ROYAL GLOBAL UNIVERSITY
GUWAHATI

Diabetes Prediction Using Machine Learning

A Project Report submitted
in fulfillment of the requirements for the degree of
Bachelors of Technology in
Computer Science and Engineering

Submitted by

Levid Kumar (202025017)
Surjyadeep Baruah (202025031)
Tanzimur Rohman (202025034)
B. Tech CSE, 7th Semester
Royal School of Engineering and Technology

Under the guidance of

Mrs. Nilakshi Deka

Assistant Professor

Royal School of Engineering and Technology
THE ASSAM ROYAL GLOBAL UNIVERSITY
GUWAHATI: 781035
Session: 2023-24

ROYAL GLOBAL UNIVERSITY

GUWAHATI

CERTIFICATE OF APPROVAL

It is certified that the work contained in the report entitled **" Diabetes Prediction Using Machine Learning"** by **Levid Kumar** bearing Roll No 202025017, **Surjyadeep Baruah** bearing Roll No 202025031 and **Tanzimur Rohman** bearing Roll No 202025034 of **B. Tech CSE, 7th Semester** under the **Department of Computer Science and Engineering**, **Royal School of Engineering and Technology**, The Assam Royal Global University, Guwahati, Assam for the fulfilment of the degree of **Bachelors of Engineering** has been carried out under my supervision and that work has not been submitted elsewhere for a degree.

Project Guide: | **Signature of the External Examiner**

Mrs. Nilakshi Deka | **Name of the External Examiner**

(Assistant Professor)

Date: 26.12.2023

Place: Guwahati

ROYAL GLOBAL UNIVERSITY
GUWAHATI

FORWARDING CERTIFICATE

It is certified that the work contained in the report entitled "**Diabetes Prediction Using Machine Learning**" by **Levid Kumar** bearing Roll No 202025017, **Surjyadeep Baruah** bearing Roll No 202025031 and **Tanzimur Rohman** bearing Roll No 202025034 of **B. Tech CSE, 7th Semester** under the **Computer Science and Engineering**, **Royal School of Engineering and Technology**,under the guidance of **Mrs. Nilakshi Deka, Assistant Professor** has been presented in a manner satisfactory to permit its acceptance as a prerequisite to the degree for which has been submitted.

Date:26.12.2023
Place: Guwahati

Dr. Ishita Chakraborty
(Associate Professor)
Head of Department, Royal School
Of Engineering and Technology

DECLARATION

We, **Levid Kumar** bearing Roll No 202025017, **Surjyadeep Baruah** bearing Roll No 202025031 and **Tanzimur Rohman** bearing Roll No 202025034 hereby declare that this project work entitled “**Diabetes Prediction Using Machine Learning**” was carried out by us under the guidance & supervision of **Mrs. Nilakshi Deka, Assistant Proffesor**. This project work is submitted during the academic session 2023-24. This work or no part of it has been submitted elsewherefor any other purpose till date.

Date: 26.12.2023
Place: Guwahati

Signature

Levid Kumar	**Surjyadeep Baruah**	**Tanzimur Rohman**
Roll No. 202025017	**Roll No. 202025031**	**Roll No. 202025034**

ACKNOWLEDGEMENT

We take the opportunity to express our sincere gratitude to all those who supported us throughout this project/dissertation work. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work.

We convey our special thanks to Mrs. Nilakshi Deka for their support and guidance at during our project/dissertation work.

Our special thanks also go to our Faculty Guide and other faculty members for their kind support for the successful completion of our project/dissertation.

Thank you

Levid Kumar (202025017)
Surjyadeep Baruah (202025031)
Tanzimur Rohman (202025034)

ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or imply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy of different machine learning techniques. This project aims to predict diabetes via five different machine learning methods including: SVM, Logistic regression, KNN, Naïve Bayes and Random Forest This project also aims to propose an effective technique for earlier detection of the diabetes disease using Machine learning algorithms.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter-1

Introduction

1.1Diabetes

Diabetes is a chronic medical condition characterized by elevated blood sugar levels, either due to insufficient insulin production or the body's inability to effectively use insulin. Early detection and timely intervention are crucial in managing diabetes and preventing complications. Machine learning techniques have shown promise in predicting the risk of diabetes based on various factors.

Diabetes directly affects the pancreas, and the body is incapable of producing insulin. Insulin is mainly responsible for maintaining the blood glucose level. Many factors, such as excessive body weight, physical inactivity, high blood pressure, and abnormal cholesterol level, can cause a person get affected by diabetes. It can cause many complications, but an increase in urination is one of the most common ones. It can damage the skin, nerves, and eyes, and if not treated early, diabetes can cause kidney failure and diabetic retinopathy ocular disease. According to IDF (International Diabetes Federation) statistics, 537 million people had diabetes around the world in 2021.

1.2Types Of Diabetes

Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2 diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

Gestational diabetes appears in pregnant women who suddenly develop high blood sugar. In two thirds of the cases, it will reappear during subsequent pregnancies. There is a great chance that type 1 or type 2 diabetes will occur after a pregnancy affected by gestational diabetes.

1.3Symptoms Of Diabetes

- Frequent Urination
- Increased Thirst
- Tired/Sleepiness
- Weight Loss
- Blurred Vision
- Mood Swings
- Confusion and Difficulty Concentrating
- Frequent Infections

1.4 Cause Of Diabetes

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

1.5 Motivation

Early detection of diabetes risk allows for timely intervention and preventive measures. Machine learning models can identify individuals at higher risk before clinical symptoms manifest, enabling healthcare providers to implement lifestyle modifications, education, and medical interventions to prevent or delay the onset of diabetes. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy of different machine learning techniques. This project aims to predict diabetes via five different machine learning methods including: SVM, Logistic regression, KNN, Naïve Bayes and Random Forest This project also aims to propose an effective technique for earlier detection of the diabetes disease using Machine learning algorithms.

1.6 Objective

In this project, the objective is to predict whether the person has Diabetes or not based on various features like Glucose level, Pregnancy, Insulin, Age, BMI and also Develop a Predictive System to determine the possibility of an individual of having diabetes.

Chapter-2
Literature Review

2.1 Datasets

DeeptiSisodia, Dlip Singh Sisodia[1] this paper, work is diabetes prediction at early stage. Decision tree, SVM and Naive Bayes classification algorithms are used to prediction. Accuracy is evaluated using WEKA tool. The Naïve Bayes gave the highest accuracy.

Xue-hui menget al.in [2] this study was comparison of three algorithms for predict the diabetes or prediabetes using common risk factors. The logistic algorithm, ANN and decision tree algorithms are compared. To test 735 patients, they came from two communities in Guangzhou, China. The highest accuracy (77.87%) level is achieved by classification algorithm (C5.0).

Monisha.A et al. in [3] in machine learning, different classifiers are used for predicting and diagnosing diabetes. Like Naive Bayes statistical modelling, logistic regression, Extreme Gradient Boosting. Pima Indians diabetes datasets are experimented. The accuracy for Extreme Gradient Boosting algorithm is 81%. which is greater than other two algorithms.

S.Selvakumar et al. in [4] Discussed about diabetes challenges. Data mining methods are used to predict the peoples whether diabetic or not. Binary Logistic Regression, Multilayer perception and K-Nearest neighbor algorithms are classified. The accuracy level in Binary Logistic Regression is 0.69, Multilayer Perception is 0.71 and K-Nearest neighbor is 0.80. K-Nearest neighbour is the highest accuracy than Binary Logistic Regression, Multilayer perception.

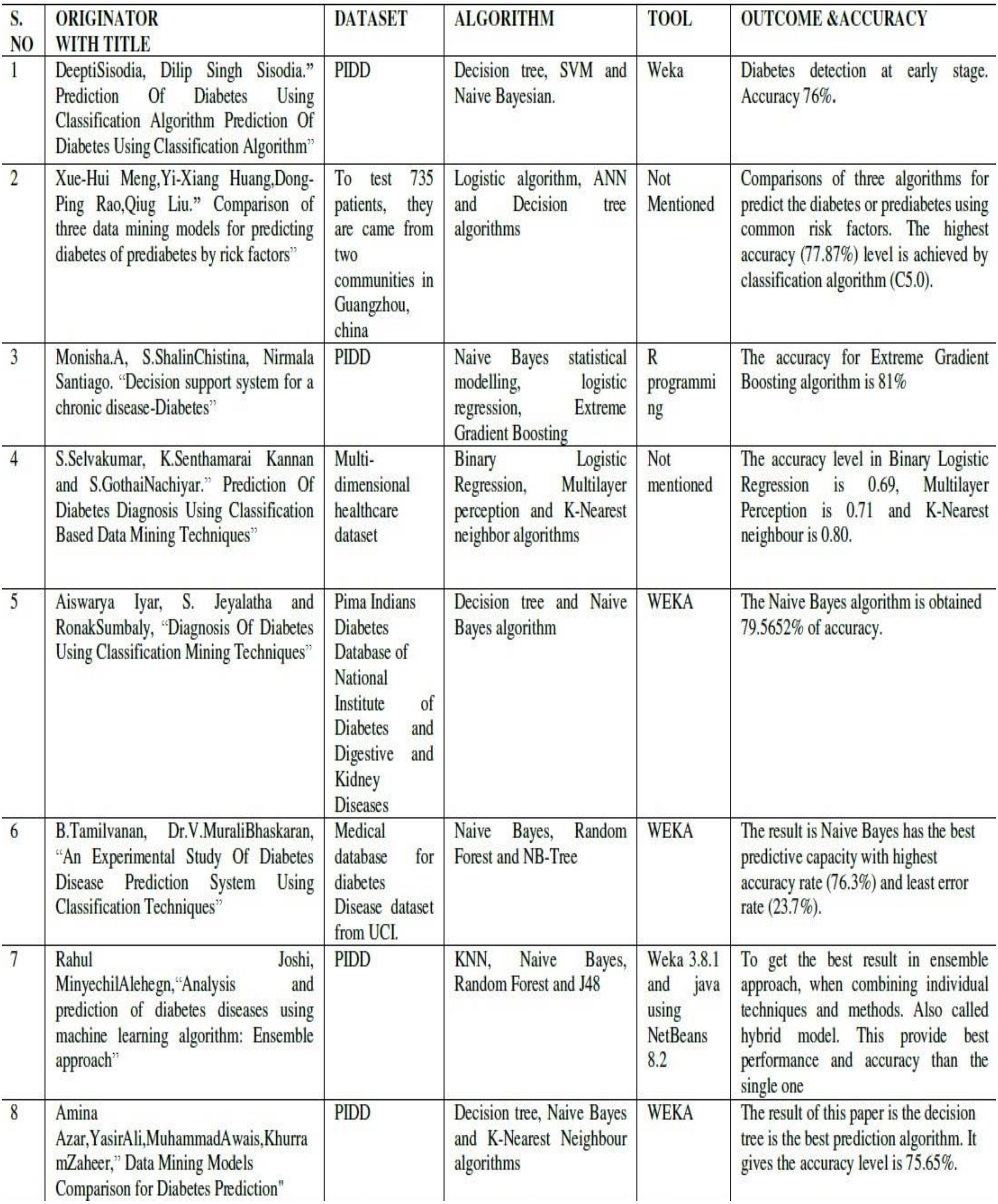
Aiswarya Iyar et al. in [5] worldwide 246 million peoples are affected by diabetes. These are increased over 380 million in 2025 reported by WHO. This paper in aim is finding solution to diagnosis the disease. Using decision tree and Naive Bayes algorithms. Weka tool is used for implementation. The naive Bayes algorithm is obtained 79.5652% of accuracy.

B.Tamilvanan et al. in [6] this paper in objective is predict diabetes with more accuracy. The three classification algorithms are compared for accuracy rate, namely Naive Bayes, Random Forest and NBTree. Implementation using weka tool. The result is Naive Bayes has the best predictive capacity with highest accuracy rate (76.3%) and least error rate(23.7%).

Rahul Joshi et al. in [7] prediction of medical datasets at an early stage is safe for human life using machine learning techniques. To test the Pima Indians diabetes dataset. The applied algorithms are KNN, Naive Bayes, Random Forest and J48. We get the best result to ensemble approach, when combining individual techniques and methods. It is also called hybrid model. This provides the best performance and accuracy than the single one. Weka and java tools are used to predict diabetes.

Amina Azar et al. in [8] Diabetes affected among young peoples and ancient peoples. These are increased day by day and it does not curable. Data mining is used to early stage prediction. This paper in main aim is gives the differentiation and suggest best algorithm. The PID datasets are used. The Decision tree, Naïve Bayes and K-Nearest neighbor algorithms are compared and used for predict the diabetes diagnosis at early stage with highest accuracy and efficiency. The WEKA is used for testing and validation using rapid miner. The result is the decision tree is the best prediction algorithm. It gives the accuracy level is 75.65%.

2.2 Classification of Datasets



S. NO	ORIGINATOR WITH TITLE	DATASET	ALGORITHM	TOOL	OUTCOME &ACCURACY
1	DeeptiSisodia, Dilip Singh Sisodia." Prediction Of Diabetes Using Classification Algorithm Prediction Of Diabetes Using Classification Algorithm"	PIDD	Decision tree, SVM and Naive Bayesian.	Weka	Diabetes detection at early stage. Accuracy 76%.
2	Xue-Hui Meng,Yi-Xiang Huang,Dong-Ping Rao,Qiug Liu." Comparison of three data mining models for predicting diabetes of prediabetes by rick factors"	To test 735 patients, they are came from two communities in Guangzhou, china	Logistic algorithm, ANN and Decision tree algorithms	Not Mentioned	Comparisons of three algorithms for predict the diabetes or prediabetes using common risk factors. The highest accuracy (77.87%) level is achieved by classification algorithm (C5.0).
3	Monisha.A, S.ShalinChistina, Nirmala Santiago. "Decision support system for a chronic disease-Diabetes"	PIDD	Naive Bayes statistical modelling, logistic regression, Extreme Gradient Boosting	R programming	The accuracy for Extreme Gradient Boosting algorithm is 81%
4	S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar." Prediction Of Diabetes Diagnosis Using Classification Based Data Mining Techniques"	Multi-dimensional healthcare dataset	Binary Logistic Regression, Multilayer perception and K-Nearest neighbor algorithms	Not mentioned	The accuracy level in Binary Logistic Regression is 0.69, Multilayer Perception is 0.71 and K-Nearest neighbour is 0.80.
5	Aiswarya Iyar, S. Jeyalatha and RonakSumbaly, "Diagnosis Of Diabetes Using Classification Mining Techniques"	Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases	Decision tree and Naive Bayes algorithm	WEKA	The Naive Bayes algorithm is obtained 79.5652% of accuracy.
6	B.Tamilvanan, Dr.V.MuraliBhaskaran, "An Experimental Study Of Diabetes Disease Prediction System Using Classification Techniques"	Medical database for diabetes Disease dataset from UCI.	Naive Bayes, Random Forest and NB-Tree	WEKA	The result is Naive Bayes has the best predictive capacity with highest accuracy rate (76.3%) and least error rate (23.7%).
7	Rahul Joshi, MinyechilAlehegn,"Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach"	PIDD	KNN, Naive Bayes, Random Forest and J48	Weka 3.8.1 and java using NetBeans 8.2	To get the best result in ensemble approach, when combining individual techniques and methods. Also called hybrid model. This provide best performance and accuracy than the single one
8	Amina Azar,YasirAli,MuhammadAwais,KhurramZaheer," Data Mining Models Comparison for Diabetes Prediction"	PIDD	Decision tree, Naive Bayes and K-Nearest Neighbour algorithms	WEKA	The result of this paper is the decision tree is the best prediction algorithm. It gives the accuracy level is 75.65%.

Fig2.0. Classification of Datasets

2.3Comparative study of Machine Learning Algorithms

Let's summarize the machine learning methods used in the mentioned papers and make a brief comparison:

- **Naïve Bayes:** Found to be effective in several studies, providing high accuracy in [1],[5] and [6].
- **Decision Tree:** Showed competitive performance with the highest accuracy in [8].
- **K-Nearest Neigbour:** Demonstrated high accuracy in [4] and was part of the ensemble approach in [7].
- **Extreme Gradient Boosting (XGBoost):** Achieved the highest accuracy in [3].
- **Random Forest:** Performed decent level of accuracy in [6].

While Naive Bayes and Decision Tree methods were frequently employed and showed good performance, the choice of the best method may vary depending on the dataset, features, and specific characteristics of the study population. The ensemble approach also appeared promising, combining the strengths of multiple methods to achieve improved accuracy in certain cases.

Chapter-3
Proposed System

This section describes the working procedures and implementation of various machine learning techniques to design the proposed diabetes prediction system. Figure 2.0 shows the different stages of this research work. First, the dataset was collected and preprocessed to remove the necessary discrepancies from the dataset. Then the dataset was separated into the training set and test set using the holdout validation technique. Next, different classification algorithms were applied to find the best classification algorithm for this dataset. Finally, the best-performed prediction model is deployed into the proposed system.

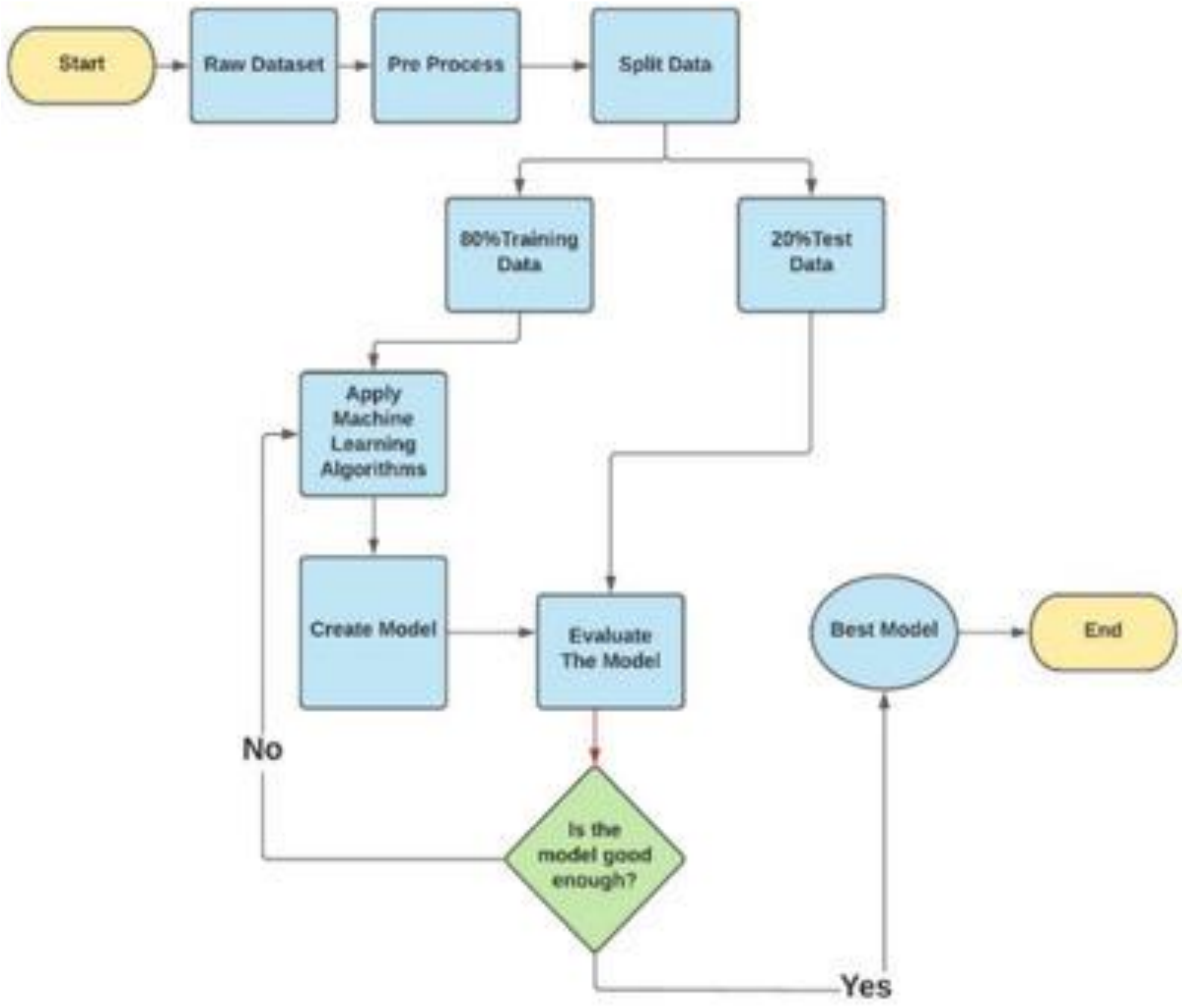


Fig3.0

3.1 Dataset

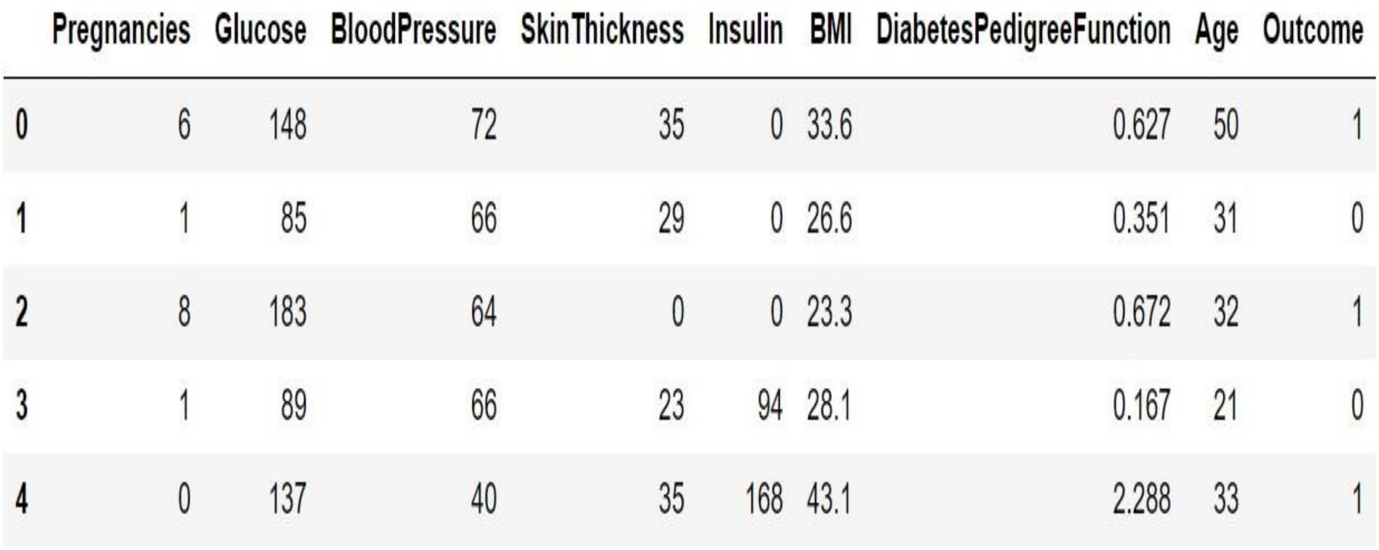
The Pima Indian dataset is an open-source dataset that is publicly available for machine learning classification, which has been used in this work along with a private dataset. It contains 768 patients' data, and 268 of them have developed diabetes.
There are 8 features in the Dataset which are **Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction,** and **Age.**
The Target Variable here is the Outcome that shows whether a patient is diabetic or not.

0- Represents that the patient does not have diabetes

1- Represents that the patient is diabetic

3.2 Top 5 records of the Dataset

These are the top five records of the Dataset that we have used.

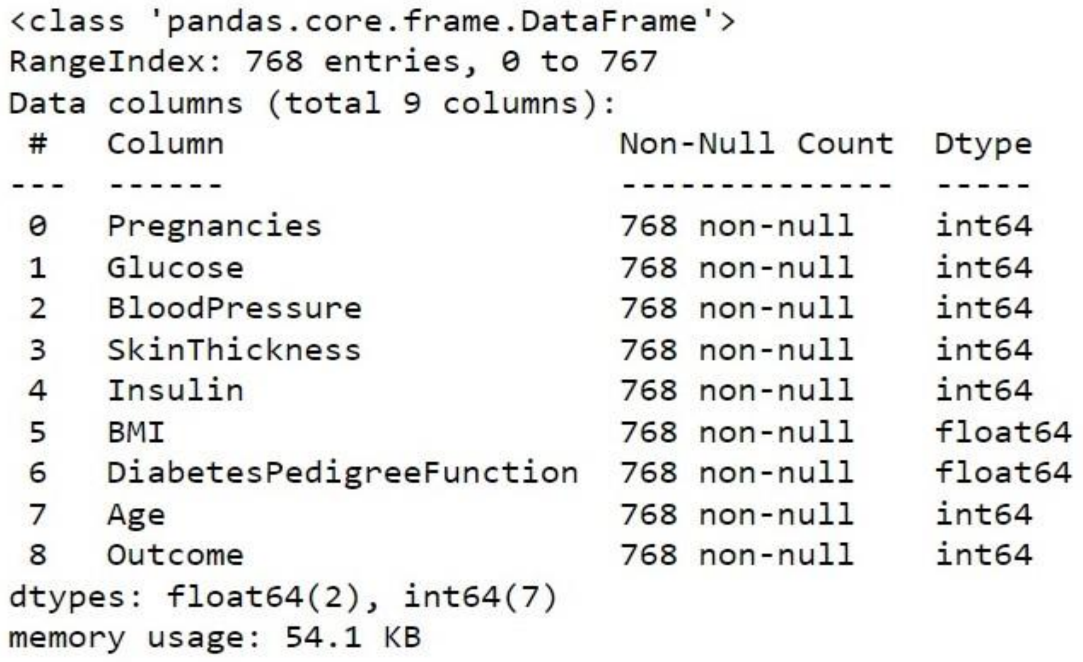


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig3.1

3.3 Information about the Dataset

This provides details about the information about the data and data type of each attribute.

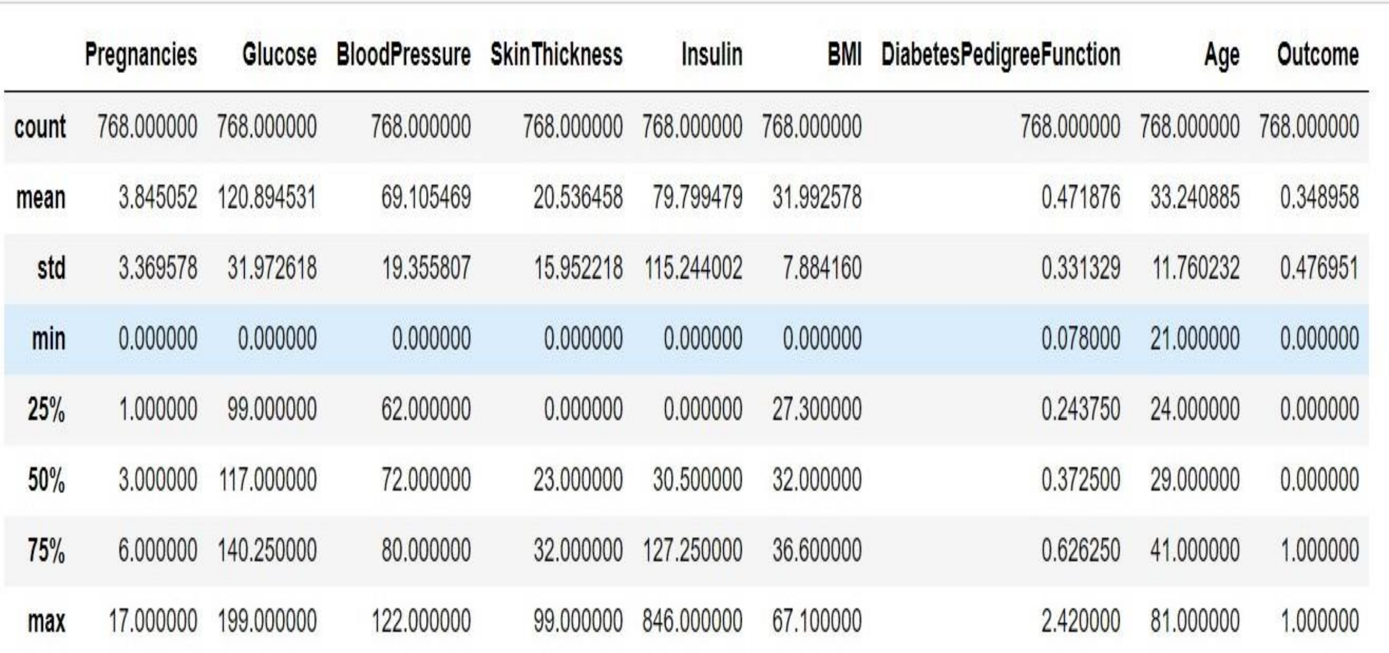


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Fig3.2

3.4 Summary of The Dataset

In this figure we can see how data has been spread for numerical values. We can clearly see the minimum value, mean value, different percentile values and maximum values.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Fig3.3

3.5 Ratio of People having Diabetes

```
Negative (0):    500
Positive (1):    268
```

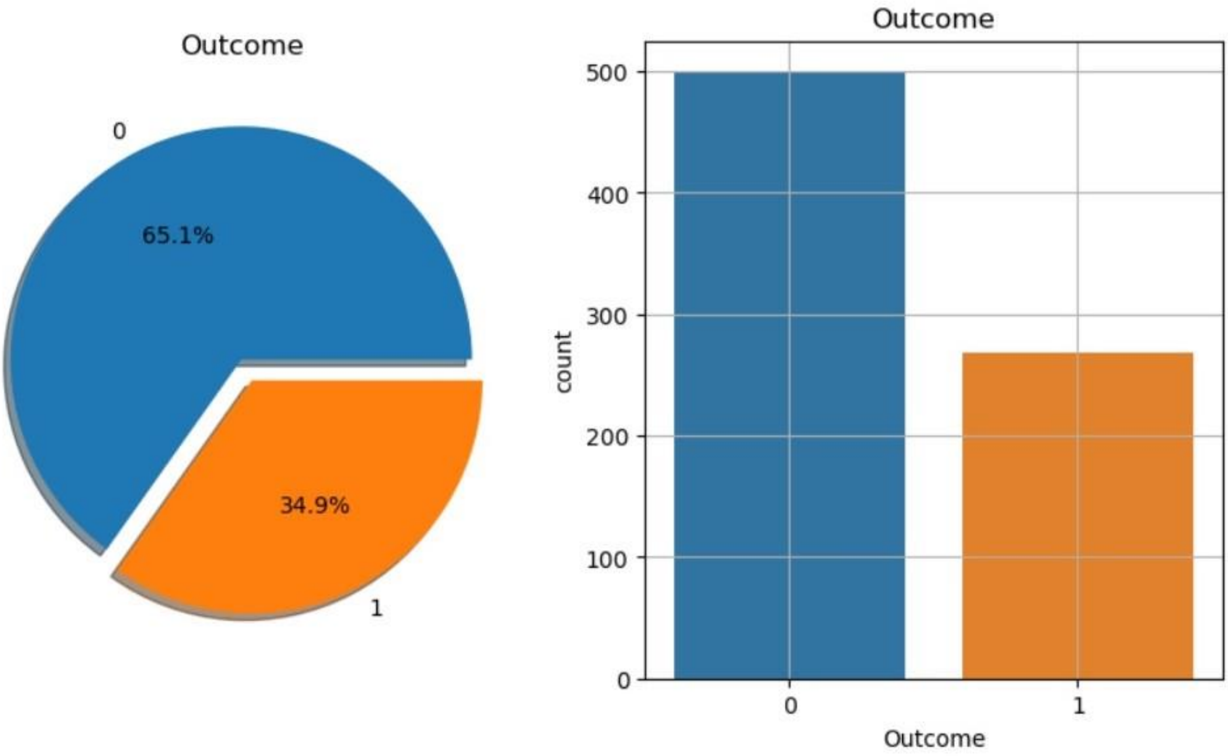


Fig3.4

3.6 Histogram

It is used for display numeric data in form of graphs. It shows the distribution of data.

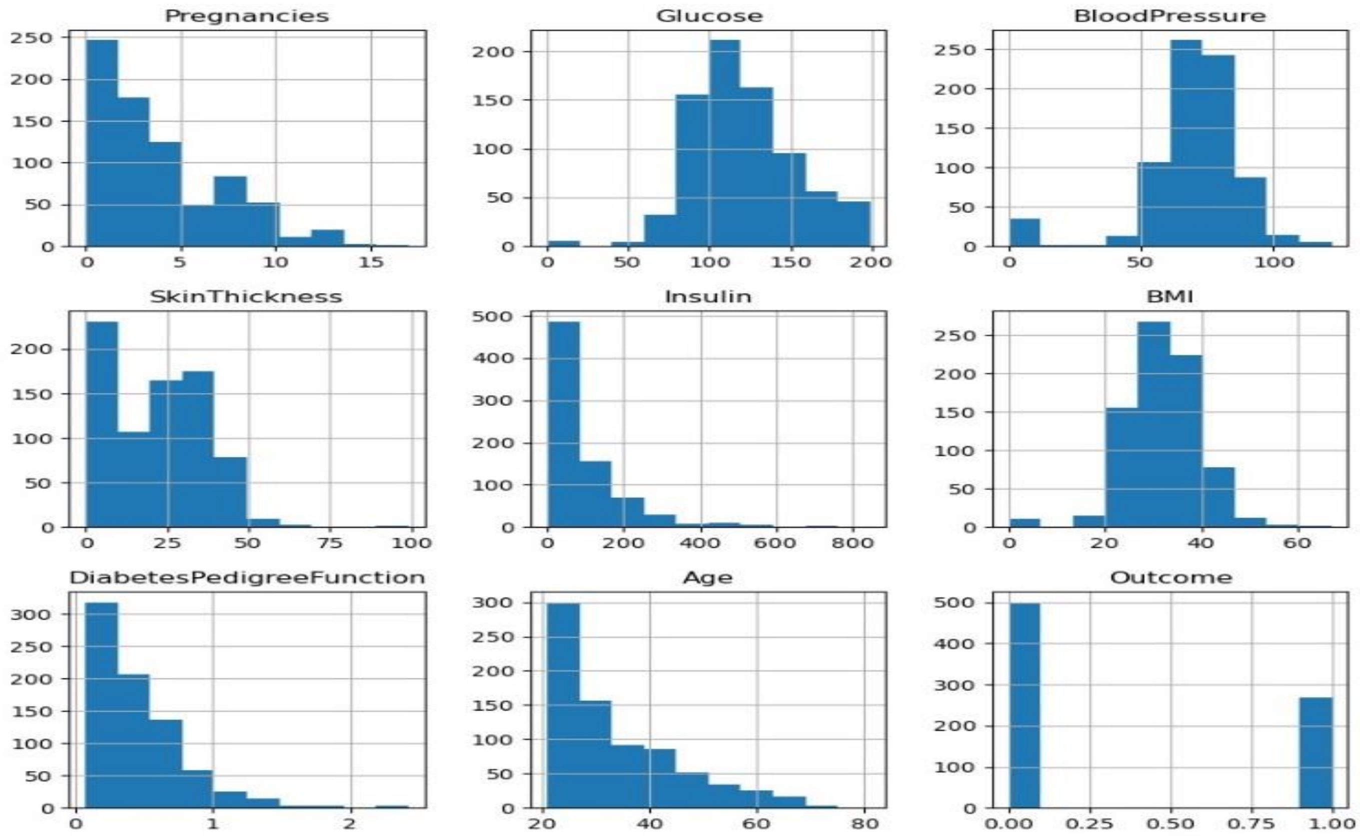


Fig3.5

3.7 Correlation Analysis

Correlation analysis is used to quantify the degree to which two variables are related. It tells us how much one variable change when the other one does.

Fig3.6

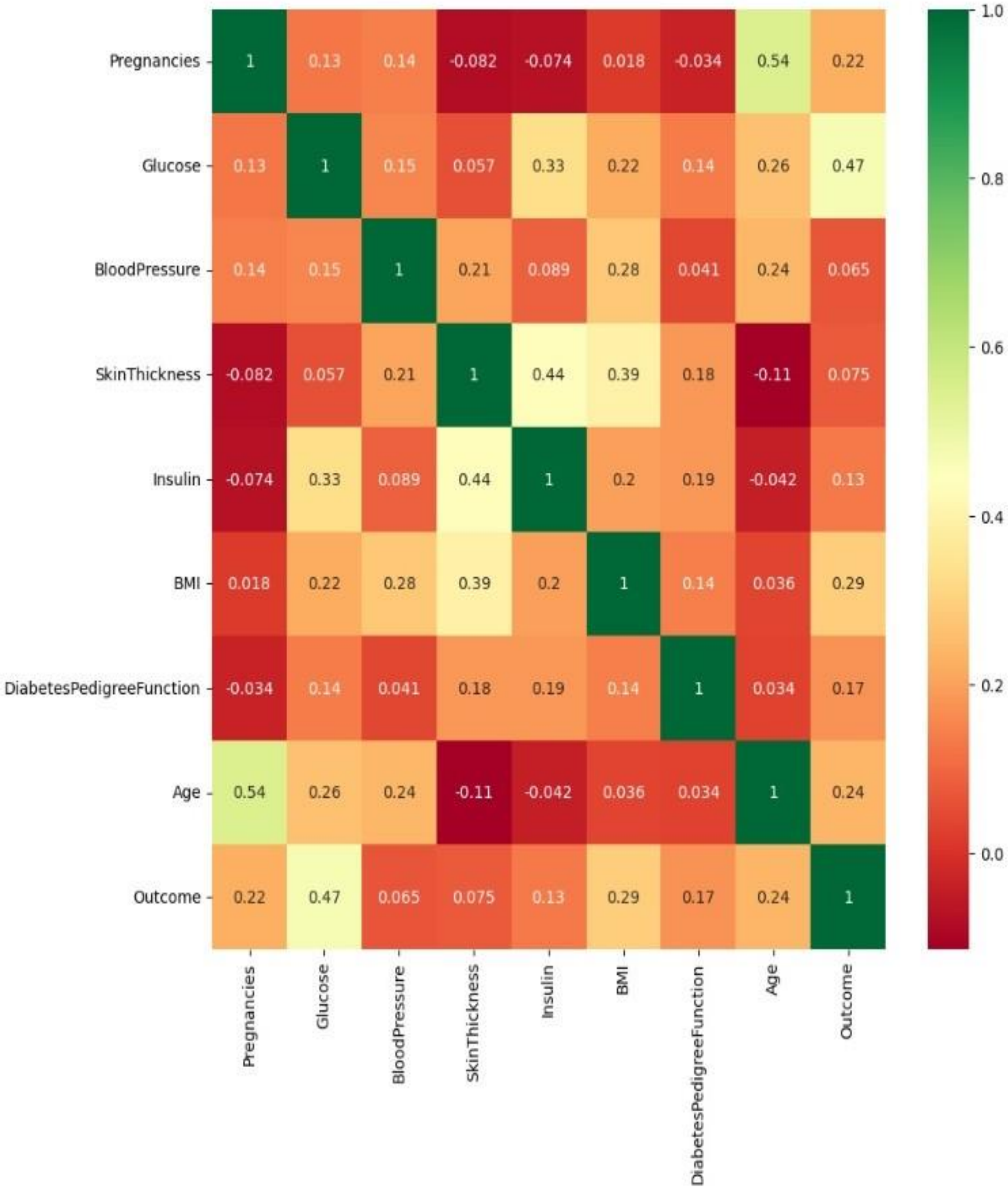


Fig3.6

Chapter 4
Algorithms Used

In this project we have made use of five different machine learning algorithms for the dataset we

have tested.
Those are:

4.1 K-Nearest Neighbors Algorithm (KNN)

The K-Nearest Neighbors (KNN) algorithm is a versatile and straightforward machine learning algorithm used for classification and regression tasks. It falls under the category of instance-based learning or lazy learning algorithms, as it doesn't explicitly build a model during the training phase. Instead, it memorizes the training dataset and makes predictions based on the proximity of new instances to existing data points.
KNN algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.
Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances.

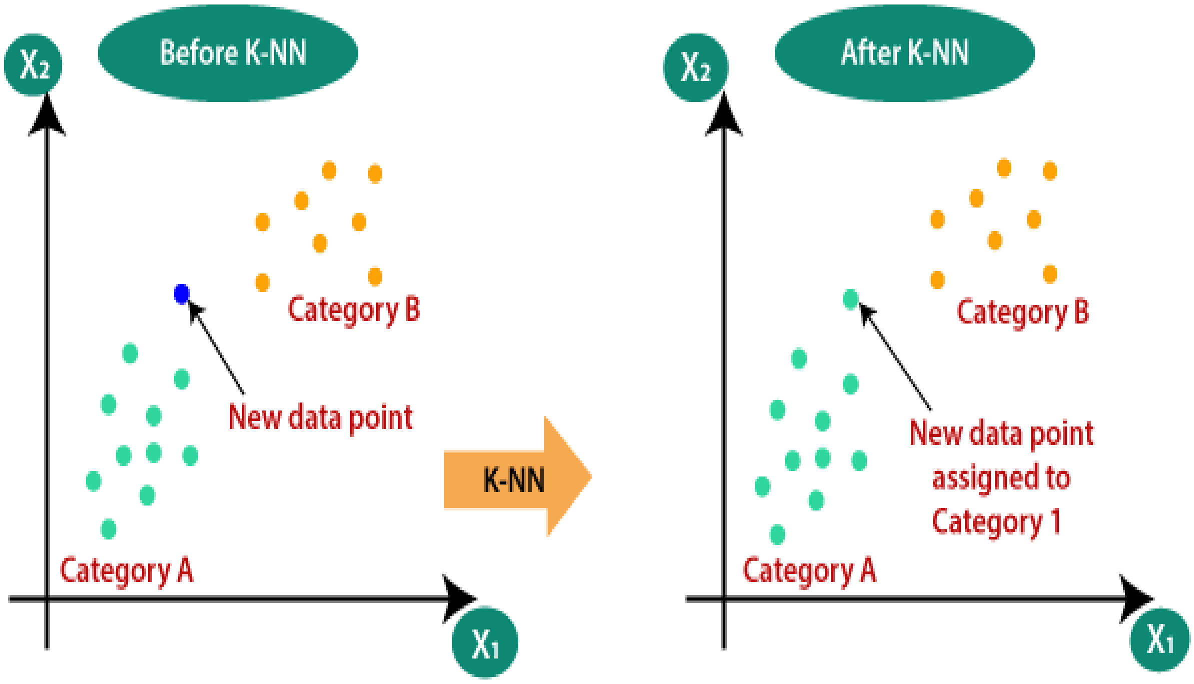


Fig4.0

Advantages Of KNN Algorithm

- Simple to implement and understand

- No training phase, making it particularly useful for online learning
- Effective for datasets with a small number of features or instances.

Disadvantages Of KNN Algorithm

- Can be computationally expensive, especially with large datasets.
- Sensitive to irrelevant or redundant features.
- Performance may degrade when the dataset has varying densities.

4.2 Logistic Regression Algorithm

Logistic Regression is a widely used machine learning algorithm for binary and multi-class classification problems. Despite its name, it is a classification algorithm rather than a regression one. Logistic Regression models the probability that an instance belongs to a particular class.
Logistic Regression models a relationship between predictor variables and a categorical response variable. Logistic regression helps us estimate a probability of falling into a certain level of the categorical response given a set of predictors.

We can choose from three types of logistic regression, depending on the nature of the categorical response variable.

Binary Logistic Regression: Used when the response is binary (i.e., it has two possible outcomes).

Nominal Logistic Regression: Used when there are three or more categories with no natural ordering to the levels.

Ordinal Logistic Regression: Used when there are three or more categories with a natural ordering to the levels, but the ranking of the levels does not necessarily mean the intervals between them are equal.

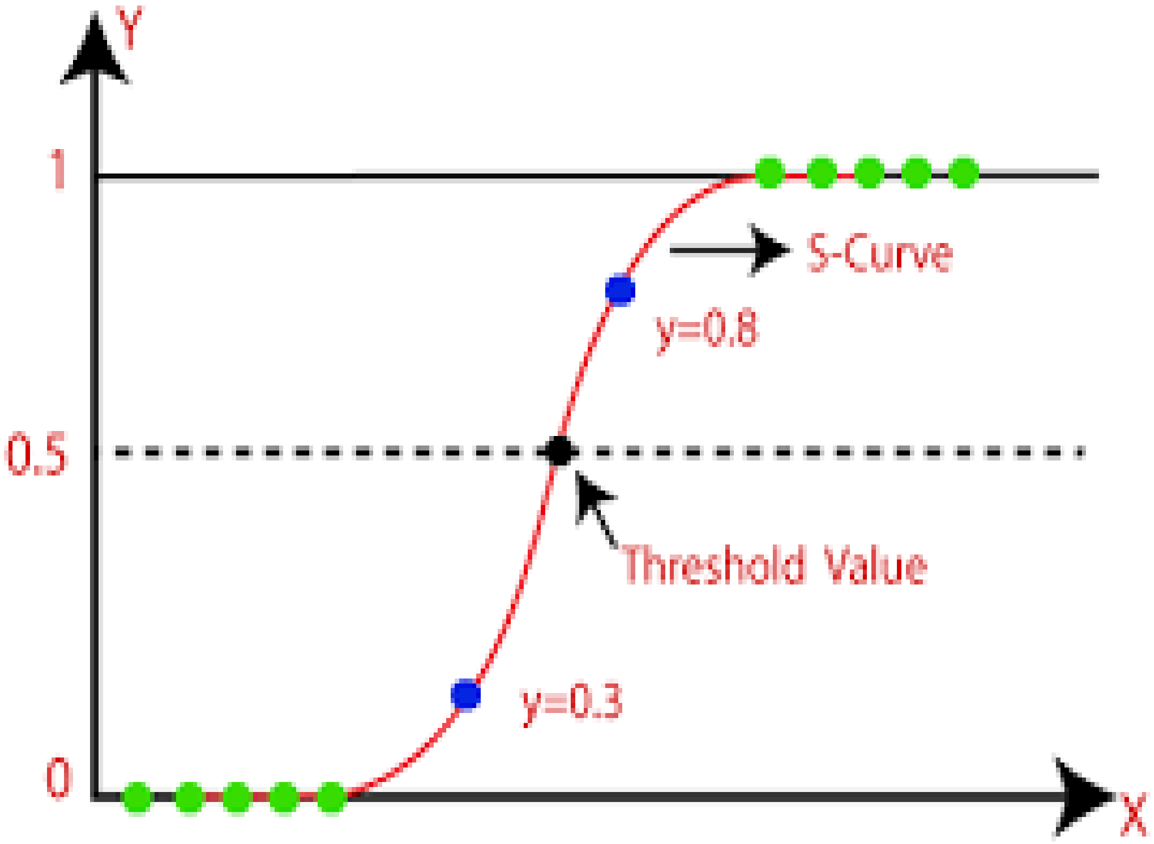


Fig4.1

Advantages Of Linear Regression Algorithm

- Simplicity and ease of interpretation.
- Efficient for binary and multi-class classification.
- Outputs probabilities, making it useful for ranking predictions

Disadvantages Of Linear Regression Algorithm

- Assumes a linear relationship between features and the log-odds of the response.
- Sensitive to outliers.
- May not perform well with highly non-linear data.

4.3 Support Vector Machine (SVM) Algorithm

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for

classification and regression tasks. It's particularly effective in high-dimensional spaces and is widely used for tasks such as image classification, text classification, and bioinformatics.
SVM is supervised learning algorithm used for classification. In SVM we have to identify the right hyper plane to classify the data correctly.
In this we have to set correct parameters values. To find the right hyper plane we have to find right margin for this we have choose the gamma value as 0.0001 and rbf kernel. If we select the hyper plane with low margin leads to miss classification.

Types Of SVM: -

1. Linear SVM:
Suitable for linearly separable data, it aims to find the optimal hyperplane that maximizes the margin between classes.

2. Non-Linear SVM:
Utilizes kernel functions to map the input data into a higher-dimensional space, making it possible to find a hyperplane for non-linearly separable data.

3. Support Vector Regression (SVR):
Extends SVM to regression tasks, where the goal is to predict a continuous variable rather than a categorical one.

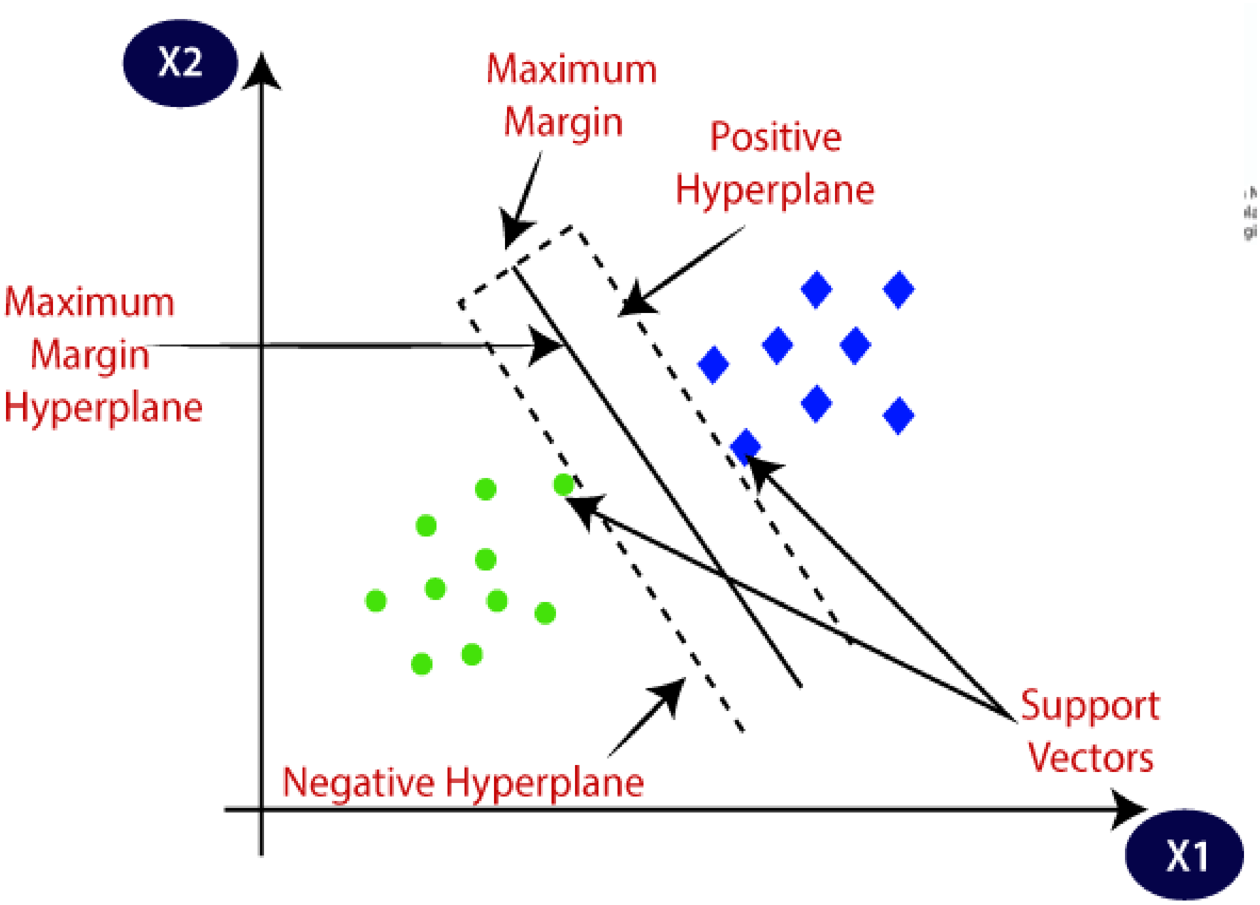


Fig4.2

Advantages Of SVM Algorithm

- Effective in high-dimensional spaces.

- Robust to outliers due to the focus on support vectors.
- Versatile with different kernel functions for handling non-linear

Disadvantages Of SVM Algorithm

- Computationally intensive, especially with large datasets.
- Sensitivity to the choice of kernel and parameters.
- Limited interpretability compared to some other models.

4.4 Naïve Bayes Algorithm

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem. Despite its simplicity and the "naive" assumption of feature independence, it is surprisingly effective for a wide range of classification tasks. Naive Bayes is particularly popular for text classification and spam filtering.
Naive Bayes is built on Bayes' theorem, which describes the probability of an event based on prior knowledge of conditions related to the event. The formula is expressed as:
P(A/B) = P(B/A) x P(A)/P(B)
Naïve 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.

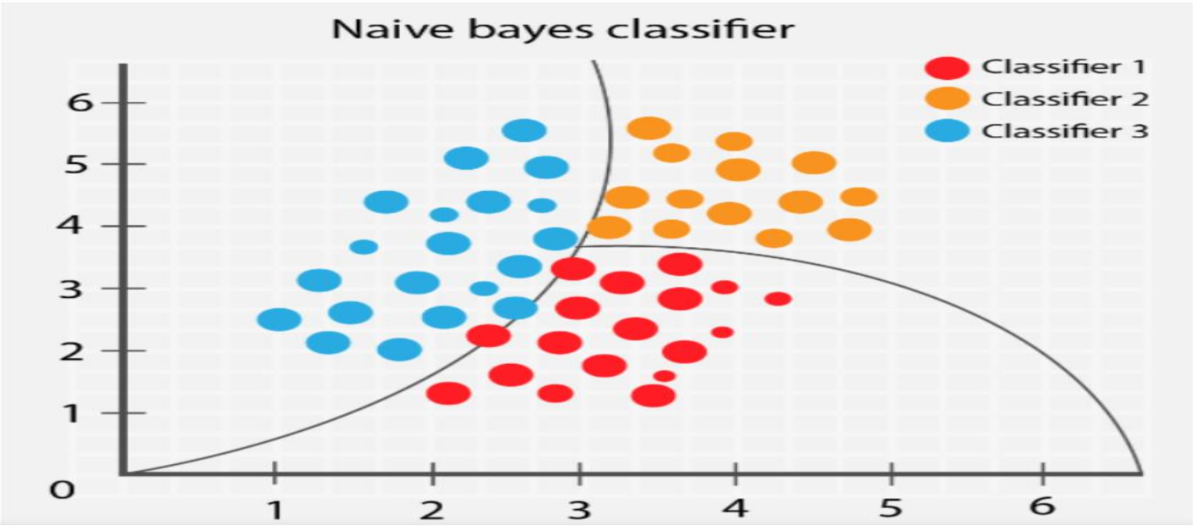


fig4.3

Advantages Of Naïve Bayes Algorithm

- Simple and easy to implement.

- Requires a small amount of training data.
- Can handle a large number of features.

Disadvantages Of Naïve Bayes Algorithm

- Assumes independence of features, which may not hold in some cases.
- Sensitivity to irrelevant features.
- Limited expressiveness compared to more complex models.

Random Forest Algorithm

Random Forest is an ensemble learning algorithm that belongs to the family of bagging methods. It builds multiple decision trees during training and merges their predictions to obtain a more accurate and stable result. Random Forest is widely used for classification and regression tasks and is known for its robustness and ability to handle complex datasets.

The Random Forest Classifier Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It is one of the widely used algorithms, which perform well with any kind of dataset, be it classification or regression.

It is based on the concept of ensemble learning. which is a process of combining multiple classifiers to solve a complex problem, and at the end, the results are either made an average of all the classifiers or mode of all the classifiers.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Random Forest introduces randomness in two ways:

Random Sampling of Data: Each tree is trained on a random subset of the
training data (with replacement).

Random Subset of Features: At each node in a tree, only a random subset of features is considered for splitting.

Random Forest

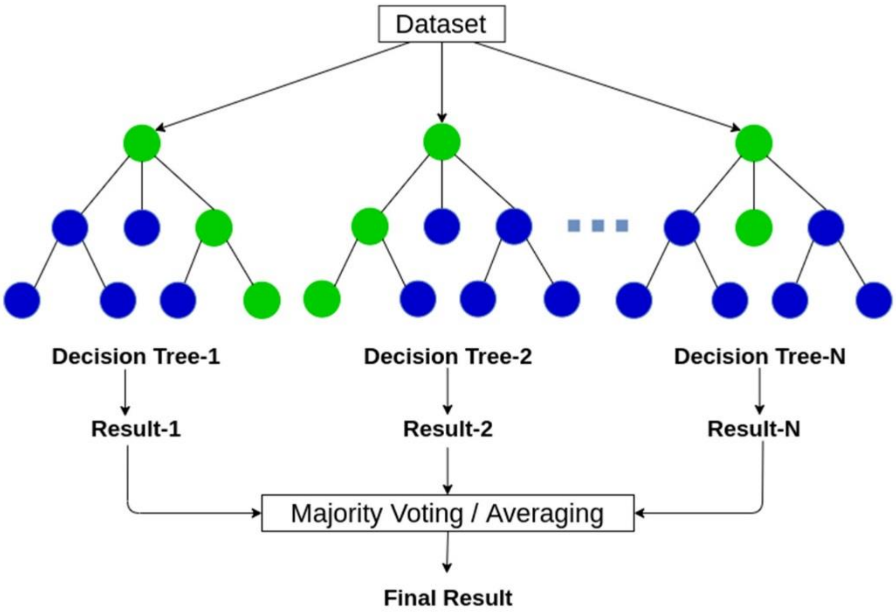


Fig4.4

Advantages Of Random Forest Algorithm

- Robust performance on a variety of datasets.
- Handles high-dimensional data well.
- Provides feature importance rankings.
- Reduces the risk of overfitting compared to individual decision trees.

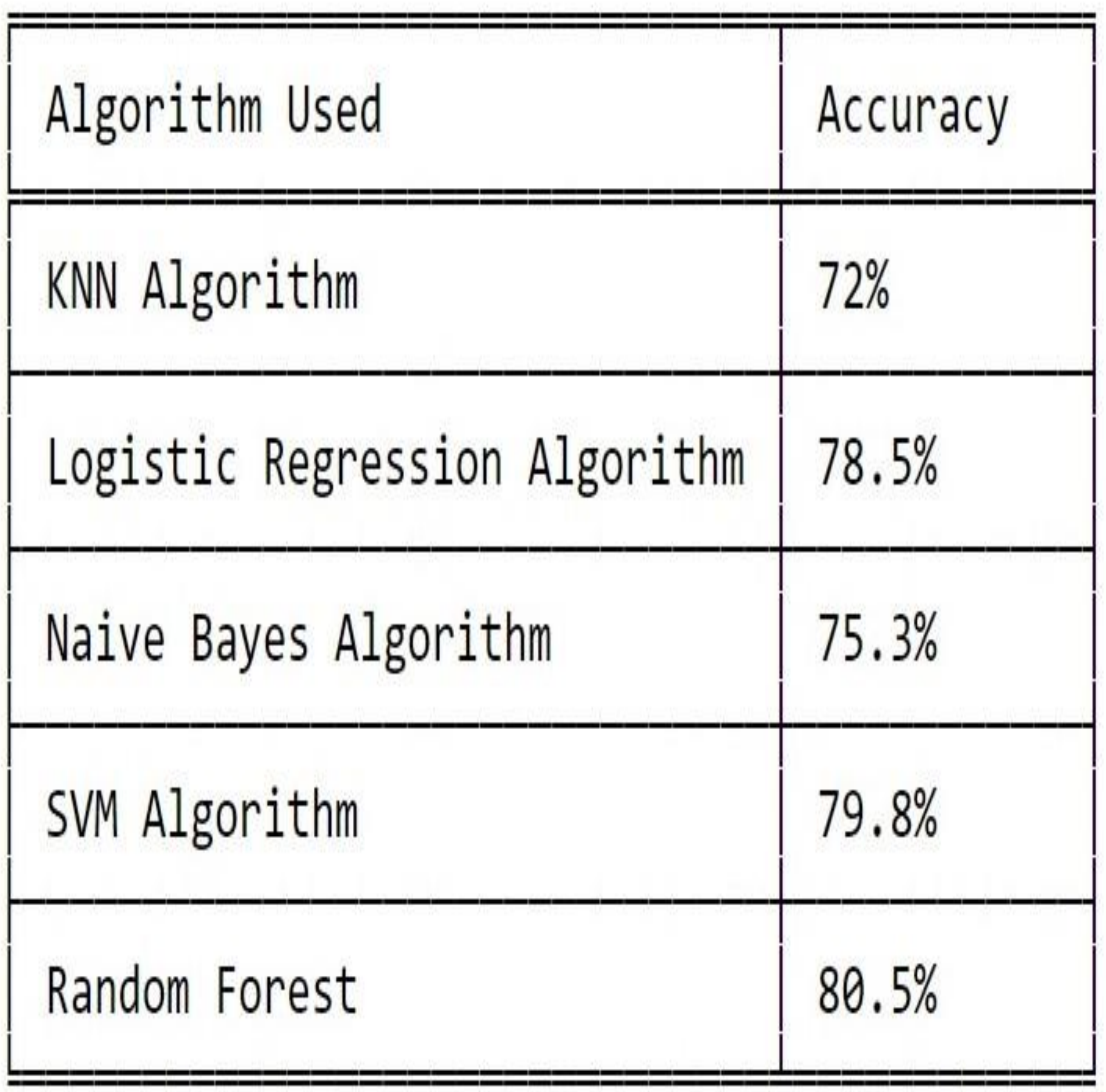
Disadvantages Of Random Forest Algorithm

- May be computationally expensive.
- Less interpretable compared to individual decision trees.
- Not suitable for real-time applications where prediction speed is critical.

Chapter-6
Results and Discussion

6.1 Results

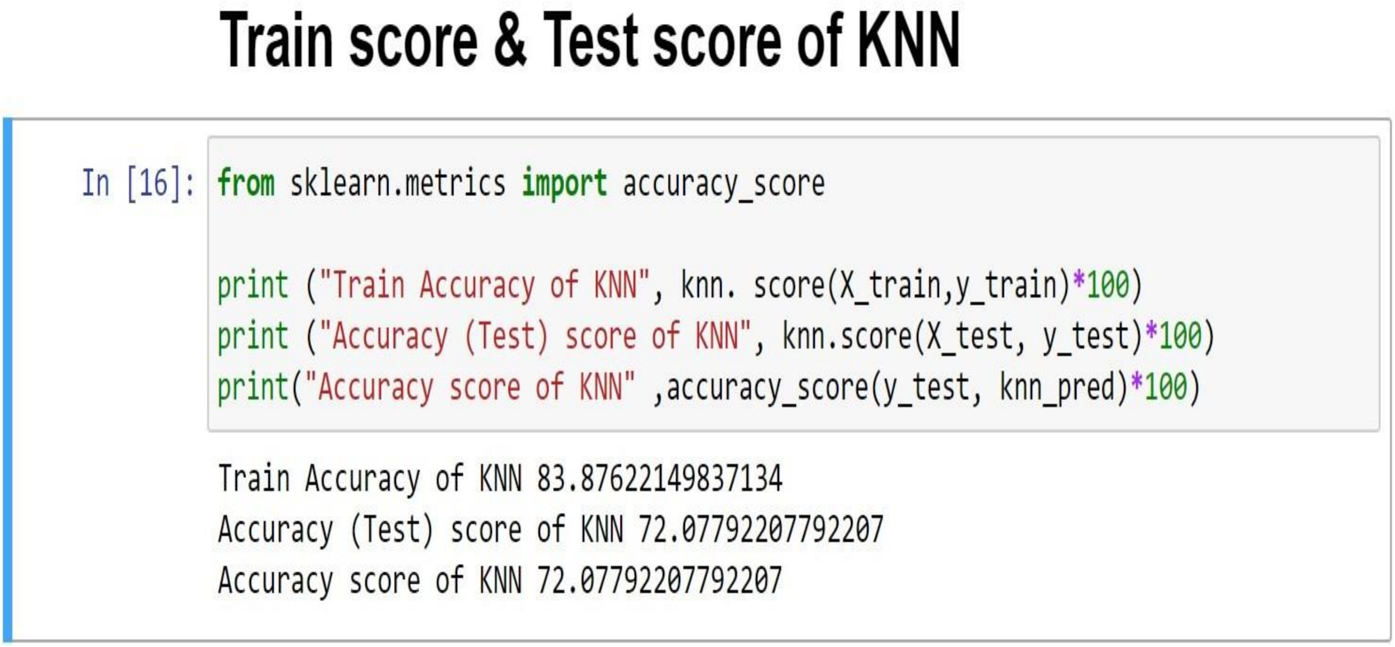
The performance of five different machine learning algorithms, K-Nearest Neighbor (KNN) Algorithm, Logistic Regression, Support Vector Machine (SVM) Algorithm, Naïve Bayes Algorithm and Random Forest Algorithm was evaluated on the dataset. The evaluation of the models' performance was based on the accuracy.
The results are summarized as follows:



Algorithm Used	Accuracy
KNN Algorithm	72%
Logistic Regression Algorithm	78.5%
Naive Bayes Algorithm	75.3%
SVM Algorithm	79.8%
Random Forest	80.5%

Fig6.0

6.1.1 Train Score & Test Score of KNN Algorithm- 72% Accuracy



Train score & Test score of KNN

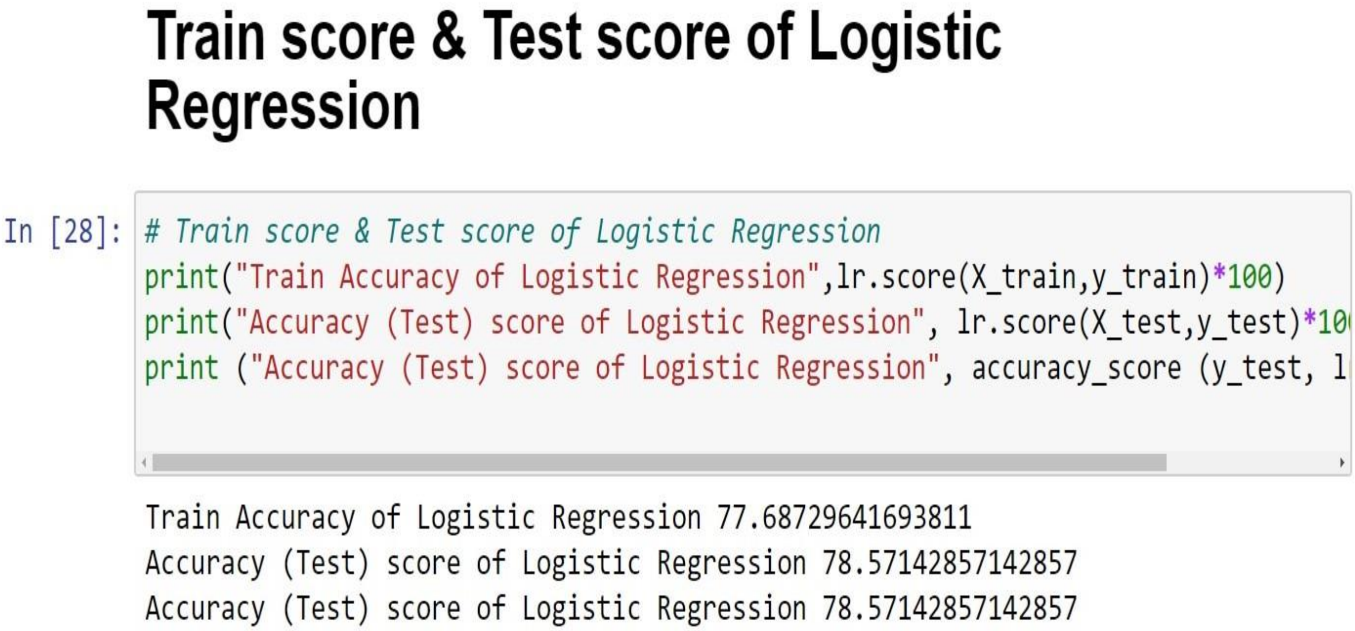
```
In [16]: from sklearn.metrics import accuracy_score

         print ("Train Accuracy of KNN", knn. score(X_train,y_train)*100)
         print ("Accuracy (Test) score of KNN", knn.score(X_test, y_test)*100)
         print("Accuracy score of KNN" ,accuracy_score(y_test, knn_pred)*100)

         Train Accuracy of KNN 83.87622149837134
         Accuracy (Test) score of KNN 72.07792207792207
         Accuracy score of KNN 72.07792207792207
```

Fig6.1

6.1.2 Train Score & Test Score of Logistic Regression Algorithm- 78.5% Accuracy



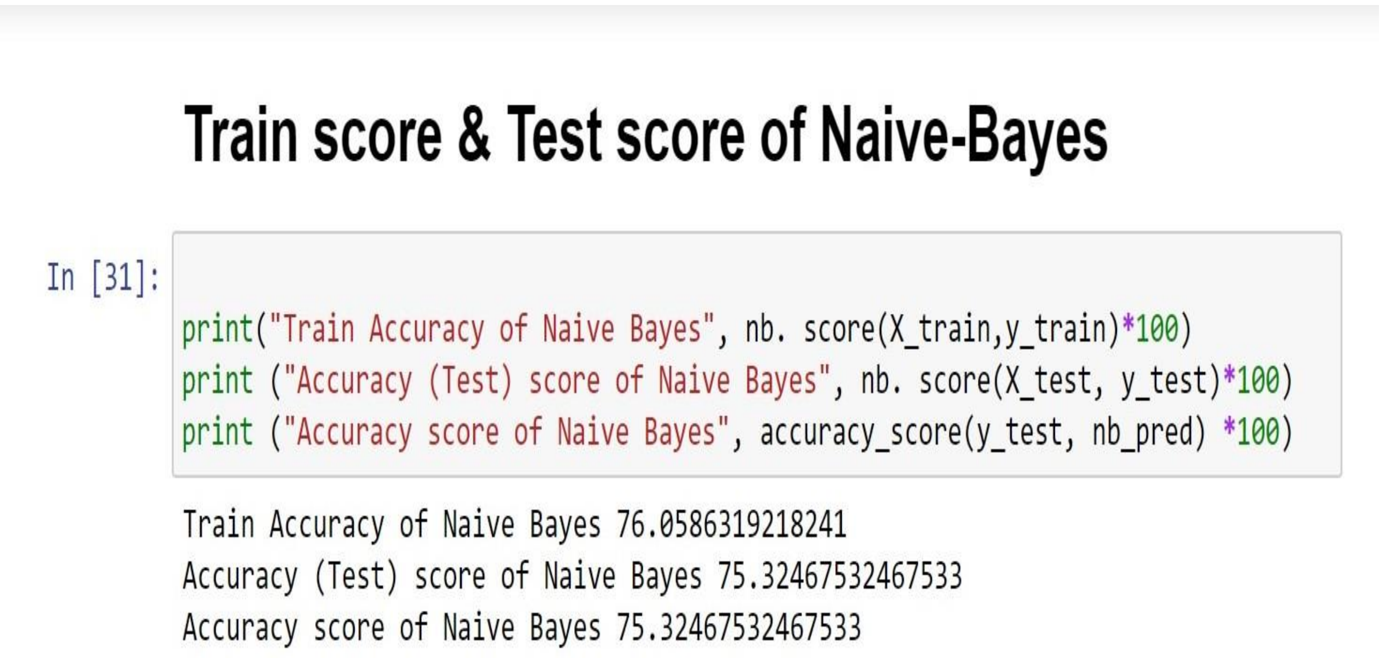
Train score & Test score of Logistic Regression

```
In [28]: # Train score & Test score of Logistic Regression
         print("Train Accuracy of Logistic Regression",lr.score(X_train,y_train)*100)
         print("Accuracy (Test) score of Logistic Regression", lr.score(X_test,y_test)*10
         print ("Accuracy (Test) score of Logistic Regression", accuracy_score (y_test, l

         Train Accuracy of Logistic Regression 77.68729641693811
         Accuracy (Test) score of Logistic Regression 78.57142857142857
         Accuracy (Test) score of Logistic Regression 78.57142857142857
```

Fig6.2

6.1.3 Train Score & Test Score of Naïve Bayes Algorithm- 75.3%



Train score & Test score of Naive-Bayes

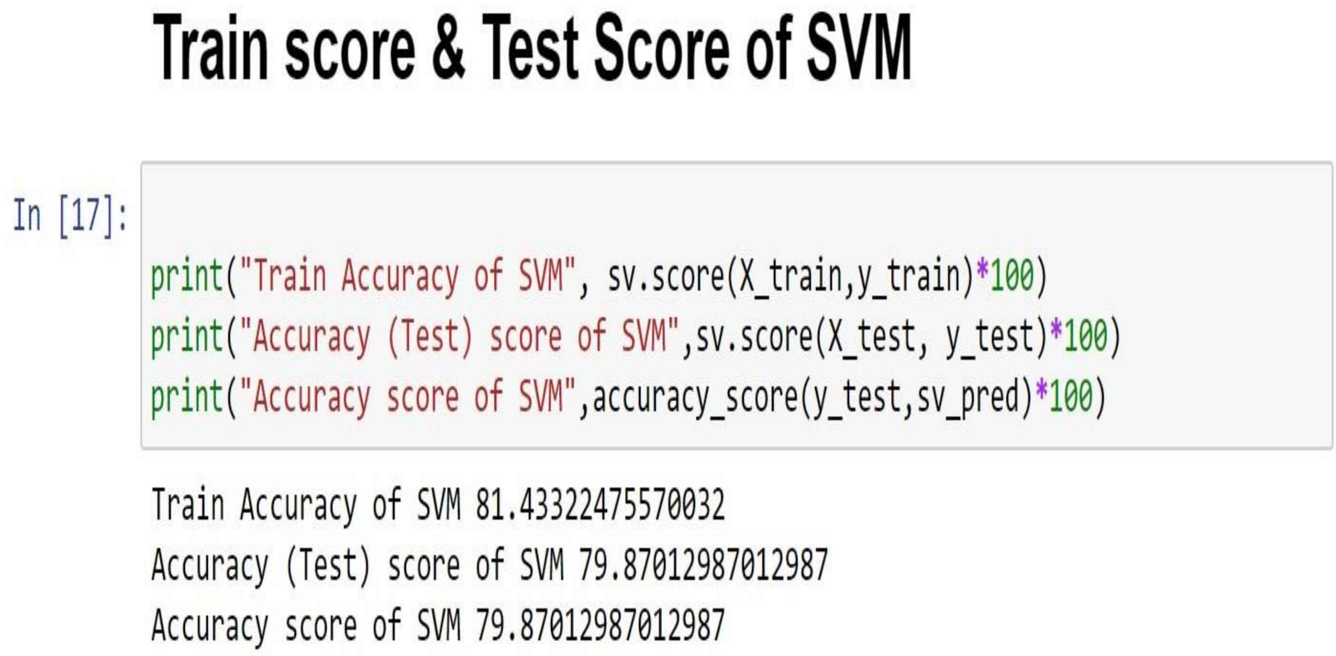
In [31]:

```
print("Train Accuracy of Naive Bayes", nb. score(X_train,y_train)*100)
print ("Accuracy (Test) score of Naive Bayes", nb. score(X_test, y_test)*100)
print ("Accuracy score of Naive Bayes", accuracy_score(y_test, nb_pred) *100)
```

```
Train Accuracy of Naive Bayes 76.0586319218241
Accuracy (Test) score of Naive Bayes 75.32467532467533
Accuracy score of Naive Bayes 75.32467532467533
```

Fig6.3

6.1.4 Train Score & Test Score of Support Vector Machine (SVM) Algorithm- 79.8%



Train score & Test Score of SVM

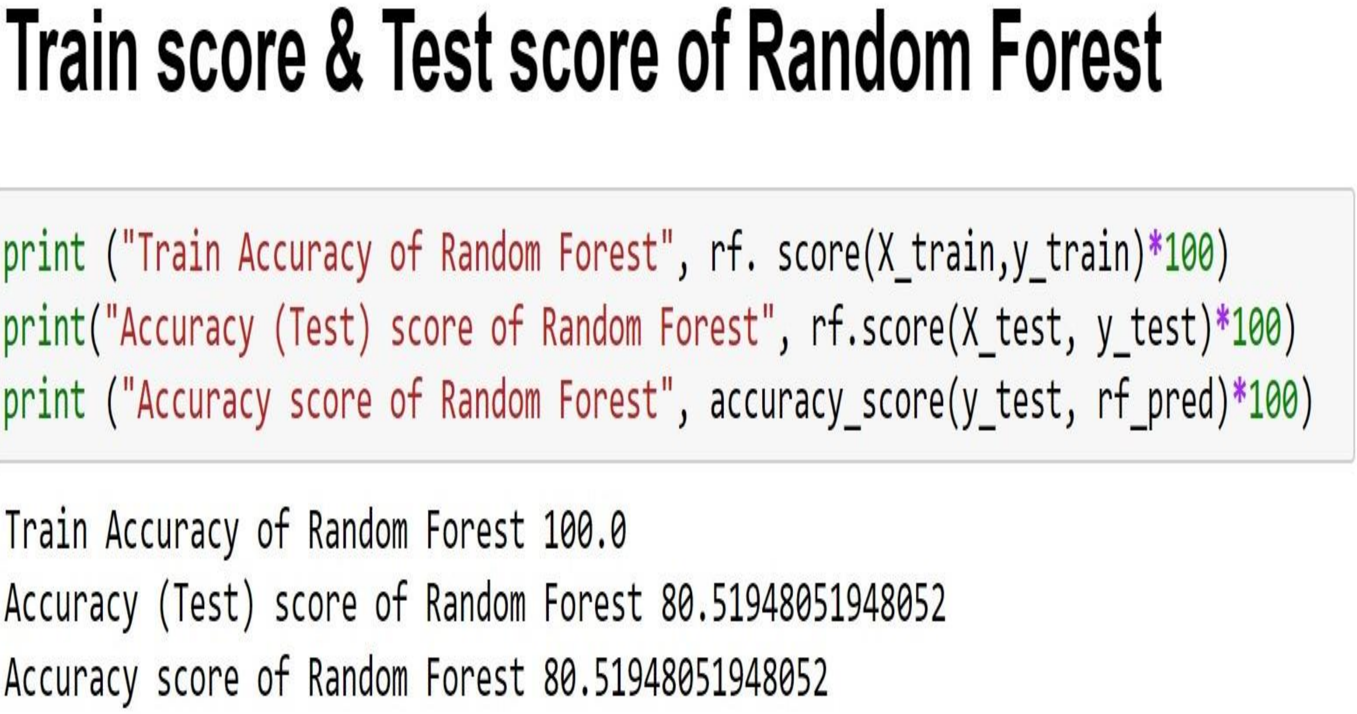
In [17]:

```
print("Train Accuracy of SVM", sv.score(X_train,y_train)*100)
print("Accuracy (Test) score of SVM",sv.score(X_test, y_test)*100)
print("Accuracy score of SVM",accuracy_score(y_test,sv_pred)*100)
```

```
Train Accuracy of SVM 81.43322475570032
Accuracy (Test) score of SVM 79.87012987012987
Accuracy score of SVM 79.87012987012987
```

Fig6.4

6.1.5 Train Score & Test Score of Random Forest Algorithm- 80.5% (BEST)



Train score & Test score of Random Forest

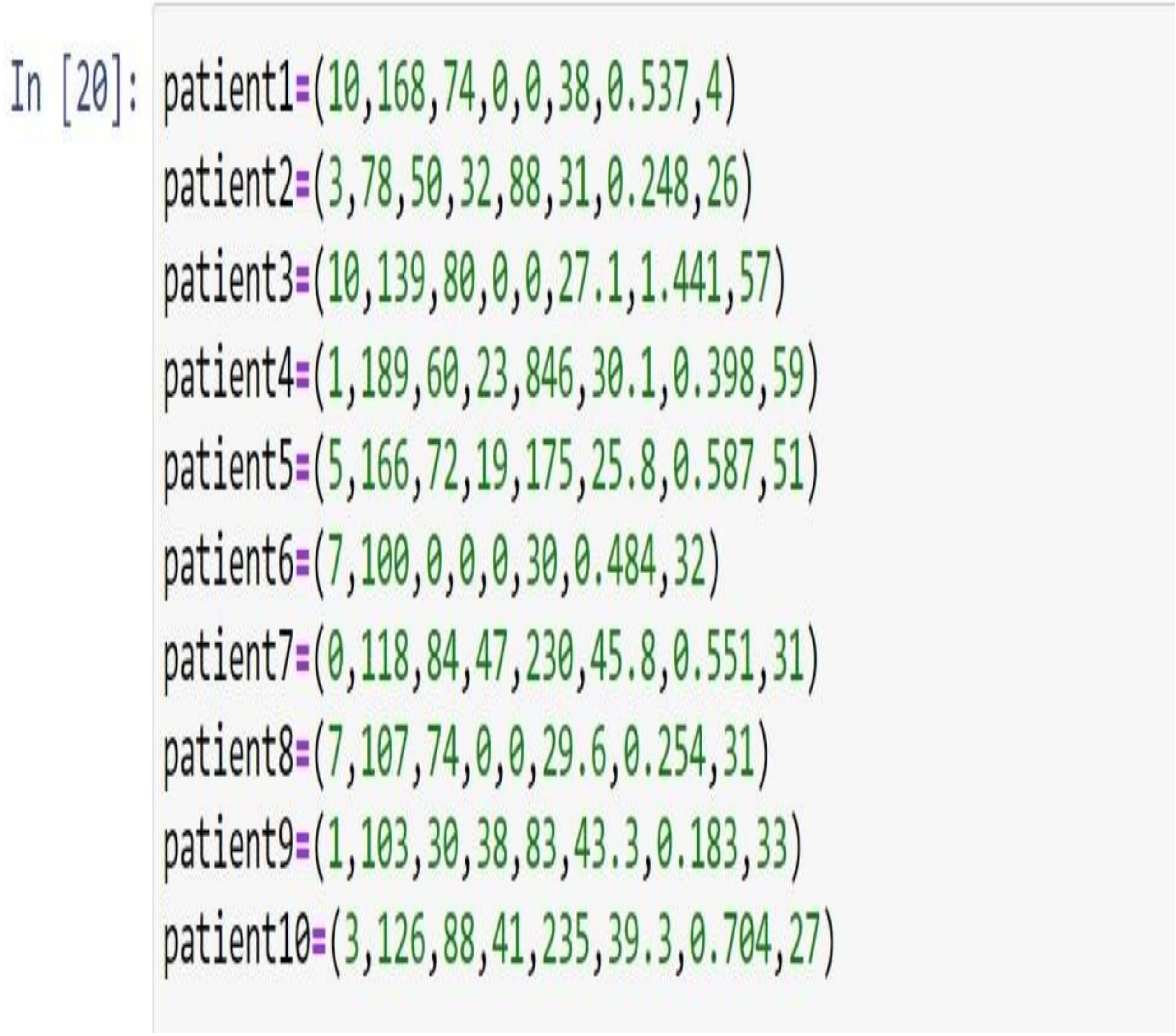
```
print ("Train Accuracy of Random Forest", rf. score(X_train,y_train)*100)
print("Accuracy (Test) score of Random Forest", rf.score(X_test, y_test)*100)
print ("Accuracy score of Random Forest", accuracy_score(y_test, rf_pred)*100)
```

```
Train Accuracy of Random Forest 100.0
Accuracy (Test) score of Random Forest 80.51948051948052
Accuracy score of Random Forest 80.51948051948052
```

Fig6.5

6.2 Predictive System

We have built a system where we have taken input of 10 different people’s attributes and predicted their possibility of having diabetes using Random Forest Algorithm.

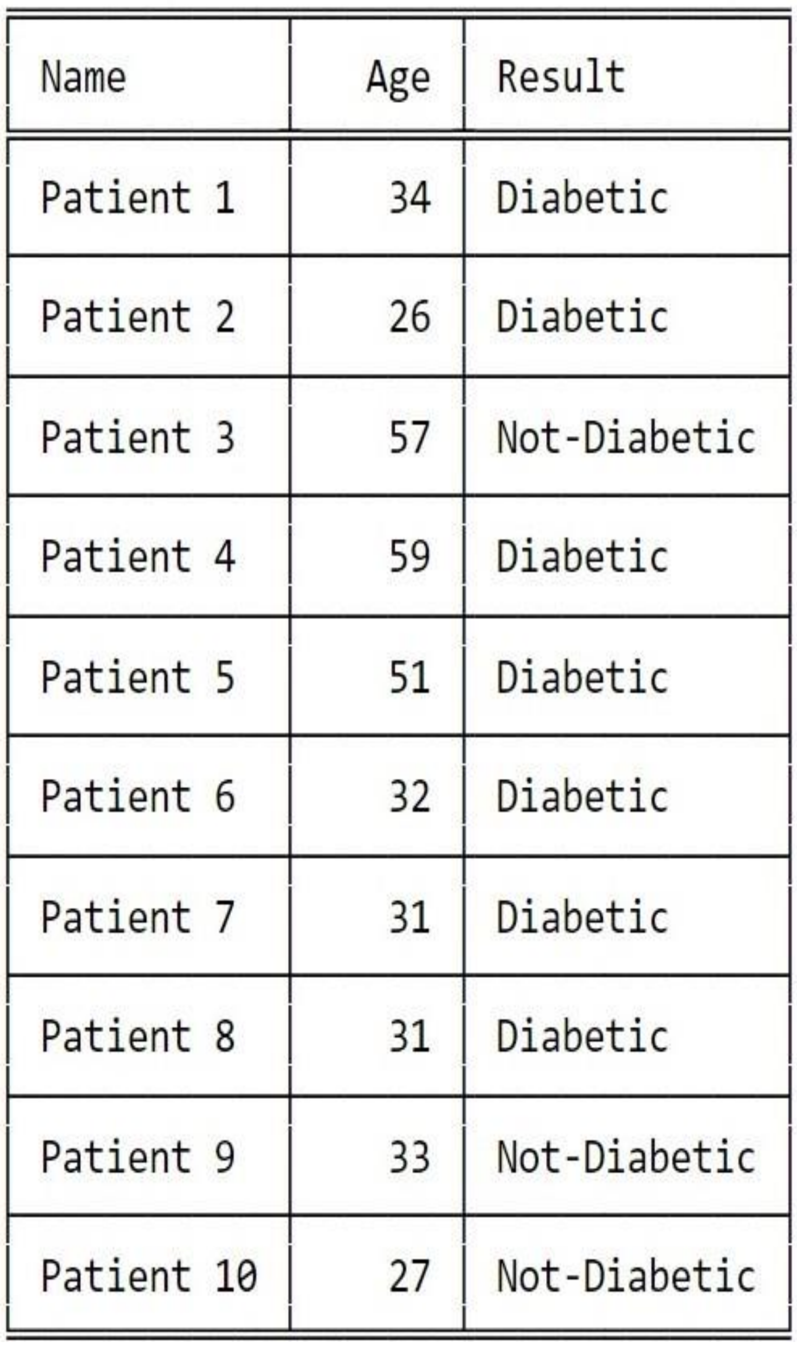


```
patient1=(10,168,74,0,0,38,0.537,4)
patient2=(3,78,50,32,88,31,0.248,26)
patient3=(10,139,80,0,0,27.1,1.441,57)
patient4=(1,189,60,23,846,30.1,0.398,59)
patient5=(5,166,72,19,175,25.8,0.587,51)
patient6=(7,100,0,0,0,30,0.484,32)
patient7=(0,118,84,47,230,45.8,0.551,31)
patient8=(7,107,74,0,0,29.6,0.254,31)
patient9=(1,103,30,38,83,43.3,0.183,33)
patient10=(3,126,88,41,235,39.3,0.704,27)
```

Fig6.6

6.3 Predicted Outcome

The predicted outcome is based on Random Forest Algorithm whose accuracy is 80.5%.



Name	Age	Result
Patient 1	34	Diabetic
Patient 2	26	Diabetic
Patient 3	57	Not-Diabetic
Patient 4	59	Diabetic
Patient 5	51	Diabetic
Patient 6	32	Diabetic
Patient 7	31	Diabetic
Patient 8	31	Diabetic
Patient 9	33	Not-Diabetic
Patient 10	27	Not-Diabetic

Fig6.7

6.4 Discussion

Diabetes prediction using machine learning has been a subject of extensive research, aiming to develop accurate models for early detection and management of diabetes. In this project we have used several different machine learning algorithms including K-Nearest Neighbor (KNN), Logistic Regression, Naïve Bayes, Support Vector Machine (SVM) and Random Forest where each of them has given high accuracy of above 70% with good training.

The choice of the algorithm often depends upon the dataset characteristics and the goals of the prediction task. We have consistently compared the performance of different algorithms to identify the most effective ones.

Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and Random Forest have demonstrated high accuracy in this study. Accuracy levels reported in this study ranges from approximately 70% to 81%, these accuracy levels indicate that the ability of the models to correctly classify individuals as diabetic or non-diabetic based on the given features.

The predictive model lets users to take input of 10 different individuals at a same time and predict their possibility of having diabetes at the same time with accuracy up to 81%.

Also, it proves that the model is more robust to noise and variations in input data. Robust models are valuable in real-world scenarios where the input data may contain uncertainties or errors.

Overall, we can say that low training and validation loss provide confidence in the model's ability to generalize, make accurate predictions, prevent overfitting, handle noisy data, and converge efficiently.

Chapter-7
Conclusion and Future Scope

7.1 Conclusion

Diabetes can be a reason for reducing life expectancy and quality. Predicting this chronic disorder earlier can reduce the risk and complications of many diseases in the long run. In this paper, we have created a diabetes prediction system using various machine learning approaches. The open-source Pima Indian dataset have been used in this work. This paper reported accuracy of various different machine learning techniques. The Random Forest Algorithm achieved the best performance with 80.5% accuracy. Next, the technique has been applied to demonstrate the versatility of the proposed prediction system. Finally, the best-performed Random Forest Algorithm has been deployed into a system to predict diabetes instantly.

The proposed model can be a valuable tool in assisting in the early detection and management of diabetes, leading to better patient outcomes and reduced risk of diabetes. These findings highlight the importance of dataset-specific analysis and model selection for classification tasks. It is crucial to consider the characteristics, class distributions, and complexities of the dataset when choosing an appropriate model. The results provide valuable insights into the strengths and limitations of the evaluated models and emphasize the need for further research and development in the field of machine learning for classification tasks.

7.2 Future Scope

As a future direction, further refinements and optimization of the model can be made to improve the model’s performance, such as Integration of Wearable Devices, the incorporation of wearable devices, such as continuous glucose monitors and fitness trackers, can provide real-time data for more dynamic and personalized predictions. Machine learning models could leverage streaming data from wearables to enhance prediction accuracy and offer timely interventions.

Additionally, the model can be extended for Advanced Deep Learning Models, exploring advanced deep learning architectures, including neural networks and recurrent neural networks (RNNs), could lead to more sophisticated models capable of capturing intricate patterns in health data. Deep learning techniques may uncover hidden relationships in data that traditional machine learning models might overlook.

Reference

1. DeeptiSisodia, Dilip Singh Sisodia,"Prediction of Diabetes Using Classification Algorithm", www.elsevier.com/locate/procedia, Procedia computer science 132(2018) 1578-1585.

2. Xue-Hui Meng,Yi-Xiang Huang,Dong-PingRao,Qiug Liu,2013,"Comparison of Three Data Mining Models For Predicting Diabetes of Prediabetes By Rick Factos",Kaohsiung journal of medical science(2013) 29,93-99.

3. V.AnujaKumari, R.Chithra."Classification of Diabetes Disease Using Support Vector Machine".vol3.,Issue 2,March-April 2013,pp.1797-1801.www.ijera.com.

4. Monisha.A, S.ShalinChistina, Nirmala Santiago, "Decision support system for a chronic disease Diabetes". International Journal of Computer &Mathematical Science(IJCMS),ISSN 2347-8527 Volume 7,Issue 3,March 2018.

5. S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar, "Prediction of Diabetes Diagnosis

6. Using Classification Based Data Mining Techniques", International Journal of statistics and Systems,ISSN 0973-2675 Volume 12,Number 2(2017),PP.183-188.http://www.ripublication.com.

7. Aiswarya Iyar, S. Jeyalatha and RonakSumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

8. B.Tamilvanan, Dr.V.MuraliBhaskaran, "An Experimental Study of Diabetes Disease Prediction System Using Classification Techniques", IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 39-44, www.iosrjournals.org.

9. Rahul Joshi, MinyechilAlehegn,"Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble

approach",International Research Journal of Engineering and Technology (IRJET),Volume: 04 Issue:10 | Oct -2017,e-ISSN: 2395-0056,p-ISSN: 2395-0072. www.irjet.net.

10. Amina Azar,Yasir Ali, Muhammad Awais, KhurramZaheer,"Data Mining Models Comparison for Diabetes Prediction", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No.8, 2018.

11. VeenaVijayan.V, Anjali.C,"Decision Support Systems for Predicting Diabetes Mellitus –A Review", Proceedings of 2015 Global Conference on Communication Technologies(GCCT 2015), 978-1-4799-8553-1/15/$31.00 © s2015 IEEE.

12. DeepikaVerma , Dr.Nidhi Mishra, "Analysis and prediction of breast cancer and diabetes disease dataset using data mining classification techniques",Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS2017),IEEE Xplore Compliant – Part Number:CFP17M19-ART, ISBN:978-1-5386-1959-9.

13. VeenaVijayan V, Anjali c, "Prediction of diagnosis of diabetes mellitus –A machine learning approach ", 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10-12 December 2015 | Trivandrum.

14. Ayman Mir, SudhirN.Dhage, "Diabetes Disease Prediction using MachineLearning on Big Data of Healthcare", 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).

15. S M Hasan Mahmud, MdAltabHossin, Md. Razu Ahmed, SheakRashedHaiderNoori, MdNazirul Islam Sarkar, "Machine Learning Based Unified Framework for DiabetesPrediction", BDET 2018, August 25–27, 2018, Chengdu, China. © 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-6582 6/18/08…$15.00.DOI:https://doi.org/10.1145/3297730.3297737.

16. AakanshaRathore, Simran Chauhan, SakshiGujral, "Detecting and Predicting Diabetes Using Supervised.