

DIABETES DETECTION IN WOMEN USING MACHINE LEARNING

**A MINI PROJECT
REPORT**

Submitted by

**SHRUTI RAJ VANSI SINGH
ENTRY No. 17BCS049
SANYA NEGI
ENTRY No. 17BCS046**

**BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE**



SHRI MATA VAISHNO DEVI UNIVERSITY

NOVEMBER 2019

SHRI MATA VAISHNO DEVI UNIVERSITY

CERTIFICATE

Certified that this project report **“DIABETES DETECTION IN WOMEN USING MACHINE LEARNING”** is the work of **“SHRUTI RAJ VANSI SINGH AND SANYA NEGI”** who carried out the mini project work under my supervision.

MR. SANJAY SHARMA

Assistant Professor
School of Computer Science

Submitted to the Viva voce Examination held on 29 November 2019

**INTERNAL EXAMINER
EXAMINER**

EXTERNAL

ACKNOWLEDGEMENT

I would like to thank my Project Mentor, Mr. Sanjay Sharma under whose guidance the successful completion of this project has been possible. The project could not have been implemented without constant inputs and encouragement from him. I would also like to thank all the faculty of Computer Science Department for their continuous evaluation of our projects and constant support. At the end my class mates who were always available for help and support whenever needed.

TABLE OF CONTENTS

Chapters Number	Content	Page Number
	Abstract	iii
	List of Tables	iv
	List of Figures	v
 1.	 INTRODUCTION	
	1.1 WHAT IS DIABETES?	
	1.1.1 Diabetes in Women	
	1.2 MACHINE LEARNING AND PREDECTIVE MODELLING	
	1.2.1 Supervised Machine Learning	
	1.2.2 Unsupervised Machine Learning	
	1.2.3 Reinforcement Learning	
	1.3 DIABETES AND MACHINE LEARNING	
 2.	 MATERIALS AND METHODS	
	2.1 DATA	
	2.2.2 Microorganism and culture condition	
	2.2.3 Effect of carbon sources	
	2.2.4 Effect of nitrogen sources	
	2.2.5 Effect of minerals	
	2.2.6	
	2..27	

2.3 CLASSIFICATION

2.2.1 Logistic Regression

2.2.2 K Nearest Neighbors

2.2.3 Linear SVM

2.2.4 RBF SVM

2.2.5 Naive Baye's Classifier

2.2.6 Decision Tree Classifier

2..27 Random Forest Classifier

2.3 Ensembling

2.4 Feature Extraction/Selection

2.5 Standardization

2.6 Graphical User Interface

4. RESULTS AND DISCUSSIONS

5. CONCLUSION

6. REFERENCES

ABSTRACT

This analysis aims to observe which features are most helpful in predicting if a woman is diabetic and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the a women is suffering from diabetics or not, based on certain diagnostic measurements included in the data set. This data set is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. It contains 9 attributes. To achieve this we have used Machine Learning classification methods to fit a function that can predict the discrete class of new input AND MAKING A HYBRID MODEL. In order to verity the universal applicability of the methods, we chose some methods that have the better performance to conduct independent test experiments. The results showed that prediction with random forest could reach the highest accuracy ($ACC = 0.81$) when all the attributes were used.

LIST OF TABLES

1. Statistical report of Pima Indian Dataset.
2. Accuracy of the classification algorithms used.
3. Accuracy of the classification algorithms used before Feature Extraction and Standardization.
4. Accuracy of the classification algorithms used after Feature Extraction and Standardization.

LIST OF FIGURES

1. Essential Learning process to develop a predictive model.
2. Graphical representation
3. Splitting of Data
4. Phases
5. Logistic regression
6. k Nearest Neighbour
7. Representation of Support vector machine.
8. Radial basis function (RBF) kernel Support Vector Machine.
9. Gaussian Naive Baye's
10. Decision tree for an XOR operation involving three operands
11. Result of Feature Extraction
12. Graphical User Interface
13. Result of the Graphical User Interface for the details entered by user

1. INTRODUCTION

1.1 WHAT IS DIABETES?

Diabetes is a common chronic disease and poses a great threat to human health. The characteristic of diabetes is that the blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects, or both .

Over time, high blood glucose leads to problems such as

- heart disease
- stroke
- kidney disease
- eye problems
- dental disease
- nerve damage
- foot problems

1.1.1 DIABETES IN WOMEN

Between 1971 and 2000, the death rate for men with diabetes fell, according to a study in *Annals of Internal Medicine*. This decrease reflects advances in diabetes treatment.

But the study also indicates the death rate for women with diabetes didn't improve. In addition, the difference in death rates between women who had diabetes and those who didn't more than doubled.

The death rate was higher among women, but there has been a shift in sex distribution of type 2 diabetes showing higher rates in men.

The findings emphasize how diabetes affects women and men differently. The reasons included the following:

- Women often receive less aggressive treatment for cardiovascular risk factors and conditions related to diabetes.
- Some of the complications of diabetes in women are more difficult to diagnose.
- Women often have different kinds of heart disease than men.
- Hormones and inflammation act differently in women.

The most current reported statsTrusted Source from 2015 found that in the United States 11.7 million women and 11.3 million men were diagnosed with diabetes.

Global reports from 2014 by the World Health Organization state that there were an estimated 422 million adults living with diabetes, up from 108 million reported in 1980.

Diabetes affects women and men in almost equal numbers. However, diabetes affects women differently than men.

Compared with men with diabetes, women with diabetes have:[¹²](#)

- A higher risk for heart disease. Heart disease is the most common complication of diabetes.
- Lower survival rates and a poorer quality of life after heart attack
- A higher risk for blindness
- A higher risk for depression. Depression, which affects twice as many women as men, also raises the risk for diabetes in women.

1.2 MACHINE LEARNING AND PREDICTIVE MODELLING

Machine Learning is concerned with the development of algorithms and techniques that allows the computers to learn and gain intelligence based on the past experience. It is a branch of Artificial Intelligence (AI) and is closely related to statistics. By learning it means that the system is able to identify and understand the input data, so that it can make decisions and predictions based on it.

The learning process starts with the gathering of data by different means, from various resources. Then the next step is to prepare the data, that is pre-process it in order to fix the data related issues and to reduce the dimensionality of the space by removing the irrelevant data (or selecting the data of interest). Since the amount of data that is being used for learning is large, it is difficult for the system to make decisions, so algorithms are designed using some logic, probability, statistics, control theory etc. to analyze the data and retrieve the knowledge from the past experiences. Next step is testing the model to calculate the accuracy and performance of the system. And finally optimization of the system, *i.e.* improvising the model by using new rules or data set. The techniques of machine learning are used for classification, prediction and pattern recognition. Machine learning can be applied in various areas like: search engine, web page ranking, email filtering, face tagging and recognizing, related advertisements, character recognition, gaming, robotics, disease prediction and traffic management. The essential learning process to develop a predictive model is given in Fig below.

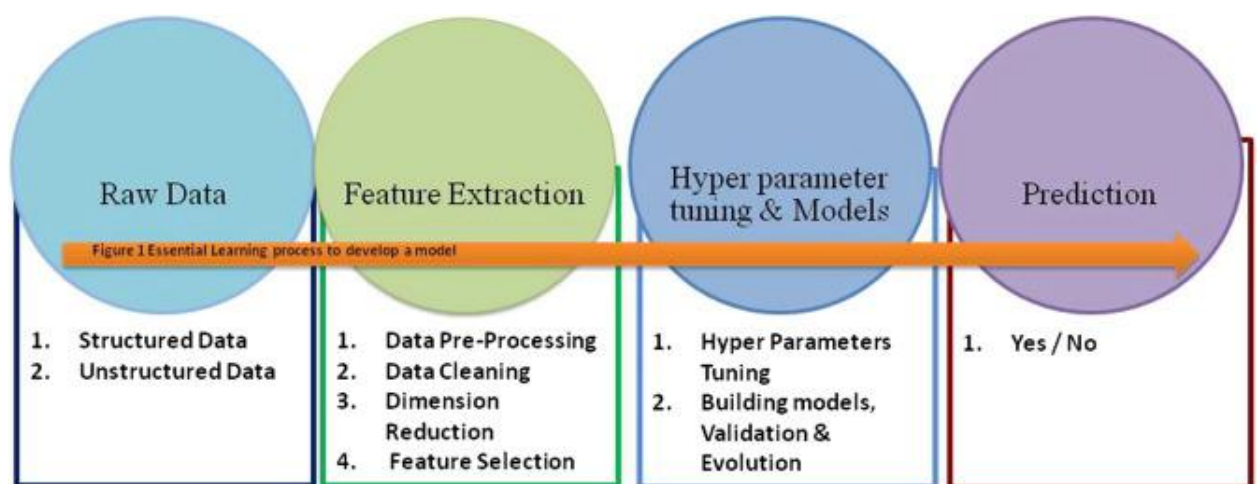


Fig. 1. Essential Learning process to develop a predictive model.

1.2.1 Supervised Learning

In supervised learning, the system must “learn” inductively a function called target function, which is an expression of a model describing the data.

In supervised learning, there are two kinds of learning tasks: classification and regression. Classification models try to predict distinct classes, such as e.g. blood groups, while regression models predict numerical values. Some of the most common techniques are Decision Trees (DT), Rule Learning, and Instance Based Learning (IBL), such as k-Nearest Neighbours (k-NN), Genetic Algorithms (GA), Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

1.2.2 Unsupervised Learning

In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels.

1.2.3 Reinforcement Learning

The term Reinforcement Learning is a general term given to a family of techniques, in which the system attempts to learn through direct interaction with the environment so as to maximize some notion of cumulative reward.

1.3 DIABETES AND MACHINE LEARNING

With the development of living standards, diabetes is increasingly common in people's daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy studying. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels. The earlier diagnosis is obtained, the much easier we can control it. Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors . For machine learning method, how to select the valid features and the correct classifier are the most important problems.

With the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treating the patients before it becomes critical. Data mining has the ability to extract hidden knowledge from a huge amount of diabetes-related data. Because of that, it has a significant role in diabetes research, now more than ever. The aim is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on three classification methods namely, Support Vector Machine, Logistic Regression and Decision tree Classifier.

Other applications of machine learning in this field include:

- **Glucose Monitoring Systems:** Machine learning algorithms help automate the process of monitoring blood sugar levels and recommend adjustments in care.
- **Nutrition Coaching:** To help recommend meal options based on the specific diet criteria of the user.
- **Early Diagnosis Tools:** Deep learning to predict the onset of diabetic retinopathy, the leading cause of vision loss among diabetics.

Diabetes management, to a large degree, involves pattern recognition thus positioning it well for applications of AI. For example, key factors such as blood glucose, weight and blood pressure must be consistently measured and monitored to inform patient care.

By automating routine processes, clinicians can devote less time to data entry and more time to patient interaction. Patients may also gain an increased sense of control over their diabetes by having coaching tools and support at their disposal.

The real potential for AI is the ability to scale up these innovations while still personalizing the user experience. Additionally, a better continuity of care may be achieved between face-to-face visits with physician and patient.

2. Material and Method

2.1 Data

Dataset of female patients with minimum twenty one year age of Pima Indian population has been taken from UCI machine learning repository. This dataset is originally owned by the National institute of diabetes and digestive and kidney diseases. In this dataset there are total 768 instances classified into two classes: diabetic and non diabetic with eight different risk factors: number of times pregnant, plasma glucose concentration of two hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, two-hour serum insulin, body mass index, diabetes pedigree function and age as in Table1.

Feature label	Variable type	Range
Number of times pregnant	Integer	0–17
Plasma glucose concentration in a 2 h oral glucose tolerance test	Real	0–199
Diastolic blood pressure	Real	0–122
Triceps skin fold thickness	Real	0–99
2 h serum insulin	Real	0–846
Body mass index	Real	0–67.1
Diabetes pedigree function	Real	0.078–2.42
Age	Integer	21–81
Class	Binary	Tested positive for diabetes = 1

Table 1. Statistical report of Pima Indian Dataset.

The following figure 2 is the graphical representation of various features helpful in detecting diabetes

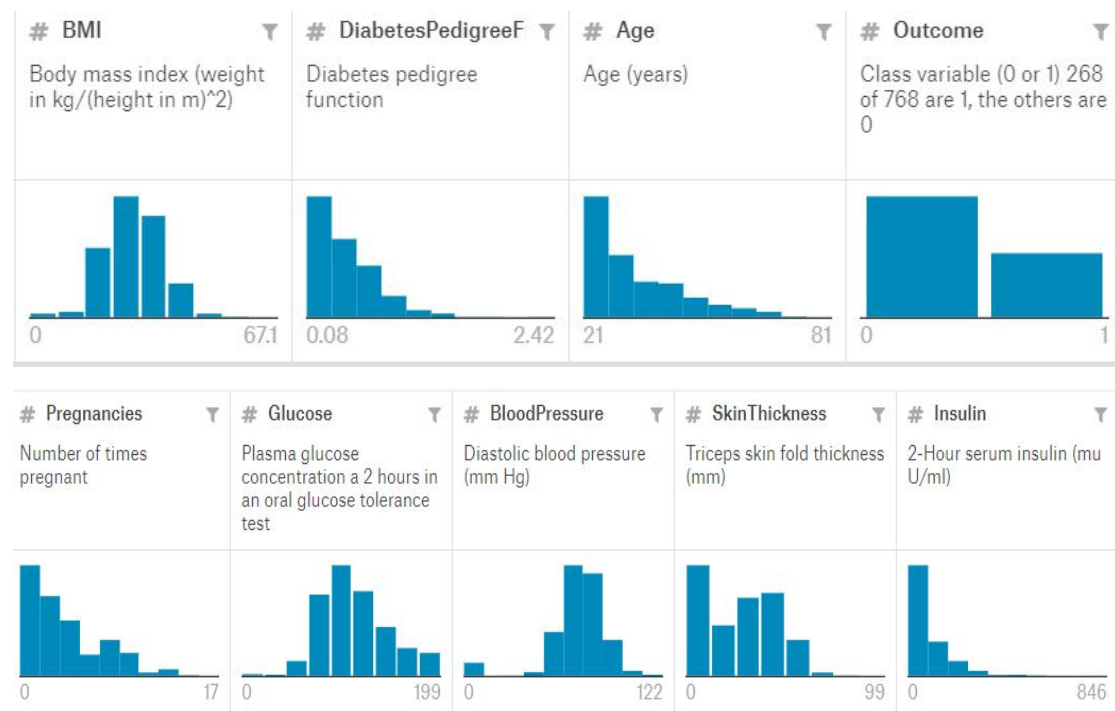


Fig 2. Graphical representation

Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set. But in the given data set we do not need encode the categorical data.

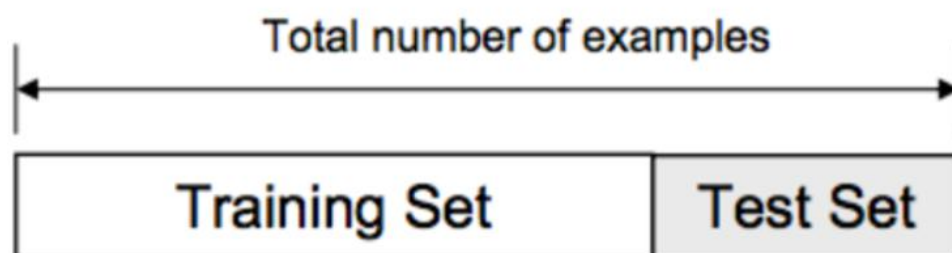


Fig. 3 Splitting of Data

2.2 Classification

The proposed model by us is of classification. Classification is one of the most important decision making techniques in many real world problem. In this work, the main objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For many classification problem, the higher number of samples chosen but it doesn't leads to higher classification accuracy. In many cases, the performance of algorithm is high in the context of speed but the accuracy of data classification is low.

The main objective of our model is to achieve high accuracy. Classification accuracy can be increase if we use much of the data set for training and few data sets for testing. This survey has analyzed various classification techniques for classification of diabetic and non-diabetic data. Thus, it is observed that techniques like Support Vector Machine, Logistic Regression, and are most suitable for implementing the Diabetes prediction .

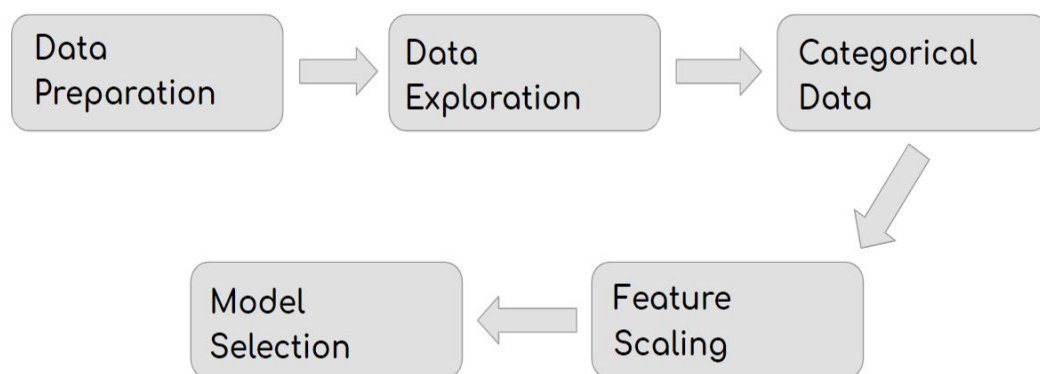


Fig. 4 Phases

Supervised machine learning algorithms are selected to perform binary classification of diabetes dataset of Pima Indians. For predicting whether a patient is diabetic or not, we have used seven different algorithms: Logistic Regression, k-nearest neighbour (k-NN), linear kernel and radial basis function (RBF) kernel support vector machine (SVM), Naives Baye's Classification, Decision Tree Classification and Random Forest Classification which details are given below:

2.2.1 LOGISTIC REGRESSION

In statistics Logistic regression is a regression model where the dependent variable is categorical, namely binary dependent variable-that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed using logistic regression.

Many other medical scales used to assess severity of a patient have been developed using logistic regression. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription. In economics it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application is about to predict the likelihood of a homeowner defaulting on a mortgage.

Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing.

In this paper, Logistic regression was used to predict whether a patient suffer from diabetes, based on nine observed characteristics of the patient.

The figure below gives a graphical representation in which Logistic Regression Model works.

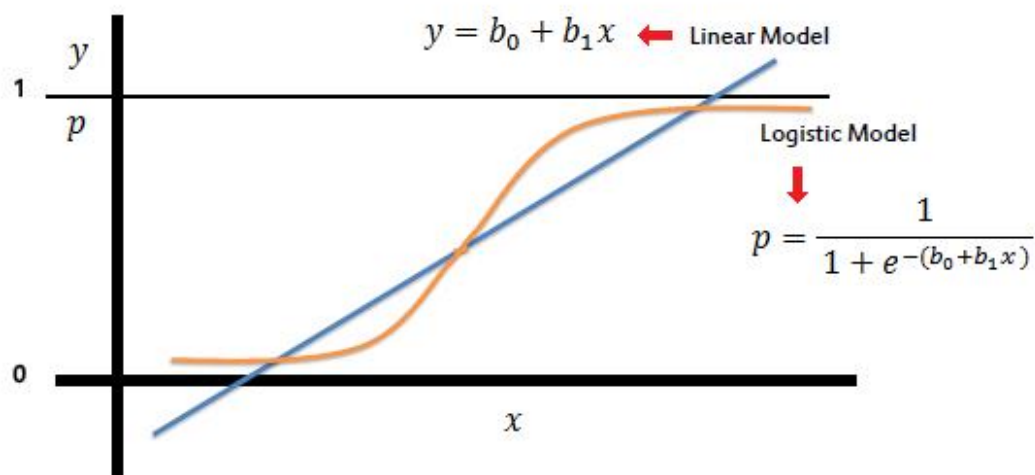


Fig 5. Logistic regression

2.2.2 *k*-Nearest neighbour (*k*-NN)

k-Nearest neighbour is a simple algorithm but yields very good results. It is a lazy, non-parametric and instance based learning algorithm. This algorithm can be used in both classification and regression problems. In classification, *k*-NN is applied to find out the class, to which new unlabeled object belongs.

For this, a ' k ' is decided (where k is number of neighbours to be considered) which is generally odd and the distance between the data points that are nearest to the objects is calculated by the ways like Euclidean's distance, Hamming distance, Manhattan distance or Minkowski distance.

After calculating the distance, ' k ' nearest neighbours are selected the resultant class of the new object is calculated on the basis of the votes of the neighbours. The k -NN predicts the outcome with high accuracy.

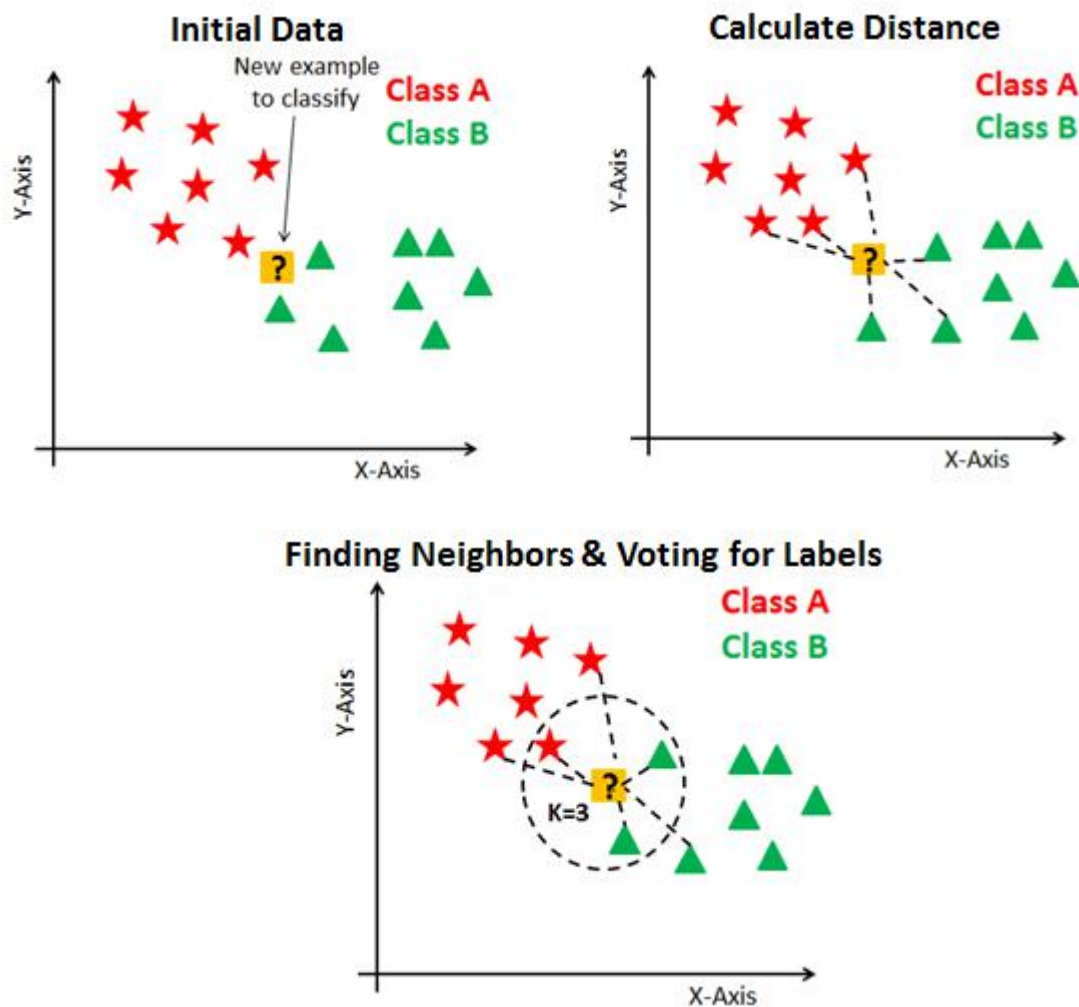


Fig. 6 k Nearest Neighbour

2.2.3 Support vector machine

Support vector machine (SVM) is used in both classification and regression. In SVM model, the data points are represented on the space and are categorized into groups and the points with similar properties falls in same group. In linear SVM the given data set is considered as p-dimensional vector that can be separated by maximum of p-1 planes called hyper-planes. These planes separate the data space or set the boundaries among the data groups for classification or regression problems as in Figure. The best hyper-plane can be selected among the number of hyper-planes on the basis of distance between the two classes it separates. The plane that has the maximum margin between the two classes is called the maximum-margin hyper-plane.

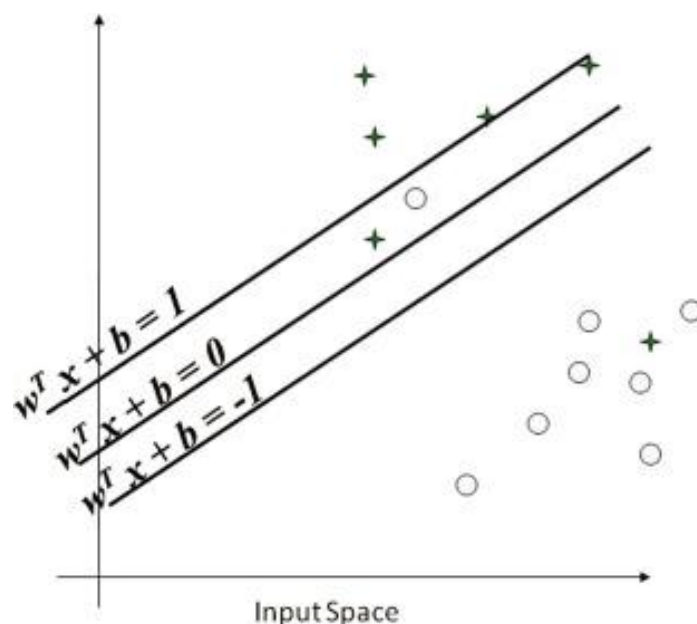


Fig. 7. Representation of Support vector machine.

For n data points is defined as:

$$(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)$$

Where \vec{x}_1 is real vector and y_1 can be 1 or -1 , representing the class to which \vec{x}_1 belongs.

A hyper-plane can be constructed so as to maximize the distance between the two classes $y = 1$ and $y = -1$, is defined as:

$$\vec{w} \cdot \vec{x} - b = 0$$

where \vec{w} is normal vector and $\frac{b}{\|\vec{w}\|}$ is offset of hyper-plane along \vec{w} .

Recently, SVM has attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVM is a technique suitable for binary classification tasks, so we choose SVM to predict the diabetes. The reason is SVM is well known for its discriminative power for classification, especially in the cases where a large number of features are involved, and in our case where the dimension of the feature is 9.

2.2.4 Radial basis function (RBF) Kernel Support Vector Machine

Support vector machine has proven its efficiency on linear data and non linear data. Radial base function has been implemented with this algorithm to classify non linear data. Kernel function plays very important role to put data into feature space.

Mathematically, kernel trick (K) is defined as:

$$K(x_1, x_2) = \exp\left(-\frac{|x_1 - x_2|^2}{2\sigma^2}\right)$$

A Gaussian function is also known as Radial basis function (RBF) kernel.

In Figure, the input space separated by feature map (Φ). By applying equation (1), (2) we get:

$$f(\mathcal{X}) = \sum_i^N \alpha_i y_i k(\mathcal{X}_i, \mathcal{X}) + b$$

Function Φ mapping the idea into another space is defined as:

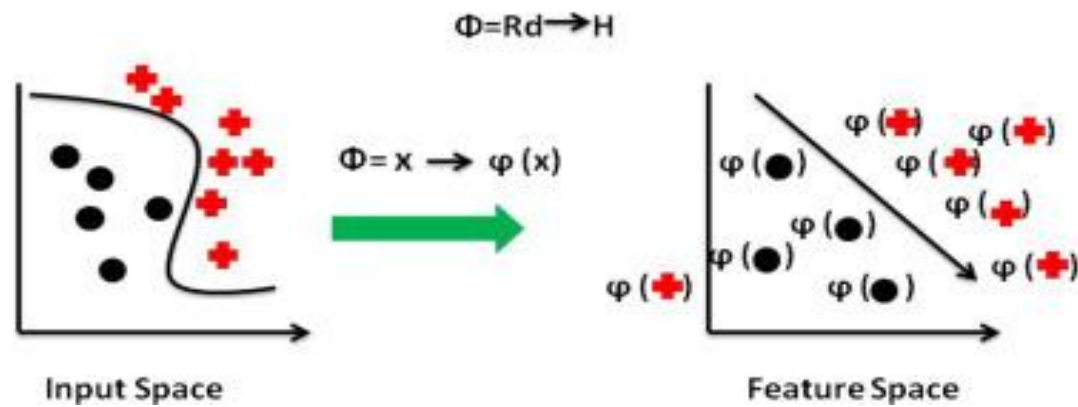


Fig. 8. Representation of Radial basis function (RBF) kernel Support Vector Machine.

By applying equation (3) in 4 we get new function, where N represents the trained data.

$$f(\mathcal{X}) = \sum_i^N \alpha_i y_i \exp\left(-\frac{|x_1 - x_2|^2}{2\sigma^2}\right) + b$$

2.2.5 Naives Bayes Classification

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of

algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

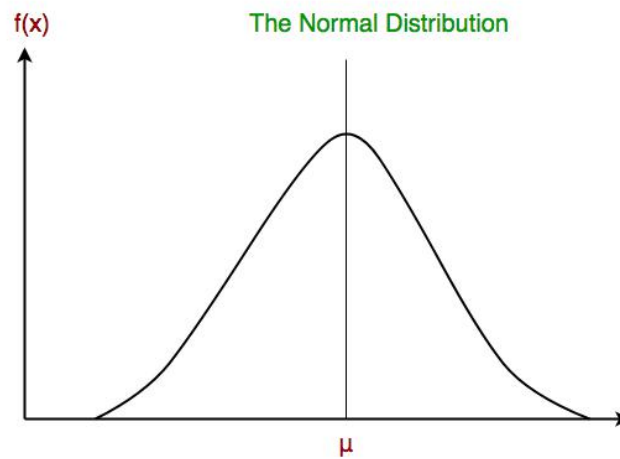


Fig 9. Gaussian Naive Bayes

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers.[\[7\]](#) Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.

2.2.6 Decision Tree Classification

A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers to the question; and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surface.

Decision trees classify the examples by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the example. Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is repeated for every sub tree rooted at the new nodes.

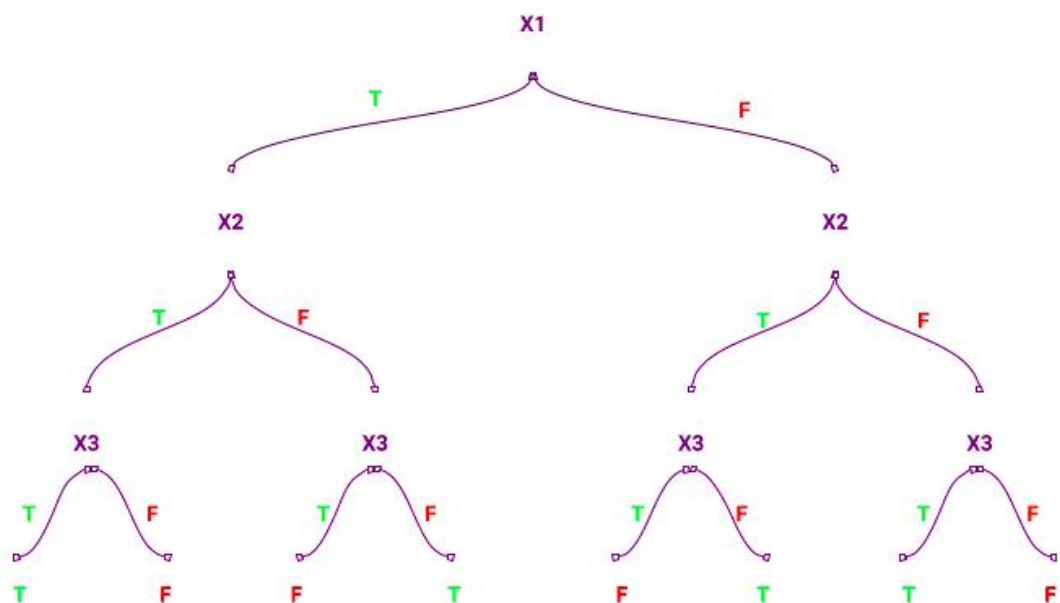
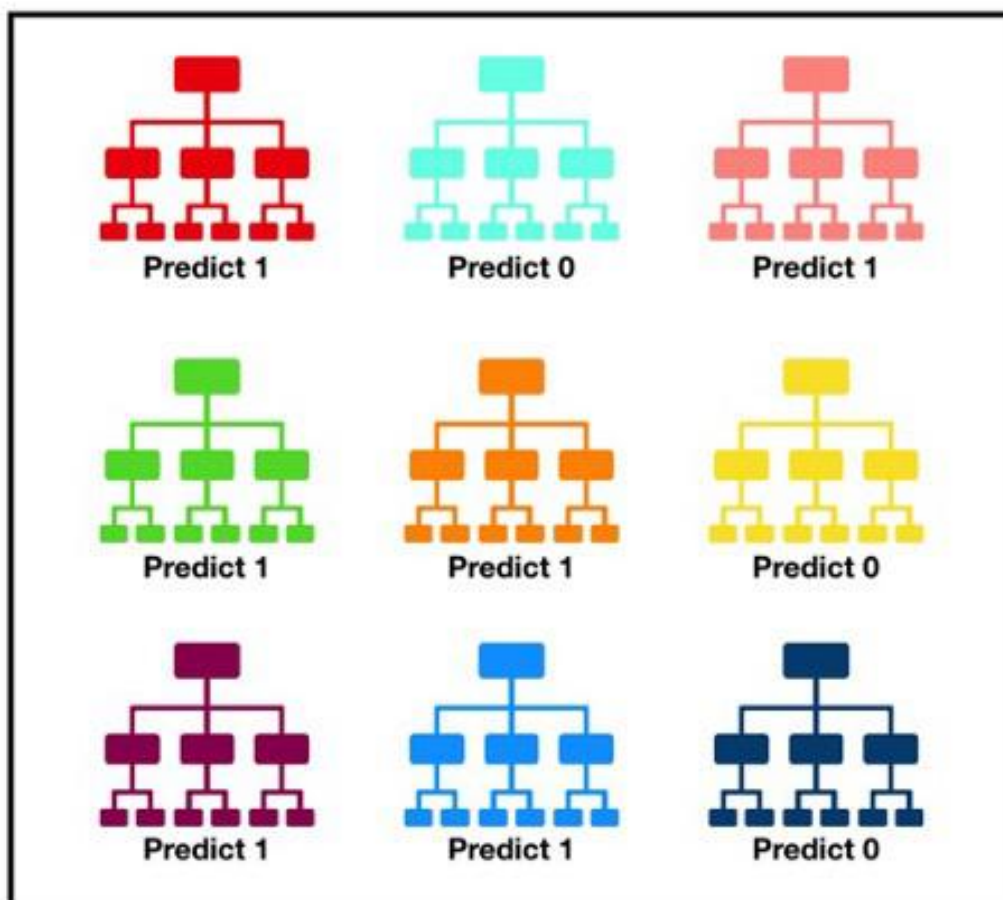


Fig 9. Decision tree for an XOR operation involving three operands

2.2.7 Random Forest Classifier

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an [ensemble](#). Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below).



Tally: Six 1s and Three 0s
Prediction: 1

Fig 10. Random Forest Classifier

Visualization of a Random Forest Model Making a Prediction

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is:

A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. **The reason for this wonderful effect is that the trees protect each other from their individual errors** (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So the prerequisites for random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.
2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

2.3 Ensembling

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. The models used to create such ensemble models are called ‘**base models**’.

We will do ensembling with the **Voting Ensemble**. Voting is one of the simplest ways of combining the predictions from multiple machine learning algorithms. It works by first creating two or more standalone models from your training dataset. A Voting Classifier can then be used to wrap your models and average the predictions of the sub-models when asked to make predictions for new data.

We will be using weighted Voting Classifier. We will assign to the classifiers according to their accuracies. So the classifier with single accuracy will be assigned the highest weight and so on.

In our case, we will use the Top 3 classifiers i.e Logistic Regression , Radial (rbf) SVM and Naive Baye’s classifiers.

Our aim is to increase the accuracy of the model for better predictions. In order to do that , we tried out various combinations of the classification models mentioned above and checked their accuracy to find the best ensemble averaging method model.

We used Hard Voting and found that Decision tree, Support Vector Machine and Logistic Regression when grouped together to form the ensemble method give the best results.

The table below shows the accuracy of the models made by using the seven classification algorithm explained above-

Algorithm	Accuracy
Linear SVM	0.791666
RBF Kernel SVM	0.770833
Logistic Regression	0.791666
KNN	0.791666
Decision Tree	0.755208
Random Forest	0.776041
Naive Bayes	0.796875
Voting Method	0.786458

Table 2 Accuracy of the classification algorithms used.

2.4 Feature Extraction/ Selection:

After testing the above classification algorithms, the accuracies were not up to the mark. This can be improved by using Feature Selection and using only relevant features.

A lot many features can affect the accuracy of the algorithm. Feature Extraction means to select only the important features in-order to improve the accuracy of the algorithm. It reduces training time and reduces over fitting.

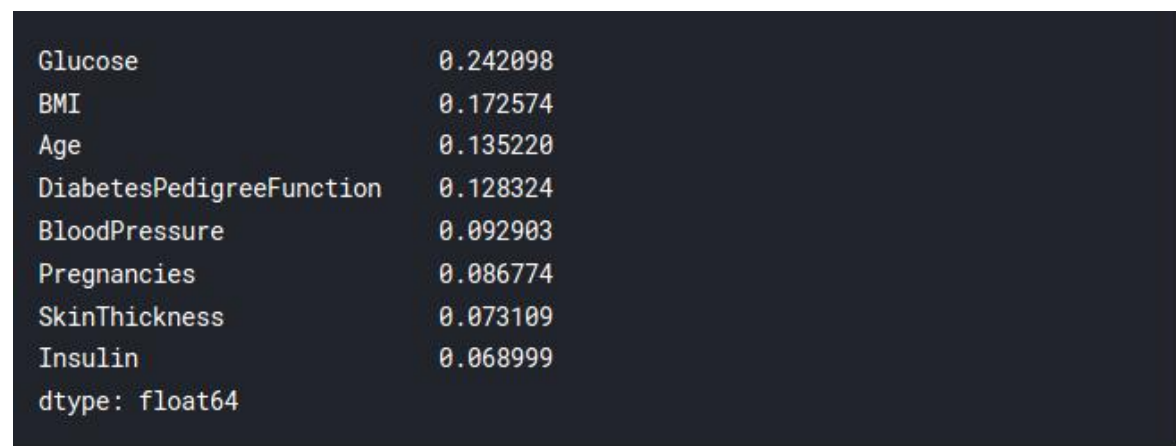
We can choose important features in 2 ways:

a) Correlation matrix--> selecting only the uncorrelated features

b)Random Forest Classifier--> It gives the importance of the features

We carried out the second method- “Random Forest Classification” for determining which features are more important in predicting the result.

The following image shows the important features are: Glucose, BMI, Age, DiabetesPedigreeFunction.



Glucose	0.242098
BMI	0.172574
Age	0.135220
DiabetesPedigreeFunction	0.128324
BloodPressure	0.092903
Pregnancies	0.086774
SkinThickness	0.073109
Insulin	0.068999
dtype: float64	

Fig 11. Result of Feature Extraction

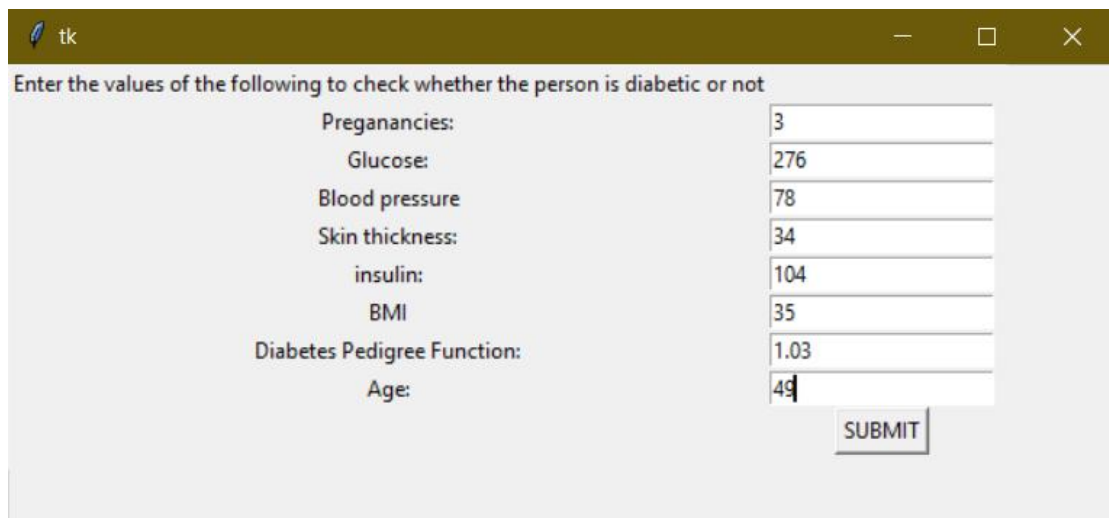
2.5 Standardization:

There can be a lot of deviation in the given dataset. An example in the dataset can be the BMI where it has 248 unique values. This high variance has to be standardised. Standardization is a useful technique to transform attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1.

We used standard scaler from the sklearn preprocessing library to get all the values of features between -1 and 1 to make them comparable.

2.6 Graphical User Interface

We also built a GUI to support our model using tkinter in python. The GUI requires the user to enter the values of features to get the prediction whether the user is Diabetic or not.



Feature	Value
Preganancies:	3
Glucose:	276
Blood pressure	78
Skin thickness:	34
insulin:	104
BMI	35
Diabetes Pedigree Function:	1.03
Age:	49

Fig 12.Graphical User Interface

```
The Patient is Diabetic
In [8]:
```

Fig 13. Result for the data user entered

3. Results and Discussion

Initially, we checked the accuracy of all the seven classification algorithms. As we can see in the table below, the three best performance and accuracy was given by Linear SVM, Decision Trees and Logistic Regression.

Algorithm	Accuracy
Linear SVM	0.791666
RBF Kernel SVM	0.770833
Logistic Regression	0.791666
KNN	0.791666
Decision Tree	0.755208
Random Forest	0.776041
Naive Bayes	0.796875
Voting Method	0.786458

Table 3 Accuracy of the classification algorithms used before Feature Extraction and Standardization.

After applying standardisation and feature extraction, we saw a considerable increase in the accuracy of the models as shown in the table below and the three best accuracy were given by Logistic Regression, Naive Baye's Classifier and RBF Support Vector Machine.

Algorithm	Accuracy Before	New Accuracy	Increase
Linear SVM	0.789191	0.789191	0.0000000
RBF Kernel SVM	0.770833	0.796875	0.026042
Logistic Regression	0.791666	0.791666	0.0000000
KNN	0.791666	0.786458	-0.005208
Decision Tree	0.755208	0.729166	0.026042
Random Forest	0.765625	0.776041	0.010416
Naive Bayes	0.796875	0.79166666666667	-0.005209
Voting	0.786458	0.802083	0.015625

Table 4 Accuracy of the classification algorithms used after Feature Extraction and Standardization.

We can clearly see from the table that the accuracy of RBF Kernel SVM increases by 3.2%, and no increase in case of Logistic Regression and a decrease in case of Naive Baye's Classifier.

So the maximum Accuracy which we could get by using ensemble models is **80.2083%**.

4. Conclusion:

Early diagnosis of diabetes can be helpful to improve the quality of life of patients and enhancement of their life expectancy. Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decisions about the disease status. Also, this work has the capacity to be extended further by investigating different deep learning methods like neural networks to improve the performance of the model.

5. References

Research Papers:

- i. Predicting Diabetes Mellitus With Machine Learning Technique
Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang
(<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232260/>)
- ii. Diabetes Prediction Using Machine Learning Techniques
Tejas N. Joshi*, Prof. Pramila M. Chawan
(http://www.ijera.com/papers/Vol8_issue1/Part-2/C0801020913.pdf)
- iii. Predictive modelling and analytics for diabetes using a machine learning approach
Harleen Kaur Vinita Kumari
(<https://www.sciencedirect.com/science/article/pii/S221083271830365X>)
- iv. Dataset
Kagel
<https://www.kaggle.com/uciml/pima-indians-diabetes-database>
- v. Domain Knowledge

National Institute for Diabetes, Digestive and Kidney Diseases

<https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes>

vi. Office on Women's Health (US)

<https://www.womenshealth.gov/a-z-topics/diabetes>

vii. Healthline

<https://www.healthline.com/health/diabetes>