Diabetes Prediction using Machine Learning

by

Subham Sinha (1900320100166)

Suyash Pratap Singh (1900320100173)

Shashank Pratap Singh (1900320100147)

Under the guidance of

MS. SHWETA ROY

Department of Computer Science and Engineering

ABES Engineering College

19th Km Stone, NH-09, Ghaziabad (U.P)

May, 2023

Diabetes Prediction using Machine Learning

by

Subham Sinha (1900320100166)

Suyash Pratap Singh (1900320100173)

Shashank Pratap Singh (1900320100147)

Submitted to the Department of Computer Science and Engineering

in partial fulfillment of the

requirementsfor the degree of

Bachelor of
Technology in

Computer Science and Engineering

Under the guidance of

MS. SHWETA ROY

Estd. 2000

ABES Engineering College, Ghaziabad

Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh Lucknow

May, 2023

STUDENT'S DECLARATION

We hereby declare that the work being presented in this report entitled *" **DIABETES PREDICTION USING MACHINE LEARNING** "* is an authentic record of ***our*** own work carried out under the supervision of Ms. **" SHWETA ROY " .**

The matter embodied in this report has not been submitted by *us* for the award of any other degree.

Date: 25/05/2023

Signature of students(s)

Subham Sinha

Suyash Pratap Singh

Shashank Pratap Singh

Department: Computer Science and Engineering

This is to certify that the above statement made by the candidate(s) is correct to the best of my knowledge.

Signature of HOD	Signature of Project Coordinator	Signature of Supervisor
Prof. (Dr.) Divya Mishra		Name: Ms.Shweta Roy
Designation: HOD - CSE		Designation: Professor
Computer Science and Engineering		Computer Science and Engineering

Date: 25/05/2023

CERTIFICATE

This is to certify that Project Report entitled " **DIABETES PREDICTION USING MACHINE LEARNING** " which is submitted by **Subham Sinha, Suyash Pratap Singh, Shashank Pratap Singh** in partial fulfillment of the requirement for the award of degree Bachelors of Technology in Department of Computer Science and Engineering of Dr. A.P.J. Abdul Kalam Technical University, formerly Uttar Pradesh Technical University is a record of the candidate's own work carried out by them under my supervision.

The plagiarism percentage evaluated for the content presented is 14 %.

The matter embodied in this Major Project Report is original and has not been submitted for the award of any other degree.

Supervisor Signature

Name: Shweta Roy

Designation: Professor

Date: 25/05/2023

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe a special debt of gratitude to Professor Ms. Shweta Roy, Department of Computer Science & Engineering, ABESEC, Ghaziabad for her constant support and guidance throughout the course of our work. Her sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only her cognizant efforts that our endeavors have seen the light of day.

We also take the opportunity to acknowledge the contribution of Professor (Dr.) Divya Mishra, Head, Department of Computer Science & Engineering, ABESEC Ghaziabad for her full support and assistance during the development of the project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature:

Name : SUBHAM SINHA

Signature:

Name : SUYASH PRATAP SINGH

Signature:

Name : SHASHANK PRATAP SINGH

ABSTRACT

Low insulin levels and high blood glucose levels in the body are the causes of diabetes. The symptoms of this raised blood sugar level include increased thirst, appetite, and frequency of urinating. Diabetes shouldn't be neglected since, if left untreated, it can have serious effects for a person, such as damage to the kidneys, heart, eyes, blood pressure, and other bodily organs. Diabetes may be controlled if it is discovered early. However, diabetes makes this procedure inefficient. The most prevalent types of diabetes are type 1 and type 2, but there are other varieties as well, including gestational diabetes, which appears during pregnancy. For a higher degree of accuracy, we will use a variety of machine learning techniques to predict early onset diabetes in a human body or patient. Machine learning techniques build models using patient datasets to improve the accuracy of predictions. A recent branch of data science called "machine learning" studies how computers pick up knowledge via knowledge.The goal of this effort is to create a system that can more accurately identify early diabetes in a patient by merging the findings of different machine learning approaches. K-Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression are some of the methodologies employed. The model's and each method's accuracy are calculated. The diabetes prediction model with the highest accuracy is then chosen. Random forest algorithm has the maximum efficiency among all the machine learning algorithms on which the data has been tested.

Keywords : *K-Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression*

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

[x]	Integer value of x.
$\neq$	Not Equal
$\in$	Belongs to

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
KNN	K - Nearest Neighbour
ML	Machine Learning
CNN	Convolutional Neural Network
HRV	Heart Rate Variability
WHO	World Health Organization
UCI	Unique Client Identifier

CHAPTER 1

INTRODUCTION

Medical experts and other healthcare providers typically refer to diabetes, a chronic illness, as diabetes mellitus. Insufficient insulin synthesis, inappropriate insulin cell response, or a combination of the two can all lead to high blood sugar levels, which is what this word refers to as a set of metabolic illnesses. As a result, the level of glucose in the blood will increase. Although certain cases of diabetes are difficult to classify, type 1 and type 2 cases of the disease may be loosely classified into two groups. If diabetes is not treated, it causes several negative side effects.

As a result, not only does it harm individuals, but it also causes heart failure, kidney difficulties, and blindness. Due to inadequate pancreatic insulin synthesis or a body's inability to use the insulin that is produced, diabetes develops when blood glucose levels rise. Elevated amounts of glucose (sugar) in the blood and urine are signs of diabetes mellitus.

Types of Diabetes

Type 1 : Type 1 diabetes patients have a weakened immune system and decreased insulin synthesis in their cells. There are presently no proven preventative measures or therapies for type 1 diabetes, nor is it known with certainty what the causes are.

Type 2 : Either inadequate insulin production by cells or incorrect insulin usage by the body are characteristics of diabetes. 90% of people with diabetes have this kind, making it the most common type. Genetics and dietary habits both have a role in its occurrence.

Gestational Diabetes: High fasting blood sugar levels in pregnant women are a risk factor for gestational diabetes. In two-thirds of the cases, it will return during subsequent pregnancies. There is a considerable chance that type 1 or type 2 diabetes will manifest after a pregnancy in which gestational diabetes was present.

Symptoms of Diabetes -

• Frequently Urination • Increased thirst • Tired / Sleepiness • Weight loss

• Blurred vision • Confusion and difficulty concentrating • Mood Swings

Causes of Diabetes

Genetics is primarily to blame for diabetes. It is caused by at least two chromosome 6 genes that are faulty and change how the body responds to certain antigens. Viral infection has the ability to influence the development of type 1 and type 2 diabetes. According to studies, carrying viruses such the CMV, mumps, rubella, or hepatitis B virus increases the likelihood of acquiring diabetes.

Diabetes is now one of the leading causes of illness and death in the vast majority of countries. This number is expected to approach 642 million by 2040, according to the International Diabetes Federation; as a result, early screening and identification of diabetes patients is essential for early detection and effective treatment. Due to the nonlinear, atypical, correlation-structured, and complex character of the majority of medical data, analysing diabetic data can be difficult.

1.1. Problem Introduction

1.1.1. Motivation

By 2020, 463 million people globally, including 88 million in Southeast Asia, are expected to get diabetes, according to the International Diabetes Federation (IDF).
These 88 million individuals include 77 million Indians. According to the IDF, 8.9% of people have diabetes. In terms of the prevalence of type 1 diabetes among children, India is second only to the United States, according to IDF estimates.

India now has 65 million diabetics, up from 26 million in 1990. The prevalence was determined to be 11.8% among those over 50, according to the Ministry of Health and Family Welfare's report on the 2019 National Diabetes and Diabetic Retinopathy Survey. According to the DHS study, 6.5% of those under 50 have diabetes, and 5.7% have prediabetes. Both the male (12%) and female (11.7%) groupings were equally frequent.

It was higher in cities. Testing revealed that the sight-threatening condition diabetic retinopathy was present in 16.9% of diabetics up to the age of 50. According to the survey, people aged 60 to 69, 70 to 79, and those beyond the age of 80 were most likely to have diabetic retinopathy (18.6%, 18.3%, and 18.4%, respectively). For people aged 50 to 59, the incidence was 14.3% lower.

In India, type 2 diabetes patients fall into four subgroups or clusters, of which two are peculiar to that nation. These categories may face varying levels of problem risk and require different treatments.

Women are the ones who are most affected, however children and young people account for the bulk of instances that have been reported. We have decided to work on a machine learning-based diabetes detection tool in view of these worrying figures.

1.1.2. Project Objective

The objective is to transform the desired outcome into a measurable and manageable goal.

Find answers by coming up with machine learning concepts (how to address the issue and accomplish the desired result). First comes divergent reasoning, then follows convergent reasoning.

The main goal is to develop and test many machine learning models, assess their precision, and choose the best and most precise one among them to recognize diabetes in a person based on particular traits and attributes.

1.1.3. Scope of the Project

The process of scoping involves detailing a project and choosing the resources that will be utilized to finish it. There is more to planning than just that, though. Additionally, you must formulate the right queries, determine the objectives of your business, and then match those objectives with machine learning solutions. The first and generally regarded as the most important stage of a machine learning project's overall process is scoping.

Around 350 million people will have diabetes globally by 2030, and 642 million will by 2040, predicts the World Health Organization (WHO).

In order to minimize the diabetes pandemic that has befallen humanity, the scope involves creating machine learning models and testing them to see which ones are the most accurate to utilize in real-world circumstances.

1.2. Related Previous Work

A great deal of research has been conducted on the non-invasive automated detection of diabetes using machine learning approaches. Utilizing the procedures of feature extraction, feature selection, and classification, machine learning was put into practise. There were various studies that differed in the classifiers used and the extracted characteristics. Additionally, it was shown that standard machine learning algorithms performed poorly on important AI tasks like speech recognition and object identification, mostly due to the amount of the data they had to handle.

The inadequacies of machine learning encouraged the development of deep learning research. Deep learning has further uses in the medical field. A considerable number of new studies have been published recently, particularly in the field of healthcare anomaly detection. Deep learning methods were used to make the diabetes diagnosis, and the accuracy level that resulted was about equivalent to the highest level of automated diabetes detection accuracy at the time. In the aforementioned study, we classified diabetes with a 95.7% accuracy rate. The most significant studies on the automated, noninvasive diagnosis of diabetes using HRV are compiled in Table 1.

Table 1: Works on the automated non-invasive detection of diabetes using HRV

Methods	**Accuracy obtained (in %)**
Nonlinear	86.0
Higher order spectrum	90.5
Higher order spectrum	79.93
Nonlinear	90.0
Discrete wavelet transform	92.02
Empirical mode decomposition	95.63
Deep learning (CNN - LSTM)	95.1
Deep learning	95.7

CHAPTER 2

LITERATURE SURVEY

K. Vijiya Kumar [1] To more precisely predict a patient's risk of acquiring diabetes early on, K. developed a machine learning system that makes use of the Random Forest algorithm. The results demonstrated the prediction system's ability to accurately, quickly, and most importantly, efficiently anticipate the diabetes condition. Following the usage of five commonly utilized classifiers for the ensembles, the findings were integrated using a meta-classifier. The outcomes are displayed and contrasted with findings from earlier studies that made use of the same dataset. It has been demonstrated that the proposed method can more precisely predict when diabetes would begin.

Aishwarya [2] attempts to create techniques to diagnose diabetes by researching and analysing the patterns that appear in the data through classification analysis using Decision Tree and Naive Bayes algorithms.. The study's goal is to develop a faster and more accurate means of disease diagnosis, which will aid in patients' quick treatment. Using a 70:30 split, the PIMA dataset, and a cross validation procedure, the study revealed that the J48 method achieves an accuracy rate of 74.8%, while the naive Bayes method achieves an accuracy rate of 79.5%.

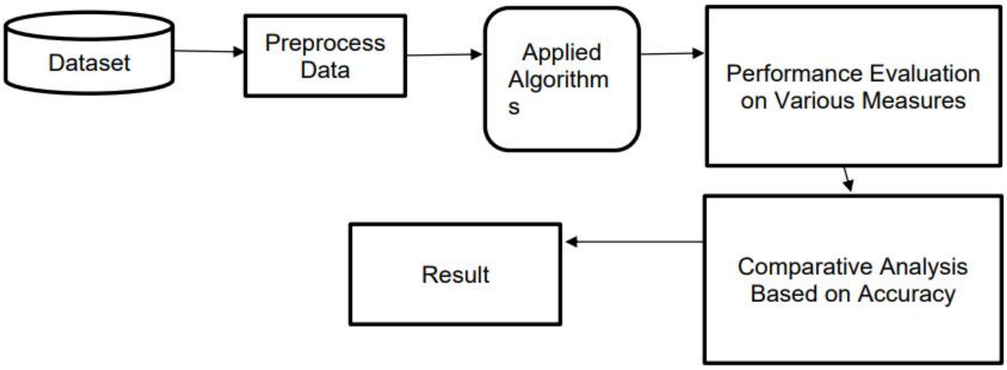
Lee et al. [3] Focus on using the CART decision tree algorithm on the diabetes dataset after the data has been processed using the resample filter. The author emphasises the need of fixing the class imbalance problem before using any technique to increase accuracy rates. Class imbalance is more common in datasets with dichotomous values, which demonstrate the existence of a class variable with two alternative outcomes. If this imbalance is identified earlier during the data pre-processing stage, the prediction model's accuracy will increase.

CHAPTER 3

SYSTEM DESIGN AND METHODOLOGY

3.1. System Design

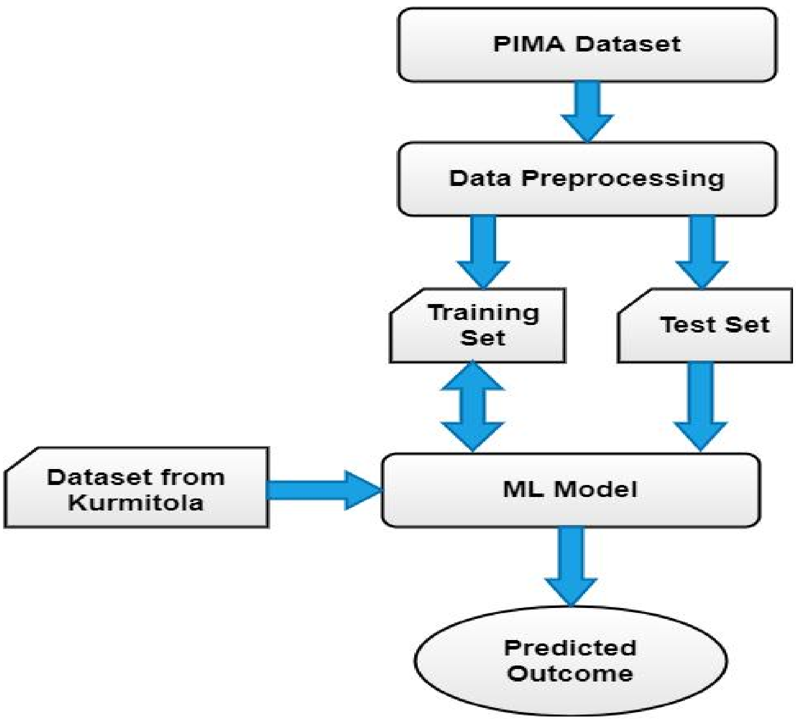
3.1.1 System Architecture



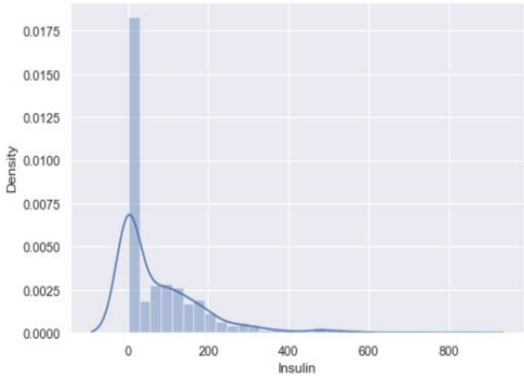
Proposed Model Diagram

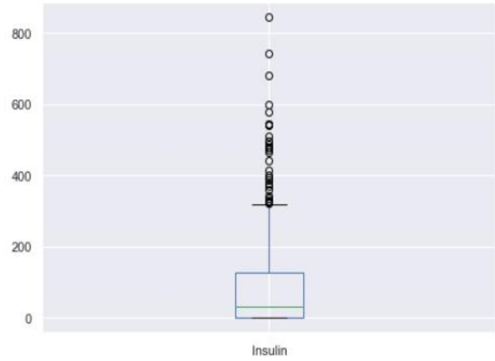
IV. RESULT & DISCUSSION

3.1.2 Data Flow Diagram

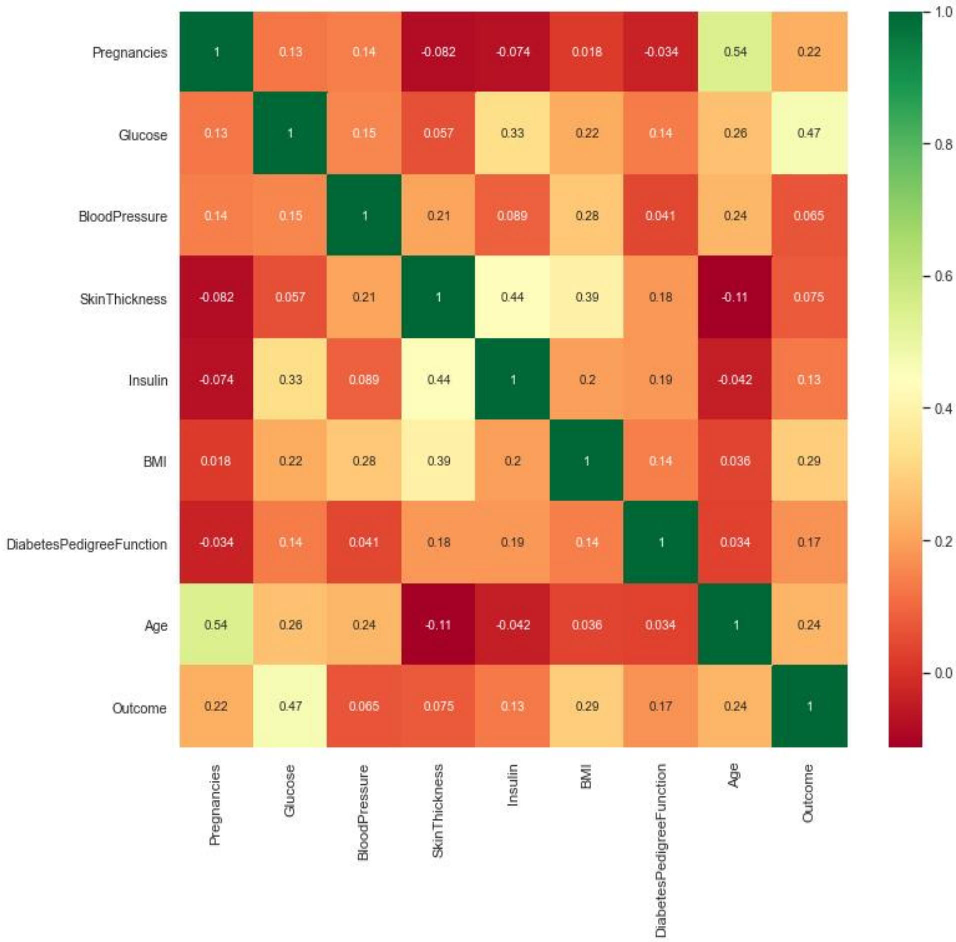


3.1.3. Box and Whiskers Plot

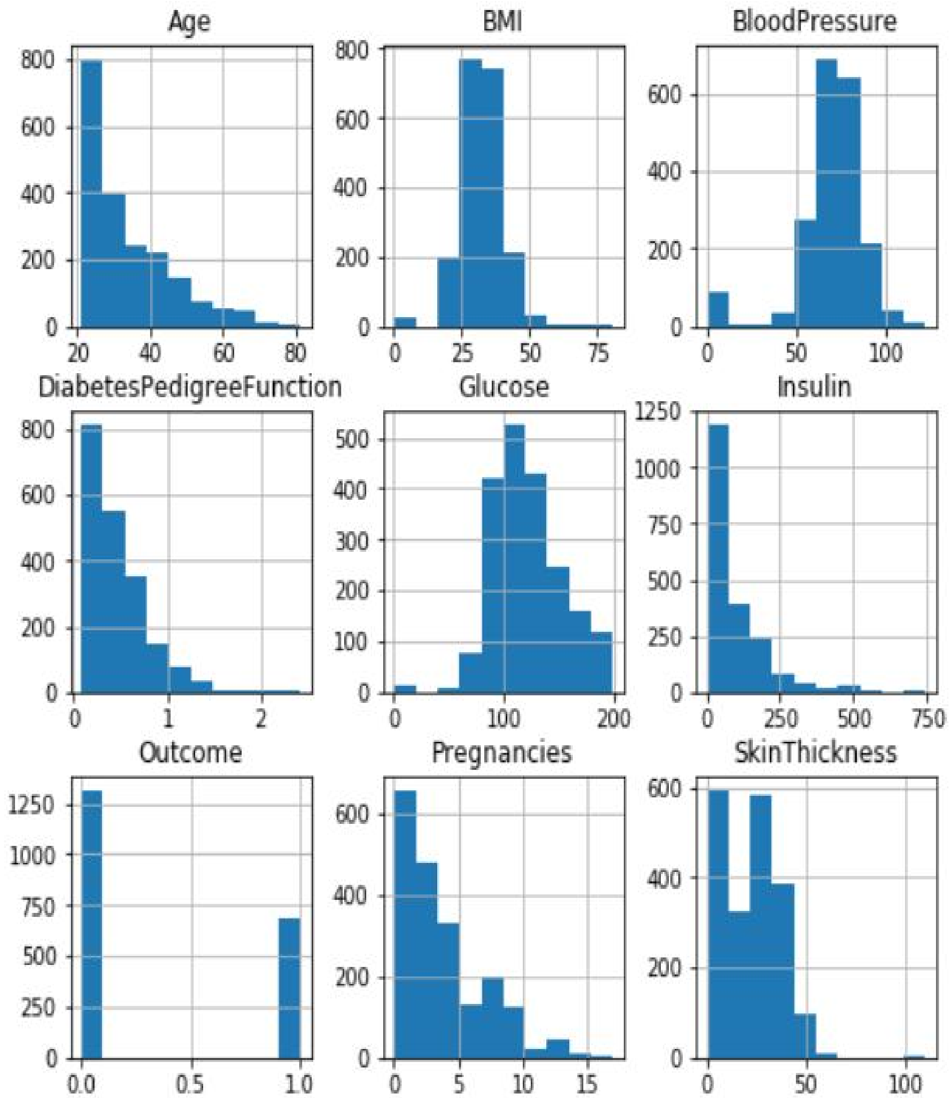




3.1.4. Corelation Matrix



3.1.5. Histogram



Dataset Description

The data was discovered in the UCI Pima Indian Diabetes Dataset repository. There is a lot of data in the collection regarding 768 patients.

The ninth characteristic for each data point is the class variable. This class variable indicates if the result is positive or negative for diabetes by showing the result for diabetics (0 or 1).

Distribution of Diabetic Patients: Despite the fact that we developed a model to predict diabetes, the dataset had 268 classes that had the label "1 indicates positive" and 268 classes that had the label " 2 means negative."

Table 2 : Description of the Dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

- There are 2000 data points in the diabetes data collection, each with nine attributes.
- We will forecast a characteristic called "Outcome," where 0 indicates no diabetes and 1 indicates diabetes.

3.2. Algorithm(s)

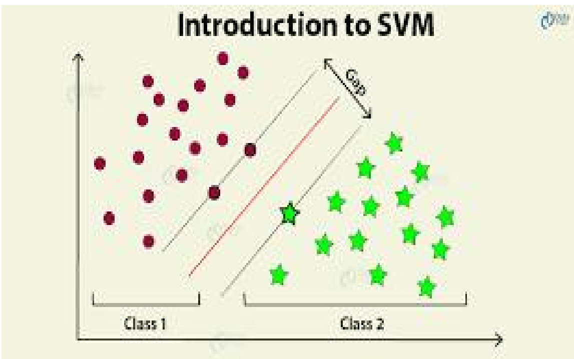
Apply machine learning: When the data is ready, the machine learning technique is used. To forecast diabetes, we employ a variety of ensemble and classification algorithms. the processes used to analyse the diabetes dataset among Pima Indians. The main goal is to use ML techniques to analyse the effectiveness of various approaches, evaluate their accuracy, and identify the critical variable that influences prediction.

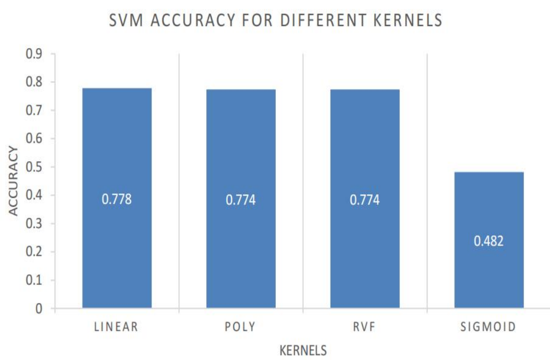
The Techniques are follows -

1. **Support Vector Machine** - SVM stands for support vector machine, which is a method of supervised machine learning. The most used classification approach is SVM. SVM creates a hyperplane that divides two classes. SVM can distinguish between samples in particular classes and categorise objects for which no supporting data is available. A hyperplane is used to locate the nearest training site for each class for separation.

Algorithm-

- Choose the hyperplane that best divides the class.
- To determine the best hyperplane, you must compute the Margin, which is the distance between the planes and the data.
- Choose the class with the highest margin. Margin equals the distance between the positive and negative points.

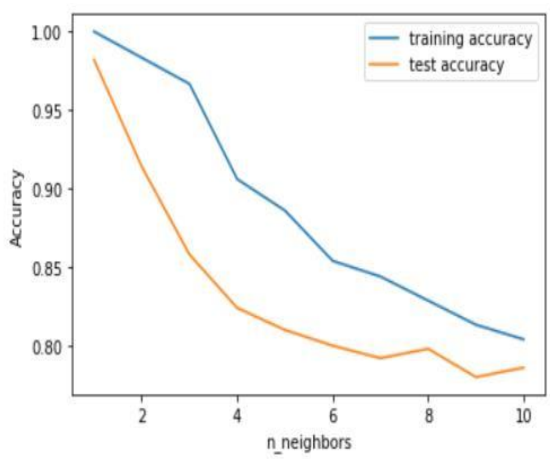




2. **K - Nearest Neighbor-** KNN is a supervised machine learning method that is distinct from others. KNN aids in the resolution of classification and regression difficulties. KNN is a slack prediction approach. According to KNN, similar objects should be found near to one another. Close proximity between similar data points is commonly seen. KNN provides assistance in categorizing new work using a similarity metric. The distance between the places is calculated using a tree-like structure. To predict a new data point, the approach determines the closest training data points. K stands for "number of near neighbours," is always a positive integer in this context. A class value is selected for neighbours from a list of class values.

Algorithm-

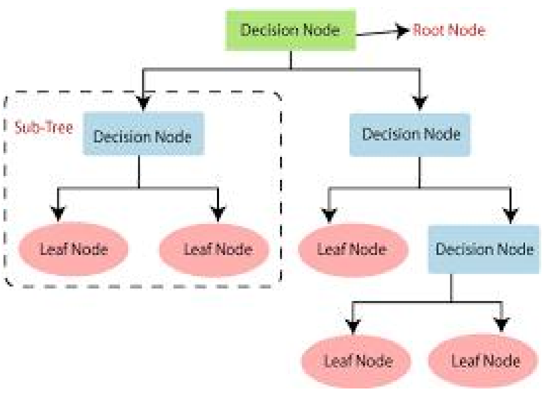
- Check out the Pima Indian Diabetes data collection, an example dataset with rows and columns.
- Think of a test dataset that has characteristics and rows.
- The number of closest neighbours, K, should then be chosen at random.
- Then, using these minimal distances and Euclidean distance, each is calculated to the nth column.
- Discover the identical output values.
- The patient is diabetic if the levels are the same; otherwise, the patient is not.

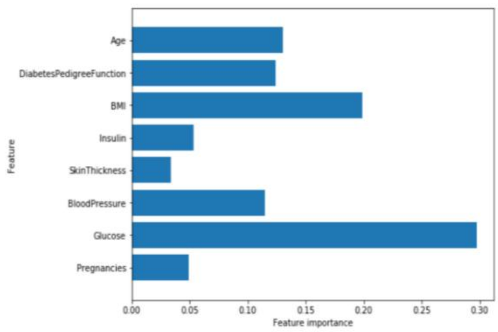


3. **Decision Tree** - A significant categorising tool is the decision tree. It is a method of supervised learning. When the response variable is categorical, a decision tree is utilised. A decision tree is a tree-like architecture that selects categorisation depending on input characteristics. Input variables might be text, discrete, continuous, or graph.

Steps for Decision Tree Algorithm-

- Build a tree using nodes as input features.
- Choose the feature with the best information gain to forecast the output from the input feature.
- For each characteristic in each tree node, the greatest information gain is determined.
- Repeat step 2 to create a sub-tree utilising the feature that was not utilised in the previous node.



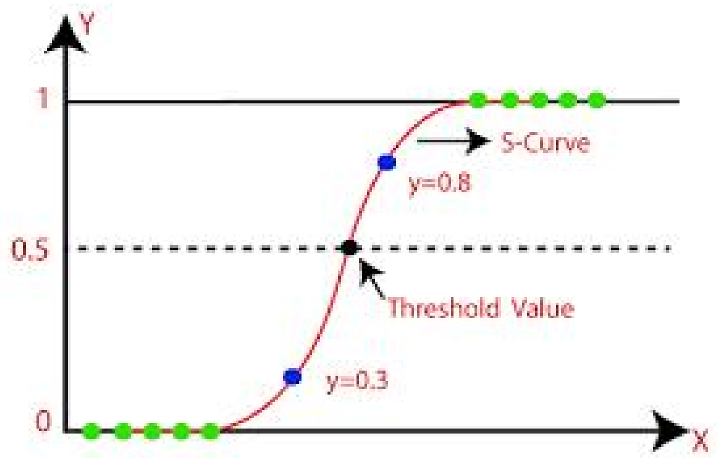


4. **Logistic Regression** -Logistic regression is yet another classification approach used in supervised learning. A binary response's propensity to be influenced by one or more predictors is assessed using this method. Discrete and continuous ones are both feasible. We utilise logistic regression to categorise or divide particular data points into groups.

Only the numbers 0 and 1 are used to classify the data in binary form, indicating whether or not a patient has diabetes.Logistic regression's main goal is to get the optimal fit, which best reflects the connection between the target and predictor variables. On top of the model for linear regression, logistic regression is constructed. To forecast the probability of the positive and negative classes, the logistic regression model uses the sigmoid function.

P = 1/1+e - (a+bx) Sigmoid function P denotes probability, a and b denotes Model parameters.

Ensembling - A machine learning strategy is being developed. Numerous learning algorithms are blended in an ensemble to accomplish a certain goal. It is utilised because it predicts more accurately than any other model. Noise bias and variation are the primary drivers of inaccuracy, and ensemble techniques assist in minimising or reducing these errors. Two popular ensemble algorithms include voting, averaging, ada-boosting, bagging, and gradient boosting. In this work, we employed the Gradient Boosting Ensemble and Bagging (Random forest) approaches to detect diabetes.



5. **Random Forest** - It is an ensemble learning approach that is used in classification and regression applications. It is more accurate than previous models. This method can easily handle large datasets. Leo Bremen created Random Forest. It is a well-liked technique for group learning. By lowering variation, Random Forest enhances Decision Tree performance. The class that reflects the average of all classes, classifications, or average predictions (regressions) of all trees is formed after a large number of decision trees have been built during training.

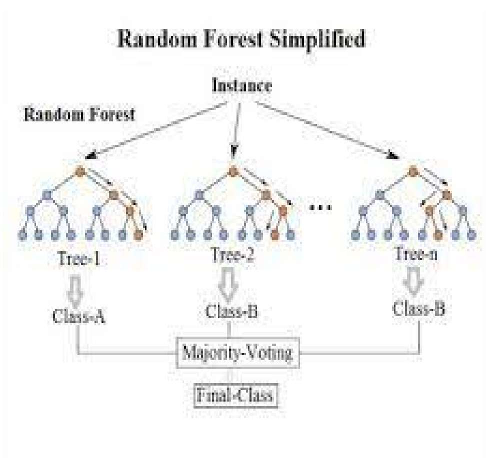
Algorithm-

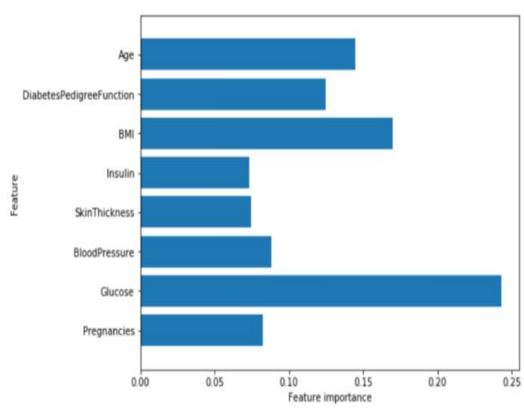
- Picking the R features where R>M from the total set of features is the first step.
- The node utilising the optimal split point among the R characteristics.
- Using the best split, divide the node into sub nodes.
- Repeat steps a through c until the l th node is reached..

Repetition of steps a through d n times produced the n trees that made up the forest.

The Gin-Index Cost Function is used by the random forest to determine the best split and is made available via:

Options are thought about, results are projected using the bases of each decision tree that was produced at random, and the projected outcomes are stored at intervals around the desired location in the first step. For a number of applications, Random Forest provides a wide range of techniques that deliver precise forecasts.

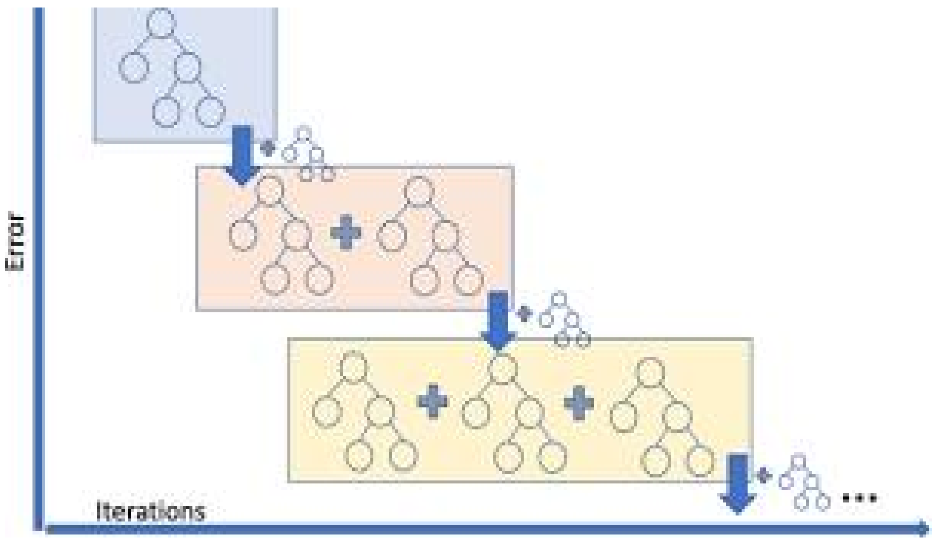




6. **Gradient Boosting** – The most effective ensemble method for prediction and classification is gradient boosting.. Weak learners are combined to produce effective learning models for prediction. It is decided to employ the decision tree model. It is a popular and commonly used method for categorising huge, complex data sets. Gradient boosting models improve with iteration.

Algorithm-

- Consider the following sample of target values: P.
- Calculate the target value error.
- To decrease mistake M, update and change the weights.
- P[x] = alpha M[x] + p[x]
- The loss function F analyses and calculates model learners.
- Repeat steps until desired and target result P is obtained.



7. **XGBoost -** It is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for “Extreme Gradient Boosting” and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression.

Algorithm-

- Make an Initial Prediction and Calculate Residuals
- Build an XGBoost Tree
- Prune the Tree
- Calculate the Output Values of Leaves
- Make New Predictions
- Calculate Residuals Using the New Predictions
- Repeat Steps 2–6

CHAPTER 4

IMPLEMENTATION AND RESULTS

4.1. Software and Hardware Requirements

The major software and hardware requirements include :

4.1.1. Python

A high-level, all-purpose programming language is Python. Code readability is prioritised in its design philosophy, which typically employs indentation.

Both dynamic typing and garbage collection are supported by Python. Procedural, structured, object-oriented, and functional programming are just a few of the programming paradigms that it supports (especially this). This language's vast standard library has given it the nickname "batteries included."

Python was developed by Guido van Rossum in the late 1980s to replace the ABC programming language. Python 0.9.0 was made public in 1991. New features including list comprehensions, reference counting, cycle-detecting garbage collection, and support for Unicode were included in Python 2.0 when it was released in 2000. 2008 saw the release of Python 3.0, a substantial change that was not entirely backwards compatible with earlier iterations. 2020 saw the end of Python 2 with version 2.7.18. Programming language Python routinely ranks among the most well-liked ones.

4.1.2. NumPy

NumPy is a Python open source project that aims to make numerical computation easier. The Numeric and Numarray libraries' initial work served as the foundation for its creation in 2005. Free, fully open source, and in line with the permissive provisions of the modified BSD licence, NumPy will always be made available.

According to the NumPy and larger science Python communities, NumPy is maintained publicly on GitHub. Please see our Governance Document for more details on our governance strategy.

4.1.3. Pandas

Pandas is a collection of data analysis and manipulation tools made especially for the Python programming language. It provides detailed instructions for utilising mathematical tables and time series data. It is free software that is released in accordance with the license's three clauses.

An econometrics term for data sets that include observations for the same people over several time periods is panel data. The name of Python data analysis is punny. In his time from 2007 to 2010 as a researcher at AQR Capital, Wes McKinney started developing the pandas that would later become well-known.

4.1.4. Matplot Lib

Python's NumPy extension for numerical mathematics, along with Matplotlib, are graphing libraries. It provides an object-oriented API for adding charts to applications that make use of a general-purpose GUI toolkit like Tkinter, wxPython, Qt, or GTK. The state machine-based procedural "pylab" interface, designed to closely resemble the MATLAB interface, should not be used (similar to OpenGL). In SciPy, Matplotlib is utilised.

Matplotlib is ascribed to its creator, John D. Hunter. Since then, a healthy development community has developed around it, and it is presently available under a BSD-like licence. Michael Droettboom and Thomas Caswell were both suggested as matplotlib's primary developers before John Hunter passed away in August 2012. The Matplotlib project receives financial support from NumFOCUS.

Python 2.7 to 3.10 are compatible with Matplotlib 2.0.x. Python 3 was initially supported by Matplotlib 1.2, whereas Python 2.6 was last supported by Matplotlib 1.4. By committing to discontinue Python 2 support after 2020, Matplotlib made a commitment to the Python 3 Statement.

4.1.5. Seaborn

For Python and its NumPy extension for numerical mathematics, Matplotlib is a graphing library. It provides an object-oriented API for adding charts to software applications that make use of a general-purpose GUI toolkit like Tkinter, wxPython, Qt, or GTK. The procedural "pylab" interface, which is built on a state machine and was designed to closely resemble the MATLAB interface, should not be used. SciPy makes use of Matplotlib.

Matplotlib is ascribed to its creator, John D. Hunter. Since then, a healthy development community has developed around it, and it is presently available under a BSD-like licence. Michael Droettboom and Thomas Caswell were both suggested as matplotlib's lead developers prior to John Hunter's dying in August 2012. The Matplotlib project receives financial support from NumFOCUS.

Python 2.7 to 3.10 are compatible with Matplotlib 2.0.x. Python 3 was initially supported by Matplotlib 1.2, whereas Python 2.6 was last supported by Matplotlib 1.4. By pledging to discontinue support for Python 2 after 2020, Matplotlib made a commitment to the Python 3 Statement.

4.1.6. Pimas Indian database

This dataset was originally stored by the National Institute of Diabetes and Digestive and Kidney Diseases. Based on key diagnostic indications that are available in the data, the dataset attempts to diagnose diabetes. Based on a number of factors, these examples were picked from a larger database. Particularly, Pima Indian women who are at least 21 years old make up the majority of the clinic's clientele.

4.2. MODEL BUILDING

The stage that involves creating a model for predicting diabetes is the most crucial. This took use of the previously stated machine learning algorithms for diabetes prediction.The proposed methodology's process-

Step 1: Import the diabetic dataset along with the necessary libraries.

Step 2: To fill in any gaps, Pre-process the data.

Step 3: Divide the dataset in half, 80% for training and 20% for testing.

Step 4: Choose from the following machine learning methods: K-Nearest Neighbor, Support Vector Machine, Gradient Boosting, Logistic Regression, Random Forest, and Decision Tree.

Step 5: For the aforementioned machine learning technique, create the classifier model based on the training set.

Step 6: To assess the Classifier model for the previously described machine learning technique, use a test set.

Step 7: Compare the experimental performance outcomes of each classifier.

Step 8: After analyzing various metrics, select the best performing algorithm.

4.3. Implementation Details

4.3.1. Snapshot of Interfaces

```
#IMPORTING LIBRARIES

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

sns.set()

from mlxtend.plotting import plot_decision_regions
import missingno as msno
from pandas.plotting import scatter_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import classification_report
import warnings
warnings.filterwarnings('ignore')
#%matplotlib inline

#READING DATASET FILE
diabetes_df = pd.read_csv(r"C:\Users\subha\OneDrive\Desktop\Major Project\Code\diabetes.csv", encoding="ISO-8859-1")
print(diabetes_df)

diabetes_df.head()
diabetes_df.columns
diabetes_df.info()
diabetes_df.describe()
diabetes_df.describe().T
diabetes_df.isnull().head(10)
diabetes_df.isnull().sum()
diabetes_df_copy = diabetes_df.copy(deep = True)
diabetes_df_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']] = diabetes_df_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']].replace(0,np
print(diabetes_df_copy.isnull().sum())
```

```
print(diabetes_df_copy.isnull().sum())

#Data Visualization
#Plotting the data distribution plots before removing null values

p = diabetes_df.hist(figsize = (20,20))

#imputing the mean value of the column to each missing value of that particular column

diabetes_df_copy['Glucose'].fillna(diabetes_df_copy['Glucose'].mean(), inplace = True)
diabetes_df_copy['BloodPressure'].fillna(diabetes_df_copy['BloodPressure'].mean(), inplace = True)
diabetes_df_copy['SkinThickness'].fillna(diabetes_df_copy['SkinThickness'].median(), inplace = True)
diabetes_df_copy['Insulin'].fillna(diabetes_df_copy['Insulin'].median(), inplace = True)
diabetes_df_copy['BMI'].fillna(diabetes_df_copy['BMI'].median(), inplace = True)

#Plotting the distributions after removing the NAN values.
p = diabetes_df_copy.hist(figsize = (20,20))

#Plotting Null Count Analysis Plot
p = msno.bar(diabetes_df)
color_wheel = {1: "#0392cf", 2: "#7bc043"}
colors = diabetes_df["Outcome"].map(lambda x: color_wheel.get(x + 1))
print(diabetes_df.Outcome.value_counts())
p=diabetes_df.Outcome.value_counts().plot(kind="bar")
plt.subplot(121), sns.distplot(diabetes_df['Insulin'])
plt.subplot(122), diabetes_df['Insulin'].plot.box(figsize=(16,5))
plt.show()

#Correlation between all the features before cleaning
plt.figure(figsize=(12,10))

# seaborn has an easy method to showcase heatmap
p = sns.heatmap(diabetes_df.corr(), annot=True,cmap ='RdYlGn')

#Scaling the Data
diabetes_df_copy.head()
```

```
sc_X = StandardScaler()
X =  pd.DataFrame(sc_X.fit_transform(diabetes_df_copy.drop(["Outcome"],axis = 1),), columns=['Pregnancies',
'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age'])
X.head()

#Model Building
#Splitting the dataset
X = diabetes_df.drop('Outcome', axis=1)
y = diabetes_df['Outcome']

#Random Forest
#Building the model using RandomForest
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.33,random_state=7)
from sklearn.ensemble import RandomForestClassifier

rfc = RandomForestClassifier(n_estimators=200)
rfc.fit(X_train, y_train)

#check the accuracy of the model on the training dataset
rfc_train = rfc.predict(X_train)

from sklearn import metrics
print("Accuracy_Score of Random Forest after Training =", format(metrics.accuracy_score(y_train, rfc_train)))

from sklearn import metrics
predictions = rfc.predict(X_test)
print("Accuracy_Score of Random Forest after Testing =", format(metrics.accuracy_score(y_test, predictions)))

#Decision Tree
#Building the model using DecisionTree
from sklearn.tree import DecisionTreeClassifier

dtree = DecisionTreeClassifier()
dtree.fit(X_train, y_train)
```

```python
from sklearn import metrics

predictions = dtree.predict(X_test)
print("Accuracy Score of Decision Tree =", format(metrics.accuracy_score(y_test,predictions)))
from sklearn.metrics import classification_report, confusion_matrix

print(confusion_matrix(y_test, predictions))
print(classification_report(y_test,predictions))

#XgBoost classifier
#Building model using XGBoost
from xgboost import XGBClassifier

xgb_model = XGBClassifier(gamma=0)
xgb_model.fit(X_train, y_train)

from sklearn import metrics

xgb_pred = xgb_model.predict(X_test)
print("Accuracy Score of XGBoost Classifier =", format(metrics.accuracy_score(y_test, xgb_pred)))

#Support Vector Machine (SVM)
#Building the model using Support Vector Machine (SVM)
from sklearn.svm import SVC

svc_model = SVC()
svc_model.fit(X_train, y_train)
svc_pred = svc_model.predict(X_test)
from sklearn import metrics

print("Accuracy Score of SVM =", format(metrics.accuracy_score(y_test, svc_pred)))
from sklearn.metrics import classification_report, confusion_matrix

print(confusion_matrix(y_test, svc_pred))
print(classification_report(y_test,svc_pred))
```

```python
#Feature Importance
rfc.feature_importances_
(pd.Series(rfc.feature_importances_, index=X.columns).plot(kind='barh'))
import pickle

# Firstly we will be using the dump() function to save the model using pickle
saved_model = pickle.dumps(rfc)

# Then we will be loading that saved model
rfc_from_pickle = pickle.loads(saved_model)

# lastly, after loading that model we will use this to make predictions
rfc_from_pickle.predict(X_test)
diabetes_df.head()
diabetes_df.tail()
rfc.predict([[0,137,40,35,168,43.1,2.228,33]]) #4th patient
rfc.predict([[10,101,76,48,180,32.9,0.171,63]])  # 763 th patient
```

4.3.2. Test Cases

	A	B	C	D	E	F	G	H	I	J
1	Pregnancies	Glucose	BloodPressu	SkinThickne	Insulin	BMI	DiabetesPe	Age	Outcome	
2	6	148	72	35	0	33.6	0.627	50	1	
3	1	85	66	29	0	26.6	0.351	31	0	
4	8	183	64	0	0	23.3	0.672	32	1	
5	1	89	66	23	94	28.1	0.167	21	0	
6	0	137	40	35	168	43.1	2.288	33	1	
7	5	116	74	0	0	25.6	0.201	30	0	
8	3	78	50	32	88	31	0.248	26	1	
9	10	115	0	0	0	35.3	0.134	29	0	
10	2	197	70	45	543	30.5	0.158	53	1	
11	8	125	96	0	0	0	0.232	54	1	
12	4	110	92	0	0	37.6	0.191	30	0	
13	10	168	74	0	0	38	0.537	34	1	
14	10	139	80	0	0	27.1	1.441	57	0	
15	1	189	60	23	846	30.1	0.398	59	1	
16	5	166	72	19	175	25.8	0.587	51	1	
17	7	100	0	0	0	30	0.484	32	1	
18	0	118	84	47	230	45.8	0.551	31	1	
19	7	107	74	0	0	29.6	0.254	31	1	
20	1	103	30	38	83	43.3	0.183	33	0	
21	1	115	70	30	96	34.6	0.529	32	1	
22	3	126	88	41	235	39.3	0.704	27	0	
23	8	99	84	0	0	35.4	0.388	50	0	
24	7	196	90	0	0	39.8	0.451	41	1	
25	9	119	80	35	0	29	0.263	29	1	
26	11	143	94	33	146	36.6	0.254	51	1	
27	10	125	70	26	115	31.1	0.205	41	1	
28	7	147	76	0	0	39.4	0.257	43	1	
29	1	97	66	15	140	23.2	0.487	22	0	
30	13	145	82	19	110	22.2	0.245	57	0	

	A	B	C	D	E	F	G	H	I	J
31	5	117	92	0	0	34.1	0.337	38	0	
32	5	109	75	26	0	36	0.546	60	0	
33	3	158	76	36	245	31.6	0.851	28	1	
34	3	88	58	11	54	24.8	0.267	22	0	
35	6	92	92	0	0	19.9	0.188	28	0	
36	10	122	78	31	0	27.6	0.512	45	0	
37	4	103	60	33	192	24	0.966	33	0	
38	11	138	76	0	0	33.2	0.42	35	0	
39	9	102	76	37	0	32.9	0.665	46	1	
40	2	90	68	42	0	38.2	0.503	27	1	
41	4	111	72	47	207	37.1	1.39	56	1	
42	3	180	64	25	70	34	0.271	26	0	
43	7	133	84	0	0	40.2	0.696	37	0	
44	7	106	92	18	0	22.7	0.235	48	0	
45	9	171	110	24	240	45.4	0.721	54	1	
46	7	159	64	0	0	27.4	0.294	40	0	
47	0	180	66	39	0	42	1.893	25	1	
48	1	146	56	0	0	29.7	0.564	29	0	
49	2	71	70	27	0	28	0.586	22	0	
50	7	103	66	32	0	39.1	0.344	31	1	
51	7	105	0	0	0	0	0.305	24	0	
52	1	103	80	11	82	19.4	0.491	22	0	
53	1	101	50	15	36	24.2	0.526	26	0	
54	5	88	66	21	23	24.4	0.342	30	0	
55	8	176	90	34	300	33.7	0.467	58	1	
56	7	150	66	42	342	34.7	0.718	42	0	
57	1	73	50	10	0	23	0.248	21	0	
58	7	187	68	39	304	37.7	0.254	41	1	
59	0	100	88	60	110	46.8	0.962	31	0	
60	0	146	82	0	0	40.5	1.781	44	0	

	A	B	C	D	E	F	G	H	I	J
61	0	105	64	41	142	41.5	0.173	22	0	
62	2	84	0	0	0	0	0.304	21	0	
63	8	133	72	0	0	32.9	0.27	39	1	
64	5	44	62	0	0	25	0.587	36	0	
65	2	141	58	34	128	25.4	0.699	24	0	
66	7	114	66	0	0	32.8	0.258	42	1	
67	5	99	74	27	0	29	0.203	32	0	
68	0	109	88	30	0	32.5	0.855	38	1	
69	2	109	92	0	0	42.7	0.845	54	0	
70	1	95	66	13	38	19.6	0.334	25	0	
71	4	146	85	27	100	28.9	0.189	27	0	
72	2	100	66	20	90	32.9	0.867	28	1	
73	5	139	64	35	140	28.6	0.411	26	0	
74	13	126	90	0	0	43.4	0.583	42	1	
75	4	129	86	20	270	35.1	0.231	23	0	
76	1	79	75	30	0	32	0.396	22	0	
77	1	0	48	20	0	24.7	0.14	22	0	
78	7	62	78	0	0	32.6	0.391	41	0	
79	5	95	72	33	0	37.7	0.37	27	0	
80	0	131	0	0	0	43.2	0.27	26	1	
81	2	112	66	22	0	25	0.307	24	0	
82	3	113	44	13	0	22.4	0.14	22	0	
83	2	74	0	0	0	0	0.102	22	0	
84	7	83	78	26	71	29.3	0.767	36	0	
85	0	101	65	28	0	24.6	0.237	22	0	
86	5	137	108	0	0	48.8	0.227	37	1	
87	2	110	74	29	125	32.4	0.698	27	0	
88	13	106	72	54	0	36.6	0.178	45	0	
89	2	100	68	25	71	38.5	0.324	26	0	
90	15	136	70	32	110	37.1	0.153	43	1	

	A	B	C	D	E	F	G	H	I	J
91	1	107	68	19	0	26.5	0.165	24	0	
92	1	80	55	0	0	19.1	0.258	21	0	
93	4	123	80	15	176	32	0.443	34	0	
94	7	81	78	40	48	46.7	0.261	42	0	
95	4	134	72	0	0	23.8	0.277	60	1	
96	2	142	82	18	64	24.7	0.761	21	0	
97	6	144	72	27	228	33.9	0.255	40	0	
98	2	92	62	28	0	31.6	0.13	24	0	
99	1	71	48	18	76	20.4	0.323	22	0	
100	6	93	50	30	64	28.7	0.356	23	0	
101	1	122	90	51	220	49.7	0.325	31	1	
102	1	163	72	0	0	39	1.222	33	1	
103	1	151	60	0	0	26.1	0.179	22	0	
104	0	125	96	0	0	22.5	0.262	21	0	
105	1	81	72	18	40	26.6	0.283	24	0	
106	2	85	65	0	0	39.6	0.93	27	0	
107	1	126	56	29	152	28.7	0.801	21	0	
108	1	96	122	0	0	22.4	0.207	27	0	
109	4	144	58	28	140	29.5	0.287	37	0	
110	3	83	58	31	18	34.3	0.336	25	0	
111	0	95	85	25	36	37.4	0.247	24	1	
112	3	171	72	33	135	33.3	0.199	24	1	
113	8	155	62	26	495	34	0.543	46	1	
114	1	89	76	34	37	31.2	0.192	23	0	
115	4	76	62	0	0	34	0.391	25	0	
116	7	160	54	32	175	30.5	0.588	39	1	
117	4	146	92	0	0	31.2	0.539	61	1	
118	5	124	74	0	0	34	0.22	38	1	
119	5	78	48	0	0	33.7	0.654	25	0	
120	4	97	60	23	0	28.2	0.443	22	0	

	A	B	C	D	E	F	G	H	I	J
121	4	99	76	15	51	23.2	0.223	21	0	
122	0	162	76	56	100	53.2	0.759	25	1	
123	6	111	64	39	0	34.2	0.26	24	0	
124	2	107	74	30	100	33.6	0.404	23	0	
125	5	132	80	0	0	26.8	0.186	69	0	
126	0	113	76	0	0	33.3	0.278	23	1	
127	1	88	30	42	99	55	0.496	26	1	
128	3	120	70	30	135	42.9	0.452	30	0	
129	1	118	58	36	94	33.3	0.261	23	0	
130	1	117	88	24	145	34.5	0.403	40	1	
131	0	105	84	0	0	27.9	0.741	62	1	
132	4	173	70	14	168	29.7	0.361	33	1	
133	9	122	56	0	0	33.3	1.114	33	1	
134	3	170	64	37	225	34.5	0.356	30	1	
135	8	84	74	31	0	38.3	0.457	39	0	
136	2	96	68	13	49	21.1	0.647	26	0	
137	2	125	60	20	140	33.8	0.088	31	0	
138	0	100	70	26	50	30.8	0.597	21	0	
139	0	93	60	25	92	28.7	0.532	22	0	
140	0	129	80	0	0	31.2	0.703	29	0	
141	5	105	72	29	325	36.9	0.159	28	0	
142	3	128	78	0	0	21.1	0.268	55	0	
143	5	106	82	30	0	39.5	0.286	38	0	
144	2	108	52	26	63	32.5	0.318	22	0	
145	10	108	66	0	0	32.4	0.272	42	1	
146	4	154	62	31	284	32.8	0.237	23	0	
147	0	102	75	23	0	0	0.572	21	0	
148	9	57	80	37	0	32.8	0.096	41	0	
149	2	106	64	35	119	30.5	1.4	34	0	
150	5	147	78	0	0	33.7	0.218	65	0	

4.3.3. Results

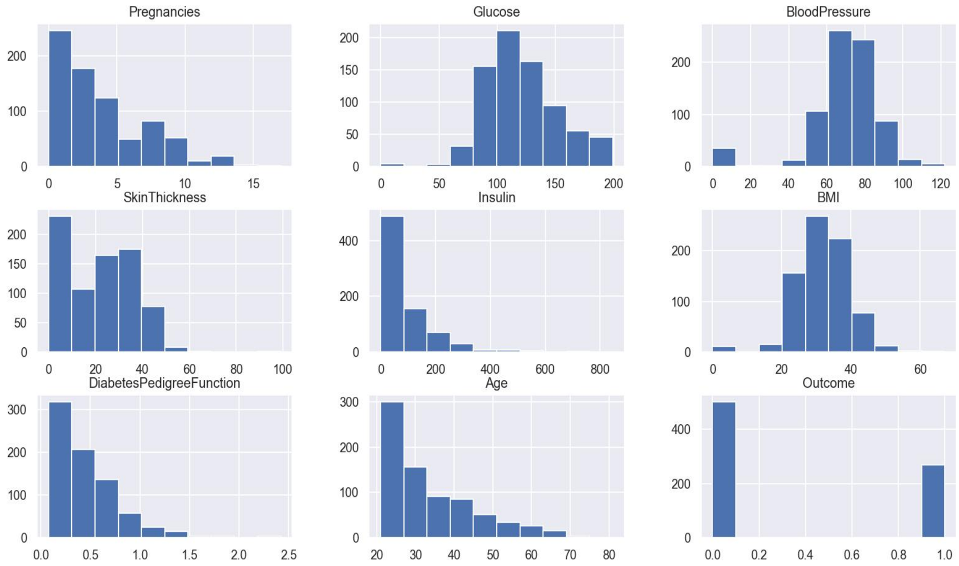
4.3.3.1. Information of the dataset

```
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
```

4.3.3.2. Showing the count of NAN

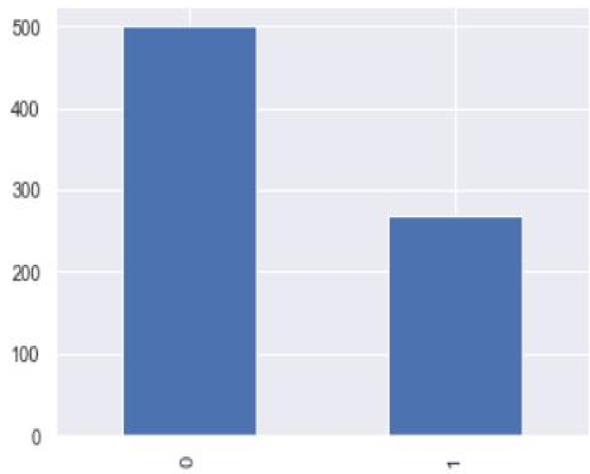
```
Pregnancies                   0
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                       0
dtype: int64
```

4.3.3.3. Data Visualization Histogram



4.3.3.4. Diabetic and Non - Diabetic count

```
0    500
1    268
Name: Outcome, dtype: int64
```



4.3.3.5. Classification report and confusion matrix of random forest model

```
[[133  29]
 [ 30  62]]
              precision    recall  f1-score   support

           0       0.82      0.82      0.82       162
           1       0.68      0.67      0.68        92

    accuracy                           0.77       254
   macro avg       0.75      0.75      0.75       254
weighted avg       0.77      0.77      0.77       254
```

4.3.3.6. Classification report and confusion matrix of the decision tree model

```
[[126  36]
 [ 32  60]]
              precision    recall  f1-score   support

           0       0.80      0.78      0.79       162
           1       0.62      0.65      0.64        92

    accuracy                           0.73       254
   macro avg       0.71      0.71      0.71       254
weighted avg       0.73      0.73      0.73       254
```

4.3.3.7. Classification report and confusion matrix of the XGBoost classifier

```
[[127  35]
 [ 31  61]]
```

	precision	recall	f1-score	support
0	0.80	0.78	0.79	162
1	0.64	0.66	0.65	92
accuracy			0.74	254
macro avg	0.72	0.72	0.72	254
weighted avg	0.74	0.74	0.74	254

4.3.3.8. Classification report and confusion matrix of the SVM classifier

```
[[143  19]
 [ 47  45]]
```

	precision	recall	f1-score	support
0	0.75	0.88	0.81	162
1	0.70	0.49	0.58	92
accuracy			0.74	254
macro avg	0.73	0.69	0.69	254
weighted avg	0.73	0.74	0.73	254

CHAPTER 5

CONCLUSION

The early diagnosis of diabetes is one of the most significant medical issues today. This strategy consciously works to create a diabetes prediction system. four machine learning categorization techniques are looked into and assessed in this study based on a number of different factors. Experiments are being carried out on the Pima Indian database.

Result:

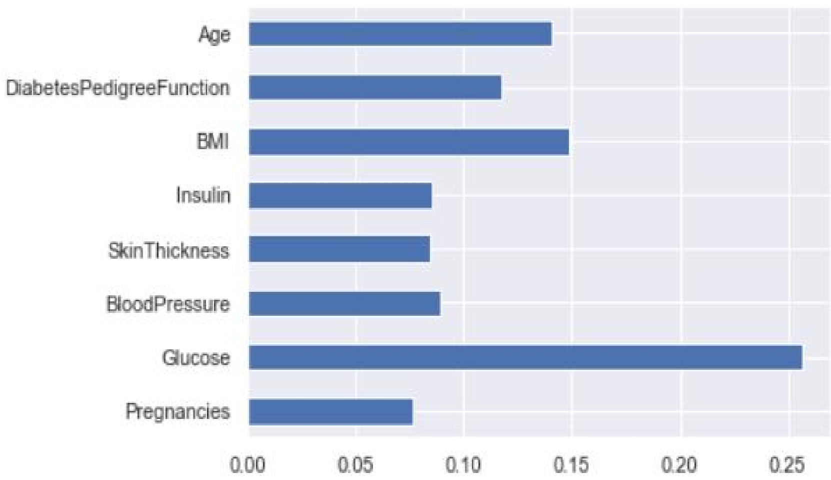
Algorithms	Accuracy (in %)
SVM	**0.7480314960629921**
Decision Tree	**0.7165354330708661**
Random Forest	**0.7677165354330708**
XGBoost Classifier	**0.7401574803149606**

Future Directions

Future research may predict or diagnose new ailments using the developed approach and ML classification techniques. The approach might be enhanced and broadened for diabetes analysis automation by including new machine learning methods and using Deep Learning methods.

Appendix

Feature Importance



Saving model Random Forest

```
array([0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0,
       1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0,
       0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1,
       0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,
       1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1,
       0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0,
       1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0,
       0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0,
       0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1,
       1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1], dtype=int64)
```

References

1. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.
2. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.
3. Benbelkacem S, Atmani B. Random forests for diabetes diagnosis. International Conference on Computer and Information Sciences. IEEE; 2019.
4. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He and Xiaoyi Wang, ”Type 2 diabetes mellitus prediction model based on data mining”, Elsevier Informatics in Medicine, vol. 10, pp. 100-107, 2018.
5. Kavakiotisab I, Tsave O, Salifoglou A, Maglaveras N, Vlahavasa I, Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal. 2017; 15: 104- 16.
6. Jia M, Tian F. Readmission prediction of diabetic based on convolutional neural networks. International Conference on Computer and Communications. IEEE; 2019.
7. M Sabibullah, V Shanmugasundaram and Priya K Raja, ”Diabetes Patient’s Risk through Soft Computing Model”, International Journal of Emerging Trends Technology in Computer Science, vol. 2, no. 6, 2013.
8. Pujianto U, Setiawan AL, Rosyid HA, Salah AMM. Comparison of naïve Bayes algorithm and decision tree C4. 5 for hospital readmission diabetes patients using hba1c measurement. Knowledge Engineering and Data Science. 2019; 2(2).
9. Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584-1589). IEEE.
10. Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59.
11. Sun YL, Zhang DL. Machine learning techniques for screening and diagnosis of diabetes: A survey. Tehnički Vjesnik. 2019; 26(3): 872-80.

12. M Sabibullah, V Shanmugasundaram and Priya K Raja, ”Diabetes Patient’s Risk through Soft Computing Model”, International Journal of Emerging Trends Technology in Computer Science, vol. 2, no. 6, 2013.

13. Warke M, Kumar V, Tarale S, Galgat P, Chaudhari D. Diabetes diagnosis using machine learning algorithms. International Research Journal of Engineering and Technology. 2019; 6(3): 1470-6.

14. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

15. https://www.researchgate.net/publication/350745659_Diabetes_Diagnosis_Using_Machine_Learning.

16. https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques.

PLAGIRISM REPORT

subham

ORIGINALITY REPORT

14%	9%	7%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ijsrcseit.com Internet Source	2%
2	Submitted to Visvesvaraya Technological University, Belagavi Student Paper	2%
3	Jana S, Bharanidharan N, ShanmukhaNagasai P, Saravan Kumar K, V Mani Nageshwar. "Diabetes Prediction Using Machine Learning Algorithms", 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), 2022 Publication	1%
4	Submitted to University of Leeds Student Paper	1%
5	www.ijirset.com Internet Source	1%
6	ijmi.ir Internet Source	1%
7	ijiemr.org Internet Source	1%

8	www.journals.scholarpublishing.org Internet Source	1%
9	Submitted to University of Sunderland Student Paper	1%
10	Submitted to Sunway Education Group Student Paper	1%
11	www.ijarem.org Internet Source	1%
12	Submitted to University of Aberdeen Student Paper	<1%
13	Submitted to University of North Texas Student Paper	<1%
14	sersc.org Internet Source	<1%
15	downloads.hindawi.com Internet Source	<1%
16	Boshra Farajollahi, Maysam Mehmannavaz, Hafez Mehrjoo, Fateme Moghbeli, Mohammad Javad Sayadi. "Diabetes Diagnosis Using Machine Learning", Frontiers in Health Informatics, 2021 Publication	<1%
17	amslaurea.unibo.it Internet Source	<1%

18	Ebrahim Mohammed Senan, Ali Alzahrani, Mohammed Y. Alzahrani, Nizar Alsharif, Theyazn H. H. Aldhyani. "Automated Diagnosis of Chest X-Ray for Early Detection of COVID-19 Disease", Computational and Mathematical Methods in Medicine, 2021 Publication	<1%
19	"Artificial Intelligence in Medicine", Springer Science and Business Media LLC, 2017 Publication	<1%
20	www.science.gov Internet Source	<1%
21	Badreddine Boudriki Semlali, El Amrani Chaker. "Towards Remote Sensing Datasets Collection and Processing", International Journal of Embedded and Real-Time Communication Systems, 2019 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		