog/supervised-learning/

[6] Jia Z, Zhou Y, Liu X, Wang Y, Zhao X, Wang Y, Liang W, Wu S. Comparison of Different Anthropometric Measures as Predictors of Diabetes Incidence in a Chinese Population. [accessed June 25, 2021]

[7] Towards data science. (2020) Diabetes prediction using logistic regression with TensorFlow [Online].Available:https://towardsdatascience.com/diabetes-prediction-using-logistic-regression-with-tensorflow-js-35371e47c49d .

[8] V., A. K. and R., C. 2013. Diabetes Disease Categorization Through Support Vector Machine. International Journal of Engineering Research and Applications. 3, (April. 2013), 1797-1801.

[9] guru99 (2021). Unsupervised machine learning [Online]. Available: https://www.guru99.com/unsupervised-machine-learning.html

[10] Morteza, M., Franklyn, S., Linying, D., Karim, K. and Aziz G. 2015. Assessing the Framingham Diabetes Risk Scoring Model's Effectiveness in Canadian Electronic Health Records. Canadian journal of diabetes 39, 30(April. 2015), 152-156.