A Comparative Study of Machine Learning Models for Diabetes Prediction

Paulo Mendoza
Computer Engineering Student
Technological Institute of the Philippines – Quezon City
Bulacan, Philippines
qpdcmendoza@tip.edu.ph

Benedick Labbao
Computer Engineering Student
Technological Institute of the Philippines – Quezon City
Rizal, Philippines
qbdlabbao@tip.edu.ph

Charlz Regala
Computer Engineering Student
Technological Institute of the Philippines – Quezon City
Metro Manila, Philippines
qcaregala@tip.edu.ph

Mark Renier Mercado
Computer Engineering Student
Technological Institute of the Philippines – Quezon City
La Union, Philippines
qmrrmercado@tip.edu.ph

***Abstract*—Diabetes is a prevalent and chronic metabolic disorder affecting millions of individuals worldwide. Early and accurate prediction of diabetes plays a pivotal role in preventing and managing this condition. Machine learning models have emerged as valuable tools for predictive healthcare applications. This study conducts a comparative analysis of various machine learning models to assess their effectiveness in predicting diabetes.**

The research employs a diverse dataset comprising demographic, clinical, and lifestyle attributes to train and evaluate a range of machine learning algorithms, including logistic regression, decision trees, random forests, support vector machines, and neural networks. These models are rigorously evaluated based on performance metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). Additionally, feature importance is examined to gain insights into the key factors influencing diabetes prediction.

Results reveal the varying predictive performance of different machine learning models, shedding light on their strengths and weaknesses in diabetes prediction. This comparative analysis provides valuable insights for healthcare practitioners and researchers seeking to implement effective predictive models for diabetes diagnosis. The findings contribute to the ongoing efforts to enhance early diabetes detection and management, ultimately improving the quality of healthcare in the context of this pervasive and significant health issue.

Keywords—Diabetes prediction, machine learning, comparative analysis, predictive modeling, healthcare, feature importance, chronic disease management.

I. Introduction

Diabetes is one kind of disease that occurs when the blood glucose/blood sugar level in the human body is very high. According to medical professionals, diabetes develops when the pancreas, a gland located in the human body, is unable to create enough insulin (Type 1 diabetes) or when the body's cells are unable to utilize the insulin that is produced (Type 2 diabetes).

When we eat food, after the digestion process, glucose gets released. Insulin is a blood hormone that moves from blood to cells and instructs cells to consume blood glucose and transform it into energy. When the pancreas cannot produce enough insulin, the cells cannot absorb glucose, and the glucose remains in the blood. Hence the blood glucose/blood sugar increases in the blood at a very unacceptable level. The healthcare sector gathers a vast amount of data, including hospital records, patient medical information, and examination findings. A doctor's experience and knowledge are used to examine the disease's forecast for early disease diagnosis, although this analysis is prone to error.

The manual judgments can therefore be concerning. The data's concealed pattern may go undiscovered, which could have an effect on how decisions are made. Consequently, patients are denied the necessary care. The early detection of diabetes requires automated identification with higher accuracy.

II. Review of related Literature

- Reference [1] conducted research about classification and comparison study of supervised machine learning algorithms. The study recommended thorough fine tuning of the hyperparameters and a sizable number of instances for the dataset.
- Reference [2] presented diabetes prediction using machine learning techniques, they use machine learning and ensemble techniques to predict diabetes. The accuracy result shows Random Forest has achieved the highest accuracy.
- Reference [3] performed a study about predictive models and feature importance for early detection of diabetes. This study shows that Age influences the predictive strength of the models towards predicting the test data.

III. METHODS

A. Design

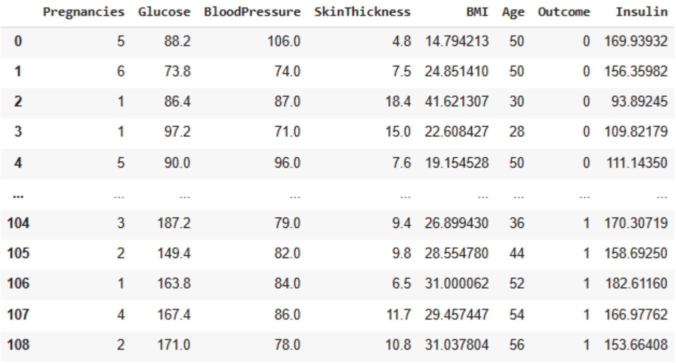
The primary goal of this research paper is to perform a comparative analysis of different machine learning models on predicting diabetes. This includes assessing various metrics scores used in evaluating machine learning models.

H0: The machine learning models will have little to no differences in performance metrics.

Ha: The machine learning models will have significant differences in performance metrics.

B. Data Collection and Preprocessing

The Dataset we gathered was obtained from 203 female individuals of Rownak Textile Mills Ltd, Dhaka, Bangladesh. Fig. 1 shows that the data consisted of pregnancy, glucose, blood pressure, skin thickness, BMI, age, insulin, and outcome of diabetes. Data was split into 2 subsets used for training, and testing.



	Pregnancies	Glucose	BloodPressure	SkinThickness	BMI	Age	Outcome	Insulin
0	5	88.2	106.0	4.8	14.794213	50	0	169.93932
1	6	73.8	74.0	7.5	24.851410	50	0	156.35982
2	1	86.4	87.0	18.4	41.621307	30	0	93.89245
3	1	97.2	71.0	15.0	22.608427	28	0	109.82179
4	5	90.0	96.0	7.6	19.154528	50	0	111.14350
...	...	...	...	...	...	...	...	...
104	3	187.2	79.0	9.4	26.899430	36	1	170.30719
105	2	149.4	82.0	9.8	28.554780	44	1	158.69250
106	1	163.8	84.0	6.5	31.000062	52	1	182.61160
107	4	167.4	86.0	11.7	29.457447	54	1	166.97762
108	2	171.0	78.0	10.8	31.037804	56	1	153.66408

Fig. 1. Dataset Preview

C. Machine Learning Algorithms

This study focused on the most used supervised machine learning algorithms for classification to see the performance of these methods. These algorithms are:

1) Logistic regression (LR)

a) Use Case: is commonly used for binary classification problems, where the outcome variable is categorical with two classes.

b) Model: Despite its name, logistic regression is used for classification, not regression. It models the probability that a given instance belongs to a particular class using the logistic function.

2) K-nearest Neighbor (KNN)

a) Use Case: is used for classification and regression tasks. In classification, it assigns a data point to the most common class among its k-nearest neighbors.

b) Model: it memorizes the training dataset and makes predictions based on the majority class or average value of its k-nearest neighbors in the feature space.

3) Support Vector Machine (SVM)

a) Use Case: is particularly effective in high-dimensional spaces and is often used for image classification, text classification, and bioinformatics.

b) Model: It aims to find the optimal hyperplane that maximally separates data points belonging to different classes.

D. Hyperparameter Tuning Techniques

Grid Search and cross-validation will be used in optimizing the machine learning hyperparameters.

1) Grid Search: is a technique used to find the optimal hyperparamaters by combining every possible combination of hyperparameters and selects the best value.

2) K-fold Cross-validation: is a technique that splits the input data into k subsets, and taking a subset to be used as testing data and the rest of the data as training data.

E. Performance Metrics

The performance metrics are used to evaluate a machine learning model to determine its performance. We will be using a combination of accuracy score, ROC Curve and ROC-AUC Score.

1) Accuracy score: score that shows the number of true positive predictions in relation to the total number of predictions.

2) Receiver Operating Characteristic (ROC) Curve: Curve that displays the performance of a model based on the rate of true positives in relation with false positives.

3) Receiver Operating Characteristic - Area Under Curve (ROC-AUC) Score: This metric shows the total area under the ROC Curve.

F. Procedure

- Using NumPy and Pandas library in Python, we performed the data collection and preprocessing.
- We used scikit-learn library in Python to create each machine learning model.
- We performed Grid Search from the scikit-learn library to get the best hyperparameter for each model.
- Using the model that was provided by Grid Search which contains the best hyperparameters, we fitted them using the training data.
- Each model was then assessed and evaluated using the performance metrics which can be found in the scikit-learn library.

IV. RESULTS

A. Logistic Regression

Logistic regression has achieved a score of 91% in accuracy and 85% in ROC-AUC score.

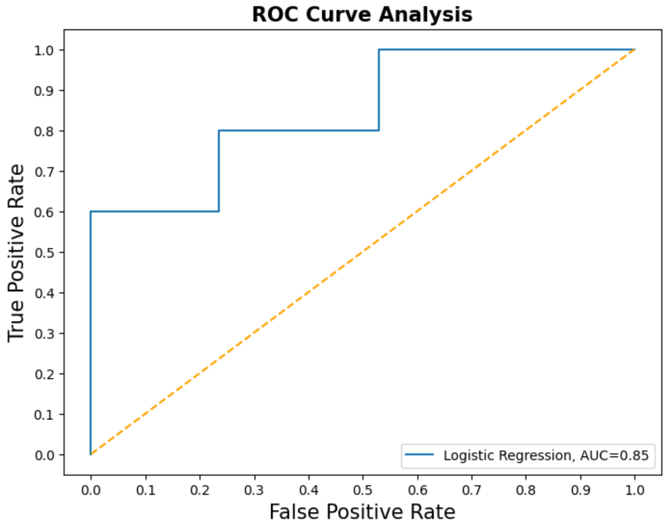


Fig. 2. ROC Curve – Logistic Regression

B. *K-Nearest Neighbor*

K-Nearest Neighbor had an accuracy score of 77% before it was tuned using hyperparameter tuning technique. In Fig. 3, it shows the result of the hyperparameter tuning that was done using the number of neighbors(K) as the hyperparameter for the model to achieve a new accuracy score of 91%.

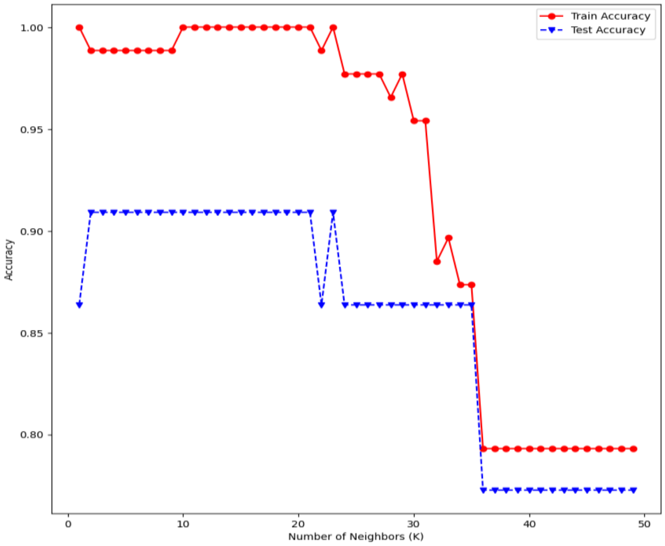


Fig. 3. Accuracy vs. Number of neighbors(K)

In Fig.4, it shows the ROC curve of the KNN model, the ROC-AUC score that was calculated from the curve was 80%.

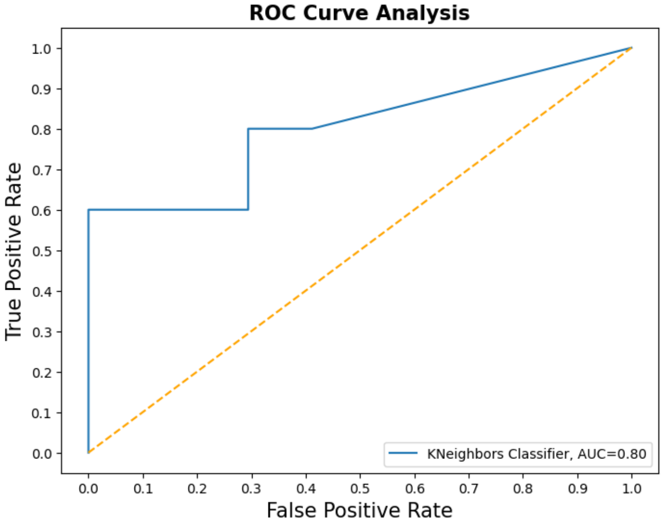


Fig. 4. ROC Curve – K-Nearest Neighbor

C. *Support Vector Machine*

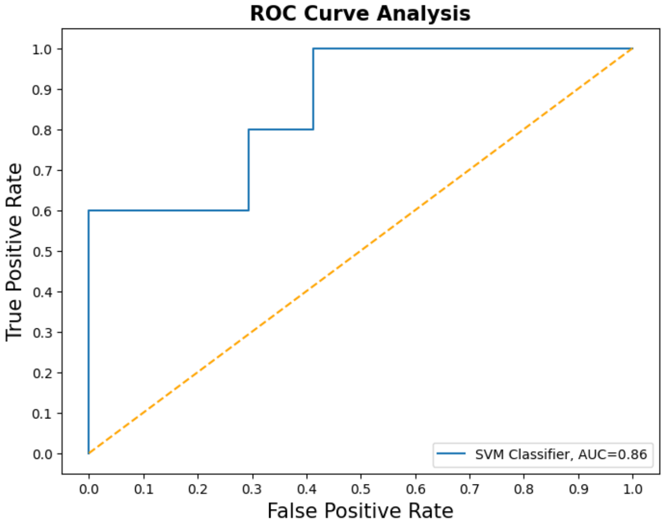
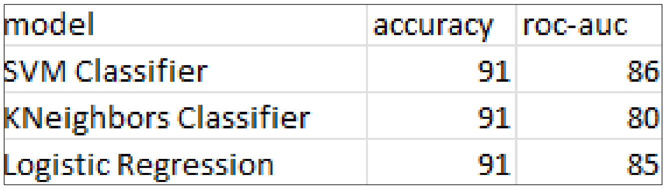


Fig. 5. ROC Curve – Support Vector Machine

D. *Side by Side Comparison*

In Table I, all classifiers have managed to achieve 91% accuracy score, this is where having more than one performance metrics matters. In the ROC-AUC score the model that managed to achieve a high performance was SVM, the most underperforming model was KNN.

TABLE I. MODELS PERFORMANCE METRICS



model	accuracy	roc-auc
SVM Classifier	91	86
KNeighbors Classifier	91	80
Logistic Regression	91	85

The results of ROC Curve in Fig. 6 show that SVM model has the highest true positive rate in relation to how false positive rate, while KNN model can be seen to not perform well.

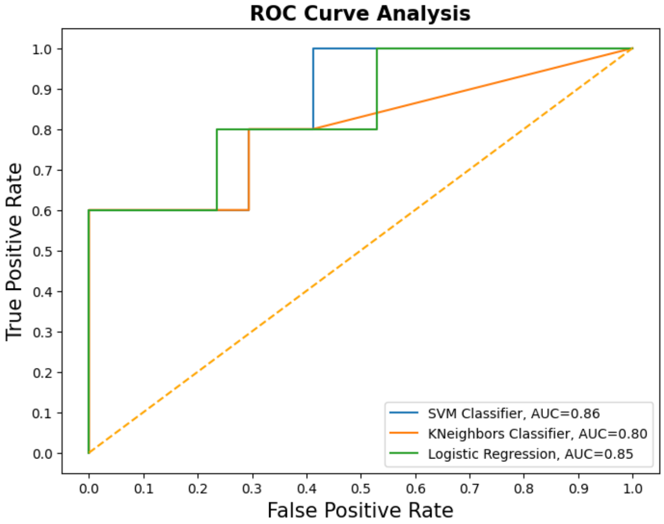


Fig. 6. ROC Curve – Support Vector Machine

All the models that we used achieved 91% in accuracy score, but that is not the only metrics that can be used to determine a model's performance, we also used ROC-AUC score as a performance metric. Of all the models we used, the SVM model had the highest ROC-AUC score.

In summary, based on the results of our experiments, the SVM model was the best-performing classifier for this study. However, these results do not consider other factors such as model complexity, computational resources, and specific requirements.

V. DISCUSSION

In 2020 diabetes is ranked fourth as the most deadly diseases in the Philippines and have killed millions around the globe.

Making a model that can predict if a person would contract diabetes is highly beneficial especially to those who might contract diabetes type II since it can be cured, and can prevent the contraction of diabetes type II for other people.

Our aim in this study is to see if there would be differences in the machine learning models that we used. In the accuracy score, the models showed no differences in each other with proves the null hypothesis, but the determining factor was the ROC-AUC score with shows differences in the score, thus, our alternative hypothesis was proven, and the null hypothesis was rejected.

ACKNOWLEDGMENT

We extend our appreciation to our instructor Engr. Valerie Peralta of the Technological Institute of the Philippines, Quezon City, for her valuable contributions to this research

REFERENCES

[1] FY, O., JET, A., Awodele, O., O, H. J., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and comparison. International Journal of Computer Trends and Technology, 48(3), 128–138.

[2] Soni, M. (2020). Diabetes Prediction using Machine Learning Techniques. IJERT.

[3] (E. Adua et al., "Predictive model and feature importance for early detection of type II diabetes mellitus," Translational Medicine Communications, vol. 6, no. 1, Aug. 2021, doi: 10.1186/s41231-021-00096-z.