

Vidhyayana - ISSN 2454-8596

An International Multidisciplinary Peer-Reviewed E-Journal

www.vidhyayanaejournal.org

Indexed in: Crossref, ROAD & Google Scholar

10

Diabetes Classification and Diet Recommendation

Radhika Thakkar,

MSc. Data Science and Big Data Analytics, (*School of Computer Science & Engineering*)

MIT-WPU, Kothrud, Pune, India,

radhikathakkar2898@gmail.com.

Bhumika Ostwal,

MSc. Data Science and Big Data Analytics, (*School of Computer Science & Engineering*)

MIT-WPU, Kothrud, Pune, India,

bhumika.ostawal2@gmail.com.

Suraj Gandhi,

MSc. Data Science and Big Data Analytics, (*School of Computer Science & Engineering*)

MIT-WPU, Kothrud, Pune, India,

surajgandhi23@gmail.com.

Dhairya Shah,

MSc. Data Science and Big Data Analytics, (*School of Computer Science & Engineering*)

MIT-WPU, Kothrud, Pune, India,

dhairshah7@gmail.com

Guide:

Dr. Sumegh Tharewal

Assistant Professor, Program Head of M.Sc. Blockchain Technology,

MIT-WPU, Kothrud, Pune, India

sumeghtharewal@gmail.com



Abstract—

The rising prevalence of diabetes has become a critical subject in healthcare development. Type 2 diabetes, which was once considered a disease of the wealthy, is now affecting millions of people worldwide. Diabetes management is a challenging task that requires patients to monitor blood glucose levels, take medicine, eat a nutritious diet, and exercise frequently. In this research, we investigate diabetes types using Machine Learning Classification algorithms, including Logistic Regression, SVM, Decision Tree, Random Forest, and KNN. Our study aims to provide insights into the effectiveness of these supervised learning algorithms in classifying diabetes types. Additionally, we offer a website that recommends diets based on the level of diabetes. This research aims to contribute to the development of effective diabetes management strategies that can improve patients' quality of life. Diabetes is associated with a significantly increased risk of developing other diseases such as heart disease, renal disease, vision issues, nerve damage, and so on. Those with uncontrolled diabetes may also have impaired circulation, which causes the blood to circulate more slowly, making it difficult for the body to carry nutrients to wounds and causing the damage to heal more slowly. [15]

Keywords—*DT (Decision Tree), SVM, KNN, Logistic Regression (LR).*

I. INTRODUCTION

Our research focuses on type 2 diabetes mellitus and adult-onset Diabetes and its dietary requirements, aiming to predict and classify diabetes into high, low, and normal categories while recommending appropriate food items. Adult-onset Diabetes is a persistent health condition in which the body cannot efficiently regulate and utilize glucose, resulting in excess glucose in the blood. While typically seen in elderly populations, the rise in childhood obesity rates has resulted in a greater incidence of type 2 diabetes among younger individuals.

10% of diabetes cases are Type 1 (Insulin-dependent diabetes mellitus), which arises from the immune system that targets the cells in the pancreas that generate insulin. In contrast, Type 2 diabetes accounts for the other 90% and is primarily caused by insulin resistance and abnormal insulin interactions in cells located in the liver, fat, and muscle that lead to



inadequate sugar absorption. In some case scenarios, the pancreas may also not be producing the required amount of insulin to regulate blood sugar levels.

Although Type-2 diabetes has no known cure, it can be managed through a combination of lifestyle changes, including weight loss, healthy eating habits, and physical activity. If these changes are insufficient to regulate blood sugar levels, diabetes medications or insulin therapy may be necessary. Our research aims to help manage Type 2 diabetes by predicting and classifying it into appropriate categories and recommending dietary changes to improve patients' quality of life.

Table 1 shows the types of diabetes.

Table 1 Diabetes Categorization.	
Type-1 Diabetes	IDDM diabetes is a long-term autoimmune disorder that results in high blood sugar levels due to insufficient insulin production. Insulin therapy is crucial for managing glucose levels. It affects mostly young people and accounts for approximately 10% of diabetes cases. Effective management strategies are crucial to mitigate complications. Further research is needed to develop better treatments.
Type-2 Diabetes	Approximately 90% of patients have Type 2 diabetes, which is a metabolic condition. Its incidence is increasing in children and adolescents, making prevention strategies and effective treatments critical.
Gestational Diabetes (Diabetes in Pregnancy)	Gestational diabetes is a health condition that arises during pregnancy in women who have not had diabetes before. Although it typically resolves itself after childbirth, it heightens the chances of both the mother and infant developing type 2 diabetes in the future. It's important to recognize and manage gestational diabetes to reduce the risk of long-term health problems for both the mother and child.
Prediabetes	Prediabetes is a medical condition characterized by elevated blood sugar levels that do not meet the threshold for a type 2 diabetes diagnosis.

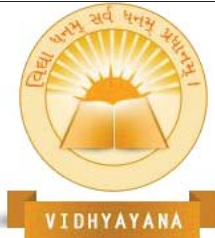


Table 1 Diabetes Categorization.

Type-1 Diabetes	IDDM diabetes is a long-term autoimmune disorder that results in high blood sugar levels due to insufficient insulin production. Insulin therapy is crucial for managing glucose levels. It affects mostly young people and accounts for approximately 10% of diabetes cases. Effective management strategies are crucial to mitigate complications. Further research is needed to develop better treatments.
	Shockingly, 96 mil US adults have prediabetes, but 84% of them are unaware. Identifying and managing prediabetes is crucial to prevent or delay the onset of type 2 diabetes and its related complications. .

II. LITERATURE SURVEY

J. Pradeep Kandhasamy [2], The study aimed to assess four distinct diabetes prediction models using eight crucial attributes. The accuracy rate of the J48 decision tree classifier was 73.82% prior to dataset pre-processing. On the other hand, the KNN where $k=1$ and Random Forest classifiers exhibited superior performance after pre-processing, attaining a flawless 100% accuracy rate when predicting diabetes, surpassing other models. This outcome suggests that the elimination of inaccurate data from the dataset can significantly improve the accuracy of diabetes prediction models. Therefore, the study's results offer valuable information for selecting the most appropriate classifier to predict diabetes.

Muhamad Soleh [3] researched the Logistic Regression method for classifying diabetes and developed software based on Streamlit. Data pre-processing, Logistic Regression, and method evaluation were the three stages involved in making predictions. The study utilized the correlation coefficient for data cleaning, and the One Point Crossover technique was used for oversampling. The Logistic Regression method produced a higher predictive accuracy of 80%, which was an improvement compared to earlier research that reported an accuracy rate of 75.97%.



Michael Onyema Edeh [4] The study proposed a diabetes diagnostic system utilizing the four machine learning algorithms in question naive Bayes, SVM, random forest, and DT. Among the methods, the random forest technique Showed the greatest level of accuracy, while the others also provided satisfactory outcomes. The primary aim of the research was to aid diabetologists in formulating more precise treatment plans and put forth potential research areas, such as constructing a diabetes database, investigating deep learning approaches, and devising an Android application for prediction.

Deepti Sisodia [5] conducted a study to create a system for diabetes prediction, employing three machine learning algorithms on the Pima Indians Diabetes Database. The researchers utilized the Naive Bayes classification method, achieving an accuracy rate of 76.30%. Other machine learning techniques can be integrated into the system for the detection and diagnosis of various ailments, and to automate diabetes analysis.

Nishat MM [6] presented a study wherein they compared and evaluated various machine learning algorithms for diabetes prediction. The Gaussian Process was proven to be the best most accurate and most efficient method, followed by Gradient Boosting, Random Forest, and Artificial Neural Networks. The study's objective is to assist healthcare providers in devising precise treatment regimens for type 2 diabetics and to explore the implications of developing an e-healthcare system.

KM Jyoti Rani [7] A study was conducted to create a system for early detection and recognition of diabetes, using the John Diabetes Database to evaluate five machine-learning classification methods based on different metrics. The Decision Tree algorithm achieved an accuracy rate of 99%, validating the system's effectiveness.

Quan Zou, [8] Machine learning techniques are chosen by the researcher to predict and make a diagnosis of diabetes. The study found that using all features and mRMR is more effective than using PCA. Fasting glucose is the most crucial indicator, but more indicators are needed for better results. Random forests perform better than decision trees and neural network classifiers in some methods. The Luzhou dataset generated the most accurate results, of 0.8084, trailed by the Pima Indians dataset, which prepared findings with an accuracy of 0.7721. The study emphasizes the importance of suitable attributes, classifiers, and data



mining methods for accurate diabetes prediction. Future work aims to predict diabetes type and explore the proportion of each indicator for improved prediction accuracy.

Max Ray, [9], Diabetes is a prevalent health issue caused by a deficiency in insulin production or inadequate insulin usage in the body. Type 1 diabetes is exacerbated by the pancreas producing suboptimal insulin, while Type 2 diabetes results from the inability to utilize insulin effectively. Gestational diabetes, on the other hand, develops during pregnancy due to increased insulin requirements. Several risk factors for Type 2 diabetes exist, including obesity, inactivity, smoking, age, family history, hypertension, Polycystic Ovary Syndrome, and a history of Gestational diabetes. Maintaining a healthy diet and engaging in regular exercise are effective preventive measures for managing weight and preventing diabetes.

Umair Muneer Butt, [10] Biosensors and advanced ICT can be utilized for real-time monitoring of glucose levels in diabetic patients, using portable devices and CGM sensors. This technology can help patients better comprehend their blood sugar changes. A system has been proposed that classifies and identifies the early stages of diabetes, using modern sensors, machine learning techniques, and IoT. Three prediction models (LSTM, MA, and LR) were used to analyze diabetes classifications made by three classifiers (random forest, multilayer perceptron, and logistic regression). Using the PIMA Indian Diabetes dataset, MLP was able to attain an accuracy of 86.083% in mellitus classification, while LSTM secured a prediction accuracy of 87.26%.

Mitushi Soni, [11] The goal of the research was to generate predictions of diabetes using multiple machines learning methods, including Decision Tree, Gradient Boosting, K-Nearest Neighbor, Logistic Regression, Random Forest, and Support Vector Machine. The Pima Indian Diabetes Dataset, which contained information about 768 patients such as pregnancy, age, BMI, BP, glucose, insulin, and diabetes pedigree function, was utilized. The study found that Random Forest was the most accurate technique for predicting diabetes compared to other methods. The class variable in the dataset represented the outcome of each data point, with 0 indicating negative and 1 indicating positive for diabetes.

Tarig Mohamed Ahmed [12] The paper introduces a hypothetical self-monitoring system for diabetes using IoT technology. The system relies on BLE devices for data collection and



real-time processing of weight and blood glucose data. It employs Apache Kafka for streaming messages and MongoDB for data storage.

Bhoia SK, [13] The objective of the study was to apply machine learning techniques such as Random-Forest (RF), K-NN, and logistic regression to forecast the likelihood of diabetes in females belonging to the Pima Indian population. Logistic regression outperformed other models based on metrics such as AUC, CA, F1, precision, and recall. The findings were derived using k-fold cross-validation and the Orange 3.24.1 platform, which uses Python open-source modules. For each approach, confusion matrices were created, with logistic regression performing the best.

Veena Vijayan V, [14] This research looks at how algorithms that mine data can be used to diagnose diabetes more precisely than traditional methods, which can be inaccurate. Using the Pima Indian Diabetic dataset, the accuracy of several methods, such as KNN, K-means, ANFIS, EM, and amalgam KNN, was compared. Amalgam KNN and ANFIS exceeded prior techniques in classification accuracy, with Amalgam KNN attaining over 80% accuracy. Symptoms of diabetes consist of heightened thirst, frequent urination, unintended loss of weight, and slow-healing infections. Among the diagnostic exams used to diagnose diabetes are urine tests, fasting blood sugar levels, random blood sugar levels, and glycosylated hemoglobin (HbA1c).

Rishab Bothra [15] The study includes comparing the accuracy of various machine learning algorithms used to classify a dataset. The Random Forest algorithm was found to be the most accurate, with a 90% prediction accuracy. To ensure that the number of false negative predictions was kept to a minimum, confusion matrices were compared as well. The authors suggest that future research could investigate whether non-diabetic individuals are likely to develop diabetes in the next few years.

Table 2: Literature analysis:

Sr. No	Year	Author	Title	Dataset Used	Techniques Used	Results (Accuracy)
1	2015	J. Pradeep Kandhasamy, et. al [2]	Performance Analysis of Classifier	UCI machine learning data repository's	J48, SVM, Decision Tree, K-Nearest	Before any pre-processing is applied, the J48



			Models to Predict Diabetes Mellitus	example dataset.	Neighbors, and Random Forest.	classifier shows an accuracy of 73.82%. However, after applying pre-processing techniques, both the KNN algorithm with k=1 and the Random Forest algorithm achieved perfect accuracy results of 100%. paraphrase this shorter
2	2021	Muhamad Soleh, et. al [3]	Website-Based Application for Classification of Diabetes Using Logistic Regression Method	Pima Indian Diabetes Dataset.	Logistic Regression	The accuracy of logistic regression was 75.97%.
3	2022	Michael Onyema Edeh, et. al [4]	A Classification Algorithm-Based Hybrid Diabetes Prediction Model	Pima Indian Diabetes from UCI Machine Learning repository and database of the hospital from Germany	SVM, Random Forest, K-means, KNN and Naïve Bayes.	SVM performs best with the highest accuracy of 78.2%.
4	2018	Deepti Sisodia, et, al [5]	Prediction of Diabetes Using Classification Algorithms	PIDD-Pima Indians Diabetes Dataset	Naive Bayes, Support Vector Machine, and Decision Tree (DT)	Naive Bayes achieved the highest accuracy rate of 76.30%, making it the best-performing method. The Decision Tree method followed in second place, with an accuracy rate of 73.82%.



5	2020	Mirza Muntasir Nishat, et. al [6]	Performance Assessment of Different Machine Learning Algorithms in Predicting Diabetes Mellitus	Kaggle Diabetes Dataset, Frankfurt Hospital, Germany	Linear Regression, Naive Bayes, Support Vector Machine, Adaboost, Stochastic Gradient Descent, Gradient Boosting, Random Forest, Gaussian Process, K-Nearest Neighbors, Artificial Neural Network	Gaussian Process performs best with the highest accuracy of 98.25% and the second is Random Forest with an accuracy of 97.25%.
6	2020	KM Jyoti Rani [7]	Diabetes Prediction Using Machine Learning	Diabetes Dataset	Linear Regression, Decision Tree, SVM, Random Forest, K-NN	The decision Tree performs best with the highest accuracy of 99% and the second is Random Forest with an accuracy of 97%.
7	2018	Quan Zou, et. al [8]	Predicting Diabetes Mellitus with Machine Learning Techniques	obtained from the hospital in China	Random Forest, J48, Neural Networks	Random Forest performs best with the highest accuracy of 80.84% and the existing J48 with an accuracy of 78.53%.
8	2021	Umair Muneer Butt, et al [10]	Machine Learning-Based Diabetes Classification and Prediction for Healthcare Applications	PIMA Indian Diabetes dataset.	RF, Multilayer perceptron model, Logistic Regression, LSTM, Moving	Multilayer perceptron = 86.08%; LSTM = 87.26%. These two models outperform the other models



					Average, and Liner Regression	used.
9	2020	Mitushi Soni, et al [11]	Diabetes Prediction Using Machine Learning Techniques	Pima Indian Diabetes Dataset, Data collected from patients	K-NN, Logistic Regression, Decision Tree, SVM, and Gradient Boost.	Achieved 77% classification accuracy.
10	2016	Tarig Mohamed Ahmed [12]	Using Data Mining to Develop a Model For Classifying Diabetic Patient Control Level Based On Historical Medical Records	Risk Factors in Saudi Arabia Collected From WHO	Naïve Bayes, Logistic and J48	Logistic regression outperformed other models with an accuracy score of 74.4%.
11	2021	Bhoia SK, et al [13]	Prediction of Diabetes in Females of PimaIndian Heritage: A Complete Supervised Learning Approach	Female Pima Indians diabetic dataset From Kaggle and UCI data repository	Classification Tree, SVM, KNN, Naïve Bayes, Random Forest, Neural Network, AdaBoost, and Logistic Regression.	The logistic regression model attained a precision rate of 76.8%, whereas the neural networks model displayed an accuracy of 75.8%.
12	2021	Rishab Bothra [15]	Diabetes Prediction Using Machine- Learning Algorithms	Sample data	Random- Forest, Logistic Regression, xgboost, SVM, and KNN.	Random Forest gives the best accuracy of 90% after which Knn and XGBoost have an accuracy of 89% and 88% respectively.



II. PROPOSED ALGORITHMS

1. Logistic Regression:

Logistic regression is a statistical Method that establishes a connection between predictor variables and a binary categorical outcome. The probability of an event occurring is calculated using a linear combination of independent factors. This method is frequently used to estimate the likelihood that a certain character will occur in a binary variable. The technique is encapsulated in the logistic regression mathematical equation.:

$$\text{logit}(P(\text{diabetes} = 1)) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \dots + \beta_kX_k$$

were,

$\text{logit}(P(\text{diabetes}=1))$ is used to determine the logarithm of the probability of a person having diabetes, which is expressed as the dependent variable.

β_0 represents the coefficient that corresponds to the constant term or intercept in the equation.

β_1 to β_k are the coefficients are linked with the independent variables X_1 to X_k , respectively.

X_1 to X_k are the independent variables, that represent various factors such as age, BMI, blood pressure, physical activity, and so on.

In logistic regression, the probability of having diabetes is calculated using a function of the independent variables. This involves adding up the products of the independent variables and their associated coefficients to get the linear combination, which is then used to predict the probability of diabetes.:

$$z = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \dots + \beta_kX_k$$

The logistic function commonly referred to as the sigmoid function is used to estimate the anticipated likelihood of diabetes after computing the linear combination of the independent variables and their coefficients:

$$P(\text{diabetes} = 1) = 1 / (1 + e^{(-z)})$$

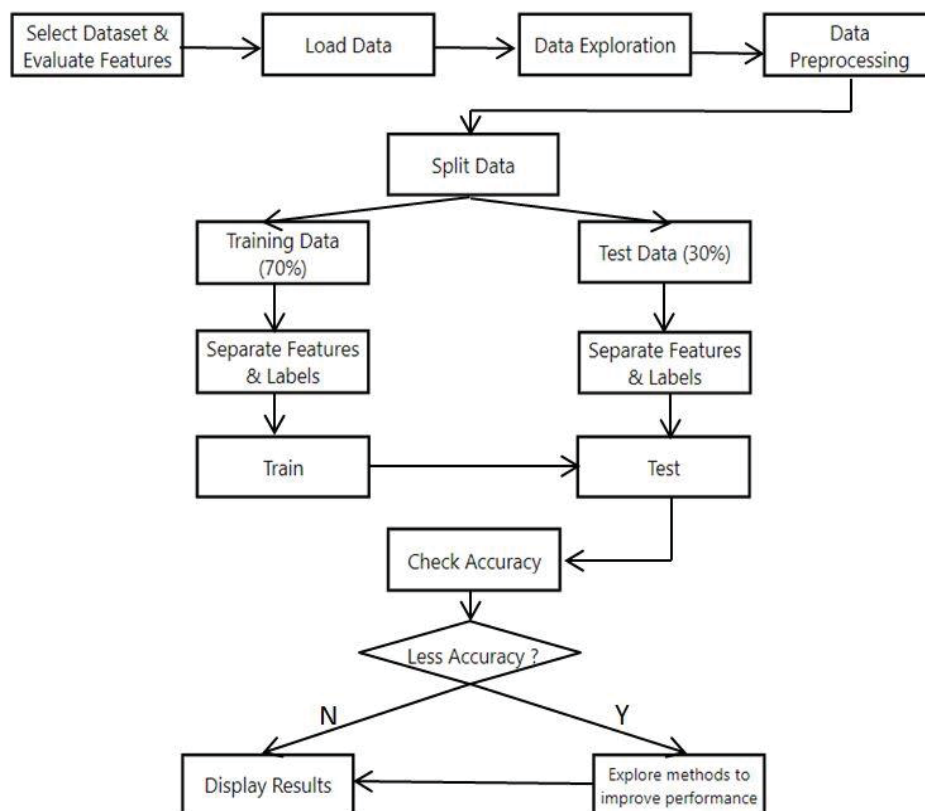
The logistic regression equation is useful for predicting the likelihood of diabetes in a patient based on their attributes, such as age, BMI, blood pressure, and physical activity. It can also

help determine which independent variables have the strongest correlation with the probability of developing diabetes.

2. *Random Forest:*

Random forest is a commonly used machine learning technique for classification tasks. This method utilizes an ensemble of decision trees that are merged to produce a more precise and dependable model. It is possible to use random forest to examine the relationship between a variety of characteristics or risk factors, such as age, BMI, and blood pressure, and the likelihood of acquiring diabetes.

Fig 1: Flow of Algorithm:



a) *Prepare the data:*

Gather and preprocess the dataset containing details of patient attributes, including age, BMI, blood sugar levels, blood pressure, family history, etc., and the binary classification indicating the presence or absence of diabetes (0 or 1).



b) ***Split the data:***

Divide the dataset into two subgroups, one for modeling and the other for the test. The performance of the model is evaluated using the test set after constructing it with the training set.

c) ***Select the desired quantity of trees.:***

Choose the best number of trees to use in the model; this is a hyperparameter that must be improved for the best results.

d) ***Train the model:***

After selecting the appropriate number of trees, the model will be trained on the training dataset. For each variable, a decision tree will be constructed by the model to evaluate its significance in forecasting diabetes.

e) ***Evaluate the model:***

The efficiency of the model can be assessed using the testing dataset by employing diverse metrics such as accuracy, precision, recall, and F1-score to evaluate its performance.

3. ***Decision Tree:***

Decision trees are a machine learning technique that categorizes or predicts a target variable using input information. The method creates a structure that resembles a tree, with each branch representing a choice made in light of the feature values. Before the final prediction is reached, the data is split into smaller subgroups by a series of judgments. Decision trees offer a flexible Method that can be utilized for classification and regression assignments without imposing any assumptions regarding the distribution of the data. Here is the basic equation for a decision tree:

$$y = f(x)$$

were,

y represents the outcome;

x represents a group of independent variables or features utilized to estimate or predict the dependent variable;



$f(x)$ is a mathematical representation that takes the input features as an input and predicts the target variable.

Decision trees can be applied to predict diabetes in patients based on input features such as age, BMI, glucose levels, and family history. This algorithm generates a tree-like structure, where each split corresponds to a decision based on the input features. This decision leads to the grouping of patients into different categories, and the final leaves of the tree represent the predicted classification for each patient (either diabetic or not diabetic). Healthcare experts can decide on patient care and treatment strategies after looking at the input characteristics and building the decision tree.

4. SVM:

Support Vector Machine (SVM) is a machine learning technique that separates data into two categories, diabetic and non-diabetic, without making any assumptions about the data distribution. It seeks to identify the hyperplane with the greatest margin of separation between the two classes. This equation expresses that the weight vector is represented by the variable W , X is the input or feature vector that contains the values of the predictor variables, and b refers to the bias term, $W \cdot X + b = 0$ is an equation that represents the hyperplane. The SVM method looks for the W and b values that will produce the greatest margin between the two classes.

In order to classify data into groups of diabetes and non-diabetes, Support Vector Machine (SVM), a machine learning method, is utilized. It looks for the hyperplane that separates the two classes by the largest amount. The hyperplane $\|W\|^2$ is represented by the equation $y_i * (W * X_i + b) \geq 1$, In which W represents the weight vector, X stands for the feature vector, and b denotes the bias factor. The W and b parameters that will result in the largest margin between the two classes are sought for by the SVM algorithm.

5. KNN:

The K-Nearest Neighbor (KNN) is a powerful machine-learning technique for solving both classification and regression problems. By comparing a patient's resemblance to other patients in the dataset, KNN may be utilized to classify patients as diabetic or not when used



for diabetes prediction. Because it is non-parametric, it does not assume anything about how the data are distributed.

The KNN algorithm selects the K value, the number of neighbors to consider, and then compares the similarity of each patient to other patients in the dataset based on their input features, including age, BMI, glucose levels, family history, and other related factors.

The KNN method calculates the predicted class by selecting the most frequent class label among those K-nearest neighbors after identifying the K-nearest neighbors for a specific patient. The algorithm will predict that the patient has diabetes, for example, if among the K closest neighbors, 4 are categorized as diabetic and 1 is not.

In mathematical notation, the KNN algorithm can be represented as follows:

$$y = \text{mode}(y_1, y_2, \dots, y_k)$$

where y is the mode function, which yields the class that is most prevalent among the patient's K closest neighbors, determines the projected class for a patient, and y_1, y_2, \dots, y_k is the classes of the K nearest neighbors, selected based on their similarity in terms of input features.

IV. PROPOSED METHODS:

Dataset Collection:

Data collection is gathering and compiling information about diabetic patients is known as data collection, which involves collecting three distinct datasets: one for sugar levels, one for diabetes, and one for food.

There are 3 types of data that we used for our research.

a. *Sugar-level:*

Data on blood sugar levels, which shows variations in diabetes patients' blood sugar levels every 15 minutes, was gathered with assistance from one of the institution's professors. The 857-row dataset may be used to study the correlation between sugar levels. Blood sugar values under 150 mg/dL (7.8 mmol/L) are regarded as normal; but, after two hours, readings over 200 mg/dL (11.1 mmol/L) are indicative of diabetes. Prediabetes is suggested by readings between 140 and 199 mg/dL (7.8 mmol/L and 11.0 mmol/L). Based on their



Glycemic Index (GI), food items were divided into three groups in order to suggest an optimal diet depending on a person's sugar-level history. Foods with a GI of 0 to 130 are classified as a LOW GI, those with a GI of 130 to 260 as MID GI, and those with a GI of 260 to 400 as HIGH GI.

b. *Diabetes data:*

The diabetes dataset contains information about people with diabetes and can be utilized for forecasting the probability of developing diabetes and pinpointing potential risk factors. It usually comprises details such as age, gender, medical history, and lifestyle choices, and comprises 768 rows of data.

c. *Food intake dataset:*

The food dataset includes data on the amounts and glycemic index ratings of various foods, as well as information about their nutritional composition, which may be used to provide customized dietary advice based on a person's health profile. The glycemic index is a measure of the speed at which a particular food causes a rise in blood sugar levels. Foods with a high glycemic index cause a rapid increase in blood sugar levels, while those with a low glycemic index are digested more slowly, resulting in a more gradual effect. With the food's glycemic index readings, nutritional advice may be more specifically tailored to a person's health requirements. 150 rows make up the food dataset.

Data Pre-Processing:

By resolving contradictory data, we hoped to increase the accuracy and precision of our findings throughout this study phase. We eliminated the ID feature from our dataset after noticing its consistency issues. Also, we found that numerous important variables, including age, blood pressure, skin thickness, BMI, and glucose level, were missing information. To make predictions, we identified key characteristics, imputed the missing data, scaled the data using StandardScaler, and identified these features. We concentrated on these aspects to provide more accurate findings for our study paper even though we did not undertake feature selection.

- During the exploratory data analysis phase, we carried out the following steps:



i. Data cleaning:

We reviewed the dataset to detect any discrepancies, errors, absent values, or extreme values and implement necessary measures to rectify them.

ii. Descriptive statistics:

Fig 2 provides derived statistical measures for the various aspects of the dataset, such as the mean, median, and standard deviation, to gain an understanding of their distribution and characteristics.

Fig 2: Descriptive Statistics:

	count	mean	std	min	25%	50%	75%	max
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.00000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.00000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.00000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.50000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.00000	36.60000	67.10
Sex	768.0	0.352865	0.478172	0.000	0.00000	0.00000	1.00000	1.00
Age	768.0	33.240885	11.760232	21.000	24.00000	29.00000	41.00000	81.00
DiabetesPF	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Smoker	768.0	0.468750	0.499348	0.000	0.00000	0.00000	1.00000	1.00
HeartDisease	768.0	0.125000	0.330934	0.000	0.00000	0.00000	0.00000	1.00
PhyActivity	768.0	0.621094	0.485431	0.000	0.00000	1.00000	1.00000	1.00
Fruits	768.0	0.580729	0.493761	0.000	0.00000	1.00000	1.00000	1.00
Alcohol	768.0	0.033854	0.180972	0.000	0.00000	0.00000	0.00000	1.00
GenHealth	768.0	2.861979	1.098399	1.000	2.00000	3.00000	4.00000	5.00
PhyHealth	768.0	5.916667	10.043495	0.000	0.00000	0.00000	7.00000	30.00
Walk	768.0	0.304688	0.460575	0.000	0.00000	0.00000	1.00000	1.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.00000	1.00000	1.00

iii. Data visualization:

Various types of charts and graphs, such as histograms, scatter plots, and box plots, were utilized to detect patterns, trends, and correlations among the attributes in the dataset. For instance, a box plot was employed to examine the outliers in the data, and these outliers were eliminated, as shown in the accompanying box plot.

Fig 3 & 4 Show the before and after boxplot for outliers:

Fig 3: Before Removing Outliers:



VIDHYAYANA

Vidhyayana - ISSN 2454-8596

An International Multidisciplinary Peer-Reviewed E-Journal

www.vidhyayanaejournal.org

Indexed in: Crossref, ROAD & Google Scholar

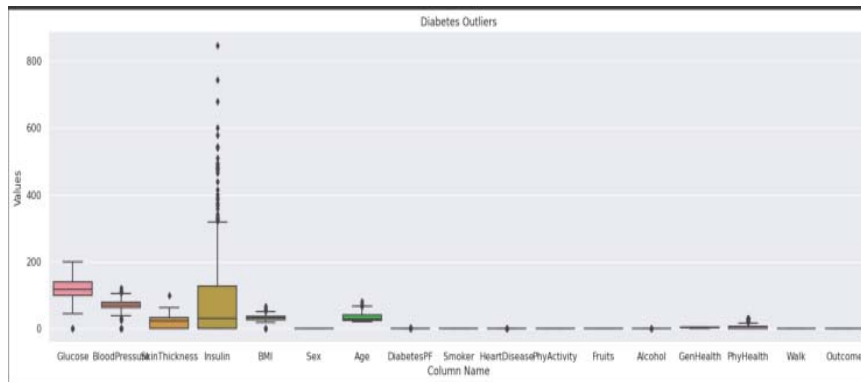


Fig 4: After Removing Outliers:

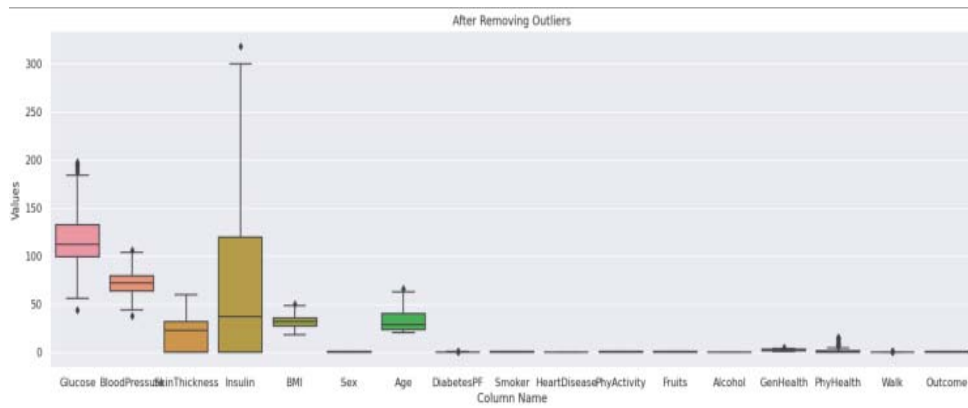


Fig 5: Code to detect outliers:

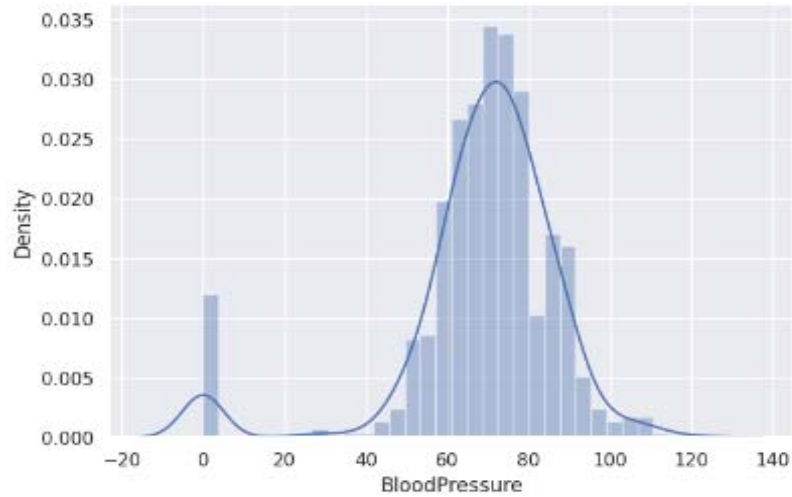
```
df = pd.read_csv('diabetes_project.csv')
# Identify columns of interest
cols = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'Sex',
        'Age', 'DiabetesPF', 'Smoker', 'HeartDisease', 'PhyActivity', 'Fruits',
        'Alcohol', 'GenHealth', 'PhyHealth', 'Walk', 'Outcome']

summary = df[cols].describe()

# Calculate outliers using the IQR method
Q1 = df[cols].quantile(0.25)
Q3 = df[cols].quantile(0.75)
IQR = Q3 - Q1
outliers = ((df[cols] < (Q1 - 1.5 * IQR)) | (df[cols] > (Q3 + 1.5 * IQR)))

# Visualize outliers using box plots
fig, ax = plt.subplots(figsize=(20,6))
sns.boxplot(data=df[cols], ax=ax)
ax.set_xlabel('Column Name')
ax.set_ylabel('Values')
ax.set_title('Diabetes Outliers')
plt.show()
```

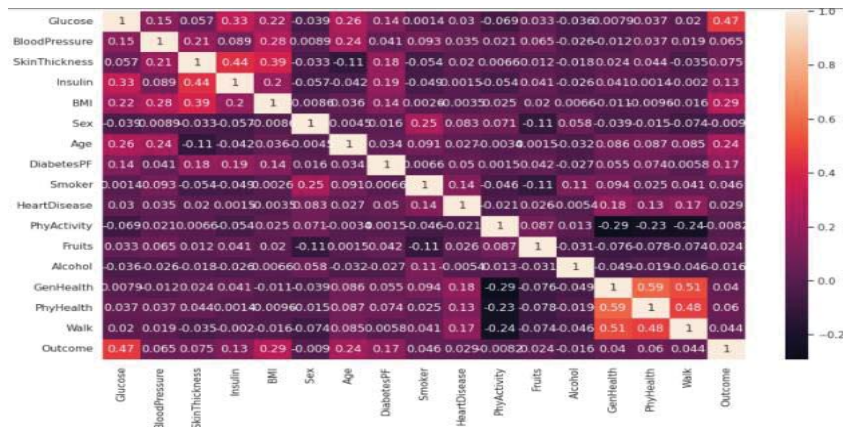
Fig 6: distplot for Blood pressure



iv. *Correlation analysis:*

We examined how various attributes in the dataset were correlated with one another to determine if there were significantly positive or negative associations between them.

Fig 7: Shows correlation between Features:



Through these procedures, we obtained a more comprehensive comprehension of the dataset, detected potential problems or tendencies, and made the data ready for predictive analysis.

v. *Model Building:*

In our research study, we used a variety of existing data models to create predictions regarding diabetes, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, Random Forest, and Logistic Regression.



KNN is a machine learning method that categorizes incoming data points based on how close they are to existing data points in the training set, without making any assumptions about the distribution of the underlying data. Another supervised machine learning technique, SVM, seeks to maximize the margin between two classes in a dataset to determine the best decision boundary dividing them. Models called decision trees categorize new data points using a sequence of binary judgments. To increase the model's accuracy, Random Forest is a type of ensemble learning method that constructs multiple decision trees and merges their forecasts. Based on input feature values, the statistical model of logistic regression calculates the likelihood that an event will occur.

We must evaluate the models' performance indicators, including precision, recall, accuracy, and F1 score, to ascertain how well they predict diabetes. These measures may be used to compare the effectiveness of several models and identify the one that operates most effectively on our dataset.

The feature importance indicates the relative importance of input variables in determining the model's outputs, providing insights into the factors that influence the prediction of diabetes.

In summary, employing various data models enabled us to conduct a comprehensive analysis of their effectiveness and determine the optimal model for predicting diabetes in our study.

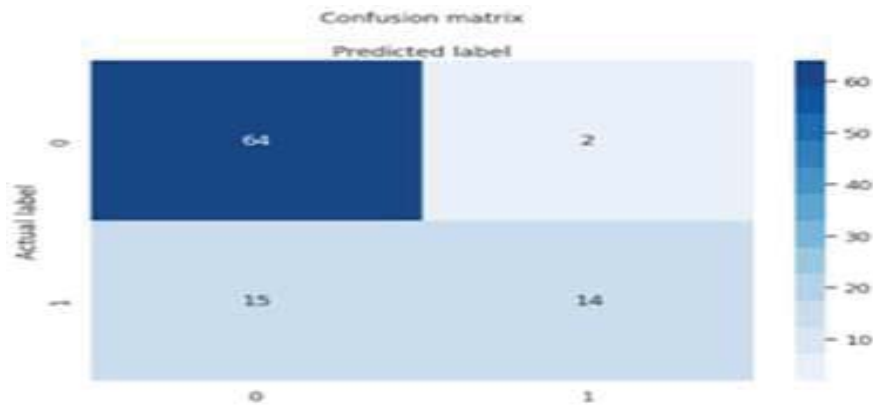
Algorithm 1: - Diabetes Prediction using Logistic Regression:

- Generate Train set and Test Set using a library from the model sklearn.
model_selection train_test_split
- Applying Logistic regression with scaler = "libliner" and then fit the logistic regression model and finally predicting the outcome.

In the end, the confusion matrix is utilized to assess the performance of a classification model by comparing its anticipated results with the actual ones. It facilitates the computation of different evaluation metrics such as accuracy, precision, recall, and F1-score, which assist in evaluating the model's effectiveness for diverse classes. The confusion matrix proves especially advantageous when handling imbalanced datasets or when the expenses associated with false positives and false negatives are dissimilar.

Logistic regression gives us the best accuracy of 82.10%.

Fig 8. Shows Confusion metrics for KNN:



Algorithm 2: - Sugar level classification using KNN and Decision Tree:

KNN:

- Create a train set and a test set using the same procedure as described earlier.
- Scale the data and then we used KNN with 5 neighbors.
- After generating the train and test sets using the same process as above, calculate the accuracy of the KNN model used to classify whether a patient has diabetes or not.

KNN gives us an accuracy score of 98.83%.

Decision Trees' Accuracy score to classify sugar levels in High, Low, and Medium viz. 100%

V. EXPERIMENTS AND RESULTS:

For Diabetes Prediction:

Logistic regression:

Fig 9.1. Accuracy Metrics for Logistic Regression:

```
Confusion Matrix
[[64  2]
 [15 14]]

Accuracy Score
```

	precision	recall	f1-score	support
0	0.81	0.97	0.88	66
1	0.88	0.48	0.62	29
accuracy			0.82	95
macro avg	0.84	0.73	0.75	95
weighted avg	0.83	0.82	0.80	95



vi. Random Forest:

Fig 9.2. Accuracy Metrics for Random Forest:

```
Confusion Metrics
[[85 15]
 [22 32]]

Accuracy Score
precision recall f1-score support
0 0.79 0.85 0.82 100
1 0.68 0.59 0.63 54

accuracy 0.76 154
macro avg 0.74 0.72 0.73 154
weighted avg 0.75 0.76 0.76 154
```

vii. Decision Tree:

Fig 9.3. Accuracy Metrics for Decision Tree Model:

```
Confusion Metrics
[[78 22]
 [25 29]]

Accuracy Score
precision recall f1-score support
0 0.76 0.78 0.77 100
1 0.57 0.54 0.55 54

accuracy 0.69 154
macro avg 0.66 0.66 0.66 154
weighted avg 0.69 0.69 0.69 154
```

viii. SVM:

Fig 9.4. Accuracy Metrics for SVM:

```
Confusion Metrics
[[91 9]
 [25 29]]

Accuracy Score
precision recall f1-score support
0 0.78 0.91 0.84 100
1 0.76 0.54 0.63 54

accuracy 0.78 154
macro avg 0.77 0.72 0.74 154
weighted avg 0.78 0.78 0.77 154
```

For Sugar Level Classification:

1) KNN:



Fig 10.1. Accuracy Metrics for KNN:

```
Confusion Matrix =  
[[115  0  0]  
 [  0 14  0]  
 [  0  0 43]]  
  
Accuracy =  
      precision  recall  f1-score  support  
High      1.00    1.00    1.00    115  
Low       1.00    1.00    1.00     14  
Normal    1.00    1.00    1.00     43  
  
accuracy          1.00    172  
macro avg        1.00    1.00    1.00    172  
weighted avg     1.00    1.00    1.00    172
```

ix. SVM:

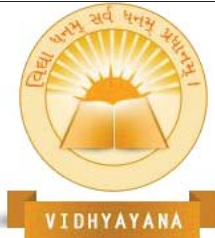
Fig 10.2. Accuracy Metrics for SVM:

```
Confusion Matrix =  
[[110  0  0]  
 [  0  8  0]  
 [  0  0 54]]  
  
Accuracy =  
      precision  recall  f1-score  support  
High      1.00    1.00    1.00    110  
Low       1.00    1.00    1.00     8  
Normal    1.00    1.00    1.00    54  
  
accuracy          1.00    172  
macro avg        1.00    1.00    1.00    172  
weighted avg     1.00    1.00    1.00    172
```

x. Decision Tree:

Fig 10.3. Accuracy Metrics for Decision Tree:

```
Confusion Matrix =  
[[110  0  0]  
 [  0  8  0]  
 [  0  0 54]]  
  
Accuracy =  
      precision  recall  f1-score  support  
High      1.00    1.00    1.00    110  
Low       1.00    1.00    1.00     8  
Normal    1.00    1.00    1.00    54  
  
accuracy          1.00    172  
macro avg        1.00    1.00    1.00    172  
weighted avg     1.00    1.00    1.00    172
```



Accuracy Table 2 for Diabetes Prediction:

ALGORITHM	ACCURACY
Logistic Regression	82.11 %
Random Forest	75.97 %
Decision Tree	69.48 %
SVM	77.92 %

Accuracy Table 3 for Sugar Level:

ALGORITHM	ACCURACY
KNN	100 %
SVM	100 %
Decision Tree	100 %

Table 4: Comparative Study Table:

Paper	Top performing Algorithm	Accuracy
[2]	KNN	100%
[4]	SVM	78.20%
[6]	Gaussian	98.25%
[7]	Decision Tree	99.00%
[10]	Multilayer Perceptron	86.08%
[12]	Logistic Regression	76.80%
[15]	Random Forest	90%
our Research	Logistic Regression	82.11%



4) Website:

Following the model-building process, we have developed a website aimed at assisting individuals with diabetes in determining the most appropriate diet to follow.

Initially, the user would input their blood glucose level in millimoles per liter (mmol/L) and then proceed to submit it.

Once the user inputs their blood sugar level, the system will categorize it as High, Medium, or Low, and provide corresponding diet plans for breakfast, lunch, dinner, and snacks, based on the user's selection.

Fig 11.1. index. Html Page

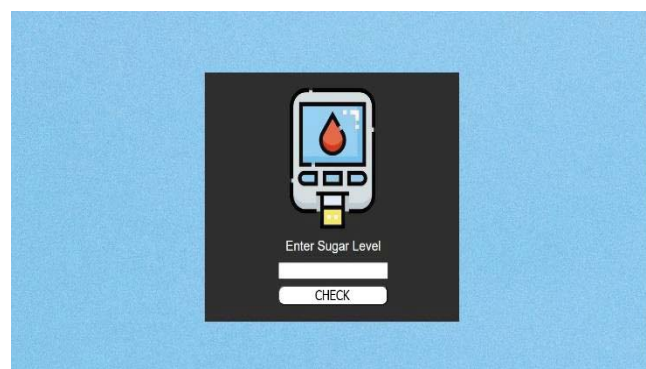


Fig 11.2. If Sugar is High

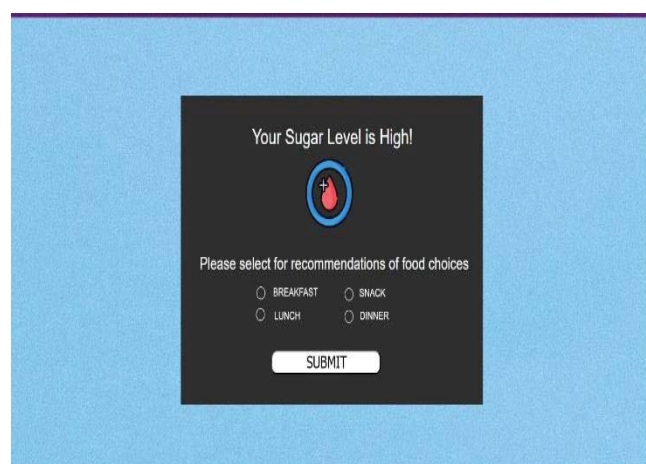


Fig 11.3. If Sugar is Low

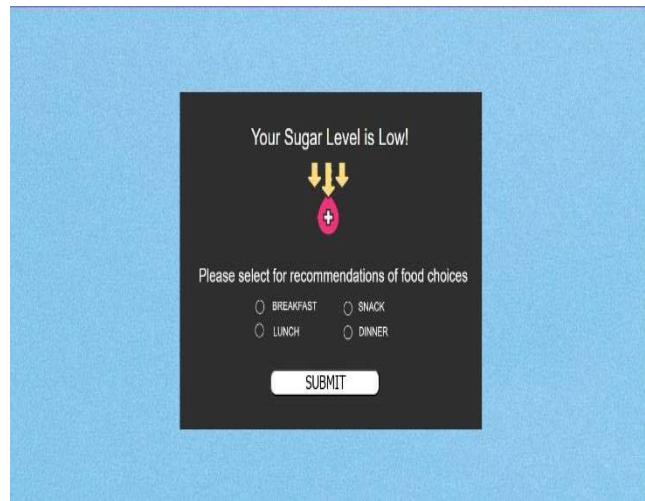
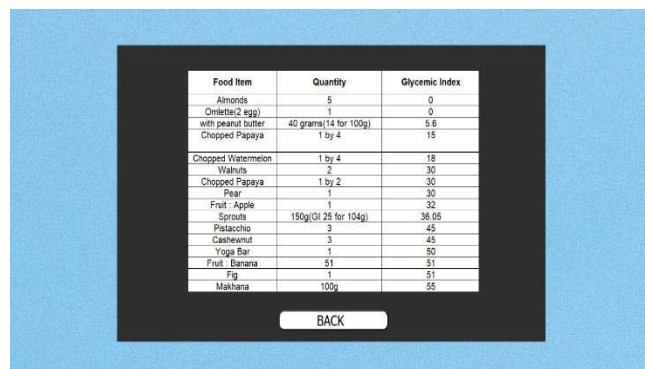


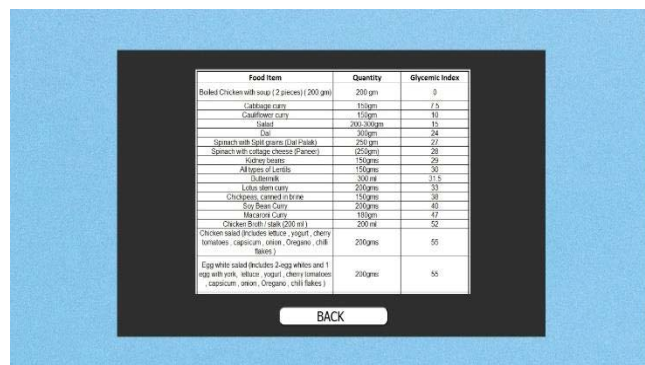
Fig 11.4. - 11.8.: Show the different food items recommended with the glycemic index of each food item.

Fig 11.4.



Food Item	Quantity	Glycemic Index
Almonds	5	0
Omelette(2 egg)	1	0
with peanut butter	40 grams(14 for 100g)	5.8
Chopped Papaya	1 by 4	15
Chopped Watermelon	1 by 4	18
Walnuts	2	30
Chopped Papaya	1 by 2	30
Pear	1	30
Fruit - Apple	1	32
Sprouts	150g(GI 25 for 104g)	38.05
Pistachio	3	45
Cashewnut	3	45
Yoga Bar	1	50
Fruit - Banana	51	51
Eg	1	51
Makhana	100g	55

Fig 11.5.



Food Item	Quantity	Glycemic Index
Baked Chicken with soup (2 pieces) (200 gm)	200 gm	0
Cabbage raita	100gm	7.5
Cauliflower curry	150gm	10
Salad	200, 300gm	15
Dal	300gm	24
Spinach with Sult cream (Use 1table)	200 gm	27
Spinach with cottage cheese (Paneer)	250gm	28
Kidney beans	150gm	29
Almonds of Lentils	150gm	30
Chickpeas	200 ml	31.5
Lentils, stem veg	200gm	33
Chickpeas, canned in brine	150gm	38
Soy Bean Curry	200gm	42
Masoori Curry	100gm	47
Chicken Breast steak (200 ml)	200 ml	52
Chicken salad (includes lettuce , yogurt , cherry tomatoes , capsicum , onion , Cresson , chilli , herbs)	200gm	55
Egg white salad (includes 2egg whites and 1 egg with onion , lettuce , yogurt , cherry tomatoes , capsicum , onion , Cresson , chilli flakes)	200gm	55

Fig 11.6.



Food Item	Quantity	Glycemic Index
Sprach	100gms	1
Cauliflower	100gms	7
Brinjal	70gms	10
Salad	250 gms	15
Porugcook	100gms	19
Okra	100gms	20
Idli	200 ml	21
Masala BundeMik	200ml	22
Kidney Beans	100gms	24
Dal	100 gms	29
Dal Fry	100gms	29
Wada	100ml	45
mixed veg	100gms	50
Jasira Rice	100gms	53
Hara Roti	2	54
Roti	2	54

BACK

Fig11.7.

Food Item	Quantity	Glycemic Index
Chickpeas curry	150 gm	56
Chicken Biryani with 2 pieces of chicken	250 gms.	56
Green peas with onion curry	150gm	61
Chh-Sonada curry	250gms	62.5
Chanaol	2,3	94
White Baked Rice	200gm	132

BACK

Fig 11.8.

Food Item	Quantity	Glycemic Index
Macha Roti	2	61
lanchor Roti	2	66
Rice Flour Roti	2	68
palak paneer	100gms	69
Lemon Rice	100gms	69
Jasira Rice	2	69
Wheat Roti	2	69
Pharosa	1	69
Phary	100gms	64
Rice	100 gms	69
masala rice	100gms	67
Chicken Biryani	100gms	68
Egg Curry	100gms	75
Chicken Curry	100gms	69

BACK

VI. CONCLUSION AND FUTURE WORK

People with low sugar levels are recommended to consume food items with a high glycemic index, while those with high sugar levels are advised to consume food items with a low glycemic index. The K-Nearest Neighbors (KNN) and Decision Tree algorithms are suitable for classifying sugar levels into high, low, and normal categories. Based on the result above, it can be inferred that the Logistic Regression algorithm is a suitable choice for predicting sugar levels, with an accuracy of 0.8211%.



Vidhyayana - ISSN 2454-8596

An International Multidisciplinary Peer-Reviewed E-Journal

www.vidhyayanaejournal.org

Indexed in: Crossref, ROAD & Google Scholar

Future Work:

- 1 We plan to enhance our website by introducing a backend and improving its user-friendliness.
- 2 We intend to incorporate fresh, modified recipes from popular cuisines such as Chinese, Mexican, and Italian. These recipes will be specifically designed to cater to diabetic patients and promote good health.
- 3 We are considering adding a new feature that predicts the amount of increase in sugar levels after consuming a particular food item. This tool will be helpful for diabetic patients, but it's important to note that individual differences in factors such as metabolism, activity levels, and stress may affect the accuracy of the predictions. Hence, we will communicate the limitations of the model and provide a disclaimer to users.

ACKNOWLEDGMENT

The success of this research would not have been possible without the excellent supervision of my distinguished professor, Prof. Dr. Sumegh Tharewal, from the School of Computer Science & Engineering at Dr. Vishwanath Karad MIT International Peace University. I would like to offer my sincere appreciation to him. His wise suggestions and meticulous instructions have been the cornerstone of this achievement.

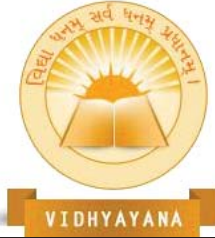
As this research was a collaborative endeavor, I also want to express my profound gratitude to the other members of my group for their continuous support and contributions throughout the project. Despite the challenges we faced during these trying times, we remained in constant communication, sharing valuable insights and perspectives that greatly enhanced the research outcome.

Since this research was conducted remotely, I would like to express our gratitude to our families for their patience and unwavering support, which enabled us to focus on our work and achieve the desired results. Throughout this period, we have gained valuable knowledge, and this has been an incredibly enriching and rewarding experience.



REFERENCES

- B. Suresh Lal, "Diabetes: Causes, Symptoms, and Treatments," Public Health Environment and Social Issues in India Edition: Chapter: 5, January 2016.
- J. Pradeep Kandhasamy, S. Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus", Procedia Computer Science 47, 2015.
- Muhamad Soleh, Naufal Ammar, and Indrati Sukmadi, "Website-Based Application for Classification of Diabetes Using Logistic Regression Method," Jurnal Ilmiah Merpati, Vol. ;, No. 1, April 2021.
- Michael Onyema Edeh, Osamah Ibrahim Khalaf, Carlos Andrés Tavera, Sofiane Tayeb, Samir Ghouali, Ghaida Muttashar Abdulsahib, Nneka Ernestina Richard-Nnabu, and AbdRahmane Louni, "A Classification Algorithm-Based Hybrid Diabetes Prediction Model", Front Public Health, 31 Mar-2022, doi: 10.3389/fpubh.2022.829519
- Deepti Sisodia a, Dilip Singh Sisodia b, "Prediction of Diabetes using Classification Algorithms", International Conference on Computational Intelligence and Data Science, 2018, <https://doi.org/10.1016/j.procs.2018.05.122>.
- Nishat MM, Faisal F, Mahbub MA, Mahbub MH, Islam S, Hoque MA "Performance Assessment of Different Machine Learning Algorithms in Predicting Diabetes Mellitus", Department of Electrical and Electronic Engineering Islamic University of Technology (IUT), Dhaka, Bangladesh, 21 Mar-2021, <http://dx.doi.org/10.21786/bbrc/14.1/10>
- KM Jyoti Rani, "Diabetes Prediction Using Machine Learning" International Journal of Scientific Research in Computer Science Engineering and Information Technology, July-2020, DOI: 10.32628/CSEIT206463
- Quan Zou,^{1,2,*} Kaiyang Qu,¹ Yamei Luo,³ Dehui Yin,³ Ying Ju,⁴ and Hua Tang⁵ "Predicting Diabetes Mellitus With Machine Learning Techniques" Pubmed Central, 6 Nov-2018, <https://doi.org/10.3389%2Ffgene.2018.00515>
- Ray Max "DIABETES -TYPE 2", divine word university faculty of Medicine and health sciences department of environmental health eh320-diseades control and Epidemiology, 17 April-2019.



Vidhyayana - ISSN 2454-8596

An International Multidisciplinary Peer-Reviewed E-Journal

www.vidhyayanaejournal.org

Indexed in: Crossref, ROAD & Google Scholar

Umair Muneer Butt, Sukumar Letchmunan, Mubashir Ali, Fadratul Hafinaz Hassan, Anees Baqir, Hafiz Husnain Raza Sherazi “Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications”, AI-Enabled Internet of Things in Sport and Public Health,01 Oct-2021, <https://doi.org/10.1155/2021/9930985>

Mitushi Soni, Dr. Sunita Varma “Diabetes Prediction using Machine Learning Techniques”, international journal of engineering research & Technology (inert),04 May-2020, doi: 10.17577/ijertv9is090496.

Tarig Mohamed Ahmed “Using data mining to develop a model for classifying diabetic patient control level based on historical medical records”, Journal of Theoretical and Applied Information Technology, 20th May-2016

Bhoia SK, Pandab SK, Jenea KK, Abhisekha PA, Sahood KS, Samae NU, etc ”Prediction of Diabetes in Females of PimaIndian Heritage: A Complete Supervised Learning Approach”, Turkish Journal of Computer and Mathematics Education,28 April 2021.

Veena Vijayan V, Aswathy Ravikumar “Prediction of Diabetes Using Data Mining Techniques”, May 2018, <https://doi.org/10.1109/ICOEI.2018.8553959>.