A Project Report on

Diabetes Prediction System
Using Machine Learning

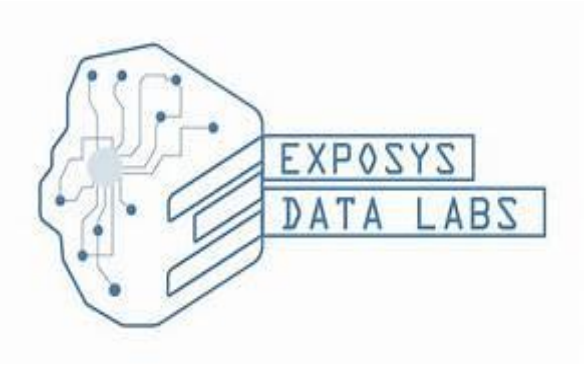
For

Data Science
Internship Project

By

V.Harshita Sai

At



CONTENTS

ABSTRACT:

Diabetes has become a common disease to the mankind from young to the old Persons nowadays. There are various reasons due to which the population of diabetic patients is increasing day by day such as obesity, bad diet, auto immune reaction, change in lifestyle, eating habits, environmental pollution etc . Hence, early prediction of diabetes is very essential to save the human life from diabetes. Data analytics is one of the branches of computer science ,which is a process of examining the large datasets and find some useful hidden patterns and draw conclusion based upon those patterns .This analytical process is carried out using machine learning algorithms in health care system.To carry out medical diagnoses,machine learning algorithms are used for analysing large medical data to build the machine learning models. This project presents a diabetes prediction system to diagnosis of diabetes.Early detection of diabetes is possible with the help of this model. More over deep learning techniques which have a significance of biomedical effect are also described.

Keywords: Deep Learning, Diabetes, Feature Extraction, Machine Learning.

INTRODUCTION:

According to the International Diabetes Federation (IDF) statistics, there were 415 million people suffering from diabetes around the world. By 2040 this number is expected to rise to over 642 million, as a consequence, diabetes has become the main cause of national disease and death in most countries. Diabetes is a group of metabolic diseases in which a person has high blood glucose, either because the body does not produce enough insulin, or because cells do not respond to the insulin that is produced. If diabetes patients cannot control blood sugar well, it is effortless to induce cardiovascular, nervous system, eye, foot and other systemic diseases. Patients whose conditions are severe can also suffer from diabetic ketoacidosis, with a high disability. Diabetes has a very great deal of harm to the human body, causing a series of complications, affecting the patient physical and mental health, bringing a heavy burden to family and society.

Diabetes can be segmented into three types: type 1 diabetes, type 2 diabetes and gestational diabetes.

- Type 1 diabetes is an autoimmune disease that occurs in childhood. In this type of diabetes, the pancreatic cells that secrete insulin have been destroyed.
- Type 2 diabetes is caused by insulin resistance in various organs, leading to a marked increase in insulin demand, which accounts for almost 90% of the diabetes cases .
- Gestational diabetes tends to occur among pregnant women, as the pancreas does not make sufficient amount of insulin.

The standards of early screening and diagnosis of diabetes are still in the exploratory stage on account of the unclear ethology and pathogenesis of diabetes. Through the continuous understanding of diabetes, the criteria of screening and diagnosis are constantly changing. Early diagnosis of diabetes mainly depends on clinical symptoms and signs. In 1965, the World Health Organization (WHO) first published diabetes diagnostic norm based on the clinical characteristics, but this criteria did not mention the diagnosis threshold of blood sugar levels. With the developing understanding of diabetes, diagnostic criteria gradually increased fasting blood glucose (FPG), oral glucose tolerance test (OGTT), glycosylated haemoglobin (HbA1c) and other physiological parameters. In 1980, the fasting blood glucose level was viewed as the main diagnostic norm. In 1997, the new standard of American Diabetes Association (ADA) increased the OGTT parameters.

Accurate screening and diagnosis of diabetes require more effective features and have a high demand on the judgment which can be closer to the nature of the disease. Some studies found that if we consider metabolic changes in diabetes from the perspective of body metabolism, doctors can better make a diagnosis of the type of diabetes and help patients with the more appropriate diabetic treatment. Metabolomics is a new discipline that has been developed in recent years to analyze all the low molecular weight metabolites of a certain organism or cell qualitatively and quantitatively. Through the change of endogenous metabolites and intermediates in diabetes and the evolution of coping rules, the metabolic status of the body can be further understood. On the basis of the study of early screening and diagnostic criteria for diabetes, diagnostic standards are increased from the initial clinical symptoms and signs to FPG, OGTT, HbA1c and other physiological parameters. Simultaneously clinical and demographic signs are also included in the diagnostic reference, such as sex, age, race/ethnicity, haemoglobin disease/anemia, body mass index (BMI), cardiovascular disease, family history/ Genetic, medication records, etc. However, there is still no way to find out the pathogenesis of diabetes from the field of biology. It is urgent to clarify the pathology and diagnostic criteria of diabetes, it has a great significance in delaying the occurrence and development of diabetes, choosing drugs, reducing the incidence of diabetic complications and extending life expectancy. With the continuous development of artificial intelligence and data mining technology, researchers begin to consider using machine learning techniques to search for the characteristics of diabetes. Machine learning techniques can find implied pathogenic factors in virtue of analyzing and using diabetic data, with a high stability and accuracy in diabetic diagnosis. Therefore, machine learning techniques which can find out the reasonable threshold of risky factors and physiological parameters provide new ideas for screening and diagnosis of diabetes.

CONVENTIONAL MACHINE LEARNING TECHNIQUES :

Diabetic diagnosis is based on a variety of epidemiology and genetic factors. Dangerous factors of epidemiology include smoking status, eating habits, physical activity, BMI and so on. Genetic factors are pathogenic genes which come from parents. Hence, doctors hope to consider all aspects of these factors and then predict and diagnose diabetes accurately, nevertheless researchers from the medical domain found that they could not explain the pathogenesis of diabetes. With the continuous development of Artificial Intelligence, it has been found that machine learning techniques are very suitable for finding the reasonable threshold of risk factors and physiological parameters affecting diabetes. Why machine learning can achieve significant achievement in the medical domain? First of all, diabetes is a kind of chronic disease, and a lot of clinical treatment information will be generated in the process of treatment. Meanwhile, machine learning has giant advantages in handling big data problems, so the machine learning techniques can be applied to the analysis and processing of diabetes data. Secondly machine learning and medical diagnosis have the uniform objective to extract the correct and valuable information from a large number of data for making decisions. At the same time, machine learning techniques can avoid the misdiagnosis of inexperienced or tired human experts, and have a high stability and accuracy in the screening and diagnosis of diabetes. Furthermore, machine learning techniques can also help patients have a clear idea of their health status as well as the situation of diabetic development, and then patients can plan their own lifestyle to slow the deterioration of disease. Therefore, we hope that we can use machine learning techniques to find pathogenesis of diabetes which cannot be found in the medical domain, which has great significance for treatment of diabetes patients early, the appropriate use of medicine and early rehabilitation. In this paper, the applications of conventional machine learning techniques in the early screening and diagnosis of diabetes mellitus will be introduced from two aspects: supervised learning and unsupervised learning.

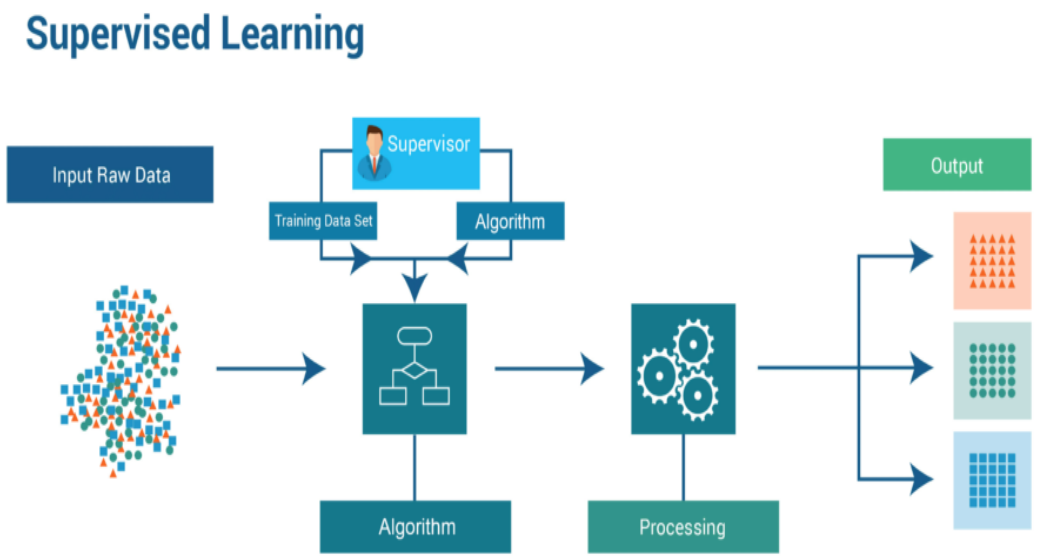
Supervised Machine Learning:

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output

data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y).**In the real-world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering,** etc.

How Supervised Learning Works?

In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.



Working of Supervised Learning

Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.

- If the given shape has four sides, and all the sides are equal, then it will be labelled as a **Square.**
- If the given shape has three sides, then it will be labelled as a **triangle.**
- If the given shape has six equal sides then it will be labelled as **hexagon.**

Now, after training, we test our model using the test set, and the task of the model is to identify the shape.

The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.

Steps involved in supervised machine learning:

- First we need to determine the type of training dataset
- Collect/Gather the labelled training data.
- Split the training dataset into training **dataset, test dataset, and validation dataset.**
- Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
- Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
- Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
- Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

Types of supervised Machine learning Algorithms:

Supervised learning can be further divided into two types of problems:

1.Regression

2.Classification

1. Regression:

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc.

Types of regression algorithms under supervised machine learning.

- Linear Regression.

- Regression Trees.
- Non linear regression.
- Bayesian Linear Regression.
- Polynomial Regression.

2. Classification:

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false etc.

Ex: Spam Filtering,

Types of classification algorithms under supervised machine learning.

- Random Forest Classifier.
- Decision Trees.
- Logistic Regression.
- Support vector Machines.

Advantages of Supervised learning:

- With the help of supervised learning, the model can predict the output on the basis of prior experiences.
- In supervised learning, we can have an exact idea about the classes of objects.
- Supervised learning model helps us to solve various real-world problems such as **fraud detection, spam filtering,** etc.

Disadvantages of supervised learning:

- Supervised learning models are not suitable for handling the complex tasks.
- Supervised learning cannot predict the correct output if the test data is different from the training dataset.
- Training required lots of computation times.
- In supervised learning, we need enough knowledge about the classes of object.

Unsupervised Machine Learning

There may be many cases in which we do not have labeled data and need to find the hidden patterns from the given dataset. So, to solve such types of cases in machine learning, we need unsupervised learning techniques.

What is Unsupervised Learning?

Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. It can be defined as: *Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.*

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to **find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.**

Example: Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups.

Why use Unsupervised Learning?

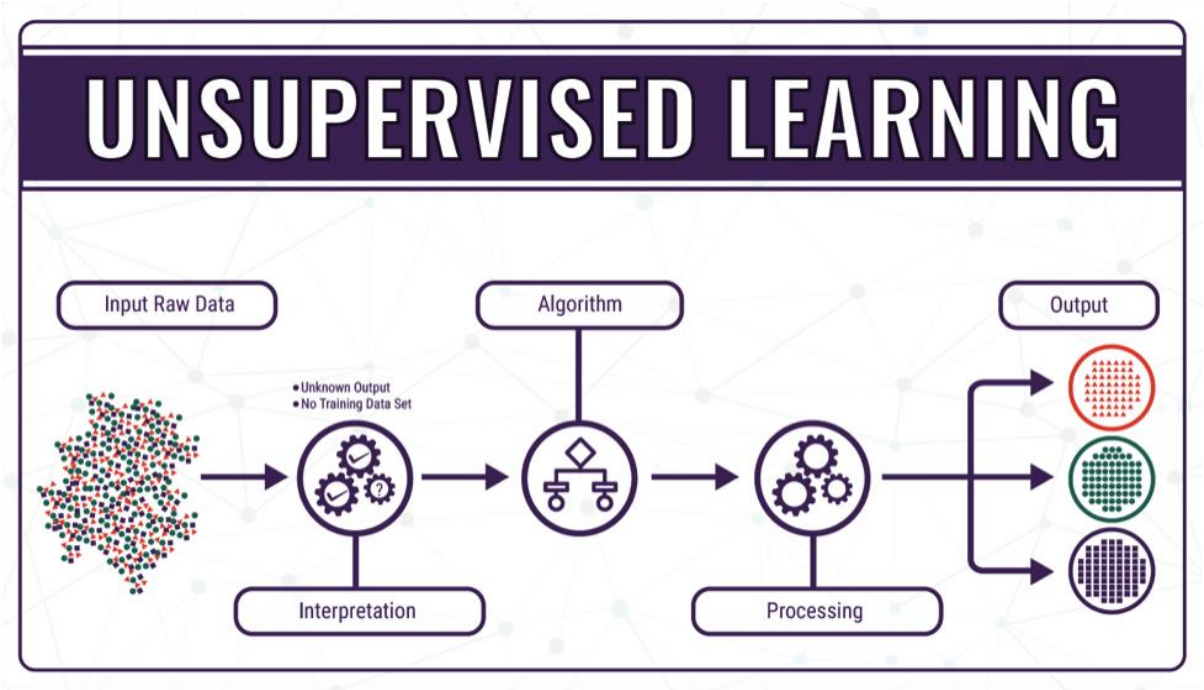
The importance of Unsupervised Learning:

- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.

- In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

How Unsupervised Learning works ?

Working of unsupervised learning can be understood by the below diagram:



Here, we have taken an unlabelled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabeled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.

Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.

Types of Unsupervised Learning Algorithm:

The unsupervised learning algorithm can be further categorized into two types of problems:

- **Clustering**: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with

the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.

- **Association:** An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset.Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical Example: Market Basket Analysis.

Unsupervised Learning algorithms:

Types of Unsupervised learning algorithms:

- **K-means clustering**
- **KNN (k-nearest neighbors)**
- **Hierarchal clustering**
- **Anomaly detection**
- **Neural Networks**
- **Principle Component Analysis**
- **Independent Component Analysis**
- **Apriori algorithm**
- **Singular value decomposition**

Advantages of Unsupervised Learning:

- Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
- Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

Disadvantages of Unsupervised Learning:

- Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.
- The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

PROPOSED METHODOLOGY :

Goal of the project is to investigate for model to predict diabetes with better accuracy, different classification algorithms to predict diabetes.

Dataset Description :

The data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.

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

- The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.
- Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabetes and 268 labeled as 1 means positive means diabetic.

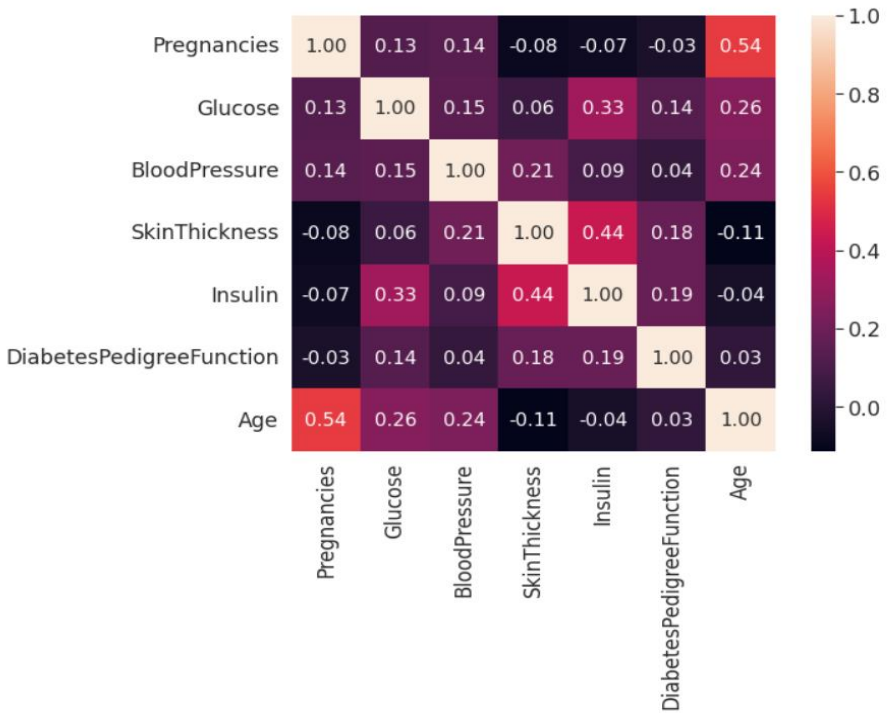
1.Data Preprocessing :

We may wind up drawing an off base surmising about the information, if the missing qualities are not dealt with appropriately. Since all the segments or columns probably won't be helpful for the model or the informational index that is accessible isn't in the structure wherein it tends to be utilized for the preparation of the machine in every one of these cases information pre-handling is a significant factor that decides the sound beginning of the model. Information pre-handling is a procedure which is utilized to turn crude information to valuable organization. Information Pre-handling is one of the significant highlights required for the preparation of the model. Data preprocessing incorporates checking for invalid values on the off chance that these invalid values are supplanted by mean of entire section. In data pre-processing straight out information can be changed into numerical information. label_encoder is object which help us in moving Categorical information into Numerical information.

Relationship shows the quality and course of the straight relationship between two quantitative factors. It takes esteems between - 1 and +1. A positive incentive for r shows a positive affiliation and a negative an incentive for r demonstrates a negative affiliation. The last step in datapre-processing is thesplitting of data into training and testing data.In our ML model we have used cross_validation object from sklearn library train_test_split.

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

correlation table



Correlation Matrix

2. Apply Machine Learning :

When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction.

3. The Techniques are follows :

1.Support Vector Machine :

Support Vector Machine also known as SVM which is a supervised machine learning algorithm. SVM is most popular classification technique. SVM creates a hyper plane that separate two classes. It can create a hyper plane or set of hyper plane in high dimensional space. This hyper plane can be used for classification or regression also. SVM differentiates instances in specific classes and can also classify the entities which are not sup- ported by data. Separation is done through hyper plane performs the separation to the closest training point of any class.

Algorithm:

- Select the hyper plane which divides the class better.
- To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
- If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to
- Select the class which has the high margin.
- Margin = distance to positive point + Distance to negative point.

2.K-Nearest Neighbor :

KNN is also a supervised ma- chine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy prediction technique.KNN assumes that similar things are near to each other. Many times data points which are similar are very near to each other.KNN helps to group new work based on similarity measure.KNN algorithm record all the records and classify them according to their similarity measure. For finding the distance between the points uses tree like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the train- ing data set its nearest neighbors.

Here K= Number of nearby neighbors, its always a positive integer. Neighbor's value is chosen from set of class. Closeness is mainly defined in terms of Euclidean distance.

The Euclidean distance between two points P and Q i.e. P (p1,p2, . ,pn) and Q (q1, q2,..qn) is defined by the following equation:-

Algorithm:

- Take a sample dataset of columns and rows named as Pima Indian Diabetes data set.
- Take a test dataset of attributes and rows.
- Find the Euclidean distance by the help of formula.
- Then, Decide a random value of K is the no. of nearest neighbors

- Then with the help of these minimum distance and Euclidean distance find out the nth column of each.
- Find out the same output values.
- If the values are same, then the patient is diabetic, other- wise not.

3.Decision Tree :

Decision tree is a basic classification method. It is supervised learning method. Decision tree is used when response variable is categorical. Decision tree has tree like structure based model which describes classification process based on input feature. Input variables are any types like graph, text, discrete, continuous etc.

Steps for Decision Tree are as follows:

Algorithm

- Construct tree with nodes as input feature.
- Select feature to predict the output from input fea- ture whose information gain is highest.
- The highest information gain is calculated for each attribute in each node of tree.
- Repeat step 2 to form a sub tree using the feature which is not used in above node.

4.Logistic Regression :

Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories. It classifies the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes. Main aim of logistic regression is to

best fit which is responsible for describing the relationship between target and predictor variable. Logistic regression is a based on Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class.

Sigmoid function P = 1/1+e – (a+bx).

Here P = probability.

b = parameter of Model.

Ensembling:

Ensembling is a machine learning technique Ensemble means using multiple learning algorithms to- gather for some task. It provides better prediction than any other individual model that's why it is used. The main cause of error is noise bias and variance, ensemble methods help to reduce or minimize these errors. There are two popular ensemble methods such as Bagging, Boosting, ada-boosting, Gradient boosting, voting, averaging etc. Here In these work we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

5.Random Forest :

It is type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is grater then compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest improve performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

Algorithm

- The first step is to select the R features from the total features m where R<<M.
- Among the R features, the node using the best split point.
- Split the node into sub nodes using the best split.
- Repeat a to c steps until l number of nodes has been reached.
- Built forest by repeating steps a to d for a number of times to create n number of trees.

The random forest finds the best split using the Gin-Index Cost Function which is given by:

The first step is to need the take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and stores the anticipated outcome at intervals the target place.

Secondly, calculate the votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula.

MODEL BUILDING :

This is most important phase which includes model build- ing for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology:

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. K- Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned ma- chine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned ma- chine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

Step8: After analyzing based on various measures conclude the best performing algorithm.

CONCLUSION :

- The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully.

- The proposed approach uses various classification and ensemble learning method in which SVM, KNN, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used.

- The Experimental results can be assist health care to predict and make early decision to cure diabetes and save humans life.