*A Mini-Project Report On*

# "DIABETES PREDICTION AND FOOD RECOMMENDATION"

*Submitted By*


**AISHWARYA KARANDIKAR**
**PRN: 1132210014**

**BHUMIKA OSTWAL**
**PRN: 1132210217**

**CHAITALI KANNURKAR**
**PRN: 1132210736**

**CHETAN BAJAJ**
**PRN: 1132210196**


**F.Y. M.Sc. (Data Science and Big Data Analytics)**


**School of Computer Science**
**Faculty of Science**
**MIT – World Peace University**
**Pune - 411038**


**Academic Year 2021-2022**
**May - 2022**

**Dr. Vishwanath Karad MIT WORLD PEACE UNIVERSITY, PUNE**
**SCHOOL OF COMPUTER SCIENCE**

## Certificate

This is to certify that
**AISHWARYA KARANDIKAR**
**PRN: 1132210014**

**BHUMIKA OSTWAL**
**PRN: 1132210217**

**CHAITALI KANNURKAR**
**PRN: 1132210736**

**CHETAN BAJAJ**
**PRN: 1132210196**

Of **M.Sc. (Data Science and Big Data Analytics)** successfully completed his/her Mini-Project in

## PROJECT TITLE : "DIABETES PREDICTION AND FOOD RECOMMENDATION"

to our satisfaction and submitted the same during the academic year 2021 - 2022 towards the partial fulfilment of degree of **Master of Science in Data Science and Big Data Analytics** of Dr Vishwanath Karad MIT World Peace University under the School of Computer Science, MIT WPU, Pune.

| | | |
|---|---|---|
| **Prof. Dr. Shubhalaxmi Joshi** | **Prof. Surabhi Thatte** | **Prof. Dr. Darshan Ruikar** |
| **Associate Dean** | **Program Head** | **Assistant Professor** |
| **Faculty of Science** | **School of Computer Science** | **School of Computer Science** |
| **MITWPU** | **MIT WPU** | **MIT WPU** |

# ACKNOWLEDGEMENT

In the accomplishment of this project, I would like to express my special thanks of gratitude to my teacher **Prof. Dr. Darshan Ruikar**, School of Computer Science, Dr. Vishwanath Karad MIT World Peace University whose valuable guidance has been the ones that helped me patch this project and make it full proof success. His/Her suggestions and instructions have served as the major contributor towards the completion of the project.

As we were working in a group, I would like to thank my group members for their fabulous support throughout the completion of the project. We learned a lot of things during this period, as it was hard to work in this time of adversity; we were in touch with each other throughout the period and shared everything which was important from the aspect of our project. As this project was completed by staying at home, I would also like to thank our families for their cooperation and for providing facilities to us.

**AISHWARYA KARANDIKAR**
**PRN: 1132210014**

**BHUMIKA OSTWAL**
**PRN: 1132210217**

**CHAITALI KANNURKAR**
**PRN: 1132210736**

**CHETAN BAJAJ**
**PRN: 1132210196**

# Contents

# Introduction

**Domain Name :** HEALTHCARE

**Introduction :**

• Diabetes is one of the most chronic medical conditions across the globe. The medical condition is characterized by high levels of sugar level in the blood of the patients. Thus, Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from your food. Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy.

• Diabetes is manifested in various forms among the victims. This is because of different types of diabetes that affect the predisposed population. The most common type of diabetes includes Type I, Type II, Gestational, and Juvenile Diabetes.

• Type 1 diabetes occurs when the body's immune system destroys the insulin-producing cells of the pancreas (called beta cells). Normally, the body's immune system fights off foreign invaders like viruses or bacteria.

• Type 2 diabetes is an impairment in the way the body regulates and uses sugar (glucose) as a fuel. This long-term (chronic) condition results in too much sugar circulating in the bloodstream. Eventually, high blood sugar levels can lead to disorders of the circulatory, nervous, and immune systems.

• Type 2 diabetes is primarily the result of two interrelated problems: Cells in muscle, fat and the liver become resistant to insulin. Because these cells don't interact in a normal way with insulin, they don't take in enough sugar. The pancreas is unable to produce enough insulin to manage blood sugar levels.

• Type 2 diabetes used to be known as adult-onset diabetes, but both type 1 and type 2 diabetes can begin during childhood and adulthood. Type 2 is more common in older adults, but the increase in the number of children with obesity has led to more cases of type 2 diabetes in younger people.

• There's no cure for type 2 diabetes, but losing weight, eating well, and exercising can help you manage the disease. If diet and exercise aren't enough to manage your blood sugar, you may also need diabetes medications or insulin therapy.

**Motivation :** Within the framework of healthcare development the diabetes is a very crucial topic. Now-a-days Diabetes is on the rise. No longer a disease of predominantly rich nations, the prevalence of diabetes is steadily increasing everywhere. Globally, rates of type 2 diabetes were 15.1 million in 2000. Type 2 diabetes impacts men and women proportionately; there are over 12 million men with diabetes and 11.5 women with diabetes. Along with the increase of diabetes, both individual concerning the management of diabetes have also increased. Diabetes is a chronic disease for which control of the condition demands patient self-management. Self-management behaviors include monitoring blood glucose levels, taking medication, maintaining a healthy diet and regularly exercising. For most patients, maintaining a healthy diet is difficult task. So we can recommend the food based on their sugar level.

**PROBLEM STATEMENT** : To predict whether the person is suffering from diabetes and classify into high, low and normal category then recommended food items accordingly.

# Literature Survey

**LITERATURE SURVEY :**

Accurate classification of diabetes is a fundamental step towards diabetes prevention and control in healthcare. However, early and onset identification of diabetes is much more beneficial in controlling diabetes. The diabetes identification process seems tedious at an early stage because a patient has to visit a physician regularly. The advancement in machine learning approaches has solved this critical and essential problem in healthcare by predicting disease. Several techniques have been proposed in the literature for diabetes prediction.

**1.RESEARCH PAPER NAME: Prediction and diagnosis of diabetes mellitus — A machine learning approach**

Researcher name/s : Vinya Vijayan and C. Anjali

Topic : Prediction and diagnosis of diabetes mellitus — A machine learning approach

Date of Conference: 10-12 Dec. 2015

Work :

Diabetes is a disease caused due of the expanded level of sugar fixation in the blood. Various computerized information systems were outlined utilizing diverse classifiers for anticipating and diagnosing diabetes. Selecting legitimate classifiers clearly expands the exactness and proficiency of the system. Here a decision support system is proposed that uses AdaBoost algorithm with Decision Stump as base classifier for classification. Additionally Support Vector Machine, Naive Bayes and Decision Tree are also implemented as base classifiers for AdaBoost algorithm for accuracy verification. The accuracy obtained for AdaBoost algorithm with decision stump as base classifier is 67.72% which is greater compared to that of Support Vector Machine, Naive Bayes and Decision Tree.

Result :

Accuracy is determined as the ratio of number of correct predictions to the total number of predictions [10][13]. Table 1 shows the performance accuracy of the classifiers after applying discretize and principal component analysis. The 10 folds cross validation is used as the test option for all the 3 classifiers. The tool under consideration is the Weka. From the table, considering the same preprocessing technique, the main inference obtained is that Decision tree and Naïve Bayes have more accuracy to that of the SVM

**2. RESEARCH PAPER NAME: Classification Algorithm-Based Hybrid Diabetes Prediction Model**

Researcher name/s **: Michael Onyema Edeh, Osamah Ibrahim Khalaf ,Carlos Andrés Tavera, Sofiane Tayeb, Samir Ghouali, Ghaida Muttashar Abdulsahib, Nneka Ernestina Richard-Nnabu,AbdRahmane Louni**

Topic : A Classification Algorithm-Based Hybrid Diabetes Prediction Model

Date : 27 December 2017

Work :

Diabetes is considered to be one of the leading causes of death globally. If diabetes is not treated and detected early, it can lead to a variety of complications. The aim of this study was to develop a model that can accurately predict the likelihood of developing diabetes in patients with the greatest amount of precision. Classification algorithms are widely used in the medical field to classify data into different categories based

on some criteria that are relatively restrictive to the individual classifier, Therefore, four machine learning classification algorithms, namely supervised learning algorithms (Random forest, SVM and Naïve Bayes, Decision Tree DT) and unsupervised learning algorithm (k-means), have been a technique that was utilized in this investigation to identify diabetes in its early stages. The experiments are per-formed on two databases, one extracted from the Frankfurt Hospital in Germany and the other from the database. PIMA Indian Diabetes (PIDD) provided by the UCI machine learning repository. The results obtained from the database extracted from Frankfurt Hospital, Germany, showed that the random forest algorithm outperformed with the highest accuracy of 97.6%, and the results obtained from the Pima Indian database showed that the SVM algorithm outperformed with the highest accuracy of 83.1% compared to other algorithms. The validity of these results is confirmed by the process of separating the data set into two parts: a training set and a test set, which is described below. The training set is used to develop the model's capabilities. The test set is used to put the model through its paces and determine its correctness.

Result :

In this study, we proposed a supportive diagnosis system based on the comparison four models of prediction algorithms to predict diabetes in two different databases. On the basis of several performance assessment methodologies like as accuracy and recall, as well as the F1 score, different machine learning algorithms are compared and assessed. Using the classification results obtained, it can be concluded that the random forest machine learning technique provides more accurate prediction and higher performance than the other methods described in this study. However, when compared to other research accessible in the current literature, some of the other approaches utilized in this study, such as naive Bayes, DT and SVM, Random Forest, and others, produce the most optimum outcomes.

The main objective of this study is to help diabetologist to establish an accurate treatment routine for their diabetic patients. Due to the high accuracy and diagnose the disease in a shorter time and the rapid treatment, this study could open a window in the development of an electronic health system for diabetic patients. There are also a few aspects in this study that could be improved or expanded in the future. In perspective term:

- Creation of diabetes database for Algerian patients
- Diabetes prediction with the deep learning approach.
- Developed a solution based on an Android application in order to help people predict if they have diabetes.

# Solution Design

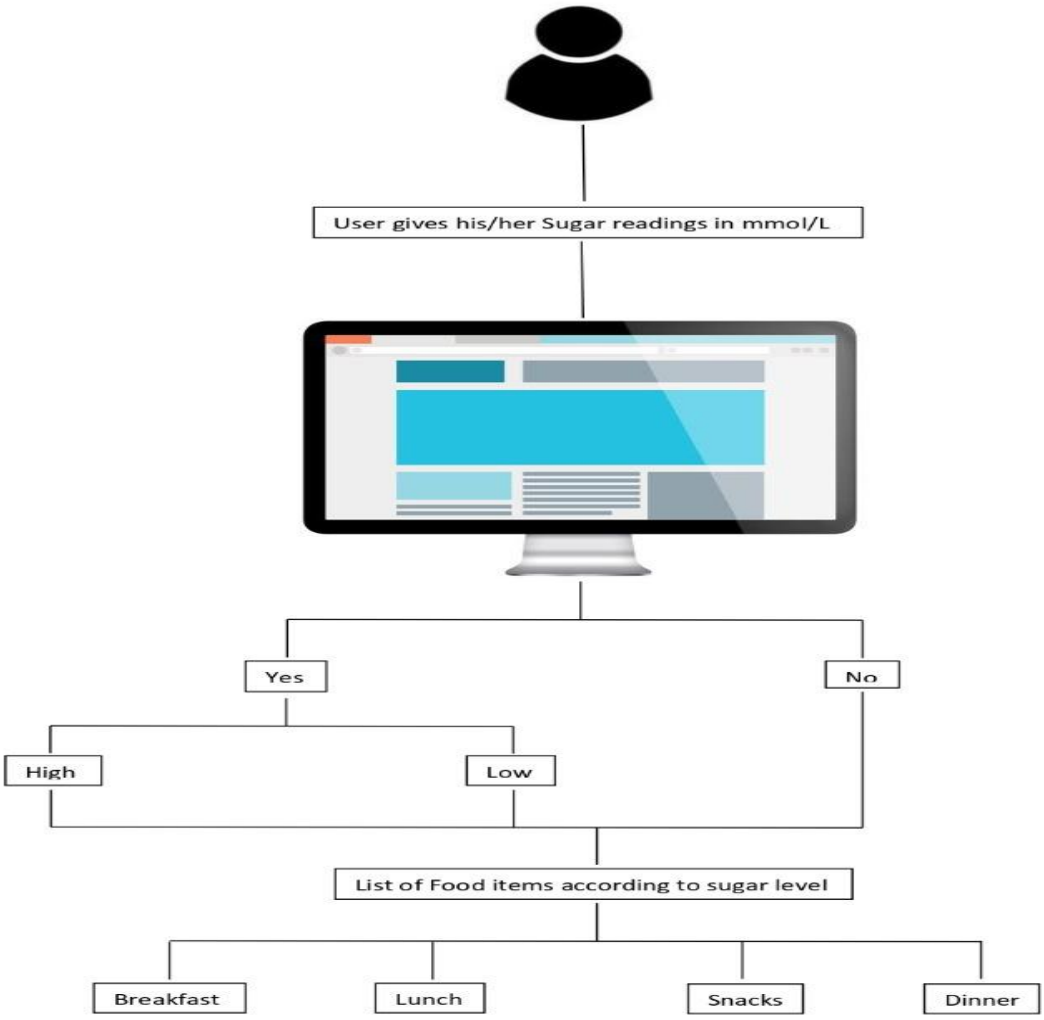
**SOLUTION APPROCH**:
**Aim**: To predict diabetes that classify into high low and normal category and recommend food accordingly

First we collected diabetes data from one of our researchers in MIT college. His name is Shantanu sir . So he had history of his own sugar levels which were measured on daily basis and we wanted to use that data as a dataset for our project which is based on diabetes prediction. We also worked on creating a list of Individual food items for breakfast , lunch , snacks , dinner . The list was specifically designed which is feasible for type II Diabetes. A blood sugar level less than 150 mg/dL (7.8 mmol/L) is normal. A reading of more than 200 mg/dL (11.1 mmol/L) after two hours indicates diabetes. A reading between 140 and 199 mg/dL (7.8 mmol/L and 11.0 mmol/L) indicates prediabetes. Then we sorted the whole data in 3 different categories so that we can use the food Items dataset as per the history dataset of sugar levels and so recommendation of Diet can be achieved . Here we decided to label each food item in 3 parts as per the Respective Glycemic index. The items for labelled as per glycemic index . We labelled the food items as LOW whose GI was ranging 0-130, MEDIUM whose GI was ranging 130-260 and HIGH whose GI was ranging 260-400 . The reason being for this step was that , we had to convert 4 files to a single food items CSV file. Now finally decided to implement our Dataset in python code for this project. Here we used SVM logistic regression and Decision tree Random forest methods using python code or the best outcome of this project.

**TECHNOLOGY STACK :**

1. jupyter notebook : coding part(execution of algorithm)
2. weka : for the mining of data, visualizing attributes and visualization of decision tree
3. powerbi : dashboard(visualization using various chart)
4. html : for front-end (webpage)
5. bootstrap: for front-end(webpage)
6. css: for front-end (webpage)

**DESIGN MODEL :**




**User : take the input of their sugar(mmol/L) from the user**
**System :After using different algorithm system predicts whether the person is diabetic or not and list of food items according to sugar level.**

# Solution Implementation and Results

**OBTAINING DATA :**

**1.** SUGAR LEVEL DATA : For the sugar level the dataset is given by prof shantanu kanade sir. This is their personal sugar level record of every 15min.

Diabetes : The diabetes dataset is referred from kaggle

1) https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset
2) https://www.kaggle.com/saurabh00007/diabetescsv

Food items : For food item we create the dataset having different food items, quantity and glycemic index

| Glucose | BloodPres | SkinThickn | Insulin | BMI | DiabetesPe | Age | Outcome | Smoker | HeartDise | PhysActivit | Fruits | HvyAlcoho | AnyHealth | NoDocbcC | GenHlth | MentHlth | PhysHlth | DiffWalk | Sex |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 5 | 18 | 15 | 1 | 0 |
| 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 0 |
| 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 5 | 30 | 30 | 1 | 0 |
| 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 0 |
| 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 3 | 0 | 0 | 0 |
| 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 0 | 2 | 0 | 1 |
| 78 | 50 | 32 | 88 | 31 | 0.248 | 26 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 14 | 0 | 0 |
| 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 3 | 0 | 0 | 1 | 0 |
| 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 5 | 30 | 30 | 1 | 0 |
| 125 | 96 | 0 | 0 | 0 | 0.232 | 54 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 1 |
| 110 | 92 | 0 | 0 | 37.6 | 0.191 | 30 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 1 |
| 168 | 74 | 0 | 0 | 38 | 0.537 | 34 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 3 | 0 | 30 | 1 | 0 |
| 139 | 80 | 0 | 0 | 27.1 | 1.441 | 57 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 15 | 0 | 0 |
| 189 | 60 | 23 | 846 | 30.1 | 0.398 | 59 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 0 | 0 | 1 | 0 |
| 166 | 72 | 19 | 175 | 25.8 | 0.587 | 51 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 4 | 30 | 28 | 0 | 0 |
| 100 | 0 | 0 | 0 | 30 | 0.484 | 32 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 5 | 0 | 0 | 0 |
| 118 | 84 | 47 | 230 | 45.8 | 0.551 | 31 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 0 |
| 107 | 74 | 0 | 0 | 29.6 | 0.254 | 31 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 1 |
| 103 | 30 | 38 | 83 | 43.3 | 0.183 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 15 | 0 | 0 | 0 |
| 115 | 70 | 30 | 96 | 34.6 | 0.529 | 32 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 | 10 | 0 | 0 | 1 |
| 126 | 88 | 41 | 235 | 39.3 | 0.704 | 27 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 3 | 30 | 0 | 1 | 0 |
| 99 | 84 | 0 | 0 | 35.4 | 0.388 | 50 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 5 | 15 | 30 | 1 | 0 |
| 196 | 90 | 0 | 0 | 39.8 | 0.451 | 41 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 7 | 0 | 1 |

FIG : SAMPLE DATASET FOR DIABETES PREDICTION

| | A | B |
|---|---|---|
| 1 | Sugar level | Label |
| 2 | 222 | High |
| 3 | 227 | High |
| 4 | 232 | High |
| 5 | 234 | High |
| 6 | 234 | High |
| 7 | 233 | High |
| 8 | 225 | High |
| 9 | 208 | High |
| 10 | 202 | High |
| 11 | 191 | High |
| 12 | 175 | High |
| 13 | 166 | High |
| 14 | 151 | High |
| 15 | 147 | Normal |
| 16 | 143 | Normal |
| 17 | 140 | Normal |
| 18 | 140 | Normal |

FIG : SUGARLEVEL CLASSIFY INTO CATEGORIES HIGH,LOW AND NORMAL

| A | B | C | D |
|---|---|---|---|
| Serial Number | Food Item | Quantity | Glycemic Index |
| 1 | Tea | 100ml | 8.9 |
| 1 | Poha | 100gms | 40 |
| 1 | Misalpav | 2 pav with 200gms | 62 |
| 1 | Coffee | 100ml | 12.8 |
| 1 | Upma | 250gms | 167.5 |
| 1 | Khakra | 1 piece | 66.8 |
| 1 | Multigrain Bread | 2 slice | 46.5 |
| 1 | Peanut Butter | 10 gms | 2.5 |
| 1 | Eggs | 2 | 0 |
| 1 | Amla juice | 150ml | 22.5 |
| 1 | Skimmed Milk | 100ml | 32 |
| 1 | Oats | 300gms | 165 |
| 1 | Green Tea | 100ml | 79 |
| 1 | Dosa | 1 | 66 |
| 1 | WheatFlakes | 1 bowl | 45 |
| 1 | Skimmed Milk | 150ml | 16 |
| 1 | Boiled Chicken | 150gms | 0 |
| 1 | Black coffee | 100nl | 0 |
| 2 | Jowar Roti | 2 | 62 |
| 2 | Spinach | 100gms | 1 |
| 2 | Rice | 100 gms | 66 |
| 2 | Dal | 100 gms | 29 |
| 2 | Salad | 250 gms | 15 |
| 2 | Buttermilk | 200 ml | 21 |
| 2 | Bajra Roti | 2 | 54 |
| 2 | Kidney Beans | 100gms | 24 |
| 2 | Fenugreek | 100gms | 19 |
| 2 | Jeera Rice | 100gms | 53 |

FIG : FOOD ITEMS WITH WITH GLYCEMIC INDEX

**EDA :** Exploratory Data Analysis (EDA) :

1) df.columns : to display all the column names in the dataset
2) df.info : printing information about the data frame(column labels, data types, number of cells, memory usage )
3) df.isnull().sum() : To check whether columns in a dataframe contain null values.
4) Plotting of histograms for all numeric attributes
5) Subplot is a matplotlib function that describe the layout of the chart.
6) Distplot is a function from seaborn package. It is use to visualize parametric distribution of a dataset.
7) Corelation analysis(heatmap) : A correlation heatmap is a graphical representation of a correlation matrix representing the correlation between different variables. The value of correlation can take any value from -1 to 1.
8) Pairplot using seaborn package : To plot multiple pairwise bivariate distributions in a dataset

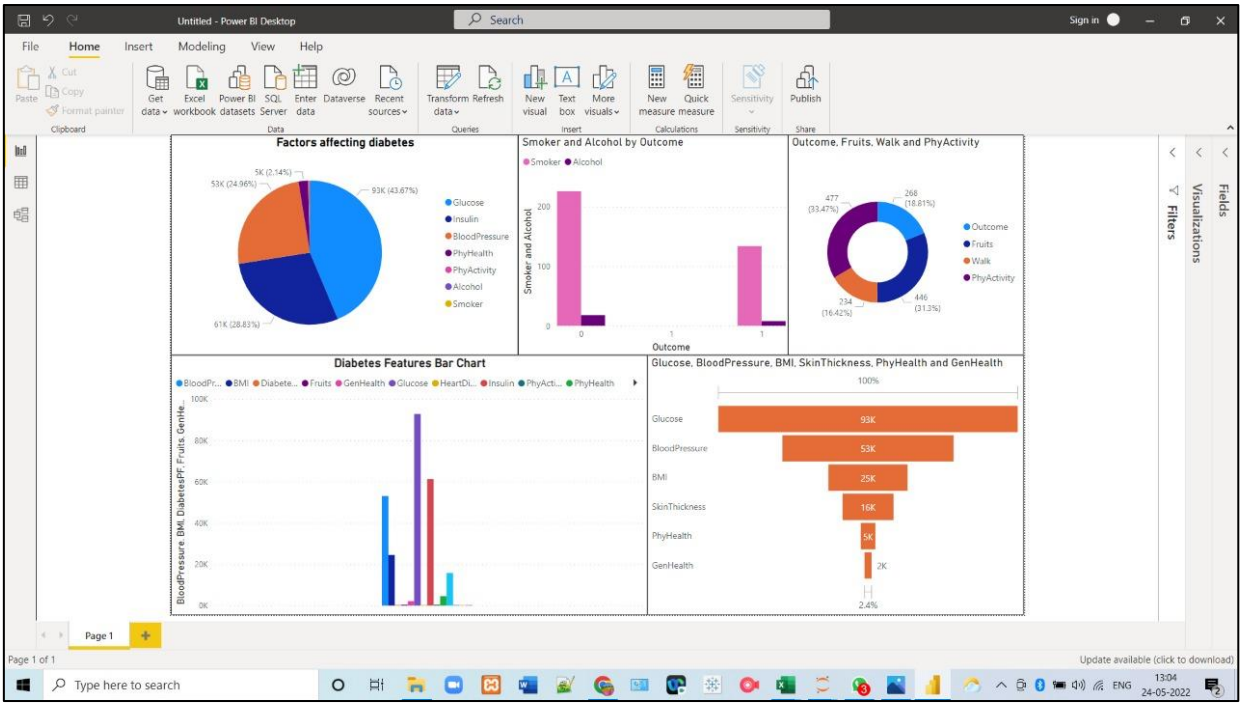



FIG: POWERBI DASHBOARD OF DATAVISUALIZATION FOR DIFFERENT ATTRIBUTES

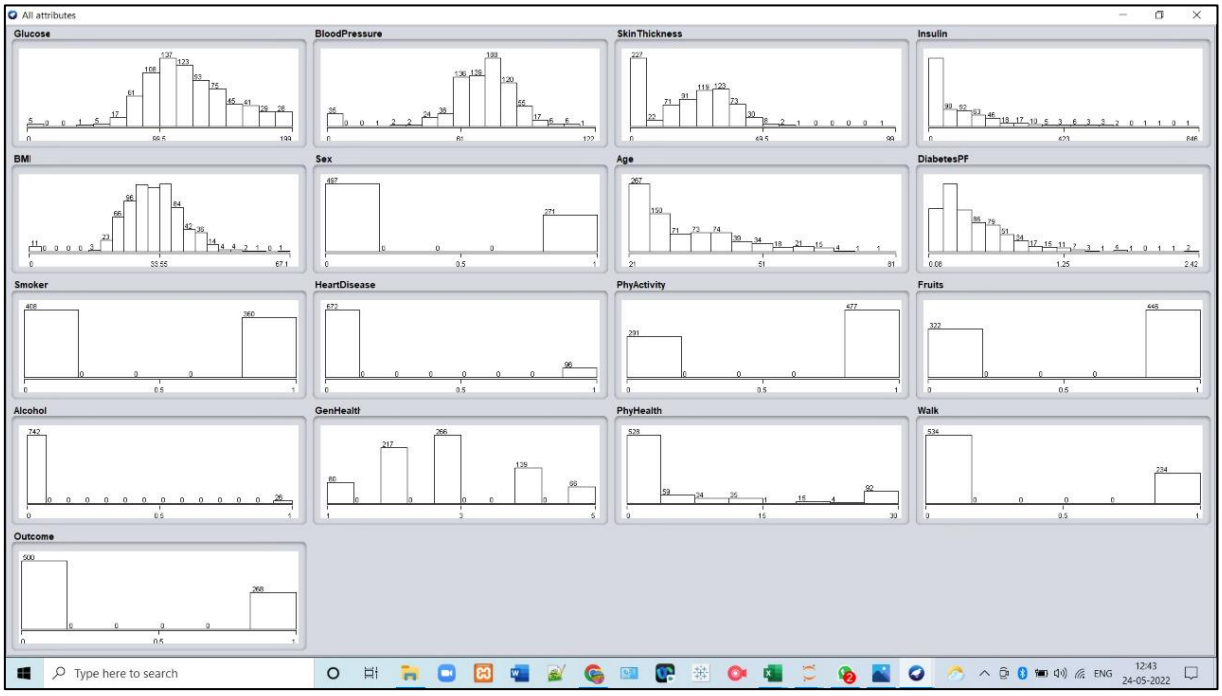



FIG : HISTOGRAM OF ALL ATTRIBUTES USING WEKA

**PRE- PROCESSING :**

Acquire the dataset: It is hard to consider any need for data quality for large acquired datasets without cleansing.so we need to acquire proper quality data.

Import all the crucial libraries: IMPORTING numpy for using array operations, pandas for using dataframe, matplotlib for graph plotting, seaborn for visualizing, sklearn for tprovides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import math
import random
import csv
sns.set()

from mlxtend.plotting import plot_decision_regions
import missingno as msno
from pandas.plotting import scatter_matrix

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
```

1) load the dataset : read csv **df = pd.read_csv('E:/Project/Sugarlevel_1.csv')**
   **df.head()**

2) Identifying and handling the missing values: for this we need to see is there is any missing values using commands
   **df.isnull().sum()**

3) Encoding the categorical data.: we need to convert categorical data into numeric data .BUT in this dataset the data is alredy converted into numeric values.

4) Splitting the dataset: splitting the data into training and testing using sklearn
   **from sklearn.model_selection import train_test_split**
   **X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=0)**

**ALGORITHM USED :**

1) **Logistic Regression** : Classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

2) **Decision tree** : Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

3) **Random forest** : Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

**Step-1:** Select random K data points from the training set.

**Step-2:** Build the decision trees associated with the selected data points (Subsets).

**Step-3:** Choose the number N for decision trees that you want to build.

**Step-4:** Repeat Step 1 & 2.

**Step-5:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

4) **Support vector machine** : Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

SVM algorithm can perform really well with both linearly separable and non-linearly separable datasets.

5) **KNN classification** : For a given data point in the set, the algorithms find the distances between this and all other **K** numbers of datapoint in the dataset close to the initial point and votes for that category that has the most frequency. A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If K = 3, then the case is simply assigned to the class of its nearest neighbour.

K-nearest neighbours (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems

**RESULT :**

DATASET1: FOR DIABETES PREDICTION(diabetes_project : dataset name)

1) **Logistic Regression**: Using logistic regression for this dataset, taking features except the target variable in x and outcome(target variable) in y, after training and testing the model the **accuracy is 0.7012**

2) **Decision tree** : In Decision tree classifier for this dataset, Decision trees classify the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node , after training and testing the model the **accuracy is 0.7099**
3) **Random Forest** : Using Random Forest for this dataset, train the model with the help of RandomForestClassifier class of sklearn we need to make prediction, after training and testing the model the **accuracy is 0.7445**
4) **Support vector machine** : Using SVM for this dataset, works by making a straight line between two classes. That means all of the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category, after training and testing the model the **accuracy is 0.7619**

DATASET2: FOR SUGARLEVEL CLASSIFICATION(Sugarlevel_1:dataset name)

For this dataset we tried to classify the data into three categories high, low, normal

1) KNN CLASSIFICATION: In KNN Classification, the accuracy after classifying the data is **1.0**
2) DECISION TREE FOR CLASSIFICATION : similarly, Decision tree for classification the accuracy is **1**
3) Support Vector Machine(SVM) : And, IN support vector machine(svm) the accuracy is **0.99**

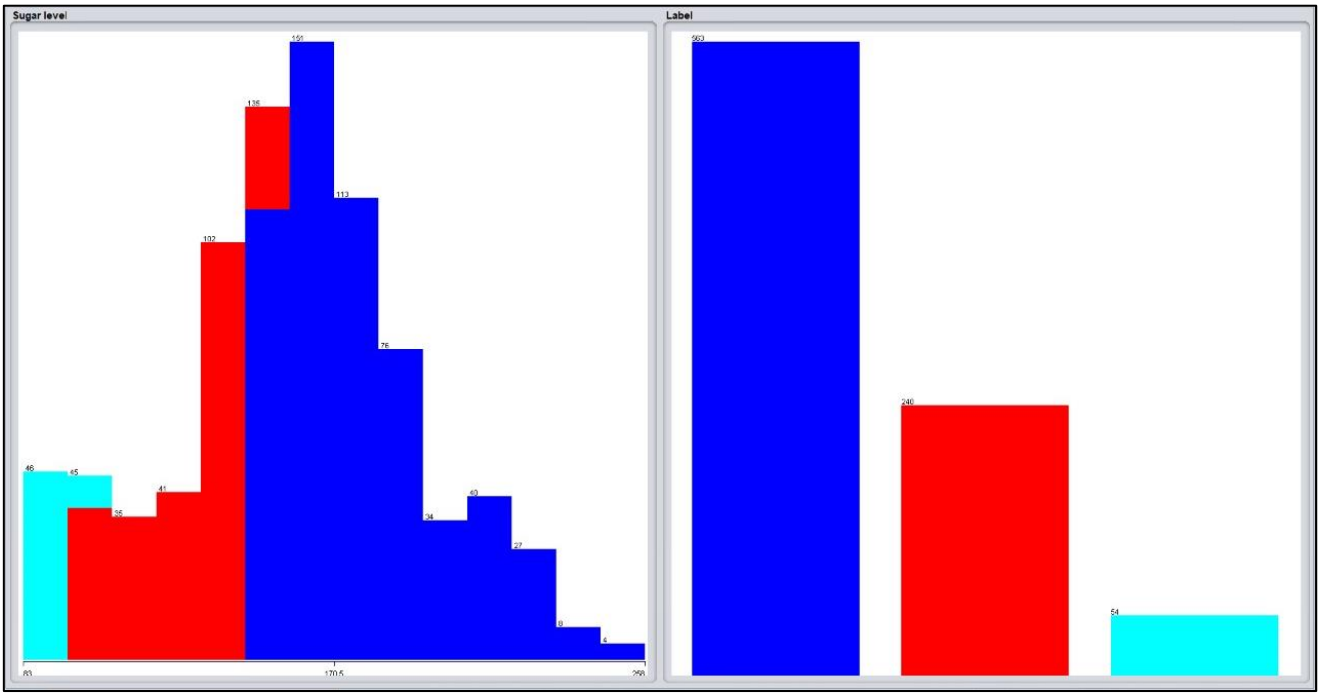



FIG : CLASSIFY DATA INTO HIGH,LOW AND NORMAL CATEGORY USING WEKA

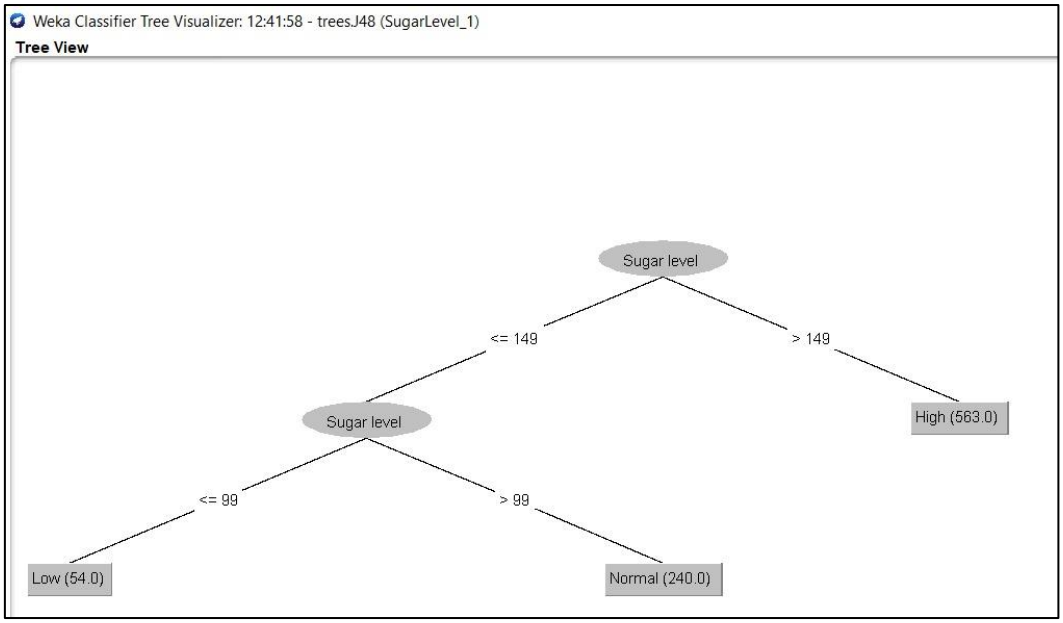
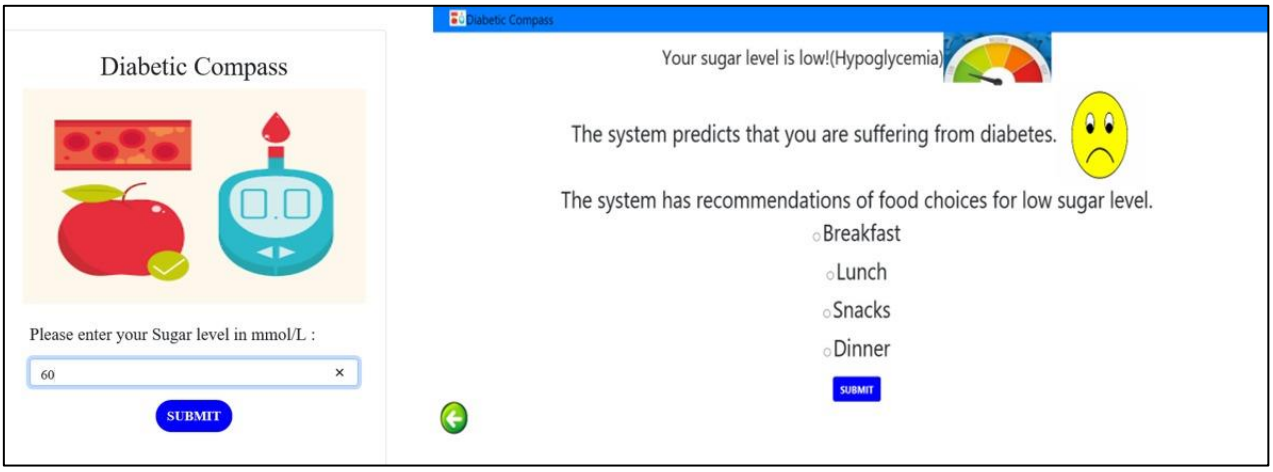
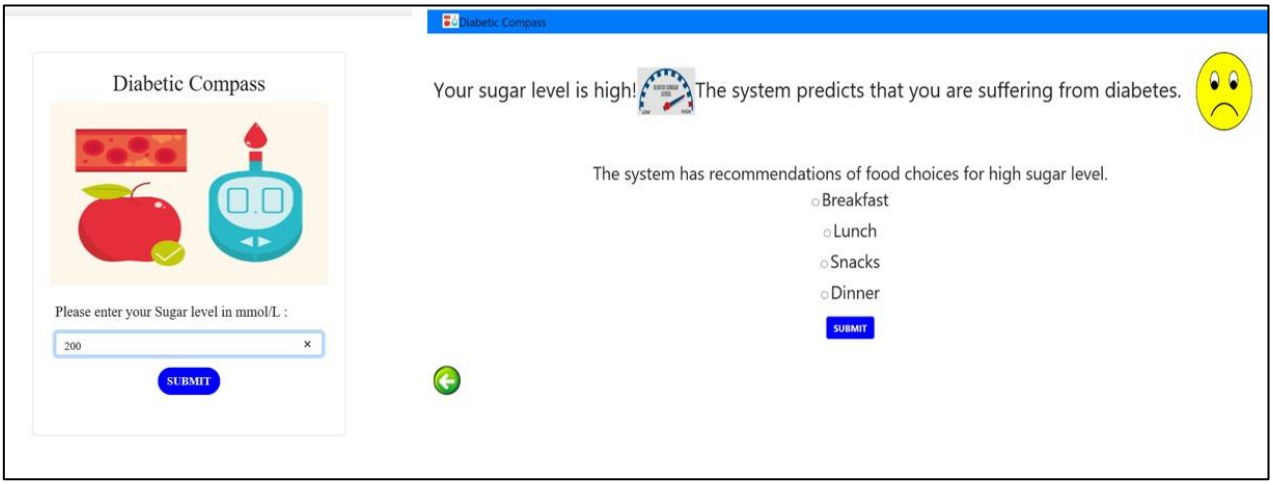
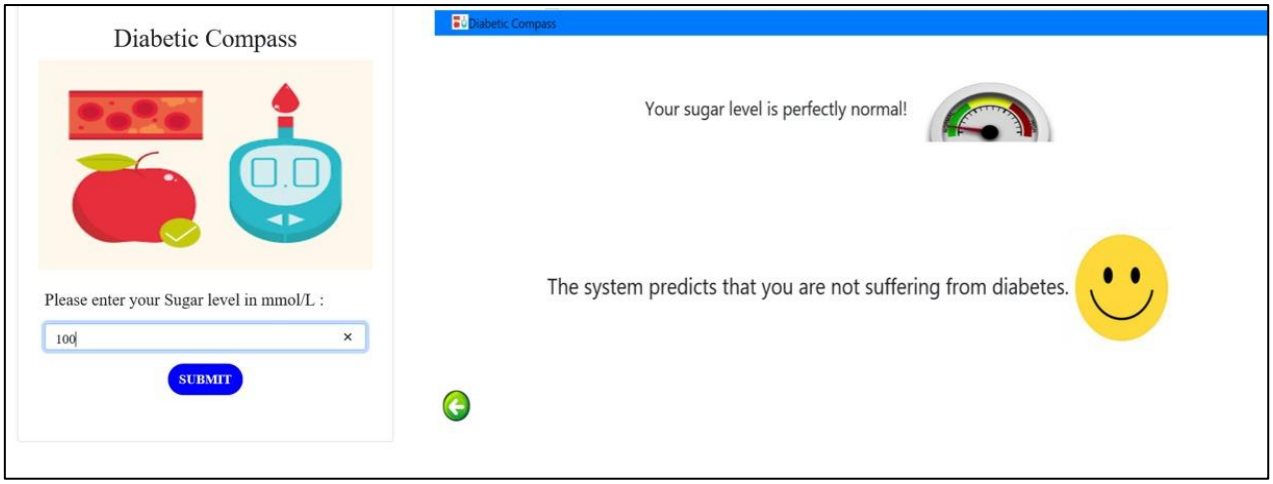


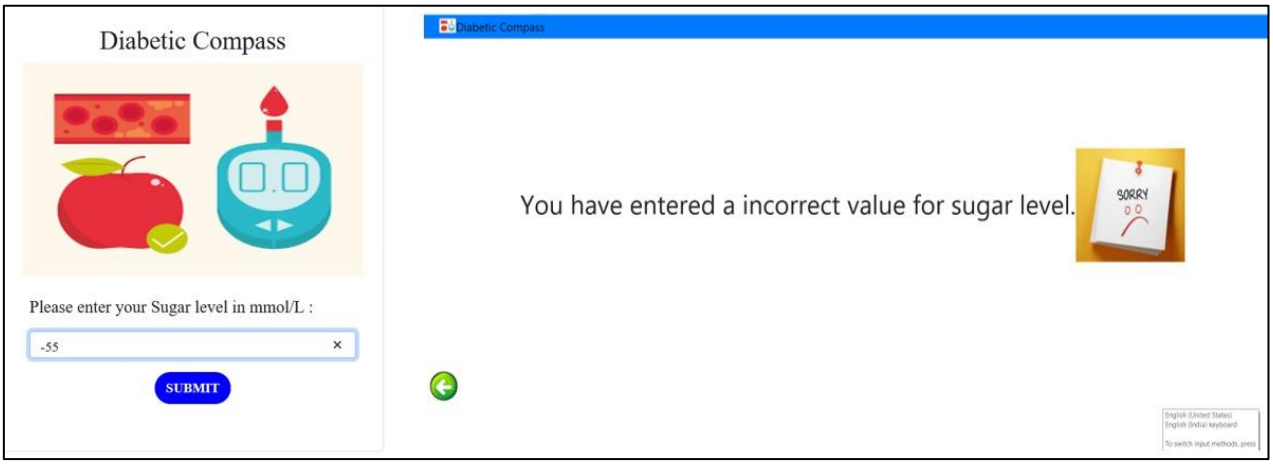

**FIG : DECISION TREE FOR CLASSIFYING SUGAR LEVEL DATA USING WEKA**

**Front End Design** : FRONT END DESIGNS ARE AS FOLLOW




Diabetic Compass
Please enter your Sugar level in mmol/L :
100
SUBMIT
Diabetic Compass
Your sugar level is perfectly normal!
The system predicts that you are not suffering from diabetes.


Diabetic Compass
Please enter your Sugar level in mmol/L :
-55
SUBMIT
Diabetic Compass
You have entered a incorrect value for sugar level.
SORRY

# Conclusion and Future Work

**Conclusion :**

1) From the above result we can conclude that Support Vector Machine(SVM) is suitable algorithm for the prediction of sugar having accuracy 0.7619

2) For classifying into high, low and normal category we can conclude KNN classification and decision tree is suitable algorithm

3) Food items with high glycemic index are recommended to people having low sugar level and food items with low glycemic index are recommended to people having high sugar level.

**Future Work**:

1) We'll add backend to our website and also make it more user-friendly.

2) We will also add new modified recipes of Cuisines like Chinese, Mexican, Italian, etc. which will also be healthy for the diabetic patients.

3) For more advancement we are also thinking to add another one model in which if you give a food item then the model will predict how much sugar level will be increased 2 or 3 hours after having that item.

# References

**References:**


1) Max Ray,. DIABETES -TYPE 2: The Review of Diabetic Studies. (2019). Available online at: https://www.researchgate.net/publication/336634065_DIABETES_-TYPE_2 (accessed September, 2021).
2) Soleh M, Ammar N, Sukmadi I. Website-based application for classification of diabetes using logistic regression method. Jurnal Ilmiah Merpati. (2021) 9. Retrieved from: https://ojs.unud.ac.id/index.php/merpati/article/view/66691 (accessed November, 2021).
3) https://www.verywellhealth.com/glycemic-index-chart-for-common-foods-1087476
4) https://www.verywellfit.com/glycemic-index-food-lists-2242513#toc-research
5) https://www.hindawi.com/journals/jhe/2021/9930985/#literature-review
6) https://www.researchgate.net/publication/347091823_Diabetes_Prediction_Using_Machine_Learning
7) https://glycemic-index.net/gi/foods-with-low-gi/
8) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232260/
9) https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques
10) https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf
11) Organisation mondiale de la santé . Rapport mondial sur le diabète. (2016). p. 88 [Google Scholar]
12) Medtronic . Le Diabète En Quelque Mots. [En ligne]. Available online at: https://www.parlonsdiabete.com/parlonsdiabete/le-diabete-en-quelques-mots (accessed November, 2021).
13) Max Ray,. DIABETES -TYPE 2: The Review of Diabetic Studies. (2019). Available online at: https://www.researchgate.net/publication/336634065_DIABETES_-TYPE_2 (accessed September, 2021).
14) Kumari VA, Chitra R. Classification de la maladie du diabète à l'aide d'une machine à vecteur de soutien. IJERA. (2013) 3:1797–801. [Google Scholar]
15) Ahmed TM. Using data mining to develop model for classifying diabetic patient control level based on historical medical records. J Theor Appl Inf Technol. (2016) 87:316–23. [Google Scholar]
16) Shetty D, Rit K, Shaikh S, Patil N. Diabetes disease prediction using data mining. In: Innovations in information, embedded and communication systems (ICIIECS) international conference. Coimbatore, India (2017). p. 1–5. [Google Scholar]
17) Bhoia SK, Pandab SK, Jenaa KK, Abhisekhc PA, Sahood KS, Samae NU, et al.. 'Prediction of Diabetes in Females of PimaIndian Heritage: A Complete Supervised Learning Approach'. Turk J Comput Math Educ. (2021) 12:3074–84. [Google Scholar]
18) Kandhasamy JP, Balamurali S. Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci. (2015) 47:45–51. 10.1016/j.procs.2015.03.182 [CrossRef] [Google Scholar]
19) Vijayanv V, Ravikumar A. Study of data mining algorithms for prediction and diagnosis of diabetes mellitus. Int J Comput Appl. (2014) 95:12–6. 10.5120/16685-6801 [CrossRef] [Google Scholar]
20) Soleh M, Ammar N, Sukmadi I. Website-based application for classification of diabetes using logistic regression method. Jurnal Ilmiah Merpati. (2021) 9. Retrieved from: https://ojs.unud.ac.id/index.php/merpati/article/view/66691 (accessed November, 2021).

**21)** Rajput DS, Basha SM, Xin Q, Gadekallu TR, Kaluri R, Lakshmanna K, et al.. Providing diagnosis on diabetes using cloud computing environment to the people living in rural areas of India. J Ambient Intell Humaniz Comput. (2021). [Google Scholar]
**22)** Deepa N, Prabadevi B, Maddikunta PK, Gadekallu TR, Baker T, Khan MA, et al.. An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. J Supercomput. (2021) 77:4–16. 10.1007/s11227-020-03347-2 [CrossRef] [Google Scholar]
**23)** Kumar DA, Govindasamy R. Performance et évaluation des techniques d'exploration de données de classification dans le diabète. IJCSIT. (2015) 6:1312–9. [Google Scholar]
**24)** Ouamri O. Contribution des arbres dirigés et les k-means pour l'indexation et recherche d'images par contenu, Mémoire de Magister en Informatique, by H. fizazi izabatene, Université des Sciences et de la Technologie d'Oran- Mohamed Boudiaf, Département d'Informatique (2011). p. 113 [Google Scholar]
**25)** Java T point. Naïve Bayes Algorithm. [En ligne]. Available online at: https://www.javatpoint.com/machinelearning-naive-bayes-classifier (accessed July, 2021).
**26)** Rokach L, Maimon. Exploration de données avec arbres de décision: théorie et applications. 2e édition, World Scientific Pub Co Inc. (2015). Available from: http://cedric.cnam.fr/vertigo/cours/ml2/coursArbresDecision.html
**27)** Kandhasamy JP, Balamurali, S,. Procedia Computer Science. Elsevier: (2015). Available online at: https://scholar.google.com/citations?user=LYSOeWMAAAAJ&hl=fr&oi=sra [Google Scholar]
**28)** Kaggle UCI Machine Learning, Base de donnée Pima Indian Diabetes. Available online at: https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed November, 2021).
**29)** Kaggle Johan Ensemble de données sur le diabète, extrait de l'hôpital de Francfort Allemagne. Available online at: https://www.kaggle.com/johndasilva/diabetes (accessed November, 2021).
**30)** Sisodia D, Sisodia DS. Prediction of Diabetes using Classification Algorithms. Procedia Comput Sci. (2018) 132:1578–85. 10.1016/j.procs.2018.05.122 [CrossRef] [Google Scholar]
**31)** Amel S, Karima R. The prediction of diabetes using machine learning algorithms, Master's thesis in Computer Science, supervised by Brahimi Farida, University AMO of Bouira Faculty of Sciences and Applied Sciences, Department of Computer Science (2019). p. 85. [Google Scholar]
**32)** Nishat MM, Faisal F, Mahbub MA, Mahbub MH, Islam S, Hoque MA. Performance assessment of different machine learning algorithms in predicting diabetes mellitus. Biosc Biotech Res Comm. (2021) 14. 10.21786/bbrc/14.1/10 [CrossRef] [Google Scholar]
**33)** Daanouni O, Cherradi B, Tmiri A. Predicting Diabetes Diseases Using Mixed Data and Supervised Machine Learning Algorithms. SCA2019 (2019). [Google Scholar]
**34)** Onyema EM. Opportunities and challenges of use of mobile phone technology in teaching and learning in Nigeria-a review. IJREI. (2019) 3:352–8. 10.36037/IJREI.2019.3601 [CrossRef] [Google Scholar]
**35)** Onyema EM, Elhaj MAE, Bashir SG, Abdullahi I, Hauwa AA, Hayatu AS. Evaluation of the Performance of K-Nearest Neighbor Algorithm in Determining Student Learning Styles. Int J of Innovative Sci, Eng & Techn. (2020) 7:91–102. [Google Scholar]
**36)** Shariq AB, Muhammad WA, Syed AH, Arindam G, Onyema EM. Smart Health Application for Remote Tracking of Ambulatory Patients. In: Hafizul Islam SK, Samanta D, editors. Smart Healthcare System design: Security and Privacy Aspects. Wiley Online Library (2021). p. 33–55. [Google Scholar]
**37)** Jo O, Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, et al.. COVID-19 patient health prediction using boosted random forest algorithm. Front Public Health. (2020) 8:357. 10.3389/fpubh.2020.00357 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

**38)** Celestine I, Suresh P. An efficient and unique TF/IDF algorithmic model-based data analysis for handling applications with big data streaming. Electronics. (2019) 8:28. 10.3390/electronics8111331 [CrossRef] [Google Scholar]
**39)** Celestine I, Mueen U, James A, Pascal N, Joseph HA, Ali KB. On detection of Sybil attack in large scale VANETs using spider-monkey technique. IEEE Access. (2018) 6:47258–67. 10.1109/ACCESS.2018.2864111 [CrossRef] [Google Scholar]
**40)** Jalil P, Celestine I, Praveen K, Gadekallu TR, Lakshmanna K, Bashir AK. A metaheuristic optimization approach for energy efficiency in the IoT networks. Softw Pract Exp. (2020) 14:2–9. 10.1002/spe.2797 [CrossRef] [Google Scholar]
**41)** Rajendran S, Khalaf OI, Alotaibi Y, Alghamdi S. MapReduce-based big data classification model using feature subset selection and hyperparameter tuned deep belief network. Sci Rep. (2021) 11:24138. 10.1038/s41598-021-03019-y [PMC free article] [PubMed] [CrossRef] [Google Scholar]
**42)** Rajalakshmi M, Saravanan V, Arunprasad V, Romero CAT, Khalaf OI, Karthik C. Machine learning for modeling and control of industrial clarifier process. IASC. (2022) 32:339–59. 10.32604/iasc.2022.021696 [CrossRef] [Google Scholar]
**43)** Alsubari SN, Deshmukh SN, Alqarni AA, Alsharif N, Aldhyani THH, Alsaade FW, et al.. Data Analytics for the Identification of Fake Reviews Using Supervised Learning. Comput Mater Contin. (2022) 70:3189–204. 10.32604/cmc.2022.019625 [CrossRef] [Google Scholar]
**44)** Khalaf OI, Abdulsahib GM. Optimized dynamic storage of data (ODSD) in IoT based on blockchain for wireless sensor networks. Peer-to-Peer Netw Appl. (2021) 14:2858–73. 10.1007/s12083-021-01115-4 [CrossRef] [Google Scholar]
**45)** Surendran R, Khalaf OI, Andres C. Deep learning based intelligent industrial fault diagnosis model. Comput Mater Contin. (2022) 70:6323–38. 10.32604/cmc.2022.021716 [CrossRef] [Google Scholar]