# Diabetes Prediction Using Machine Learning

Ahmed Emad
*Faculty of Biomedical Engineering*
*Cairo University*
ahmed.elzayat077@gmail.com

Nourhan Ahmed
*Faculty of Biomedical Engineering*
*Cairo University*
nourhanahmed0505@gmail.com

Mayar Ahmed
*Faculty of Biomedical Engineering*
Cairo University
mayar694.ma@gmail.com

Ziad Meligy
*Faculty of Biomedical Engineering*
Cairo University
ziadmeligy@gmail.com

***Abstract*—In recent years, the integration of machine learning techniques into the healthcare domain has been a promising approach to revolutionizing patient care and disease management., fueled by abundant healthcare data and the urgent demand for more efficient diagnostic tools. Our paper focuses on diabetes prediction, a global health concern characterized by high blood sugar levels, which presents a significant challenge for both individuals and healthcare professionals due to its wide prevalence and complicated nature. With current practices relying on conventional diagnostic tests, they often fail to achieve high classification and prediction accuracy. In response, we propose a novel diabetes prediction model that combines the strengths of machine learning algorithms and Big Data Analytics. Our aim is to categorize individuals into three groups: non-diabetic, pre-diabetic and diabetic classes. Identifying pre-diabetic people allows for early intervention and lifestyle modifications to potentially prevent progression to diabetes. In this paper, several machine learning classifier models were built, including Support Vector Machine (SVM), Random Forest, Decision Tree, Logistic Regression, XGBoost, AdaBoost, and Linear Discriminant Analysis (LDA). The SVM classifier demonstrated the best overall performance, achieving an accuracy of 65%, a recall of 0.65, and an F1-score of 0.70. Performing best in diabetes state prediction.**

**Keywords—Healthcare, Diabetes, Disease Prediction, Big Data, Machine Learning**

## I. Introduction

Diabetes is one of the most populated diseases in the world according to the World Health Organization (WHO). Diabetes occurs when blood glucose, also called blood sugar, is high. This condition is called Hyperglycemia, which can be caused by metabolic problems, or when your body doesn't make enough —or any—insulin. Insulin is a hormone produced by the pancreas that helps glucose get into the body cells to be used for energy, accordingly, it lowers the blood glucose level. Over time, this elevated blood glucose level can lead to serious health problems and raises the risk for damage to many body organs, like eyes, heart, kidneys and nerves [1]. However, diabetes has no cure, taking steps to a healthy lifestyle helps manage or prevent diabetes and may lower the risk of developing diabetes health problems. There are three major types of diabetes: type 1, type 2, and gestational diabetes. Type 1 diabetes is a condition where your immune system destroys insulin-making cells in your pancreas. Type 2 diabetes is a lifelong disease that keeps your body from using insulin the way it should. Gestational diabetes is high blood glucose which occurs during pregnancy. Pre-diabetic individuals have higher-than-normal blood glucose levels, but not enough to be diagnosed as diabetes [2].

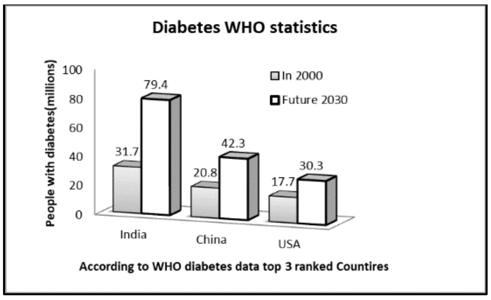


Fig. 1. WHO Diabetes statistics

The importance of early prediction of diabetes lies in its potential to reduce its dangerous effects on people's health and well-being. Early diabetes detection remains a challenge in the healthcare domain. The motivation behind pursuing this problem lies in the desire to address a critical gap in current healthcare practices. By harnessing the power of machine learning algorithms, we aim to develop predictive models capable of accurately predicting individuals at risk of diabetes, leading to the facilitation of timely interventions, and adopting proactive strategies to prevent or reduce its risk.

We pursue this problem by using a dataset with various demographic, clinical and lifestyle features as [High Blood Pressure, High Cholesterol, Cholesterol Check, BMI, Smoker, Stroke, Heart Disease or Attack, Physical Activity, Fruits, Veggies, Heavy Alcohol Consumption, General Health, Mental Health, Difficulty Walking, Sex, Age, Education, Income, Any Healthcare, No Doctor because of cost], our output is to predict whether this case is non-diabetic, diabetic, or pre-diabetic.

## II. Related Work

The study by Ayman mir investigates the use of machine learning algorithms within WEKA toolkit to predict

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

diabetes based on the Pima Indians Diabetes Database. This dataset contains 768 instances with nine attributes related to glucose levels, blood pressure, and other clinical indicators. The researchers compare the performance of Naive Bayes, Support Vector Machine (SVM), Random Forest, and Simple CART classifiers. Results show that SVM achieves the highest accuracy at 79.13%, outperforming other algorithms. Metrics like precision, recall, and F-measure further demonstrate SVM's effectiveness in predicting diabetes from clinical data. This research highlights the potential of machine learning in medical diagnosis and underscores SVM as a promising tool for diabetes prediction [3].

This study developed a Type 2 Diabetes prediction model using a six-year dataset from Hanaro Medical Foundation, Seoul, South Korea, with 253,395 subjects. Data preprocessing addressed missing values and class imbalance (normal: 68.1%, diabetes: 4.3%, prediabetes: 37.6%) using methods like SMOTE. Feature selection identified key predictors like glucose levels and BMI using ANOVA and chi-squared tests, followed by Recursive Feature Elimination (RFE). Prediction models (RF, SVM, XGBoost) were trained, optimized, and evaluated using accuracy, precision, recall, and F1-score metrics, showing promising results for diabetes prediction. Increasing training years improved model accuracy, highlighting the value of historical medical data for T2D prediction [4].

This research in diabetes classification has explored various machine learning techniques for predictive modeling. Li et al. proposed an ensemble method combining support vector machines and Naïve Bayes without preprocessing, achieving 58.3% accuracy. Deng used self-organizing maps with 10-fold cross-validation, achieving 78.4% accuracy. Sisodia et al. applied decision trees, support vector machines, and Naïve Bayes, with Naïve Bayes performing best at 76.3% accuracy. Then the paper focused on feature selection, achieving improved accuracy with six and four selected features. This prior work underscores the importance of preprocessing techniques and feature selection in enhancing diabetes classification performance [5].

Wilson et al. developed the Framingham Diabetes Risk Scoring Model (FDRSM) using Logistic Regression to predict the risk of developing diabetes mellitus (DM) in middle-aged American adults (45 to 64 years of age). This simple clinical model incorporates risk factors such as parental history of DM, obesity, high blood pressure, low levels of high-density lipoprotein cholesterol, elevated triglyceride levels, and impaired fasting glucose. The study included 3140 subjects, and the area under the receiver operating characteristic curve (AROC) was reported to be 85.0% [6].

## III. Dataset and Features

### A. Dataset Description

The dataset used in our case was obtained from the Centers for Disease Control and Prevention's (CDC) Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The BRFSS is an annual health-related telephone survey conducted by the CDC, collecting responses from a large sample of Americans on various health-related risk behaviors, chronic health conditions, and the use of preventative services

The dataset consists of responses from 253,680 individuals surveyed during the 2015 BRFSS. Each individual's response contributes to an observation in the dataset. The dataset contains a total of 21 feature variables, which include both direct survey questions and calculated variables based on individual participant responses.

The dataset includes a **target variable**, Diabetes_012, which categorizes individuals into **three** classes based on their diabetes status:

***0: No diabetes or diabetes only during pregnancy***

***1: Prediabetes***

***2: Diabetes***

TABLE I. DATASET FEATURES DESCRIPTION

| No. | Feature | Meaning |
|---|---|---|
| **1** | **Diabetes_012** | Target indicating whether the individual has diabetes (2: Yes,1: Pre-diabetic, 0: No) |
| **2** | **HighBP** | Indicates whether the individual has been diagnosed with high blood pressure (1: Yes, 0: No) |
| **3** | **HighChol** | Indicates whether the individual has been diagnosed with high cholesterol (1: Yes, 0: No) |
| **4** | **CholCheck** | Indicates whether the individual has had a cholesterol check within the past five years (1: Yes, 0: No) |
| **5** | **BMI** | Body Mass Index (a measure of body fat based on height and weight) |
| **6** | **Smoker** | Indicates whether the individual has smoked at least 100 cigarettes in their lifetime (1: Yes, 0: No) |
| **7** | **Stroke** | Indicates whether the individual has ever been told they had a stroke (1: Yes, 0: No) |
| **8** | **HeartDiseaseorAttack** | Indicates whether the individual has ever reported having coronary heart disease (CHD) or myocardial infarction (MI) (1: Yes, 0: No) |
| **9** | **PhysActivity** | Indicates whether the individual reported doing physical activity or exercise during the past 30 days other than their regular job (1: Yes, 0: No) |
| **10** | **Fruits** | Indicates whether the individual consumes fruit one or more times per day (1: Yes, 0: No) |
| **11** | **Veggies** | Indicates whether the individual consumes vegetables one or more times per day (1: Yes, 0: No) |
| **12** | **HeavyAlcoholConsumption** | Indicates whether the individual is a heavy drinker (1: Yes, 0: No) |
| **13** | **AnyHealthcare** | Indicates whether the individual has any kind of health care coverage (1: Yes, 0: No) |
| **14** | **NoDoctorDueToCost** | Indicates whether the individual could not see a doctor in the past 12 months due to cost (1: Yes, 0: No) |
| **15** | **GeneralHealth** | Self-rated general health status (1 to 5 rating) |
| **16** | **MentalHealth** | Number of days during the past 30 days when mental health was not good (0 to 30 days) |
| **17** | **PhysicalHealth** | Number of days during the past 30 days when physical health was not good (0 to 30 days) |

| | | |
|---|---|---|
| **18** | **DifficultyWalking** | Indicates whether the individual has serious difficulty walking or climbing stairs (1: Yes, 0: No) |
| **19** | **Sex** | Indicates the sex of the respondent (1: Female, 0: Male) |
| **20** | **Age** | Fourteen-level age category (1 to 14 representing different age groups) |
| **21** | **Education** | Highest grade or year of school completed (1 to 6 representing different education levels) |
| **22** | **Income** | Annual household income category (1 to 8 representing different income levels, with "Refused" if respondent did not disclose) |

### B. *Class Imbalance Problem*

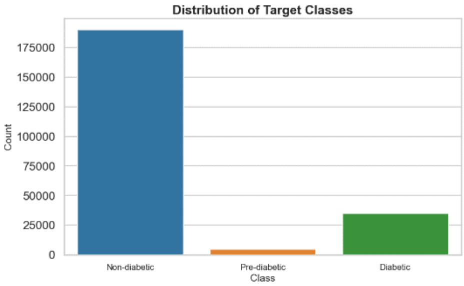


Fig. 2. Classes Distribution Count Plot

The dataset exhibits a significant class imbalance, with most non-diabetic instances (190,055) compared to diabetic (35,097) and pre-diabetic (4,629) instances. This imbalance poses a challenge as models trained on such data tend to favor the majority class, potentially overlooking important patterns in minority classes. To address this, careful data preprocessing and sampling strategies are essential to mitigate bias and ensure the model's ability to generalize well across all classes. Neglecting to apply appropriate resampling techniques during model training could lead to skewed predictions favoring the majority class, compromising the model's effectiveness in accurately predicting minority classes. Therefore, implementing effective resampling methods is crucial to improve the model's performance and reliability on imbalanced datasets.

### C. *Data Preprocessing*

□ **Handling Outliers**

Given that our dataset includes minority classes (Class 1 and Class 2), outliers were not removed. Performance evaluations indicated that models performed better with the outliers included. Therefore, all data points, including outliers, were retained for training and evaluation.

□ **Feature Selection**

Feature selection was carried out to eliminate redundant and highly correlated features, enhancing the models' efficiency and performance. The following features were dropped: 'Smoker', 'Sex', 'Fruits', 'Veggies', 'NoDocbcCost', 'PhysActivity', 'AnyHealthcare', 'Education', 'Stroke', 'CholCheck', 'HvyAlcoholConsump'. Removing these columns helped reduce noise and improve the overall performance of the predictive models.

## IV. METHODS

### A. *Our Learning Algorithms*

We employed a variety of machine learning algorithms to tackle the Diabetes problem using several Health Indicators. Each of them was chosen very carefully to deal with certain aspects of the problem and deliver varied data perspectives. Here is a small-scale summary of all models we applied in the study:

□ Support Vector Machine:

SVM algorithm finds a hyperplane that distinguishes one class from the other classes in a feature space. SVM is especially important when working with complex data and has found extensive application across different fields owing to its resilience and adaptability.

□ Logistic Regression:

Logistic Regression is a fundamental algorithm for classification tasks. Despite its simplicity, it works well with linearly separable data. We employed logistic regression to establish a baseline performance and assess the impact of more complex models on our problem [7].

□ Decision Tree:

Decision trees are intuitive models that partition the feature space recursively based on attribute values. Resistant to outliers, they can capture complex relationships within data. They are so simple and easy to interpret.

□ Random Forest:

A learning technique that generates multiple decision trees during training. It determines class predictions by aggregating the majority vote of all trees which aids in reducing overfitting.

□ Extreme Gradient Boosting:

XGBoost is an optimized and distributed library designed for Gradient Boosting that is fast and effective. It is highly scalable and has become a popular choice due to its exceptional speed and performance [8].

□ AdaBoost:

Adaptive Boosting is an ensemble learning technique that combines multiple weak classifiers into a strong classifier [9]. It sequentially trains a series of classifiers, each focusing on instances misclassified by the previous ones.

□ Linear Discriminant Analysis:

LDA is a dimensionality reduction technique commonly used for classification problems. It projects the data into a lower dimensional space while maximizing the separation between classes. LDA assumes that the data follows a gaussian distribution and estimates the parameters of the distribution for each class.

## *B. Methodology*

The input dataset containing relevant features for diabetes prediction is fed into the model. We utilize seven machine learning models which are shown in Fig 3. Each model is individually trained and tested on the dataset. 80% of the dataset is trained while 20% are tested. The performance results from each model are observed and evaluated in comparison to the rest of the models to determine the best-performing model for accurate prediction. The block diagram, as shown in Fig 3., outlines the methodology employed for conducting the comparative analysis and identifying the optimal algorithm for diabetes prediction.

Here is a brief description about the flow of proposed methodology.

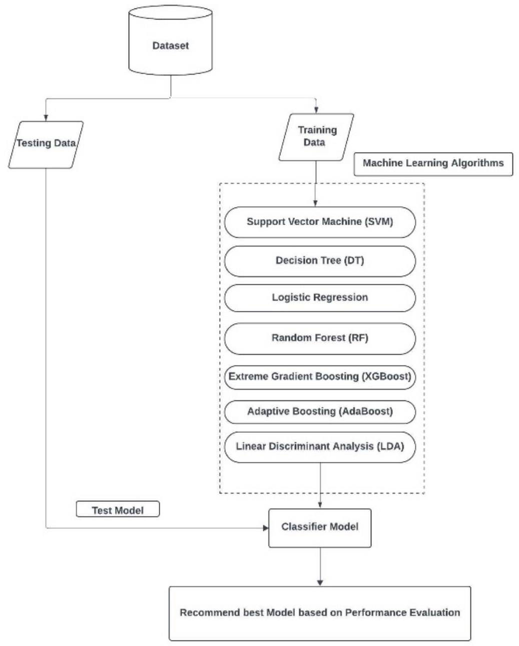

Dataset
Testing Data
Training Data
Machine Learning Algorithms
Support Vector Machine (SVM)
Decision Tree (DT)
Logistic Regression
Random Forest (RF)
Extreme Gradient Boosting (XGBoost)
Adaptive Boosting (AdaBoost)
Linear Discriminant Analysis (LDA)
Test Model
Classifier Model
Recommend best Model based on Performance Evaluation

Fig. 3. Proposed Methodology Block Diagram

## *C. Mathematical Notations*

*Objective functions:*

*a)* ***Support Vector Machine****:*

$$\min(w, b2) : \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} \xi i$$

***b) Logistic Regression****:*

$$P(y = 1 \mid x; w) = \frac{1}{1 + e^{-wTx}}$$

c) **Extreme Gradient Boosting**:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \bar{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

d) **Linear Discriminant Analysis**:

$$P(Y = k \mid X = x) = \frac{\pi_k N(x \mid \mu_k, \Sigma)}{\sum_{l=1}^{K} \pi_l N(x \mid \mu_l, \Sigma)}$$

e) **Adaptive Boosting:**

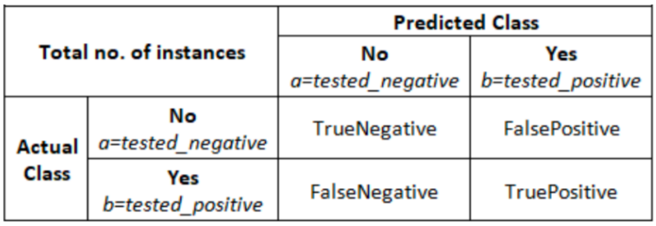
$$\epsilon_m = \frac{\sum_{i=1}^{N} w_{m,i} \big(y_i \neq Gm(x_i)\big)}{\sum_{i=1}^{N} w_{m,i}}$$

# V. Experimental Results

This section describes the experimental results obtained after training Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, Extreme Gradient Boosting, Adaptive Boosting and Linear Discriminant Analysis classifiers on the diabetes patients' dataset. The purpose of these experimental results is for performance evaluation of all seven classifiers and to recommend the best algorithm suited for prediction regarding our problem.

## *A. Confusion Matrix*

TABLE II. Confusion Matrix Structure



| **Total no. of instances** | | **Predicted Class** | |
|---|---|---|---|
| | | **No**<br>*a=tested_negative* | **Yes**<br>*b=tested_positive* |
| **Actual Class** | **No**<br>*a=tested_negative* | TrueNegative | FalsePositive |
| | **Yes**<br>*b=tested_positive* | FalseNegative | TruePositive |

In machine learning, a confusion matrix is used to analyze the performance of the classification algorithm. The confusion matrix is a tabular structure where the rows represent the actual class labels and columns represent the predicted class labels.

To interpret the confusion matrix, there is some certain terminology widely utilized in the general confusion matrix structure. This terminology will be further used for performance evaluation of each classifier and is described below:

(a) **Actual Class:** Class label representing the actual class before building the classifier.
(b) **Predicted Class:** Class label representing the predicted class after building the classifier.
(c) **True Positives:** No. of instances predicted positive and are positive.
(d) **True Negatives:** No. of instances predicted negative and are positive.
(e) **False Positives:** No. of instances predicted positive but are negative.
(f) **False Negatives:** No. of instances predicted negative but are positive.

In the case of our multi-class classification problem, the confusion matrix is comprehended by the 3 classes plotted against each other in rows and columns, where each row corresponds to actual classes and each column corresponds to predicted classes. The diagonal elements represent correctly classified instances, where predicted and actual classes match, while off-diagonal elements represent misclassifications. Higher values along the diagonal indicate better performance for individual classes. Below is the confusion matrix for our SVM classifier trained on the data:

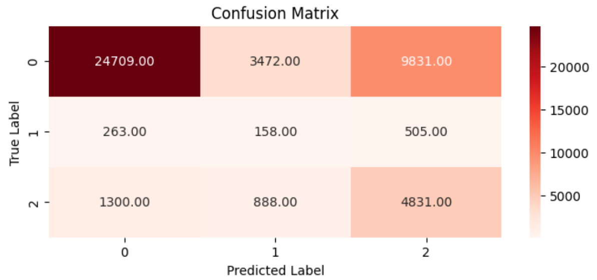


Fig. 4. Confusion Matrix for SVM classifier

### *B. Classification Accuracy*

Classification accuracy is one of the performance evaluation measures. Accuracy represents how well the classifier performs prediction of the instances based on the testing data.

**Accuracy:** It is the ratio of the no. of true predicted instance both positive and negative to the total no. of instances.

$$\textbf{Accuracy} = \frac{TruePos + TrueNeg}{Total\ number\ of\ samples}$$

The following table represents the experimental classification accuracy results of our algorithms:

TABLE III. DIFFERENT ALGORITHMS ACCURACIES

| *Algorithm* | *Accuracy (%)* |
|---|---|
| ***SVM*** | 65 |
| ***DT*** | 60 |
| ***RF*** | 64 |
| ***LR*** | 63 |
| ***XGBoost*** | 63 |
| ***AdaBoost*** | 64 |
| ***LDA*** | 63 |

### *C. Accuracy Measure Values*

These are the following classifier accuracy measures description:

(a) **TP - Rate:** It is the ratio of the no. of predicted positive instances to the actual total no. of positive instances.

$$\textbf{TP-Rate} = \frac{TruePos}{TruePos + FalseNeg}$$

**(b) FP - Rate:** It is the ratio of the no. of predicted negative instances to the actual total no. of negative instances.

$$\textbf{FP-Rate} = \frac{FalsePos}{FalsePos + TrueNeg}$$

**(c) Precision:** It is the ratio of no. of predicted positive instances to the total of all predicted positive instances.

$$\textbf{Precision} = \frac{TruePos}{TruePos + FalsePos}$$

(d) **Recall:** It is the ratio of the no. of predicted positive instances to the actual total no. of positive instances.

$$\textbf{Recall} = \frac{TruePos}{TruePos + FalseNeg}$$

**(e) F1-Score:** Used to represent overall performance. It is weighted harmonic mean of the precision and recall.

$$\textbf{F1-Score} = \frac{2 \cdot \mathrm{Pr}\,ecision \cdot Recall}{\mathrm{Pr}\,ecision + Recall}$$

TABLE IV. DIFFERENT ALGORITHMS ACCURACY MEASURES

| *Algorithm* | *TP Rate* | *FP Rate* | *Precision* | *Recall* | *F1-Score* |
|---|---|---|---|---|---|
| ***SVM*** | *0.503* | *0.186* | *0.83* | *0.65* | *0.70* |
| ***RF*** | *0.470* | *0.205* | *0.81* | *0.64* | *0.69* |
| ***DT*** | *0.424* | *0.245* | *0.78* | *0.60* | *0.66* |
| ***LR*** | *0.514* | *0.187* | *0.83* | *0.63* | *0.71* |
| ***XGBoost*** | *0.468* | *0.209* | *0.81* | *0.63* | *0.69* |
| ***AdaBoost*** | *0.517* | *0.181* | *0.85* | *0.64* | *0.72* |
| ***LDA*** | *0.515* | *0.182* | *0.85* | *0.63* | *0.71* |

### *D. Hyperparameters*

For hyperparameters selection, we directly focused on implementing the widely used Grid Search accompanied with 3 Cross-Validation folds with all the models. The 'macro recall' was the scoring method chosen, to focus on getting the highest TP rates due to the medical nature of our problem. For our proposed 'best' model, which is SVM, a **linear** kernel was chosen meaning that the decision boundary between the classes is a straight line (or a hyperplane in higher dimensions).

## VI. CONCLUSION AND FUTURE WORK

In this study, various machine learning algorithms were applied to predict diabetes, including Support Vector Machine (SVM), Random Forest, Decision Tree, Logistic Regression, XGBoost, AdaBoost, and Linear Discriminant Analysis (LDA). The **SVM** classifier demonstrated the **best overall performance**, achieving an accuracy of 65%, a recall of 0.65, and an F1-score of 0.70. The superior performance of the SVM classifier can be due to its effectiveness in handling high-dimensional data and finding an optimal hyperplane that maximizes the margin between classes. This characteristic makes SVM particularly suitable for complex datasets with overlapping features. The confusion matrix for

SVM showed superior accuracy in correctly classifying instances across all classes. Although AdaBoost achieved slightly higher precision and recall, SVM provided a better balance across all metrics and exhibited a more reliable confusion matrix. Logistic Regression and LDA also performed well, but their overall metrics were slightly lower than those of SVM.

In our case, we specifically focused on optimizing the recall score to minimize the false negatives, where individuals with diabetes are incorrectly classified as not having the disease. While this approach may result in more false positives, where individuals without diabetes are classified as having the disease, it is a less dangerous trade-off in the medical field. Ensuring that those with diabetes are identified and recognized immediately aligns with the medical principle of prioritizing patient safety and minimizing harm.

For future work, given additional time, team members, or computational resources, we would focus on exploring advanced modeling techniques and enhancing feature engineering. By experimenting with more sophisticated models, including deep learning architectures like neural networks, we could potentially capture complex patterns in the data more effectively. Simultaneously, investing more time in feature engineering could help uncover more nuanced relationships within the data, potentially improving the predictive performance of our models. These efforts would aim to refine our analysis and provide deeper insights into the factors influencing diabetes.

## VII. Acknowledgment

We would like to extend our heartfelt thanks to **Dr. Inas A. Yassine** and **Eng. Merna Bibars** for their invaluable guidance and support throughout this course. Their expertise and insights were crucial in the writing of our paper on this topic. We deeply appreciate their dedication and the time they invested in helping us achieve our goals.

## VIII. Contribution

In the development of this project, each member of our team contributed equally and collaboratively. We worked together to leverage the best of our abilities and the data available to us, ensuring a comprehensive approach to the research and analysis. Our collective efforts spanned both phases of the project, from initial data collection and preprocessing to in-depth analysis and the final preparation of our paper. This cooperative spirit was pivotal in addressing the challenges encountered and in achieving the insightful outcomes presented in our work.

## IX. References

[1] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)

[2] Diabetes Resource Center (WebMD)

[3] A. Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697439.

[4] Deberneh HM, Kim I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. Int J Environ Res Public Health. 2021 Mar 23;18(6):3317. doi: 10.3390/ijerph18063317. PMID: 33806973; PMCID: PMC8004981.

[5] Roshi Saxena, Sanjay Kumar Sharma, Manali Gupta, G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods", Computational Intelligence and Neuroscience, vol. 2022, Article ID 3820360, 11 pages, 2022. https://doi.org/10.1155/2022/3820360

[6] Peter W F Wilson 1, James B Meigs, Lisa Sullivan,.Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study

[7] Xiuping Jia, Bor-Chen Kuo, Melba M. Crawford, "Feature Mining for Hyperspectral Image Classification", Proceedings of the IEEE, vol.101, no.3, pp.676-697, 2013.

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.

[9] Z. Zheng and Y. Yang, "Adaptive Boosting for Domain Adaptation: Towards Robust Predictions in Scene Segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 5371-5382, 2022. DOI: 10.1109/TIP.2022.3195642.