Predicting Diabetes: A Comparative Study of Machine Learning Algorithms for Classification

Group 7

Hemlatha Kaur Saran, George David Asirvatharaj, Raminder Singh

Module: Introduction to Artificial Intelligence (AAI-501-IN2)

Course: Master in Applied Artificial Intelligence

Institution: University Of San Diego

Professor: Ms. Azka

Date – 11/Aug/2025

Contents

Predicting Diabetes: A Comparative Study of Machine Learning Algorithms for Classification

Overview of Diabetes

Diabetes is an ongoing disease that afflicts millions of people all over the planet who have high levels of sugar in their blood. According to the World Health Organization (WHO), over 422 million of the world's population have diabetes, and this figure keeps on increasing (World, 2024). The condition comes with considerable complications: heart diseases, kidney failures, and loss of vision, imposing a great cost to the health systems and economy of the population at large.

Machine Learning in Healthcare

Healthcare has been changed due to the inclusion of machine learning (ML) that has allowed predictive models to help diagnose and predict a disease. More specifically, ML algorithms are capable of detecting trends in complicated healthcare-related information, including medical records and patient history (Olalekan Kehinde, 2025). These models have been strong in disease prediction of such illnesses as diabetes, empowering early diagnosis, individualized treatment approaches, and patient outcomes, becoming one of the sources of change in medical practice across the world.

Purpose and Objectives

This report aims to use machine learning to predict diabetes with the help of a dataset consisting of several health characteristics. Such models as logistic regression, random forest, XGBoost, and K-nearest neighbors (KNN) will be applied to cluster patients into diabetic and non-diabetic. Its primary goals would be data preprocessing, metrics assessment of the model

performance, and the choice of the most adequate model depending on the accuracy and interpretability.

Exploratory Data Analysis (EDA)

Table 1

Basic Statistics for Numerical Features

Variable	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Age	100,000	41.89	22.52	0.00	24.00	43.00	60.00	95.00
Hypertension	100,000	0.07	0.26	0.00	0.00	0.00	0.00	1.00
Heart Disease	100,000	0.04	0.19	0.00	0.00	0.00	0.00	1.00
BMI	100,000	27.32	6.64	10.01	23.63	27.32	29.59	95.60
HbA1c Level	100,000	5.53	1.08	3.50	4.00	5.00	6.20	9.00
Blood Glucose Level	100,000	138.06	40.71	80.00	100.00	140.00	159.00	300.00
Diabetes	100,000	0.09	0.28	0.00	0.00	0.00	1.00	1.00

Overview of Diabetes

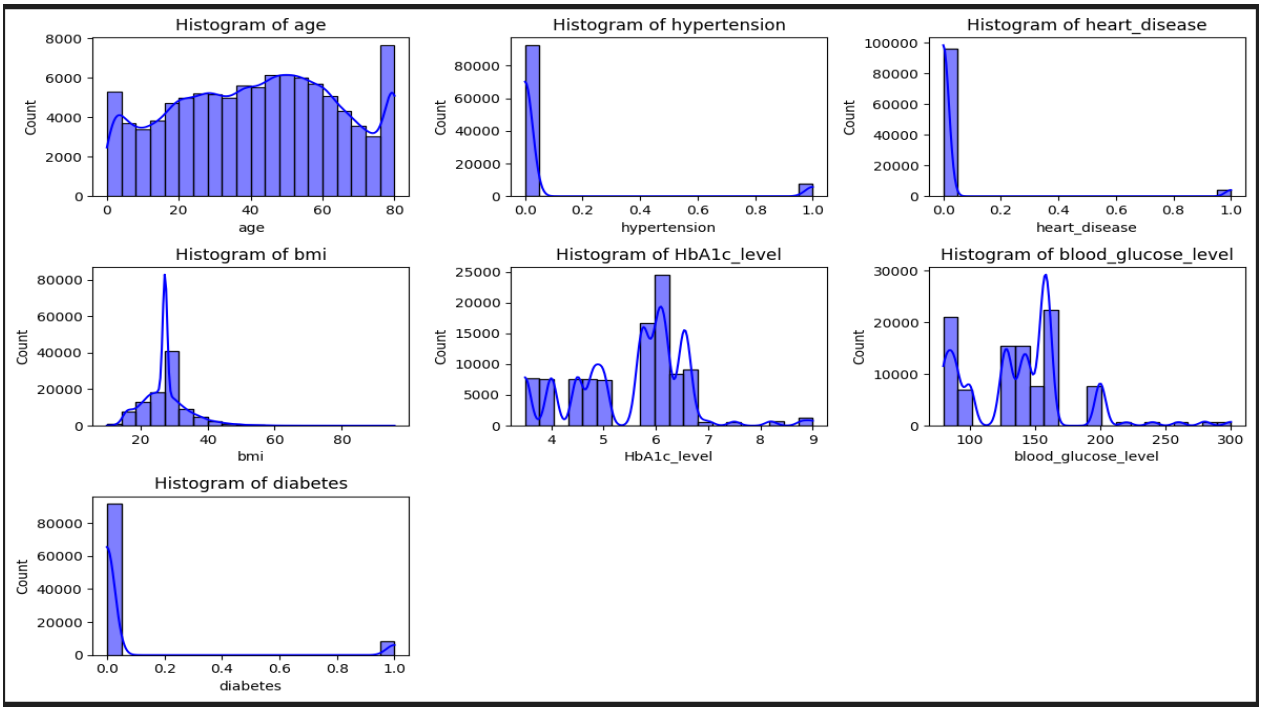
Diabetes is a serious health problem in the world with a total population of more than 420 million, and more than 50 percent of them are known to have type 2 diabetes (Basith et al.,

2019). The condition is linked to dire health risks, such as heart diseases, damage in kidneys, and blindness. The average age of people in the data is 41.89 years, and the mean BMI equals 27.32, and the mean blood glucose is 138.06, with the shown effect of lifestyle-related factors and reasons to be diagnosed and treated as early as possible.

Univariate Analysis

Figure 1

Histograms of numerical features in the diabetes dataset.



Note. The histograms illustrate the distribution of numerical variables such as age, BMI, HbA1c level, and blood glucose level. The target variable, diabetes, shows a highly imbalanced distribution with more non-diabetic cases than diabetic ones.

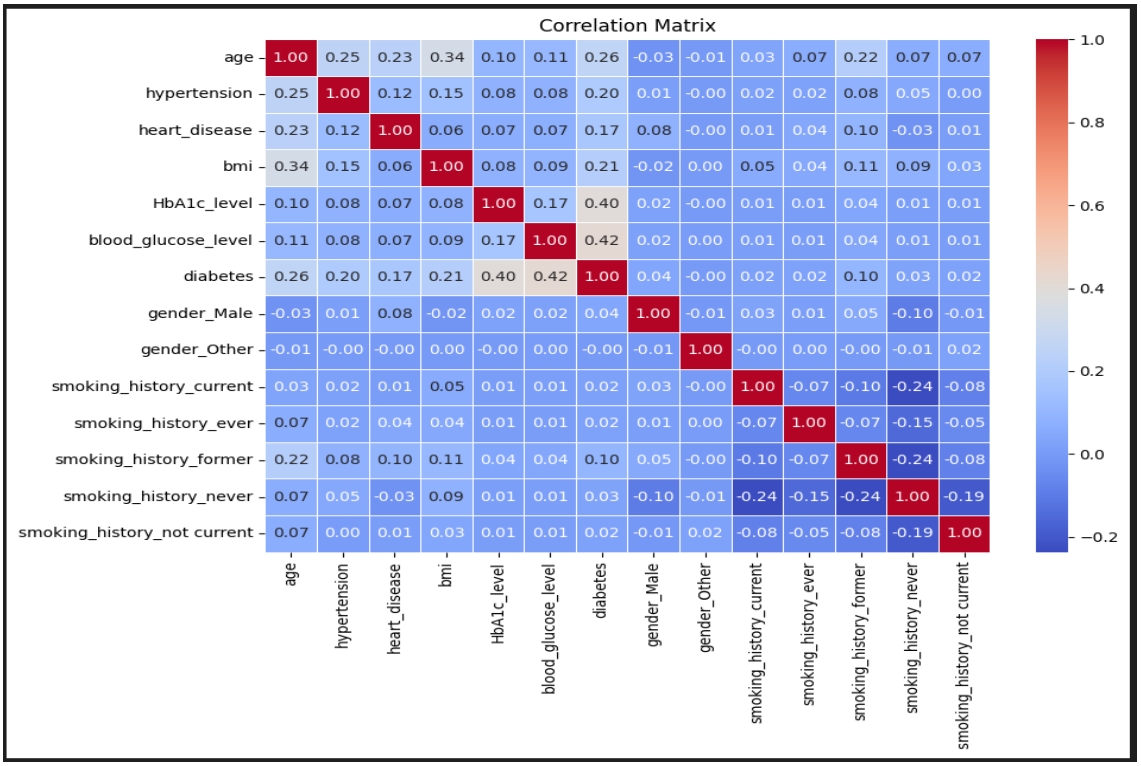
The histograms show graphically the number of occurrences of numerical data like age, BMI, HbA1c, and blood glucose levels. The age attribute has a maximum at 50 years, signifying that many of the people would be in middle age. The BMI shows a high peak of 30 that indicates

more people are overweight or obese. The concentration of HbA1c level is maximized at 6 with some elevated measurements. The concentration of blood glucose has been highly concentrated between 100 and 150 mg/dL. The distribution of diabetes is very skewed, as the cases of non-diabetes are bigger in number.

Correlation and Feature Relationships

Figure 2

Correlation Matrix of Numerical and Categorical Features in the Diabetes Dataset.



Note. The heatmap visualizes the correlation between features in the dataset. Stronger correlations are indicated by red, and weaker correlations are shown in blue. Notably, blood glucose and HbA1c levels have a significant positive correlation, and diabetes is moderately correlated with age and BMI.

The above heatmap shows the pairwise feature correlations. HbA1c level also has a strong positive relationship with blood glucose level of 0.42, which shows that an increase in blood glucose level implies an increase in HbA1c level. The correlations between age and BMI with diabetes are also positive but moderate (0.26 and 0.21, respectively), which implies that diabetes is more likely to occur in older people and in people who have higher BMI. Other features have looser ties to other features and the target variable.

Data Preprocessing

Explanation of Data Collection and Dataset Characteristics

The dataset utilized in this study was an open public one related to the prediction of diabetes. It contains 100,000 rows and 8 columns featuring age, gender, the presence of hypertension, heart disease, body mass index (BMI), level of HbA1c, level of blood glucose, and history of smoking. The target variable is the presence or absence of diabetes (1) or not (0). The features act as predictors when determining the existence or the nonexistence of diabetes.

Handling Missing Values, Outliers, and Data Normalization

To address the incompleteness, mean imputation was implemented on numerical characteristics. In the case of categorical variables, missing values were replaced by mode. The outliers were determined by the IQR method, such as capping or removing, depending on the influence of the outliers on the model. To standardize the numerical variable, the StandardScaler was used, and to adjust for the sensitivity of the models to feature scaling, such as logistic regression and KNN, we required all features to have a mean of 0 and a standard deviation of 1.

Feature Engineering Techniques

The categorical variables gender and smoking history were transformed by one-hot encoding (using pd.get_dummies) into binary form. Take the example of the gender variable,

which was transformed to two columns; one was for Male, and the other was Other (without Female as the reference group). Smoking history also underwent such a transformation and resulted in a binary feature such as current smoker, former smoker, and never smoker. These ones enable the model to work with categorical variables well in the training process.

Overview of Machine Learning Models

Logistic Regression

Table 2

Evaluation Metrics for Logistic Regression and Random Forest Models

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.9590	0.8646	0.6171	0.7202	0.8040
Random Forest	0.9699	0.9467	0.6868	0.7961	0.8416

Note. Logistic Regression and Random Forest evaluation metrics are based on their performance in predicting diabetes using the provided dataset.

Logistic regression is an efficient but easy model of binary classification. It avails the interpretability of estimating the probability of an outcome on the basis of input features. In this diabetic prediction model, the logistic regression had a performance accuracy of 95.90%, a precision of 86.46%, and a recall of 61.71%, showing that although the model performs sufficiently in distinguishing between the non-diabetic instances, it does not perform well when

it comes to recognizing the diabetic instances. F1 was 0.7202, and ROC AUC was 0.8040, which displays a satisfactory result.

Feature Importance (Random Forest)

Table 3

Feature Importance (Random Forest)

Feature	Importance
HbA1c_level	0.397138
blood_glucose_level	0.329611
bmi	0.121995
age	0.100404
hypertension	0.014600
heart_disease	0.010685
gender_Male	0.007010
smoking_history_never	0.005169
smoking_history_former	0.004357
smoking_history_current	0.003233
smoking_history_not current	0.003071
smoking_history_ever	0.002724
gender_Other	0.000003

Note. Feature importance values from the Random Forest model show the contribution of each feature to the diabetes prediction task. Higher values indicate greater importance in the model's decision-making process.

The Random Forest algorithm also gives good clues as to which features are the most influential following the predictions of diabetes. The level of HbA1c and the level of blood glucose have the most significance, having the relative importance of 0.3971 and 0.3296,

respectively, and thus are the most relevant ones in predicting diabetes. Other significant characteristics are the BMI (0.1220) and age (0.1004) that also significantly contribute towards the decisions made by the model. Such features as heart disease (0.0107) and gender (0.0000) are of very low importance as compared to the others.

XGBoost and KNN

Table 4

Evaluation Metrics for XGBoost and K-Nearest Neighbors (KNN)

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
XGBoost	0.9717	0.9568	0.6996	0.8083	0.8483
K-Nearest Neighbors (KNN)	0.9607	0.8960	0.6107	0.7263	0.8020

Note. Evaluation metrics for XGBoost and K-Nearest Neighbors (KNN) models, showcasing their performance on predicting diabetes. The XGBoost model outperforms KNN, particularly in accuracy and ROC AUC.

XGBoost is a powerful model that runs on gradient boosting, and it is also great with imbalanced data. XGBoost presented an accuracy of 97.17 percent, a precision of 95.68 percent, and a recall of 69.96 percent in the given diabetes prediction problem. This showed a good precision versus recall balance F1 of 0.8083. The ROC AUC value of 0.8483 also shows that it is able to, in a strong way, separate the classes compared to having simpler models such as the logistic regression.

Description of Model Evaluation Metrics

Accuracy is among the most frequently used evaluation metrics, as it assesses the percentage of correct predictions conducted by a model (Owusu-Adjei et al., 2023). It is defined as the fraction of correct predictions and the total number of predictions. Although accuracy gives a rough measure of the model performance, it is a wrong measure in the imbalanced datasets, wherein a type of class might dominate the estimates. On the other hand, precision is concerned with the accuracy of positive predictions by the model. It is calculated as the proportion of the true positive predictions to the number of the true positives plus false predictions. When the high cost comes in the form of false positives, the exactitude is essential, as in the case of diagnosing a disease like diabetes, where a false classification of someone healthy as diabetic would result in him or her receiving unwarranted care. Recall, also referred to as sensitivity, is defined as the ratio of the true positives to the actual positives. It should be considered where it is more costly to miss a positive (false negative) than a false positive, e.g., medical diagnosis where there are serious consequences due to failure to identify a diabetic person. The F1 score is the harmonic mean of precision and recall, and it provides a balanced value of both when false positives and false negatives are of interest. Lastly, the AUC-ROC curve measures the discrimination of a model in terms of classes, where a high value (2 is near-perfect) denotes higher performance, particularly when dealing with imbalanced data.

Justification for Algorithm Selection

The four algorithms (logistic regression, random forest, XGBoost, and K-nearest neighbor (KNN)) selected to undertake this test of prediction of diabetes were arrived at based on their capability of handling the profile of the dataset, as well as their effectiveness in binary classification. Logistic regression was selected because of its ease of interpretation and the

ability to understand the relationship between features and diabetes; this is of importance in the medical field. It chose Random Forest because it is robust, accurate, and purchasable in a linear and non-linear relationship, besides its ability to assign feature relevance, which is useful in the determination of the most influential factors in the prediction of diabetes (Afolabi et al., 2024). It was decided that the model should be XGBoost because it performs best when facing an imbalanced dataset, e.g., the population of non-diabetic people tends to exceed the numbers of diabetic ones. Due to its high accuracy and speed, it is good with large datasets. At last, a simple baseline model, KNN, was also added as the possibility to compare a model's performance with more complex models and see how the developed model complexity affects the prediction accuracy.

Discussion of Hyperparameter Tuning and Cross-Validation

GridSearchCV was used to tune the hyperparameter, where an exhaustive search is undertaken on the hyperparameter values provided to determine the optimal combination that can enhance model performance. The best hyperparameters of Random Forest were max_depth=10 and n_estimators=100, which means that the depth of trees needs to be limited and the number of trees should be set to 100. And the optimal hyperparameter in K-Nearest Neighbors (KNN) is n_neighbors=7 since this is the value that balances the complexity of the model and its performance that considers 7 nearest neighbors to classify. K-fold cross-validation guarantees that the performance of the model is tested against various subsets of the data, and the results would be robust and generalized, and issues of overfitting are minimized, and also the reliability of the model is improved with the results obtained.

Results and Discussion

Summary of Performance Metrics

The performance of each model based on their evaluation results is impressive, with Random Forest ranking ahead, as it achieved an accuracy of 96.99% and a precision of 94.67, and XGBoost almost followed with an accuracy of 97.17% and a precision of 95.68. Both models have high recall scores, which are 69.96 percent and 68.68 percent, respectively, and thus demonstrate that they are effective in identifying positive cases of diabetics. Logistic regression with the accuracy of 95.90 also proved to be very accurate but had a lower recall, 61.71, and thus, it was not as good at identifying diabetic patients as the ensemble models. KNN was a good baseline, where it achieved a 96.07% rate of accuracy but with a low rate of recall of 61.07 and a rate of precision of 89.60.

Analysis of Strengths and Weaknesses

Random forest and XGBoost are good models at a high level in terms of accuracy, precision, and recall, but XGBoost shone when working with imbalanced data. Logistic regression is easy to understand and interpret, and its low recall rate does not seem to fit well for the purpose of medical diseases when one wants to detect all the positive cases. KNN is simple to apply but not good on huge data and complex separation boundaries. It also has a low recall score, which means that it is not as reliable for making healthcare predictions and that it is also worse with diabetic patients.

Comparison of Models

XGBoost and Random Forest perform better than Logistic Regression and KNN, both according to accuracy and recall as they were directly compared, which makes the former more suitable as an approach to this diabetes predictive forecast. XGBoost takes the first step with a

bit better AUC-ROC and accuracy, which means better performance. The upside of Random Forest is that it gives feature importance, which comes in handy when it comes to comprehending the influence of each feature. But for logistic regression and KNN, there are simpler models that are likely to be more explainable but cannot be as predictive as the use of the ensemble methods.

Conclusion and Future Directions

The findings show that XGBoost and Random Forest models are the best predictive models of diabetes since they give high values in accuracy, precision, and recall, and XGBoost performs better in overall predictive power. Such models are best suited for the handling of imbalance data and provide good performance. Logistic regression and KNN can be applied in low complex tasks yet cannot be utilized in the detection of all cases of diabetes which is a prerequisite in the performance of healthcare.

Further improvement of the model performance can be made by considering other features such as genetic information or lifestyle parameters to model observation with greater accuracy. It may be considered to use cutting-edge algorithms that can process large and complex information, including deep learning. Additionally, the entire set of hyperparameters can be optimized further, and a more sophisticated technique should also be employed, e.g., an ensemble learning or feature selection, which could be helpful in terms of optimization of its performance. Regarding the idea of cross-validation by stratified samples, it would also offer leads on the aspects of correcting the class imbalances in being more accurate on the class predictions.

References

Afolabi, S., Nurudeen Ajadi, Jimoh, A., & Ibrahim Adenekan. (2024). Predicting Diabetes Using Supervised Machine Learning Algorithms. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-4527374/v1

Basith, A., Hashim, M. J., King, J. K., Govender, R. D., Mustafa, H., & Juma Al Kaabi. (2019). Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. *Journal of Epidemiology and Global Health*, *10*(1), 107–107. https://doi.org/10.2991/jegh.k.191028.001

Olalekan Kehinde. (2025). Machine Learning in Predictive Modelling: Addressing Chronic Disease Management through Optimized... *International Journal of Research Publication and Reviews*, *6*(1), 1525–1539. https://www.researchgate.net/publication/388123356_Machine_Learning_in_Predictive_Modelling_Addressing_Chronic_Disease_Management_through_Optimized_Healthcare_Processes

Owusu-Adjei, M., Ben Hayfron-Acquah, J., Frimpong, T., & Abdul-Salaam, G. (2023). Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*, *2*(11), e0000290. https://doi.org/10.1371/journal.pdig.0000290

World. (2024, November 14). *Diabetes*. Who.int; World Health Organization: WHO. https://www.who.int/news-room/fact-sheets/detail/diabetes

Appendix

Python based Jupyter Notebook Code Details

GitHub Repo url: https://github.com/ramindersinghusd/aai-501-in2-project-group7