# Diabetes Prediction

DSCI3415 Project

Mona Ibrahim
Mathematics and Actuarial Science
American University in Cairo
900212749
monamahmoud@aucegypt.edu

Maya ElShweikhy
Mathematics and Actuarial Science
American University in Cairo
900204233
mayahanyy1@aucegypt.edu

## ABSTRACT

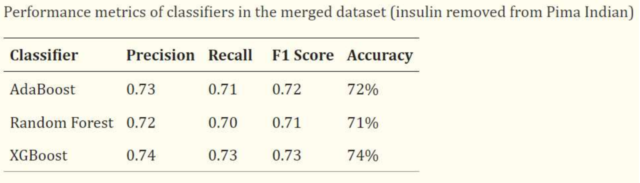
This paper addresses the global health crisis of diabetes, a chronic illness affecting millions worldwide. If accurate early prediction of the illness is achieved, the risk factor and severity of diabetes can be considerably decreased. The objective of this project is to create a system capable of early diabetes prediction for patients, with increased accuracy through the integration of outcomes from diverse machine learning techniques. Employing algorithms such as K-nearest neighbor, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree, the model's accuracy is assessed for each algorithm. Subsequently, the algorithm demonstrating notable accuracy is selected as the predictive model for type 2 diabetes. Through delving into the literature and comparing datasets, the paper explores the current landscape of employing machine learning techniques in predicting type 2 diabetes and highlights previous datasets and efforts in this domain.

## INTRODUCTION

Diabetes is a chronic illness that is characterized by inadequate insulin production by the pancreas or inefficient utilization of the insulin produced by the body [13]. According to the World Health Organization, diabetes is particularly dangerous as Hyperglycemia, also known as elevated blood glucose or sugar levels, is a common consequence of uncontrolled diabetes, leading to substantial damage to various bodily systems, particularly the nerves and blood vessels [13]. As of 2014, 8.5% of adults aged 18 and above were affected by diabetes. In 2019, diabetes directly contributed to 1.5 million fatalities, with 48% of these diabetes-related deaths occurring before the age of 70. Notably, in lower-middle-income countries, the mortality rate linked to diabetes witnessed a 13% increase [13]. The chronic illness is growing even among young people. According to statistics, 451 million individuals globally had diabetes in 2017, and by 2045, that number is expected to rise to 693 million [7]. There are two types of diabetes, type 1 diabetes; a chronic condition characterized by the immune system's attack on and destruction of the insulin-producing cells within the body. On the other hand, type 2 diabetes occurs when the body either fails to produce sufficient insulin or when the body's cells do not respond adequately to insulin [10]. Despite the absence of a long-term cure, diabetes can be managed and prevented with accurate early prediction strategies. This paper aims to explore the landscape of predicting type 2 diabetes using machine learning techniques.

## LITERATURE REVIEW

In recent years, plenty of diabetes prediction methods have been proposed and published. According to an article published by the American National Institute of Health, an automatic diabetes prediction system has been attempted using multiple machine learning models [12]. The researchers used a private dataset of female Bangladeshi patients as well as an open-source dataset named the Pima Indian Diabetes dataset to produce their models. The study applied SMOTE and ADASYN techniques to address imbalanced class problems, evaluating various metrics for machine learning and ensemble methods. XGBoost with ADASYN achieved the highest performance, displaying 74% accuracy, 0.73 F1 score, and 0.73 recall. Additionally, domain adaptation was demonstrated, and the optimized XGBoost model was deployed in a website and smartphone app for instant diabetes prediction.



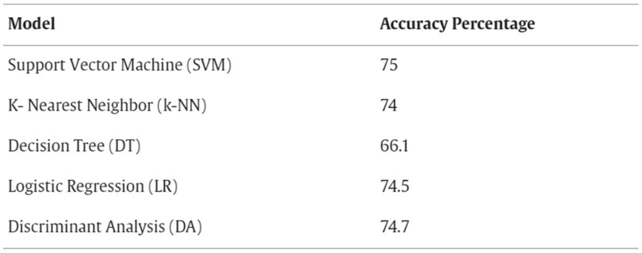
Performance metrics of classifiers in the merged dataset (insulin removed from Pima Indian)

| Classifier | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| AdaBoost | 0.73 | 0.71 | 0.72 | 72% |
| Random Forest | 0.72 | 0.70 | 0.71 | 71% |
| XGBoost | 0.74 | 0.73 | 0.73 | 74% |

**Figure 1: Table depicts various performance metrics of the merged dataset [12].**

In addition, a study by the Institute of Electrical and Electronics Engineers (IEEE) [7] proposed a different pipeline for diabetes prediction using the Pima Indians Diabetes dataset. With a focus on reprocessing, the proposed models consisted of outlier rejection, filling missing values, data standardization, feature selection, and K-fold cross-validation. The study considered the mean value in the missing position of attribute rather than median value, as it has a more central tendency toward the mean of that attribute distribution. The folding of the dataset for cross-fold validation was performed thoughtfully to preserve the percentage of class proportion, as same as in the original dataset. Different machine learning classifiers (k-nearest Neighbour (k-NN), Random Forest, Decision Tree, Naïve Bayes, AdaBoost, and XGBoost, also Multilayer Perceptron were implemented in the study's proposed pipeline. The final pipeline showed that XGBoost outperformed the other models.

Another study by Chatrati et al.[1], highlighted that the support vector machine classification algorithm (SVM) was found to be the most accurate and thus chosen to train the model for diabetes prediction. The model also used the famous Pima Indian Diabetes dataset, predicting the hypertension and diabetes status using the patient's glucose and blood pressure readings.



| Model | Accuracy Percentage |
|---|---|
| Support Vector Machine (SVM) | 75 |
| K- Nearest Neighbor (k-NN) | 74 |
| Decision Tree (DT) | 66.1 |
| Logistic Regression (LR) | 74.5 |
| Discriminant Analysis (DA) | 74.7 |

**Figure 2: Table shows a comparison between the accuracy of different models implemented by Chatrati et.al, [1].**

Comparing different machine learning models to predict diabetes using Pima Indian Diabetes dataset was also done by Anshi Gupta [5]. She implemented Logistic Regression, K-Nearest Neighbors Algorithm, Support Vector Classifier, Decision Tree Classifier, Gradient Boosting Classifier, XG Boost and Stacking, and compared their accuracy score showing that the Stacking and Logistic Regression models had the highest accuracy scores.

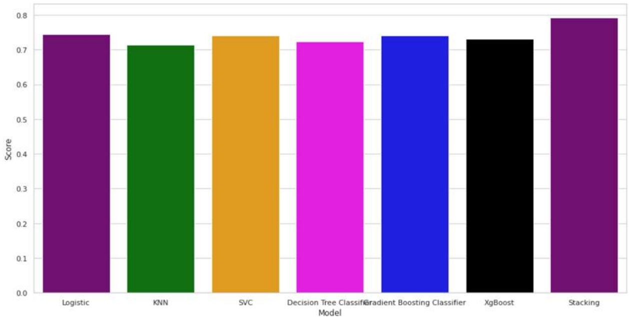


**Figure 3: Barchart comparing the accuracy of different models [5].**

Furthermore, a study by Deberneh et. al [2] used a six-year electronic medical record to predict diabetes. This study developed a machine-learning model to predict the occurrence of type 2 diabetes (T2D) for the following year using current-year variables from a medical record collected from 2013 to 2018 at a private medical institute called Hanaro Medical foundation in Seoul, South Korea. Key features were selected, including fasting plasma glucose (FPG), HbA1c, triglycerides, BMI, gamma-GTP, age, uric acid, sex, smoking, drinking, physical activity, and family history. Logistic regression, random forest, support vector machine, and XGBoost were employed. The model demonstrated good performance, offering valuable predictive information for clinicians and patients. Ensemble models, particularly CIM, ST, and SV, showed superior cross-validation performance to that of the single models, including Logistic Regression. The study incorporated medical history to improve prediction accuracy with Figure 4 showing a comparison between models.

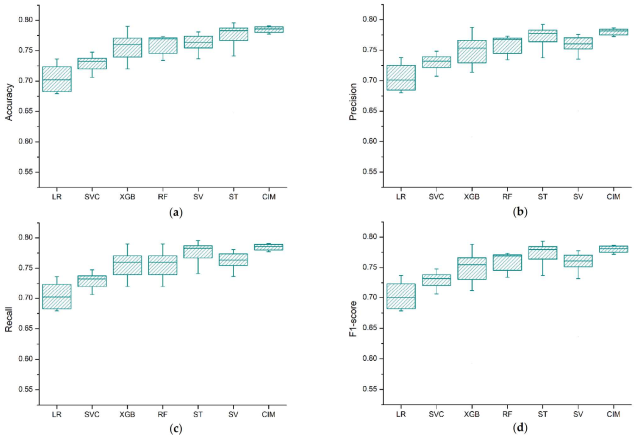


**Figure 4: Box plot for the CV score of the prediction models (LR = logistic regression, RF = random forest, XGB = XGBoost, SVM = support vector machine, ST = stacking classifier, CIM = confusion matrix-based classifier integration approach): (a) accuracy, (b) precision, (c) recall, (d) F1-score [2].**

## PROJECT DESCRIPTION AND ATTEMPTED SOLUTION

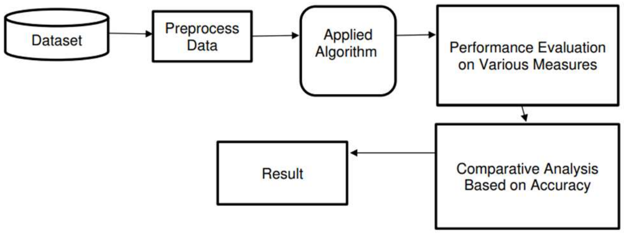

Dataset
Preprocess Data
Applied Algorithm
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Result

**Figure 5: Project Roadmap**

In this project, we aim to develop a predictive model for type 2 diabetes disease using machine learning techniques. We want to select a dataset that contains a variety of features such as demographic information, medical history, lifestyle factors, and possibly genetic markers to predict the likelihood of an individual developing diabetes.

Our attempted solution involves several steps:

- **Data Collection and Preprocessing:** we will choose an appropriate dataset that is relevant to our problem and consider its limitations. Then we will preprocess the data by dealing with outliers, missing values (adding the mean value to any missing data points), encoding categorical variables, imbalanced classes, and re-evaluating any incorrect entries.
- **Feature Selection:** We will select important and meaningful features and create new features if necessary to build a predictive model using these features.
- **Model Development:** We will split our data into training and testing sets. We will experiment with different machine learning algorithms such as Logistic Regression, K-Nearest Neighbors, Support Vector machines, Decision Tree Classifier, Gradient Boosting Classifier, and Random Forests.
- **Model Evaluation:** We will evaluate our models using performance metrics, such as accuracy, precision, recall, F1 score and other measures to compare the accuracy of each algorithm and determine the most accurate and suitable model.
- **Comparative Analysis**: We will compare models that had high accuracy and analyze which model performs better.
- **Application**: We will explore the potential application(s) of our predictive model once its validated. One of the applications is that we can deploy the model in hospitals database systems or create mobile applications to assist doctors for early prediction in identifying individuals at high risk of develop diabetes.

## OVERVIEW ON EXISTING DATASETS

The most popular dataset used in the literature is the Pima Indian Diabetes Dataset. This dataset is donated by the National Institute of Diabetes and Digestive and Kidney diseases [3], collected from the Pima Indian population near Phoenix, Arizona. It contains 768 observations with no missing data. It has 9 main features:

- Age
- Body Mass Index (weight in kg/(height in m)^2)
- Number of Pregnancies
- Plasma Glucose Concentration (in 2 hours in an oral glucose tolerance test)
- Triceps skin fold thickness
- Diastolic blood pressure
- Insulin levels
- Diabetes pedigree function (measures the patient's diabetic family history).

The target variable is the class variable Outcome showing the patient's status, 0 if the patient doesn't have diabetes and 1 if the patient has diabetes. This dataset is relevant as it shows multiple features that determine if a person develops diabetes, and it has a label variable aiding us in applying machine learning models, however it has some limitations. In particular, all patients in the dataset are females, at least 21 years old of Pima Indian heritage. In addition, it has a limited number of observations (768 only) which affects model training. Such limitations prevent us from producing accurate predictions for patients of ages below 21, patients of other ethnicities and male patients.

Moving on to another noteworthy dataset [9], it shows a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes 9 features:

- Age
- Gender
- Body Mass Index
- Hypertension
- Heart Disease
- Smoking History
- Hemoglobin A1c (HbA1c) level
- Blood glucose level

It has a target variable labeled diabetes that shows the status of the patient, which helps in model training for classification. The dataset has 100k observations with no missing data making it suitable for model training. Moreover, it collects data about heart disease which is associated with an increased risk of developing diabetes making it a relevant dataset to predict diabetes. However, the dataset does not disclose the country nor continent it has been collected from, preventing us from determining if the produced predictions are relevant to a specific population.

In a different context, the Diabetes dataset, sourced from real patient data collected in October 2023, focuses on rural African-American patients [8]. This dataset contains 403 observations and 19 variables. These variables are:

- Age
- Gender
- Height
- Weight
- chol (Cholesterol),
- stab.glu (Stabilized Glucose),
- hdl (High Density Lipoprotein Cholesterol),
- Ratio (Standard = Less than 5.7, Prediabetes = Between 5.7 and 6.5, Diabetes = More than 6.5),
- Frame,
- bp.1s (First diastolic blood pressure),
- bp.1d (First systolic blood pressure),
- bp.2s (Second diastolic blood pressure),
- bp.2d (Second systolic blood pressure),
- Waist, hip (Hybrid insulin peptides),
- time.ppn (Partial parenteral nutrition)

Glyhb is the target variable. The target variable contains 390 observations and is created using the glyhb variable in which if their hemoglobin A1c was 6.5 or greater they were labelled with diabetes = yes. 65 patients were found as diabatic and 325 were not diabatic. An advantage of this dataset is that it includes information about patients and using these features we can calculate other features such as the Body mass index (BMI) which is a measure of body fat based on height and weight. Some limitations are that this dataset has some missing values and class imbalance in the target variable.

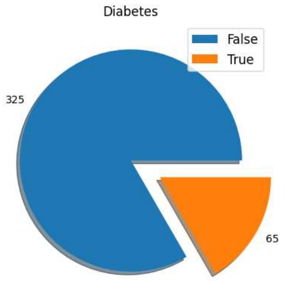


**Figure 6: Number of diabatic and non diabatic patients [7].**

Early Stage Diabetes Risk Prediction Dataset [4] is another interesting dataset that was collected in 2021 using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. This dataset contains 520 observations and 17 variables. There are 16 categorical variables:

- Sex
- Polyuria
- Polydipsia
- Sudden weight loss
- Weakness
- Polyphagia Genital thrush
- Visual blurring
- Itching
- Irritability
- Delayed healing
- Partial paresis
- Muscle stiness
- Alopecia
- Obesity

Class is the target variable (Positive or Negative), and it has one numeric variable which is the Age. An advantage of this dataset is that the class variable is balanced in which 200 patients were tested negative for diabetes indicating that the patient is non diabatic and 320 were tested positive indicating that the patient is diabatic. There are some limitations of this dataset as it mainly contains categorical variables; these limitations include limited information which can limit the model's performance, imbalanced classes, where one category is higher than the other. Lastly having mostly categorical variables affects Model Selection as some machine learning algorithms may be more suited to handling binary variables than others.

## REFERENCES

[1] Chatrati, S.P. , Hossain, G. , Goyal, A. , et al. 2020. Smart home health monitoring system for predicting type 2 diabetes and hypertension. J. King Saud Univ. Comput. Inf. Sci. 34(3), 862–870

[2] Deberneh, Henock M., and Intaek Kim. 2021. "Prediction of Type 2 Diabetes Based on Machine Learning Algorithm" International Journal of Environmental Research and Public Health 18, no. 6: 3317. https://doi.org/10.3390/ijerph18063317

[3] Dua, D., & Taniskidou, E. K.,2018,Kaggle Machine Learning Repository, "Pima Indians Diabetes Dataset", [Online]: Retrieved from: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data

[4] Dutta, Ishan. 2021. Early Stage Diabetes Risk Prediction Dataset. Retrieved from https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset/data

[5] Gupta, Anshi. 2021. Diabetes Prediction(EDA + Models). Retrieved from https://www.kaggle.com/code/anshigupta01/diabetes-prediction-edamodels/notebook

[6] Hosseini, Mojtaba. Diabetes. 2023. Prediction | 99% acc. Retrieved from https://www.kaggle.com/code/mojtbwa/diabetes-prediction-99-acc

[7] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, (2020) "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, doi: 10.1109/ACCESS.2020.2989857.

[8] Momeni, Mohamadreza. 2023. Diabetes. Retrieved from https://www.kaggle.com/datasets/imtkaggleteam/diabetes/data

[9] Mustafa, Mohammed. 2023. Diabetes prediction dataset. Retrieved from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data

[10] NHS. (n.d.-a). NHS choices. https://www.nhs.uk/conditions/diabetes/

[11] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

[12] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2022). Diabetes prediction using machine learning and explainable AI techniques. Healthcare technology letters, 10(1-2), 1–10. https://doi.org/10.1049/htl2.12039

[13] World Health Organization. (n.d.). Diabetes. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes