Sigma J Eng Nat Sci, Vol. 44, No. 5, pp. 3209–3222, October, 2026

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.2154

Research Article

Explainable artificial intelligence for proactive management of diabetes

Kavita PABREJA[1], Nishant SINGH[2], Rohan BANDUNI[2]

[1]Bennett University, Greater Noida, Uttar Pradesh, 201310, India
[2]Maharaja Surajmal Institute, New Delhi, 110058, India

ARTICLE INFO

Article history
Received: 28 January 2025
Revised: 31 May 2025
Accepted: 10 July 2025

Keywords:
Diabetes Mellitus; Feature Engineering; Logistic Regression; Machine Learning; Random Undersampling

ABSTRACT

Diabetes is becoming an increasingly serious global health and economic problem affecting 589 million people as of 2024 and 643 million people by 2030. Early and proactive measures are required to prevent the spread of this illness. The current study attempts to solve this issue by implementing an explainable, machine learning-based model for predicting diabetes, which used the Behavioral Risk Factor Surveillance System dataset, with more than 250,000 entries, supplied by Center for Disease Control and Prevention. In order to train the model, the dataset was prepared by applying feature engineering, normalization, and class imbalance issues have been addressed by means of applying random undersampling. A logistic regression model was built using the seven features that were chosen according to their mutual information values. These features were General Health, Blood Pressure, Age, Body Mass Index, Cholesterol, Difficulty in walking, and Income. As a result, the proposed model obtained 74.72% accuracy, 76.63% recall and ROC-AUC score of 0.75. It has demonstrated satisfactory performance and strong clinical applicability. Additionally, a user-oriented web interface was created for applying the aforementioned model to perform a fast and interpretable diabetes risk assessment with minimal input data. The current study improved upon previous works by including model explainability, high scalability and deployment due to creating the user-oriented web interface in the process.

Cite this article as: Pabreja K, Singh N, Banduni R. Explainable artificial intelligence for proactive management of diabetes. Sigma J Eng Nat Sci 2026;44(5):3209–3222.

INTRODUCTION

Diabetes is one of the main concerns in public health all around the globe. More than 537 million adults suffered from this disease in 2021, and it is projected to be up to 643 million in 2030 [1]. Complications of diabetes include cardiovascular disease, kidney failure and neuropathy. The financial toll of the disease is equally high: the amount of global expenditures on health care connected to diabetes exceeded an estimate of USD 1 trillion in 2024 [2, 3].

There are many diagnostic tests like glycated haemoglobin (HbA1c) and fasting blood glucose. However, there is still a significant proportion of patients who remain undiagnosed or diagnosed late when intervention becomes ineffective [4]. One of the tools that may help to change this situation is machine learning (ML) which can predict the

***Corresponding author.**
*E-mail address: kavitapabreja1@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

Published by Yıldız Technical University Press, İstanbul, Turkey
© Author. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

risk of disease based on non-invasive data [5]. The problem, though, is that most of the research either relies on a complicated algorithm with hardly explainable predictions or uses small and unrepresentative datasets to train their models [6, 7]. In both cases, the model is not easy to implement in the real-world setting.

This research aims to fill this gap by developing an explainable ML model based on logistic regression and using a large real-world dataset from the CDC's "Behavioral Risk Factor Surveillance System" (BRFSS) including over 250,000 samples. Class imbalance, feature selection, and model interpretability are addressed as problems common in previous researches [8, 9]. Using feature engineering, mutual information-based feature selection, and random undersampling, the model reaches high accuracy (accuracy: 74.72%, recall: 76.63%, Receiver Operating Characteristic - Area Under the Curve (ROC-AUC): 0.75) using only seven easily accessible non-clinical features.

Most importantly, the model is implemented in the user-friendly web application allowing to assess the diabetes risk fast without laboratory tests or any medical equipment needed. The application enables early intervention and promotes personalized preventive healthcare according to the global goals of precision medicine [10, 11].

The novelty of our research is in the holistic approach to the problem. The model was created as a result of step-by-step process and then converted into an easy to use decision support tool. In contrast to most previous studies, which either remain on the theory level or focus only on improving the algorithm, we aimed to develop something practical and helpful.

The structure of the rest of the paper is as follows. Literature review, methodology (consists of data collection, exploratory data analysis and its results, feature engineering, addressing class imbalance, feature selection and predictive model development), results, and conclusion.

Literature Review

Diabetes mellitus continues to be a serious health problem all over the world. Its burden falls on patients as well as on healthcare systems, in high-income and developing countries alike. The steady rise in cases is linked mainly to sedentary lifestyles, changing food habits and an ageing population [12]. The economic burden is equally profound; the "American Diabetes Association" (ADA) estimated the total cost of diagnosed diabetes in the U.S. to be $327 billion in 2017, comprising $237 billion in direct medical expenses and $90 billion due to reduced productivity [3]. International Diabetes Federation [1] has predicted that the number of adults with diabetes will increase from 425 million in 2017 to 629 million by 2045 worldwide.

Advancements in Machine Learning For Diabetes Prediction

In the era of Artificial intelligence, Machine learning (ML) is revolutionizing as a predictive tool for diseases because of its high precision and ease of application in the risk prediction models of the diseases.[5] have highlighted ML's role in the diagnosis of diabetes by using various ML models such as "logistic regression", "decision tree", "random forest", "support vector machines" (SVMs) etc. [13] highlighted the utility of logistic regression models in identifying high-risk individuals by detecting strong predictors of diabetes. In [14], the superiority of ML algorithms: SVMs, decision trees and logistic regression have been established for diabetes prediction.

Ensemble learning algorithms have recently gained popularity in medical field. For instance, [15] employed ensemble methods including, XGBoost, LightGBM, CatBoost and achieved the best performance from CatBoost (accuracy of 95.4%, AUC-ROC of 0.99). Contrastingly, [16] applied ensemble-based methods with SHAP (SHapley Additive exPlanations) interpretability to predict fasting glucose levels and highlighted the role of features such as age and waist-to-hip ratio in the early screening of diabetes.

A limited number of authors have proposed the comparisons of one ML model against a combination of weak models. [17] have used a variety of ML techniques for numerous diseases, including diabetes, and concluded that the ensemble models are more powerful than using a single classifier. Similarly, [18] demonstrated that high performance was achieved (89.7% accuracy) by gradient boosting in comparison to single classifiers viz. Naïve Bayes and k-NN. According to a study by [19] random forest (an ensemble method in terms of bagging) was the best ensemble type of model with the highest accuracy in comparison to single ML model types.

Feature Engineering and Selection Techniques

Feature selection is an important factor that helps to improve the model's performance, saves computational cost and time, increases the model's representability, and solves multicollinearity problem. Feature selection methods can be utilised to identify important predictors and to enhance model interpretability. [9] emphasized the influence of these techniques on redundancy reduction and enhancement of the classifier. In the same vein, [20] utilized particle swarm optimization (PSO) in feature selection on three medical datasets and found that ML models (decision trees, random forests, and Naïve Bayes) were consistently accurate and with low error rates when the features were reduced.

Further, [21] integrated data pre-processing techniques with dimensionality reduction techniques to develop more robust models. The importance of feature engineering for improving the performance of Machine Learning models has been emphasized upon in this study.

Deep Learning and Hybrid Approaches

Traditional ML models are interpretable and understandable, but they struggle when faced with complex and non-linear relations among data. This is where deep learning offers better predictive performance. [7] showed

it by utilizing deep learning models that outperformed traditional ML models. In another study, the authors in [22] applied a hybrid long short-term memory (LSTM) model to the Pima Indians Diabetes dataset and obtained higher accuracy. Their results suggest that recurrent neural networks, a class of deep learning models, can also work well on clinical data.

Retinal imaging is another area where deep learning has been tried. The study in [23] combined three convolutional neural networks, ResNet50, DenseNet201 and InceptionV3, through soft voting, and used data balancing to grade the severity of diabetic retinopathy from fundus images. The ensemble raised classification accuracy to 90%, which shows the value of ensemble learning in diagnosing retinal disease.

Ethical, Practical, and Interpretability Considerations

The potential benefits of machine learning in the health sector are many. However, there are several issues that need to be taken into account. These include data protection, informed consent, and algorithmic bias. According to the findings of [11], it is necessary for a model to be interpretable to gain acceptance in clinical practice since it will be used by the doctor as well as the patient.

Similar results are provided by other works as well. For instance, the study provided in [10] demonstrates how ML models can be used for individualised treatment of diabetics. The article published in [24] describes predictive analytics as a disrupting factor in the field of medicine. Thus, all these studies have one conclusion in common - besides accuracy, the model must be ethical and accessible [16].

Practical Implementations and Interface Integration

With the recent advancements in machine learning, there has been simultaneous growth in terms of easy to use interfaces, calling on machine learning algorithms at the back end. As stated by the authors [25], the proposed work includes an Intelligent Diabetes Mellitus Prediction Framework (IDMPF). The framework uses Support Vector Machines (SVM) and Random Forest techniques to deal with imbalanced classes.

Other studies undertaken by different researchers have involved the use of machine learning algorithms in Graphical User Interfaces (GUI). This has proven very helpful especially when dealing with non-experts. As described by the researchers in [26], a GUI decision support system was designed and developed, while [27] proposed a GUI ensemble interface for screening.

Broadening the Scope of ML in Healthcare

The application of machine learning extends beyond diabetes to include broader healthcare domains. [24] reviewed ML applications in chronic disease monitoring, personalized treatment optimization, and early risk detection.

Such discoveries reflect the ability of machine learning algorithms to be flexible and develop in order to help prevent diseases.

The paper [28] provides an interesting analysis of different machine learning algorithms like CatBoost, XGBoost, and ANN to predict mortality among patients who suffer from heart failure based on their clinical data. There is a PCA and non-PCA analysis presented, and their findings reveal the superiority of CatBoost in predictive modeling, particularly in terms of accuracy and MCC (Matthews Correlation Coefficient). It is quite useful in the case when datasets are unbalanced. This offers insight into ML's potential utility for risk assessment frameworks. These findings reinforce the necessity of continually refining algorithmic approaches to match the ever-evolving landscape of digital health.

Summary and Research Direction

It becomes evident from the review of existing literature that machine learning algorithms, when properly utilized, can be highly useful in diabetes detection and control. Nevertheless, there exist several critical issues associated with the topic, including the issue of balancing algorithm accuracy and interpretability, particularly the issue of creating an interface, as well as the question of ethics and fairness related to data privacy and use.

There is definitely room for development of such models which would be light enough, easy to understand, and accessible to use in clinical settings, specifically, in the absence of invasive diagnosis procedures. This study aims to address this gap by developing a predictive logistic regression model using the technique of mutual information, and implementing it in a convenient web interface.

Research Objectives

Based on the identified gaps, this study is directed by the following objectives:

1. To investigate both linear and non-linear machine learning techniques, such as logistic regression, decision trees, and random forests, with the aim of building reliable predictive models for the early detection of diabetes.
2. To enhance model performance through feature engineering and selection and also addressing class imbalance.
3. To assess how effectively the models perform by examining key evaluation metrics, including accuracy, recall, and the area under the receiver operating characteristic curve (ROC-AUC).
4. To develop an accessible web-based interface allowing healthcare professionals and individuals to input basic health indicators and receive real-time diabetes risk assessments, promoting early intervention and personalized care.

These objectives aim to bridge the gap between technical advancement and real-world implementation, offering

a scalable, and explainable solution to the persistent challenge of diabetes prediction.

MATERIALS AND METHODS

The research framework comprises data collection, exploratory data analysis, feature engineering, exploring various machine learning models and finally deploying the model with an interactive web site. This framework has been visualized step by step in Figure 1. All stages of the framework have been explained in the following sections.

Data Collection

The dataset used for this study comprises 2,53,680 survey responses from the "Center for Disease Control and Prevention", CDC's BRFSS2015. The dataset includes 21 feature variables and the target variable, Diabetes_binary, which has two classes: 0 for no diabetes and 1 for prediabetes or diabetes.

Exploratory Data Analysis

Before implementing any machine learning or statistical modeling, exploratory data analysis, is essential for comprehending the structure and trends inside a dataset. It was found that there were no missing values. Various insights have been drawn based on univariate and bivariate analysis. These are described in the preceding sections. To facilitate feature selection, Correlation analysis has been performed.

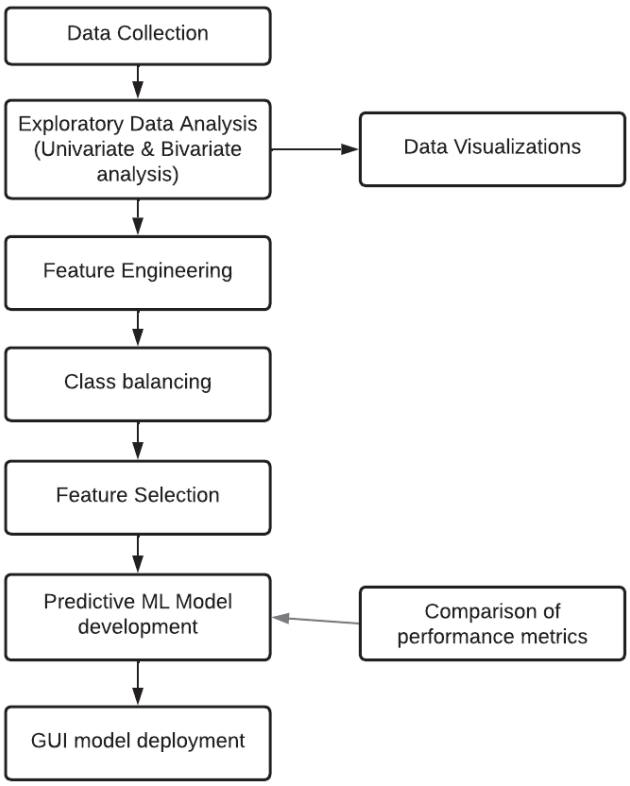


Figure 1. Research framework deployed for the study.

Correlation Heatmap provides a complete representation of the linear relationships among the features in the dataset, and the values range between -1 and +1. The positive values show direct relationships, and the negative values show inverse relationships.

As depicted in the heatmap in Figure 2, majority of the pairs of features are relatively uncorrelated, which is an ideal case for building models, as multicollinearity will be minimal. It is worth noting that there is no correlation greater than 0.8 between any pairs of predictors.

Several noteworthy insights emerge from the plot:

1. Diabetes_binary, the target variable, shows moderate positive correlation with BMI (0.22), HighBP (0.26) and GenHlth (0.29), and a modest negative correlation with Income (-0.16). These associations are both statistically and clinically plausible, reinforcing known risk factors—such as hypertension, poor self-reported health, and lower socioeconomic status—as being indicative of diabetes risk.
2. General Health (GenHlth) is strongly correlated with Physical Health (PhysHlth, 0.52) and Mental Health (MentHlth, 0.30), which suggests some overlapping information among these subjective health measures. However, the correlations remain moderate, not high enough to necessitate elimination, especially if they contribute uniquely in non-linear relationships.
3. A few additional moderate correlations that should be mentioned include DiffWalk with PhysHlth (0.48), and Age with HighBP (0.34). Both correlations seem reasonable as, for instance, older people can report difficulties in moving and problems with walking are obviously connected with poor physical health condition. At the same time, correlation of Age and High Blood Pressure (HighBP) equal to 0.34 fits well into the existing medical knowledge. With age increase, the risk of developing high blood pressure is also increasing due to changes in blood vessels and various factors of life style.

In conclusion, the heatmap has demonstrated the validity of the selected features for use in the creation of the model. With such a low level of inter-correlation between features, there is a minimal likelihood of overfitting and an independent contribution of each feature in the formation of the predictive ability of the model. Together with the feature selection using mutual information, this provides a solid foundation for the selected features.

The bar chart in Figure 3 visualizes the correlations between each feature and the binary diabetes outcome. While the heatmap shows the overall relationships among all the variables, this chart helps to show the linear influence of each variable separately on the diabetes status.

1. General Health is correlated most positively with diabetes (~0.29). That is why the self-reported health status is an essential predictor. Thus, this variable was chosen as one of the most relevant features at the stage of mutual information selection and practical use in non-invasive models.

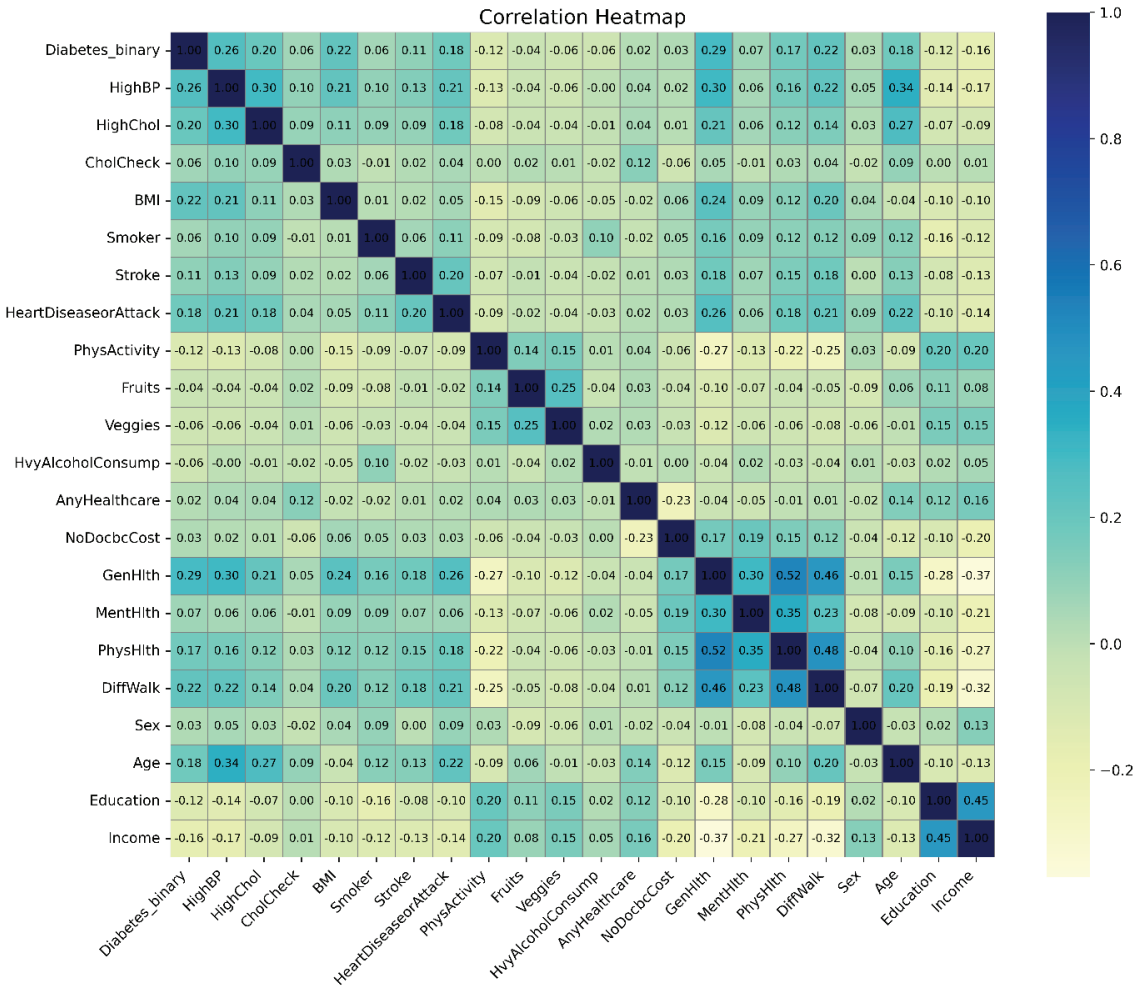


Figure 2. Heatmap depicting correlation among all features including target.

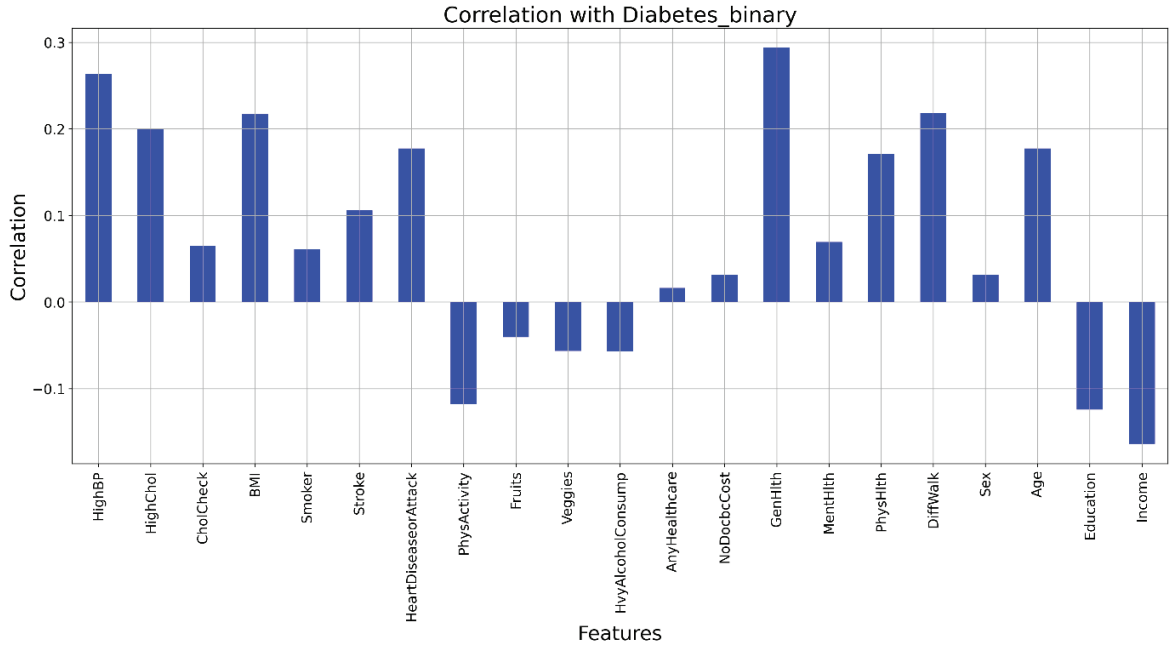


Figure 3. Correlation between independent (features) and dependent (target) variable.

2. High Blood Pressure, BMI, and High Cholesterol are associated positively with diabetes too. These are traditional risk factors for diabetes mentioned in clinical literature, and they show that the comorbidities often appear before or alongside with the diabetes
3. Some socioeconomic and behavioral features like Income and Education are negatively correlated with the diabetes status. It means that the patients of lower socioeconomic status have higher risks of developing diabetes because of access problems, health literacy, etc.
4. The weak linear relationships between features like Sex, AnyHealthcare, or HvyAlcoholConsump confirm the efficiency of your feature selection algorithm. You decided to ignore them correctly.

Overall, the barchart helps to justify the inclusion of each variable in terms of its linear contribution to the prediction of the diabetes status.

Univariate Analysis -Summarization of Dichotomous Features

To gain insights into characteristics of the data, the summarization of 14 columns that have dichotomous values, has been done as shown in in Table 1.

The descriptive statistics of the dataset as shown in Table 1 reveal important patterns relevant to diabetes prediction. However, 13.93% of the respondents have diabetes, which shows a class imbalance that needs to be considered during the development of the predictive model so that bias can be avoided. In addition, 42.9% of the people have high blood pressure and 42.41% have high levels of cholesterol, which are known risk factors for diabetes. Importantly, 96.27% have been screened for cholesterol in the last five years, implying good preventive care coverage. Also, 95.11% have healthcare coverage and 16.82% have trouble walking, which could mean presence of chronic illness such as diabetes. Socioeconomic factors may play a role in diabetes development as 8.42% of people cannot see a doctor because of costs.

Not only do the above results help understand the relevance of the features observed in correlation analysis and feature importance analysis, but also show that there are many different reasons why people have diabetes.

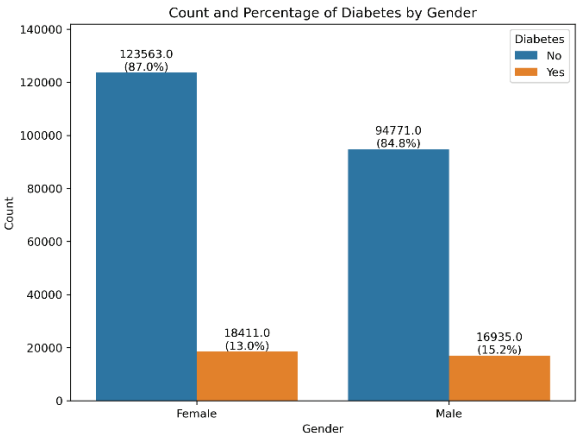
Bivariate Analysis

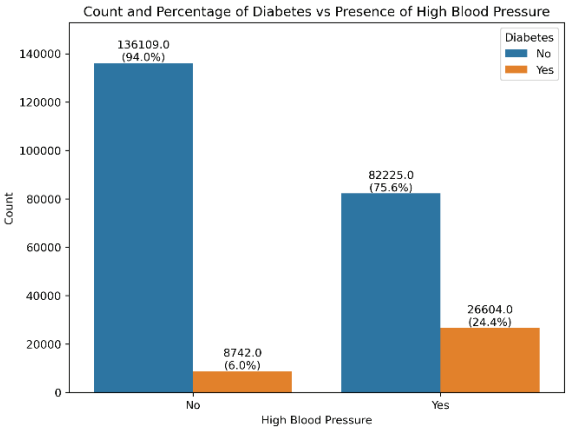
The distribution of significant features impacting the presence or absence of diabetes has been visualized in Figure 4. The inference drawn are as follows:

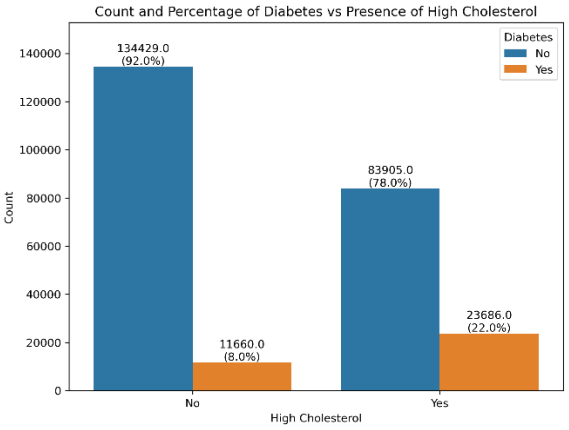
- Presence or absence of diabetes is not dependent on gender
- People with high blood pressure are more likely to be suffering from diabetes. 24.4% of those having high BP are affected by diabetes whereas only 6% of those with normal BP are having diabetes disease.
- People with high cholesterol are more likely to be suffering from diabetes. 22% of those having high cholesterol are affected by diabetes whereas only 8% of those with normal cholesterol are having diabetes disease.
- Smoking does not seem to be causing any impact on diabetes. Those who have smoked at least 100 cigarettes in their lives or less or none, seem to be having no relation with their diabetes disease.
- Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) are less affected with diabetes. As per the data,

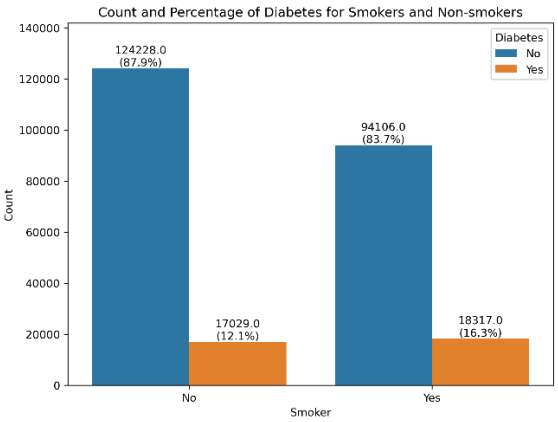
Table 1. Distribution of "Yes" and "No" responses for health and lifestyle attributes

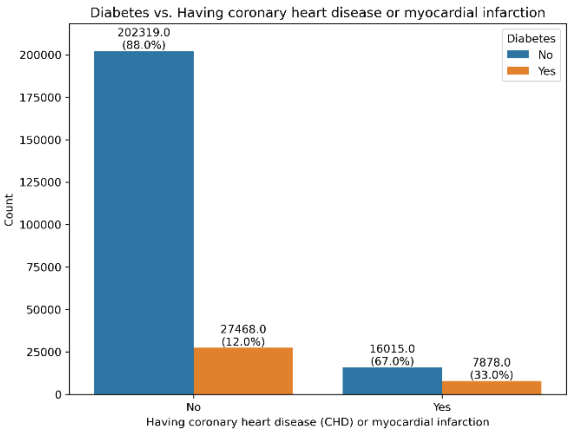
S.no.	Attribute Name	Description	No (%)	Yes (%)
1	Diabetes_binary	Suffering from diabetes	86.07	13.93
2	HighBP	Having high blood pressure	57.10	42.90
3	HighChol	Having high cholesterol	57.59	42.41
4	CholCheck	Have checked their cholesterol in 5 years	3.73	96.27
5	Smoker	Smoked at least 100 cigarettes in their lives	55.68	44.32
6	Stroke	Had a stroke in the past	95.94	4.06
7	HeartDiseaseorAttack	Having coronary heart disease (CHD) or myocardial infarction	90.58	9.42
8	PhysActivity	Having done physical activities in past 30 days - not including job	24.35	75.65
9	Fruits	Consume fruits daily	36.57	63.43
10	Veggies	Consume veggies daily	18.86	81.14
11	HvyAlcoholConsump	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)	94.38	5.62
12	AnyHealthcare	Having any healthcare coverage	4.89	95.11
13	NoDocbcCost	Couldn't see a doctor because of cost in past 12 months	91.58	8.42
14	DiffWalk	Having difficulty in walking or climbing stairs	83.18	16.82
15	Sex	Gender	55.97 (Females)	44.03 (Males)

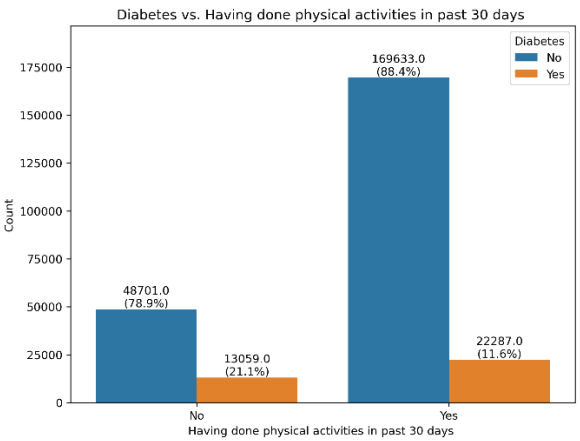

Count and Percentage of Diabetes by Gender
Diabetes
No
Yes
123563.0
(87.0%)
18411.0
(13.0%)
94771.0
(84.8%)
16935.0
(15.2%)
Count
Female
Male
Gender

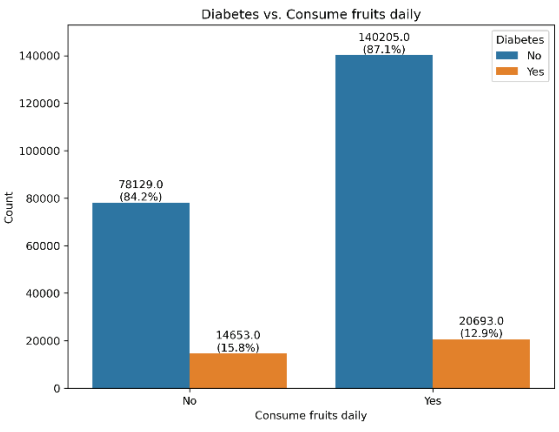

Count and Percentage of Diabetes vs Presence of High Blood Pressure
Diabetes
No
Yes
136109.0
(94.0%)
8742.0
(6.0%)
82225.0
(75.6%)
26604.0
(24.4%)
Count
No
Yes
High Blood Pressure


Count and Percentage of Diabetes vs Presence of High Cholesterol
Diabetes
No
Yes
134429.0
(92.0%)
11660.0
(8.0%)
83905.0
(78.0%)
23686.0
(22.0%)
Count
No
Yes
High Cholesterol


Count and Percentage of Diabetes for Smokers and Non-smokers
Diabetes
No
Yes
124228.0
(87.9%)
17029.0
(12.1%)
94106.0
(83.7%)
18317.0
(16.3%)
Count
No
Yes
Smoker


Diabetes vs. Having coronary heart disease or myocardial infarction
Diabetes
No
Yes
202319.0
(88.0%)
27468.0
(12.0%)
16015.0
(67.0%)
7878.0
(33.0%)
Count
No
Yes
Having coronary heart disease (CHD) or myocardial infarction


Diabetes vs. Having done physical activities in past 30 days
Diabetes
No
Yes
169633.0
(88.4%)
48701.0
(78.9%)
13059.0
(21.1%)
22287.0
(11.6%)
Count
No
Yes
Having done physical activities in past 30 days


Diabetes vs. Consume fruits daily
Diabetes
No
Yes
140205.0
(87.1%)
78129.0
(84.2%)
14653.0
(15.8%)
20693.0
(12.9%)
Count
No
Yes
Consume fruits daily

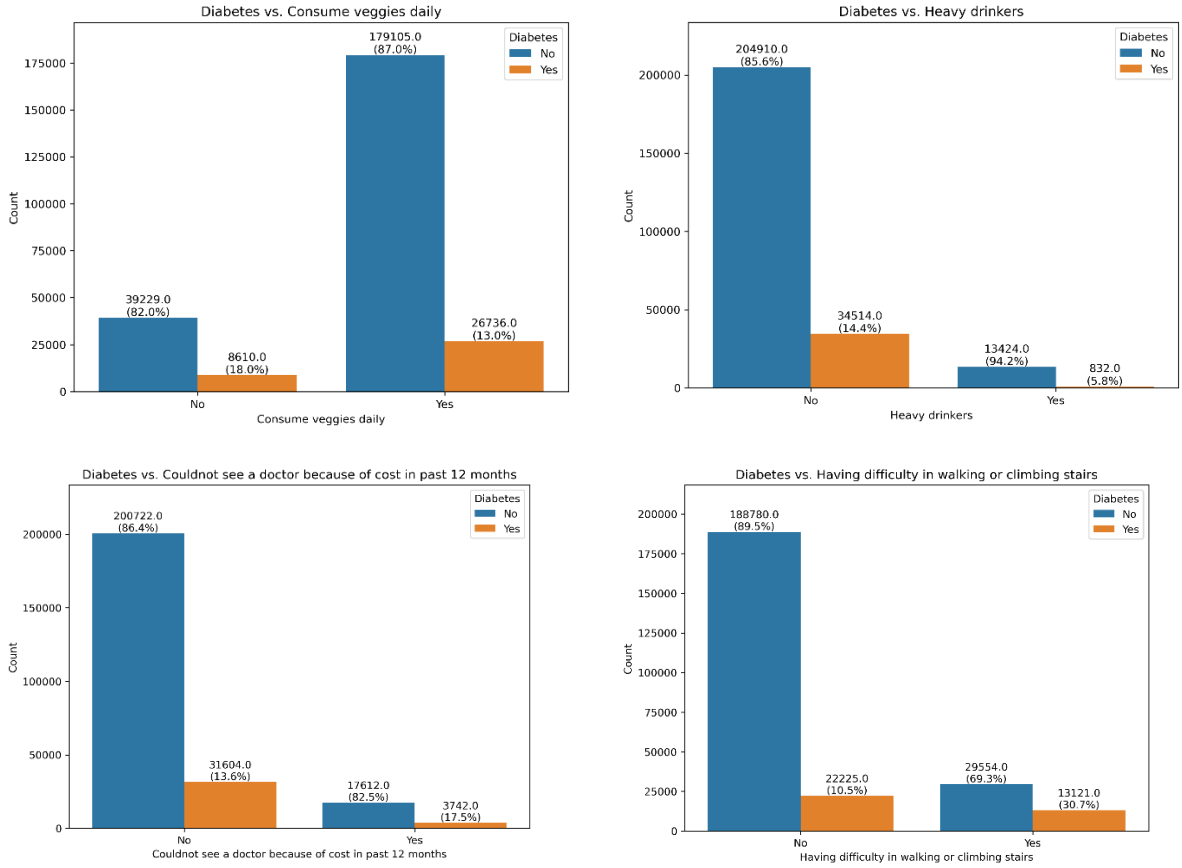


Figure 4. Bivariate analysis of features and target for diabetes dataset.

only 5.8% of those who are heavy drinkers are affected by diabetes whereas 14.4% of the non-heavy drinkers are affected.

- Past 12months of inability to see a doctor because of cost is not causing much difference in having diabetes or not.
- Having difficulty in walking or climbing stairs is a symptom depicting presence of diabetes as 30.7% of those who fall in this category are having this disease as opposed to just 10.5% of those who don't find any difficulty in walking or climbing stairs.

Life style: Physactivity, fruits and veggies effect on diabetes

- Physically active people have less chance of having diabetes. The percentage of those who have not done any physical activity in past 30 days stand 10% more than those who have done physical activity. This is apart from their job.
- Daily consumption of Fruits and vegetables is not a differentiating factor when it comes to someone affected with diabetes.

Feature Engineering

For rest of the attributes, that are multivalued, data preprocessing by applying binning has been done and new features have been constructed. These are tabulated in Table 2. A brief summary of these attributes highlights that more than 70% respondents are college graduates atleast, more than 60% never had any mental or physical health problems, more than 70% are overweight/ obese, more than 60% have annual income of less than $75,000. More than 70% are in age bracket of more than 50years.

Apart from this, categorical variables were converted to numerical format by using one-hot encoding, as the machine learning algorithm selected for the study works only on numbers. Finally, the dataset was normalized to ensure that features with different scales did not disproportionately influence the model. All development has been done using libraries of Python. A brief about their purpose for this study is mentioned here. NumPy and Pandas are basic tools used in computing numbers and managing data in a structured format, respectively. Scikit-learn enables developers to construct and evaluate traditional machine learning algorithms. Visualization tools include Matplotlib and Seaborn, which help in the visualization of data and the outcome of a model. Finally, Flask is an elegant web framework that developers can use to create web applications

Table 2. Results of feature engineering as applied on diabetes datasets

Education Level	Percentage of respondents
Never attended school or only kindergarten	0.07%
Grades 1 through 8 (Elementary)	1.59%
Grades 9 through 11 (Some high school)	3.74%
Grade 12 or GED (High school graduate)	24.74%
College 1 year to 3 years (Some college or technical school)	27.56%
College 4 years or more (College graduate)	42.31%

Number/Range of days	Percentage of respondents having suffered from mental health problems	Percentage of respondents having suffered from physical health problems
0 days	69.25%	63.09%
1-5 days	16.47%	18.45%
6-10 days	4.41%	4.91%
11-15 days	2.82%	3.24%
16-20 days	1.43%	1.44%
21-25 days	0.61%	0.87%
26-30 days	5.01%	8.00%

Age Range (in years)	Percentage of respondents
18-29	5.25%
30-49	24.01%
50-69	48.31%
70-99	22.42%

BMI Category	BMI Range	Percentage of Respondents
Underweight	less than 18.5	1.23%
Normal weight	18.5 to 24.9	27.18%
Overweight	25 to 29.9	36.96%
Obesity I	30 to 34.9	21.07%
Obesity II	35 to 39.9	8.15%
Obesity III	40 or greater	5.42%

Income range	Percentage of respondents
Less than $10,000	3.87%
$10,000 to less than $15,000	4.64%
$15,000 to less than $20,000	6.30%
$20,000 to less than $25,000	7.94%
$25,000 to less than $35,000	10.20%
$35,000 to less than $50,000	14.38%
$50,000 to less than $75,000	17.04%
$75,000 or more	35.63%

General Health	Percentage of respondents
Excellent	17.86%
Very good	35.12%
Good	29.82%
Fair	12.44%
Poor	4.76%

easily. It is especially popular for turning machine learning models into user-friendly interfaces with minimal overhead

Addressing Imbalance Nature of Data

The insights into the dataset revealed that it is highly imbalanced as 86.07% of respondents do not suffer from diabetes, and rest are either prediabetics or diabetics. Without resolving the imbalance, applying machine learning to an unbalanced dataset can result in a number of hazards and problems, including as inaccurate performance metrics (accuracy), biased predictions, and poor performance on minority classes (precision, recall, and F1-score) [29]. There is a chance that the model has suboptimal decision thresholds. The model might not adapt effectively to data from the actual world, where the minority class might be more common than in the training set. Due to inadequate representation, this would lead to underfitting the Minority Class, which would miss significant patterns and subtleties.

Many strategies, including resampling [8, 30] which involves either oversampling the minority class (like SMOTE) or undersampling the majority class; ensemble methods which are better able to handle class imbalance; and class weighting, which involves assigning higher weights to the minority class in the loss function in order to penalize misclassification of minority class instances more severely—may be taken into consideration in order to reduce these risks [31, 32]. To balance the dataset, synthetic data may also be created for the minority class.

In our research we have opted for Random Undersampling technique that randomly removes samples from the majority class to balance the dataset.The dataset is made perfectly balanced with an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes.

Feature Selection

This step aims to identify and select key features that contribute significantly to diabetes prediction and also with reduced set of features, interpretability of ML algorithm would improve. It was decided to use Mutual Information (mutual_info_classif) technique of Feature selection as the dataset comprises both categorical and continuous features and this technique finds convincing applicability under this situation [33]. Another advantage of Mutual Information Feature selection technique is that it does not assume any specific distribution of dataset. It measures not just the linear mutual dependence between the features and the target variable, but any kind of dependence and also quantifies how much information about the target variable can be obtained from each feature. This step aims to identify and select key features that contribute significantly to diabetes prediction. The accuracy was calculated by deploying Logistic Regression as a base model and iteratively selecting best features using Mutual Information technique. The plot showing variation in accuracy by adding features one at a time is visualized in Figure 5. This figure illustrates the relationship between the number of selected features and the model's prediction accuracy, with logistic regression used as the baseline classifier. The x-axis represents the number of top-ranked features selected based on mutual information scores, while the y-axis reflects the corresponding model accuracy. The plot clearly demonstrates a steep improvement in accuracy as the first few top features are added especially between 1 and 7 features. After the inclusion of the 7th feature, however, the curve plateaus, indicating diminishing returns in performance beyond this point.

This behavior provides an important insight: adding more features does not necessarily translate to better model performance. In fact, unnecessary inclusion of weakly informative or redundant features can increase model complexity, reduce generalizability, and impair interpretability, particularly in healthcare applications where transparency is critical. Thus, the decision to limit the model to the top seven features ensures a balanced trade-off between predictive power and interpretability, which is essential for medical decision support systems.

In descending order, the importance of each corresponding feature is shown in Figure 6. It can be seen from

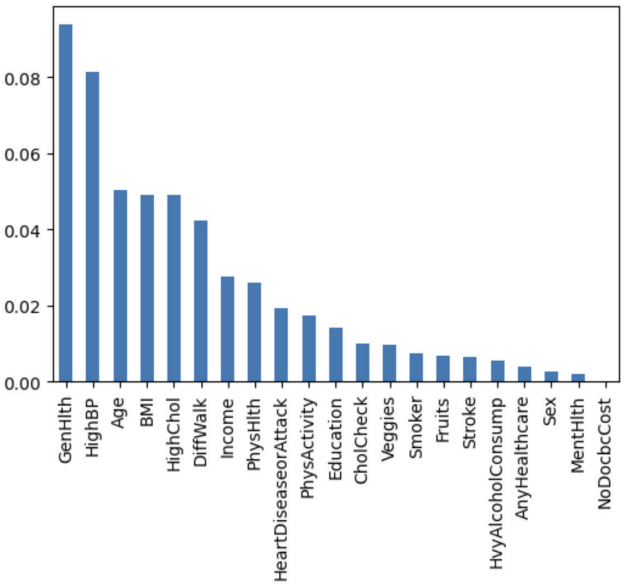


Figure 6. Feature importance rankings in predicting target variable-a bar graph of attribute contributions.

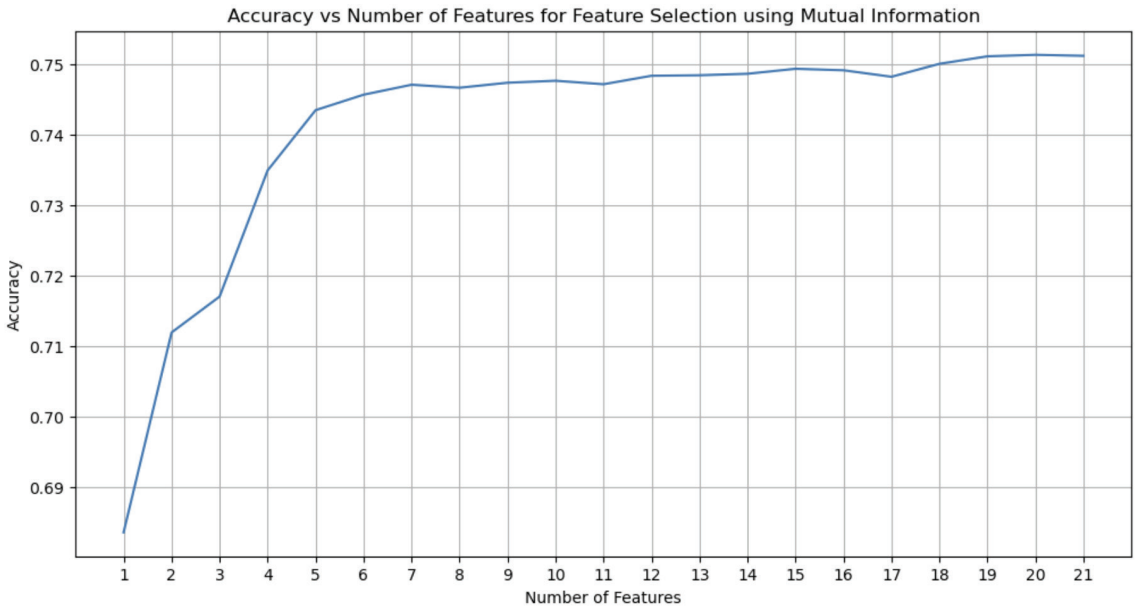


Figure 5. Variability in accuracy of model with increasing number of selected features.

this figure, that features such as General Health, High Blood Pressure, Age, and BMI rank relatively higher than others due to high statistical importance of these features in prediction of diabetes. They not only fit well with medical criteria, but also provide readily accessible input data.

Top seven features *viz.* General Health, High Blood Pressure, Age, BMI, High Cholesterol, Difficulty in Walking and Income were chosen based on mutual information score. Such a choice is quite reasonable both statistically and clinically. For example, poor general health status and comorbid factors such as hypertension and hypercholesterolemia are common risk factors in metabolic syndrome. Moreover, walking difficulties may indicate the presence of any chronic diseases, whereas income level acts as an indicator of social determinants of health. By reducing the number of features to seven critical ones, the predictive power of the model is enhanced, and its computational efficiency and interpretability improved.

Machine Learning Model Development and Evaluation

It was decided to explore a variety of ML models as different models have different strengths. Logistic regression was chosen for its simplicity and interpretability, while decision trees and random forests were selected for their ability to capture non-linear relationships between features. The models were trained on a training set comprising 80% of the dataset, with the remaining 20% used for testing. The performance of these models was evaluated using metrics such as accuracy, recall, and ROC-AUC scores.

Logistic Regression outperformed the decision tree and random forest models in terms of accuracy and interpretability as depicted in Table 3. It achieved an accuracy of 74.72%. Recall is an important metric for diagnosis of any disease as the focus is on identification of True positives out of all positives (people suffering from diabetes) and the model has achieved a value of 76.63%. Area under the Receiver operating characteristic curve is another significant metric for evaluating the performance of a classifier as it considers both the True Positive Rate (Sensitivity) and False Positive Rate and hence the model's ability to distinguish between positive and negative classes. The logistic regression has resulted in highest value of AUC that equals 0.7468 as visualized in Figure 7a. This curve highlights its effective ability to differentiate between classes, making it a reliable choice for our classification task. The Random Forest model showcases its robust classification power with an AUC of 0.7217 (Fig. 7b) and the ROC curve for Decision Tree, with an AUC of 0.7201 (Fig. 7c), demonstrates its straightforward yet effective approach to classification. So, finally Logistic regression model was selected to be used at back end of a GUI based interface as explained in next section.

Table 3. Performance metrics comparison of logistic regression, decision tree, and random forest ML models

Model	Accuracy	Recall	ROC-AUC
Logistic Regression	74.72%	76.63%	0.7468
Decision Tree	72.03%	72.71%	0.7201
Random Forest	72.23%	75.41%	0.7217

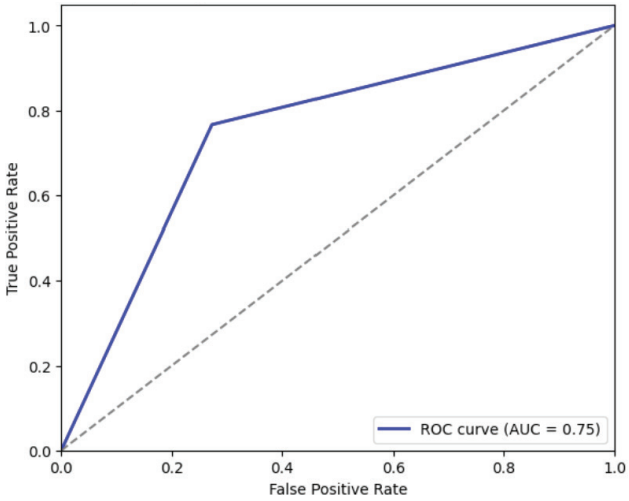


Figure 7a. ROC curve for the logistic regression model, illustrating an AUC of 0.75.

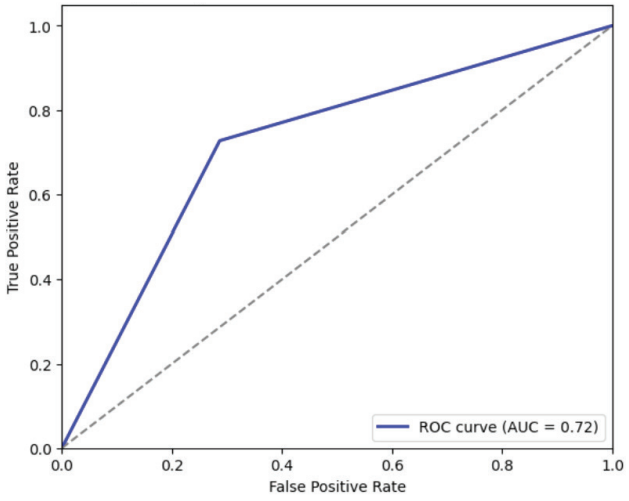


Figure 7b. ROC curve for the decision tree model, showing an AUC of 0.72.

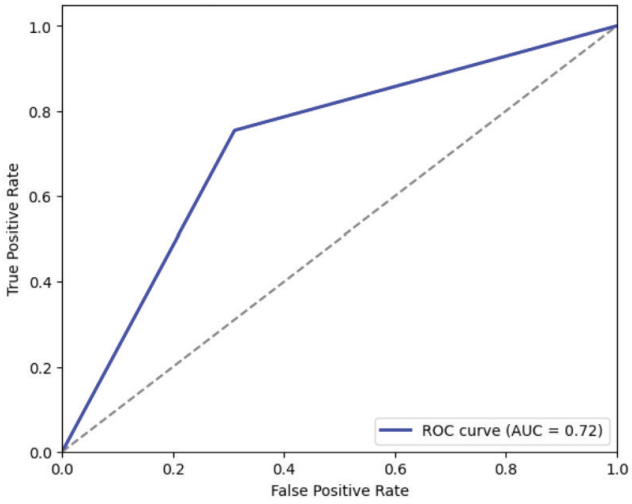


Figure 7c. ROC curve for the random forest model, demonstrating an AUC of 0.72.

User Interface and Accessibility

A simple yet intuitive graphical interface was created to enter data related to patients' health and get predictions. An accessible design was chosen to facilitate access to the system for people who have different levels of IT literacy. The graphical interface was constructed using HTML, CSS, and JavaScript, while the backend was implemented using the Python web framework Flask. The graphical interface makes it possible for the user to enter their data and immediately receive a prediction on their risk of developing type 2 diabetes. The explanation of the reasons for the predictions will be provided, along with important factors affecting the risk.

In order to implement local interpretability of a predictive machine learning model, one can use post-hoc tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). However, our project involves an integration of explainability into the process of creating the model itself in the form of Mutual Information-based feature selection. Mutual Information can help measure the level of dependency between each of the input features and the target variable; thus, it is possible to select only informative features. The reduction of the predictive model to a few interpretable features viz. general health, blood pressure, and BMI; results in natural transparency that is easier to explain through the graphical interface.

RESULTS AND DISCUSSION

The study's findings demonstrate how machine learning, and logistic regression in particular, can be used to predict diabetes. With exhaustive data pre-processing steps like feature engineering and scaling the framework for the research was augmented. Under feature engineering, binning for multivalued variables, one-hot encoding for categorical variables was done. Following this, scaling of numerical features, to normalize the impact of wider ranged features was done. Since the dataset is highly imbalanced (86% are non-diabetic and 14% are diabetic), balancing of datasets using Random Undersampling technique was executed and seven most significant features were selected using Mutual Information, an accuracy of 74.72% was achieved. With all these steps in place, the dataset was made ready to be fed to a variety of ML algorithms as each have their own strengths and weaknesses.

While previous studies—including those by [13] and [14] have employed logistic regression for diabetes prediction, they often relied on larger, more complex feature sets without emphasizing interpretability or user accessibility. In contrast, our study uniquely combines mutual information-based feature selection with logistic regression to develop a high-performing model using just seven non-invasive features, achieving an accuracy of 74.72%. This not only ensures technical rigor but also translates into a simple, interactive interface where healthcare professionals can input answers to seven straightforward questions and receive an immediate, explainable prediction—an innovation that bridges algorithmic performance with real-world usability.

The logistic regression model is a good option for diabetes prediction because of its excellent accuracy and interpretability. Logistic Regression assumes independence between features which is not the case with Decision Trees and Random Forests. Additionally, we found that in our dataset, the features have low multicollinearity and hence they contribute individually to the target, Logistic Regression is able to leverage them effectively. The numerical features have been scaled using standardscaler() that makes selected model produce better accuracy. Also, Logistic Regression is fairly robust to noise in the data as it models probabilities and avoids hard splits which is not the case with decision tree and random forest. Hence selection of Logistic regression was finalized for further development of GUI based model for convenient usage by healthcare application.

Early detection and action are encouraged by the interactive web interface used in this study, which enables users to enter personal health information that includes just the answers to seven questions viz. General Health condition (category to select), High blood pressure (Yes/No), Age (category to select), BMI (category to select), High cholesterol (Yes/No), Difficulty in Walking (Yes/No), and Income (category to select), and obtain risk evaluations. Both patients and healthcare professionals can benefit from the interface's user-friendly design, which guarantees accessibility for people in an easy manner.

CONCLUSION

This study introduces a novel, explainable machine learning framework for proactive diabetes risk assessment,

emphasizing simplicity, interpretability, and real-world usability. Unlike many prior studies that utilize large and complex feature sets or black-box models like deep learning, our work identifies just seven statistically and clinically relevant features using mutual information, a feature selection process that provides similar insights to post-hoc tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations). These features are then operationalized in a logistic regression model achieving an accuracy of 74.72%, making it both technically robust and interpretable.

The innovation lies not only in the methodological rigor but also in the practical deployment of the model through a user-friendly web interface. This interface allows patients and healthcare professionals to input responses to only seven easy questions without the need for lab tests and receive an immediate, and interpretable prediction. Predictive models have the potential to significantly improve patient care by enabling earlier diagnoses and more personalized treatment plans, while also helping to ease the growing burden on healthcare systems. This significantly enhances accessibility, particularly for underserved populations, offering a simple decision-support tool for early intervention.

In future, the effort to fight against diabetes and other chronic diseases will depend more and more on the incorporation of machine learning technologies into clinical practice. It is suggested that incorporating such predictive tools with healthcare systems for wider use should be the main goals of future study.

aIt is proposed to incorporate this model with a mobile application in Android and IOS for tapping a wider and broader client base. Longitudinal monitoring and tracking capabilities could also be included in order to enable continuous monitoring integrated with electronic health records and push notifications to alert the patient in the event of any clinically significant medical condition.

There remains a key constraint of this model that is its wider applicability. The population dataset employed is taken from the U.S. population through the BRFSS ("Behavioral Risk Factor Surveillance System") survey and could represent certain demographic, behavioral and socioeconomic characteristic that are not necessarily found in worldwide populations. Generalizing the model to other geographic or ethnic or socioeconomic context (without re-training) may yield inaccurate predictions or biased predictions. Hence, region-based fine-tuning is required before global deployment.

And finally, any use of AI applications in healthcare needs to be done cautiously. Ethical concerns especially about patient's data privacy, informed consent, fairness of algorithm and its compliance to regulatory standard such as HIPAA ("Health Insurance Portability and Accountability Act") or GDPR ("General Data Protection Regulation") must be considered as per the legal framework of country. Above all, transparent data treatment practices, explainable modeling decisions and inclusive design principles are crucial to make sure such tools not only be effective, but also equitable and trustworthy for clinical translation.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] IDF Diabetes Atlas. 2021. Available at: https://www.idf.org/privacy-policy. Accessed November 9, 2026.

[2] Renub Research. 2024. Available at: https://www.renub.com. Accessed November 9, 2026.

[3] Yang W, et al. Economic costs of diabetes in the U.S. in 2017. Diabetes Care 2018;41:917–928. [CrossRef]

[4] National Institute of Diabetes and Digestive and Kidney Diseases. Diabetes. 2024.

[5] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347–1358. [CrossRef]

[6] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017. [CrossRef]

[7] Zhou H, Myrzashova R, Zheng R. Diabetes prediction model based on an enhanced deep neural network. EURASIP J Wirel Commun Netw 2020;2020. [CrossRef]

[8] Khushi M, et al. A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access 2021;9:109960–109975. [CrossRef]

[9] Sivaranjani S, Ananya S, Aravinth J, Karthika R. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS); 2021. p. 141–146. [CrossRef]

[10] Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. J Med Internet Res 2018. [CrossRef]

[11] Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019;25:37–43. Available at: https://www.nature.com/articles/s41591–018–0272–7. Accessed November 9, 2026.

[12] World Health Organization. Global report on diabetes. 2023.

[13] Nibareke T, Laassiri J. Using big data-machine learning models for diabetes prediction and flight delays analytics. J Big Data 2020;7. [CrossRef]

[14] Khalate V, Sukeshkumar B. Diabetes prediction using machine learning algorithm. Int J Res Appl Sci Eng Technol 2024;12:1963–1969. [CrossRef]

[15] Modak SKS, Jha VK. Diabetes prediction model using machine learning techniques. Multimed Tools Appl 2024;83:38523–38549. [CrossRef]

[16] Shojaee-Mend H, Velayati F, Tayefi B, Babaee E. Prediction of diabetes using data mining and machine learning algorithms: a cross-sectional study. Healthc Inform Res 2024;30:73–82. [CrossRef]

[17] Amalarethinam G, Vignesh NA. Prediction of diabetes mellitus using data mining techniques-a survey. Int J Appl Eng Res 2015.

[18] Islam MT, Raihan M, Farzana F, Raju MGM, Hossain MB. An empirical study on diabetes mellitus prediction for typical and non-typical cases using machine learning approaches. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 2019. [CrossRef]

[19] Irmawan FN, Muliawan NH, Angky EV, Setiawan KE, Hasani MF. Comparison of machine learning algorithms for diabetes classification. In: 2023 IEEE 9th International Conference on Computing, Engineering and Design (ICCED); 2023. [CrossRef]

[20] Abdollahi J, Aref S. Early prediction of diabetes using feature selection and machine learning algorithms. SN Comput Sci 2024;5. [CrossRef]

[21] Talukder MA, et al. Toward reliable diabetes prediction: innovations in data engineering and machine learning applications. Digit Health 2025. [CrossRef]

[22] Tripathy N, Moharana B, Balabantaray SK, Nayak SK, Pati A, Panigrahi A. A comparative analysis of diabetes prediction using machine learning and deep learning algorithms in healthcare. In: Proceedings of 2nd International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC); 2024. [CrossRef]

[23] Al-Rubaye MM, Ilhan HO. The evaluation of the effect of data balancing over the classification performances of ensemble of networks for the diabetic retinopathy. Sigma J Eng Nat Sci 2024;42:1563–1574. [CrossRef]

[24] Obermeyer Z, Emanuel EJ. Predicting the future-big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216–1219. [CrossRef]

[25] Ismail L, Materwala H. IDMPF: intelligent diabetes mellitus prediction framework using machine learning. Appl Comput Inform 2021. [CrossRef]

[26] Agarwal L, Singh Y, Saxena A, Verma R. GUI-based diabetes prediction using pipeline. In: International Conference on Signal, Machines, Automation, and Algorithm. Springer; 2023. p. 363–373. [CrossRef]

[27] Qi H, Song X, Liu S, Zhang Y, Wong KKL. KFPredict: an ensemble learning prediction framework for diabetes based on fusion of key features. Comput Methods Programs Biomed 2023;231:107378. [CrossRef]

[28] Badik ST, Akar M. Machine learning classification models for the patients who have heart failure. Sigma J Eng Nat Sci 2024;42:235–244. [CrossRef]

[29] Chawla NV, Japkowicz N, Kotcz A. Editorial. SIGKDD Explor 2004;6:1–6. [CrossRef]

[30] Bagui S, Li K. Resampling imbalanced data for network intrusion detection datasets. J Big Data 2021;8. [CrossRef]

[31] Kubat M, Matwin S. Addressing the curse of imbalanced training sets. ACM SIGKDD Explor Newsl 1997.

[32] Shelke MS, Deshmukh PR, Shandilya VK. A review on imbalanced data handling using undersampling and oversampling technique. Int J Recent Trends Eng Res 2017;3:444–449. [CrossRef]

[33] Fernández-Diego M, González-Ladrón-de-Guevara F. Application of mutual information-based sequential feature selection to ISBSG mixed data. Softw Qual J 2018;26:1299–1325. [CrossRef]