Diabetes Prediction

Michael Travers
Dept. of Computer Science & Information Technology
University of the District of Columbia
Washington DC, 20008
michael.travers@udc.edu

Abstract

Diabetes is one of the most prevalent chronic diseases in the United States, affecting millions, and is a financial burden on the health care system. Early diagnosis is crucial for enabling timely lifestyle modifications and medical interventions to reduce severe complications for not just diabetes but also heart disease, vision loss, and kidney failure. This research aims to determine the most effective classification method, contributing to improved predictive analytics in healthcare. The models to be compared include Support Vector Machine (SVM), Decision Trees, Linear Regression, Random Forest Trees, and KNN, which detect complex patterns through layered computation. This study utilizes a Kaggle dataset from 2025, comprising 100000 participants and 12 health-related features, including chronic conditions and preventive measures, to compare various machine learning models for diabetes prediction.

I. Introduction

Diabetes refers to a group of conditions characterized by a high level of blood glucose/blood sugar. As carbohydrates break down into glucose, that is carried by the bloodstream to various organs in the body. When your pancreas doesn't make insulin or responds to the effects of insulin, diabetes can develop. Insulin is a hormone produced by beta cells of the pancreas and is necessary for glucose intake by the target cells. Insulin binds to its receptors on target cells and induces glucose uptake.

Diabetes causes multiple health issues, including blurred vision, being tired or weak, seizures, confusion, trouble breathing, being unresponsive, and many more. Diabetes can be broken down into 2 types: Type 1 and Type 2. Type 1, beta cells of the pancreas are destroyed by the immune system by mistake. The reasons are not clear, but the genetic factors play a major role. Insulin production is reduced, and less insulin binds to its receptor on target cells, and less glucose is taken into the cells; more glucose stays in the blood. It develops at an early ag,e under the age of 20 and is managed with insulin injection or insulin-dependent [1]

The goal of this research is to develop and compare machine learning models to determine the most accurate and efficient classification method for diabetes prediction using a large-scale health dataset. Analyze the impact of key health-related features on diabetes occurrence by leveraging predictive modeling techniques, aiding in early diagnosis and preventive healthcare strategies. And enhance public health decision-making by identifying the most effective machine learning approach for risk assessment, contributing to improved diabetes management and resource allocation.[1]

II. METHODOLOGY

For predicting diabetes using machine learning, several steps must be performed to get a accurate prediction. Prepressing and featured selection will be applied to the kaggle dataset including data cleaning and preparation by handling missing or null values, normalization, encoding of categorical values. The featured selection will be used to identify the most important health-related factors that highly influence diabetes prediction.

Model implementation and training is next, including Support Vector Machine(SVM), Decision Trees, Linear Regression, and Random Forest. These models will be trained using the preprocessed data with cross-validation to be performed to optimize performance. This process will ensure that each model is adjusted to the best configuration, maximizing prediction accuracy.

Lastly, each model will be evaluated and compared. The trained models will be assessed using the performance metrics including accuracy, recall, and precision. This comparative analysis will help identify the most effective model for diabetes classification, providing important insights into which algorithms are best suited for predicting diabetes and possible informing the future of health care.

III. RELATED WORK

As diabetes is one of the most prominent diseases, multiple studies have been performed to help find the best prediction model that can identify and reduce risk factors that cause diabetes. In December 2021, Simon Foo and Jobeda Khanam conducted a study using the Pima India Diabetes dataset with 768 patients, using 7 machine learning models to predict diabetes. Their conclusion was the Logistic Regression and Support Vector Machine worked the best with diabetes prediction.[3]

Another study was conducted in October 2023 by Ashikur Rahman and Imran Mahmud from the Daffodil International University. The study aimed to build an automated machine learning model to help predict diabetes at an early stage. 6 classification models were used, and resulted in Random Forest outperformed the other models. This study also found the highest risk factors for diabetes, including polyuria, polydipsia, and delayed healing.[4]

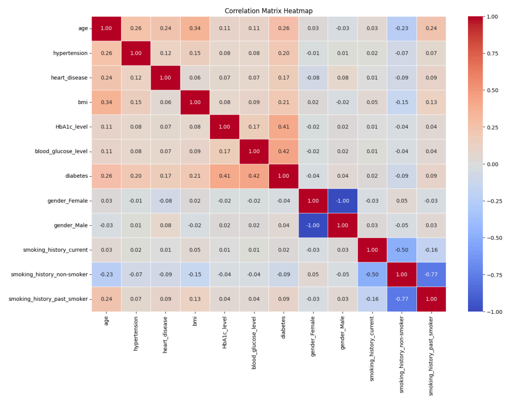
A study that had similar results is from Fadoua Kebraoui and El Mokhtar En-Naimi in November 2023. This study used the same dataset as Simon Foo and Jobeda Khanam, being the Pima India Diabetes Database. Their study compared the performance of 5 different machine learning models. Their results showed that the Random Forest was the most effective model for predicting diabetes[5].

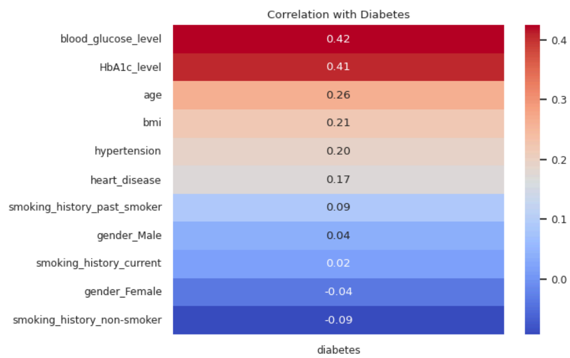
IV. PRE-PROCESSSING

A.Dataset

This dataset was obtained through a Kaggle that contains data for diabetes patients. The file for this dataset provides 100000 records for medical and demographic dat,a along with diabetes status, whether positive or negative. Consisting of numerous features including age, gender, body mass index(BMI), hypertension, heart disease, smoking, HbA1c level, and blood glucose levels. This dataset is used in the study to decide which classification model is best for predicting the likelihood of being diabetic or non-diabetic[2]

To help visualize the correlation between diabetes and the attributes contributing to diabetes, a Correlation Matrix can be used. This matrix shows the pairwise correlation coefficients between the variables in the diabetes dataset. Each cell represents the strength and direction of the relationship between two features, ranging from -1 being a strong negative correlation, 1 having a strong positive correlation ,and 0 indicating no linear relationship. This correlation matrix will help identify which health-related factors are closely related to diabetes. This matrix shows that blood_glucose and HbA1c_levels are the major factors that correlate with diabetes





B. Data Cleaning

One of the most critical steps for preparing data is data cleaning to ensure that the prediction models are accurate and reliable. The first step performed in data cleaning is handling duplicates or recurring data in each column and irrelevant entries. Removing unnecessary values is next, and in this dataset, “other” will be dropped for gender as this study only focuses on male and female. Null values will also be removed.

Normalization will be performed to scale numerical features to a standard range of 0 to 1. First separate the target value being Diabetes, and then create a new dataframe with all the other input features. Min-Max normalization will then be performed. This ensures a fair comparison among features when model training.

Categorical Encoding is used to convert categorical variables using One-Hot Encoding to make features machine-readable and suitable for modeling. One-Hot encoding turns a categorical column into multiple columns to represent a possible category. In this dataset, “gender” and “smoking” will be used.

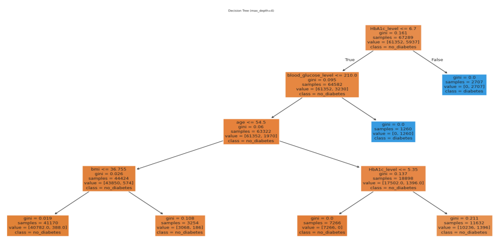
Recategorization is another step in data cleaning. In this dataset, “smoking_history” has multiple categories or inputs. These categories can be simplified into simpler groups. Converting “never” and “no Info” into Non-Smoker and “ever”, “former” and “not current” into past smoker.

Test and Training will be the last step for data cleaning to ensure the evaluation to the model performance. The data set will be split into 80 percent training and 20 percent testing, preventing overfitting and providing a more realistic estimate of accuracy. X being the input variables such as “age”, “smoker”, “BMI,” and Y will be the target value being “Diabetes”.

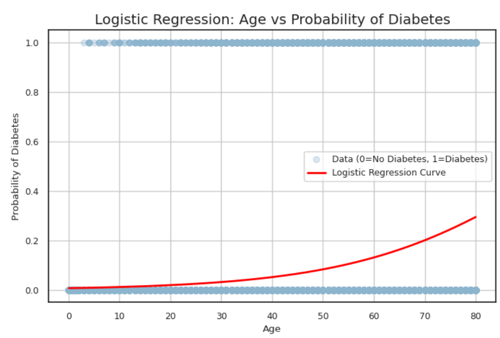
C. Models

When creating prediction models for diabetes, choosing the correct algorithms is essential for capturing patterns. Support Vector Machines(SVM) was a candidate as they can find the optimal hyperplane that best separates diabetic and non-diabetic individuals. SVM maximizes the classification margin which is the distance between the separating hyperplane and the nearest data points or support vector. This approach reduces the chances of misclassification and is effective when the data can be separated linearly.

Decision Trees is another option as it organizes data through a hierarchical tree structure, having each internal node representing a feature, and each branch is a decision rule, and each leaf is a final classification being diabetic or non-diabetic. Decision Trees are highly interpretable as the resulting path gives a clear step-by-step way a prediction is made. Especially useful in medical contexts, where transparency and interpretability are important.



Linear Regression is also going to be used for relationships between input features and diabetic occurrence. This algorithm fits a linear straight line through data points, capturing the direction and strength of the relationship between features like “heart_disease”, “age”, “smoker”, and the target variable being the diabetic status. Linear Regression can serve as a baseline for understanding the influence of individual features on diabetic risk.



KNeighborClassifier(KNN) is a distance-based algorithm that predicts the possibilities of diabetes. By examining the ‘k’ value being the most similar individuals, it assigns a new person or patient to the most common class among its nearest neighbors. The downfall of using KNN is its case sensitivity to the choice of ‘k’ and scale of the inputs, requiring careful configuration and preprocessing.

The last model to be used is the Random Forest Tree. This is an extension of the decision tree by combining multiple decision trees to form a forest of classifiers. Each tree will be trained on a random subset of data, and its outputs are combined to produce a prediction. Random Forest Trees reduce the risk of overfitting and captures complex, non-linear relationships between features, like the interaction between insulin, glucose,and hypertension.

IV. EXPERIMENTAL RESULTS

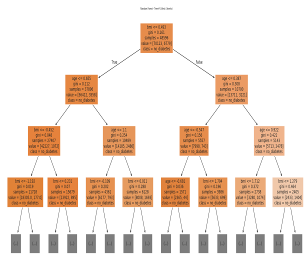
After the dataset has been preprocessed, each model will be evaluated for performance, including the accuracy, precision, and recall. KNeighborClassifier was first tested resulting in a strong overall performance with a high accuracy of 0.95 and precision of 0.87, meaning it correctly classifies most instances and reliably predicts diabetes. On the negative side, the recall was low at 0.53, meaning that it fails to identify many true diabetic cases, possibly overlooking a significant portion of at-risk patients. This imbalance shows that while the model is precise, it struggles to catch positive cases.

Support Vector Classifier(SVC) resulted in a perfect precision of 1.0 and an accuracy of 0.95, showing its ability to confidently and correctly classify non-diabetic individuals. But as precision was perfect, the recall was at a low 0.38, indicating a significant number of cases are missed. This low recall shows a high rate of false negatives limits the model's effectiveness in practical healthcare, where identifying all potential diabetic patients is critical.

Logistic Regression had a balanced but weaker performance with an accuracy of 0.93 and a precision of 0.83. This indicates that it is effective at correctly identifying non-diabetic patients and making reliable positive predictions. But the low recall being 0.39 shows that it struggles to detect a large portion of actual diabetic cases, possibly leading to underdiagnosis.

Random Forest Tree is popular for its robustness and versatility, effectively classifying non-diabetic individuals with a high precision of 0.94 and an accuracy of 0.96. The recall is 0.66, indicating that it still missed 34 percent of diabetic cases. The Decision Tree is the final model tested.

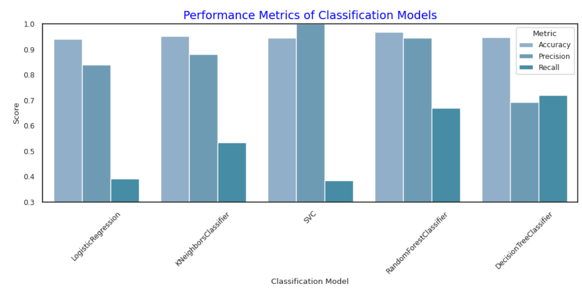
The Decision Tree was able to correctly predict most diabetes cases with an accuracy of 0.94, but it struggles to fully capture and confirm diabetic cases with a precision of 0.69 and a recall of 0.72.



V. CONCLUSION

In conclusion to this research, Random Forest Trees is the best model to use when using classification models to test prediction for Diabetes. It resulted in the most balanced at catching both diabetic and non-diabetic cases, scoring high in accuracy, precision, and recall metrics. Thai miminizes false positives and false negatives, providing the best overall prediction. Decision Tree would be the next best optio,n while its slightly less precise than Random Forest, it has a high recall, making it effective at identifying true diabetic cases, but has a slight increase in false positives.

Logistic Regression is the worst as it has a high precision, making it effective at avoiding false positives, but the low recall means that it misses many true diabetic cases. K-Nearest Neighbors had a more balanced approach, with a better recall than Logistic Regression, allowing it to catch more true cases. Support Vector Machine(SVM) is better than KNN as this model achieved a perfect precision of 1.,0 meaning every positive prediction was correct, but the low recall meant a large number of cases were missed.



	Accuracy	Precision	Recall
LogisticRegression	0.939561	0.839849	0.391534
KNeighborsClassifier	0.952356	0.879961	0.534392
SVC	0.945542	1.000000	0.384480
RandomForestClassifier	0.967336	0.945228	0.669606
DecisionTreeClassifier	0.947051	0.693265	0.720165

Reference

[1] American Red Cross, "Diabetic Emergencies," May 9, 2025.
https://www.redcross.org/take-a-class/resources/learn-first-aid/diabetic-emergencies

[2] F. Anwar, "Diabetes Prediction," Kaggle,May 9, 2025.
https://www.kaggle.com/code/fareedalianwar/diabetes-prediction/input

[3] J. Smith and R. Johnson, "Diabetes and its Management," Sci. Direct, vol. 7, no. 2, pp. 215-223, 2021.
https://www.sciencedirect.com/science/article/pii/S2405959521000205

[4] A. Gupta and M. Kumar, "Machine Learning Based Approach for Predicting Diabetes Employing Socio-demographic Characteristics," ResearchGate, 2025.
https://www.researchgate.net/publication/374449308_Machine_Learning_Based_Approach_for_Predicting_Diabetes_Employing_Socio-demographic_Characteristics

[5] Y. Chen, "Deep Learning Approaches for Diabetes Prediction," in Proc. 2025 ACM Int. Conf. on Artificial Intelligence, pp. 178-185, 2025.
https://dl.acm.org/doi/10.1145/3607720.3607764