Detecting Diabetes using different machine learning approaches

Mubashira Rahman
School of Data and Sciences
Brac University
Dhaka, Bangladesh
mubashira.rahman@g.bracu.ac.bd

***Abstract*—Diabetes is a significant worldwide health issue, and early detection is crucial for disease management as there is currently no permanent cure. Various machine learning algorithms have long been utilized for diabetes prediction. This study employed various machine-learning techniques to determine the optimal model for diabetes detection. This study evaluated the performance of several classification algorithms, including logistic regression, K-Nearest classifier, Support vector classifier, Random Forest classifier, and Gradient boosting classifier, to determine the optimal method for detecting diabetes. The logistic regression model achieved the highest accuracy of 82.46%.**

***Index Terms*—Diabetes detection, Logistic regression, K-Nearest Classifier, Support Vector Classifier, Random Forest, Gradient boosting.**

I. Introduction

Diabetes is a prevalent condition. Its rate of increase is accelerating. Identifying the condition can facilitate its management and improve quality of life, despite not invariably resulting in a negative outcome. Diabetes caused 1.5 million deaths in 2019, with 48% of those deaths occurring before the age of 70 and it is increasing as days go by [1]. Machine learning algorithms can predict diabetes risk by analyzing large patient health record datasets through predictive modeling. The models can generate precise projections by incorporating variables such as age, weight, family history, blood pressure, and blood sugar levels. This approach is efficient for detecting diabetes. Machine learning models are currently proposed to predict the detection of diabetes, which is a crucial task. Conventional machine learning models, such as Random Forest Tree, Decision Tree, and SVC, are commonly employed. This study employs various machine learning techniques on a dataset to determine the optimal algorithm for diabetes prediction.

II. Literature Review

Diabetes is a prevalent health issue. Machine learning algorithms have been utilized in numerous studies to detect this disease. One study aimed to develop a predictive model for diabetes. Aishwarya's team utilized various machine-learning models, including Random Forest, Decision Tree, KNN, SVC, and Naive Bayes, to address the issue. The models were applied to the dataset's BMI and blood pressure variables. They acquired 78.21% of accuracy [2] in predicting diabetes with the Random forest model. Valdehi conducted a study utilizing a larger dataset to forecast the occurrence of diabetes. KNN and SVC were utilized and their results were compared. The SVC model yielded a higher prediction accuracy of 0.79 ROC [3]. Hasan et al. addressed this issue by utilizing deep learning methodologies and techniques. A decision support system was developed using deep learning techniques such as Artificial Neural Networks, CNN, and LSTM, trained with a dataset [4]. Other research on this same work has been done and the 96.43% accuracy was predicted by the LSTM model. In this study, the researchers took age, BMI, and family data and used it in a decision-support framework [5]. To enhance the precision in decision-making algorithms and get better results than the low-accuracy traditional algorithms Sulistyawati and their team worked with Naive Bayes, Decision tree using a variety of accuracy like ROC, AUC, and confusion Matrix [6]. In another study using SVM for the dataset with age, and blood group information 79.36% accuracy was obtained. The AUC and ROC were used there as well.

For a change, a unique study with blockchain, IoT, and edge computing in the problem of diabetes mellitus prediction was used [7]. Their method gathers data from numerous IoT devices, including glucose sensors, blood pressure monitors, and activity trackers, utilizing a safe and privacy-preserving monitoring system. Similar to this study technique author Ismail used a decentralized and tamper-proof system, blockchain technology assures the confidentiality and privacy of patient data [8]. This approach gave them an accuracy of 91%, which most efficient result [9]. Based on a variety of demographic, behavioral, and medical characteristics, Debdari and his team attempted to predict the development of diabetes by building a model, they combined a number of machine learning techniques, including logistic regression, random forest, decision trees, K-nearest neighbor, and Naive Bayes [10]. In other work, to identify the most crucial factors that influence the development of diabetes, the scientists additionally employed feature selection approaches such as Recursive Feature Elimination (RFE) and Extra Trees Classifie to identify the most crucial factors that influence the development of diabetes [11], the scientists additionally employed feature selection approaches such as Recursive Feature Elimination (RFE) and Extra Trees Classifier. Ayon and Islam used deep learning methods to project the incidence of [12]. The scientists utilized a dataset [13] containing demographic, behavioral, and medical information to construct their prediction model. The model was developed utilizing deep learning techniques such as multilayer perceptron (MLP) and convolutional neural network (CNN) [14]. The MLP and CNN models outperformed other machine learning models with accuracy values of 88.7% and 89.8%, respectively [15]. Blood pressure, age, body mass index, and glucose level were the most important factors according to the RFE and PCA. Lastly, Khanam and Foo tried to find the most accurate algorithm for predicting diabetes utilizing various feature sets. According to the research, the SVM algorithm's highest accuracy was 78.9% [16]when all attributes were incorporated algorithm's algorithm's highest accuracy was 78.9% when all attributes were incorporated, according to the research. Another study findings suggested that the SVM model had the best accuracy, coming in at 77.9% [17]. The RFE and CFS determined that blood pressure, age, body mass index, and glucose level were the most crucial characteristics

III. Methodology

A. Dataset Description

The datasets comprise multiple medical variables that predict outcomes that are autonomous, and a single target variable, Outcome, that is dependent. The factors that are independent comprise the patient's number of pregnancies, BMI, insulin level, and age, among others [18]. This dataset also has the features of pregnancy, glucose, blood pressure, and skin thickness. The dataset is particularly clean and has useful data to be used for diabetes prediction.

B. Dataset Cleaning

The dataset underwent various data-cleaning procedures to prepare it for modeling. Initially, the dataset description was obtained for improved comprehension. The objective was to identify and remove null values, as well as to determine the count and data types of the available data.

C. Dataset Visualization

Various data analysis techniques were employed to examine datasets and identify patterns. Data visualization facilitates efficient comparison of data points and variables, thereby aiding in data analysis and informed decision-making.

Firstly, to find the correlation between all the variables heatmap was used to visualize it with clarity.

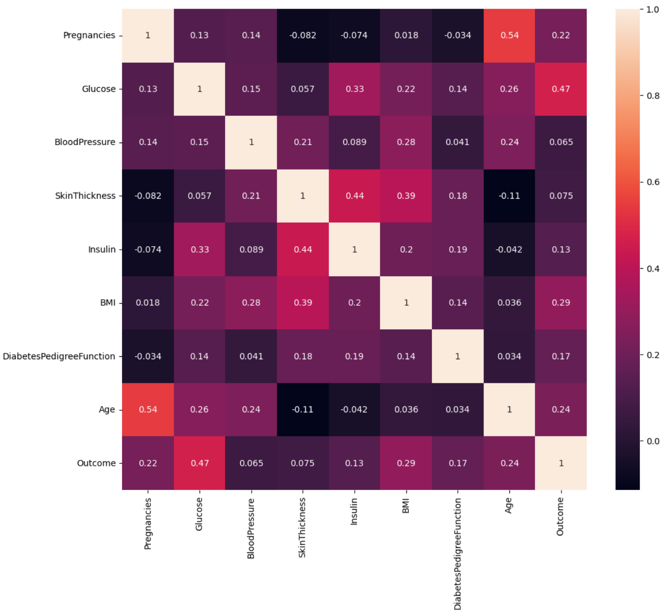


Figure 1:Correlation between all variables

A positive correlation exists between BMI and glucose levels; elevated BMI values are commonly linked to increased glucose levels. It significantly contributes to the development of diabetes by disrupting insulin function.

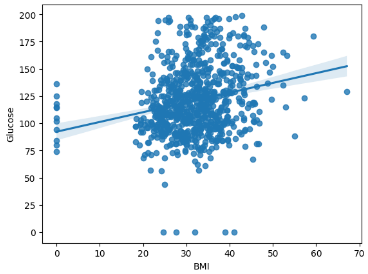


Figure 2:Correlation between BMI Glucose

Visualizing the dependent variable "Outcome" in relation to the independent variables in the dataset can enhance comprehension and facilitate decision-making.

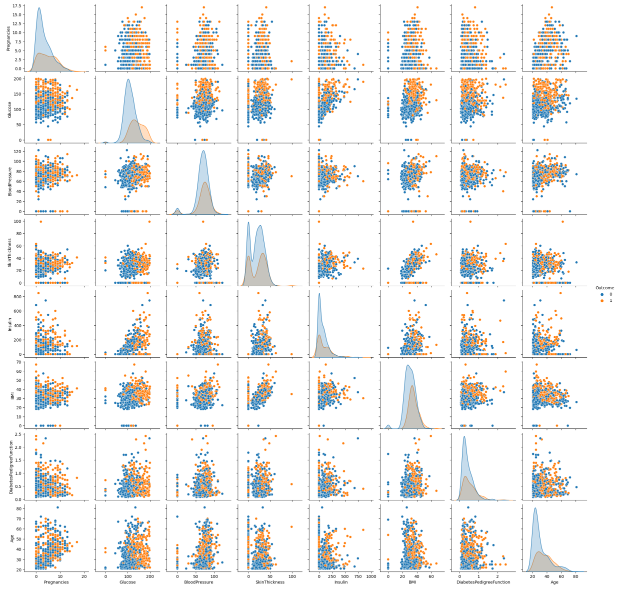


Figure 3:Correlation between all independent variable with dependent variable

IV. Models

A. Logistic regression

This statistical method analyses variables to determine the probability of diabetes presence. The method can assess various variables and offer a numerical assessment of the likelihood of illness. The model yielded an accuracy of 82.46% and a ROC AUC score of 0.76, indicating its efficacy in diabetes prediction.

B. Gradient Boosting Classifier

The learning algorithm is a robust predictive model capable of handling missing data. The tool has the ability to autonomously identify and choose the key features for classification. This model's computational efficiency yielded exceptional results when applied to the dataset. The model achieved an accuracy score of 81.81% and a ROC AUC score of 0.78 upon training the dataset.

C. Support Vector Classifier

This method is commonly utilized in classification tasks. The algorithm is a supervised learning approach that facilitates the classification of data into multiple classes based on its features. The ability to handle non-linear decision boundaries is crucial in cases where intricate relationships exist between risk factors and diabetes. The dataset yielded a satisfactory performance in classification, achieving an overall accuracy of 79.22% and a ROC AUC score of 0.71.

D. Random Forest Classifier

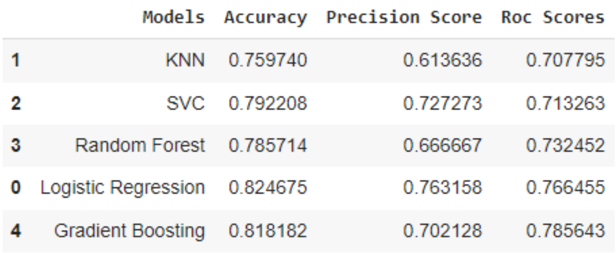
The algorithm employs an ensemble method by integrating several decision trees to produce a robust model capable of making precise forecasts. The method involves constructing several decision trees, with each tree being trained on a distinct subset of the data and features. The dataset's categorical features facilitate the identification of key variables and subsequent output generation. The model achieved an accuracy score of 78.57% and a ROC AUC score of 0.73.

E. K-Nearest Classifier

KNN is a classification method that can be applied to determine whether patients have diabetes, based on their independent variables. It operates by identifying the K closest data points in the training set to the point being classified. This approach is effective in detecting intricate patterns among different variables. This study applied KNN to a dataset, resulting in an accuracy of 75.97% and a ROC AUC score of 0.70.

V. Result

After evaluating all the models Logistic regression gave the overall best accurate result of 82.46% and a ROC AUC score of 0.76.



	Models	Accuracy	Precision Score	Roc Scores
1	KNN	0.759740	0.613636	0.707795
2	SVC	0.792208	0.727273	0.713263
3	Random Forest	0.785714	0.666667	0.732452
0	Logistic Regression	0.824675	0.763158	0.766455
4	Gradient Boosting	0.818182	0.702128	0.785643

Table 1:All models result from evaluation

The avobe table is generated to evaluate the results of all algorithms. It shows logistic regression performed quite well whereas KNN gives the lowest accuracy among all the algorithms.

VI. Discusssion

The study aimed to employ a machine learning algorithm for diabetes prediction. Logistic regression yielded the highest accuracy among all models utilized. Subsequently, Gradient boosting exhibited a satisfactory performance with a ROC score of 0.78. The ROC scores of the Support Vector Classifier and Random Forest Classifier were comparable, at 0.71 and 0.73, respectively. The KNN algorithm yielded an accuracy score of 0.75, the lowest among the tested methods.

VII. Conclusion Future Work

This study evaluated the efficacy of various machine learning algorithms, including Logistic Regression, Gradient Boosting, Support Vector Classification, Random Forest classification, and K Nearest Classifier. The algorithms were analyzed and their output was evaluated. Logistic regression outperformed KNN in terms of accuracy, yielding a higher score. Future work may involve utilizing larger datasets and alternative algorithms to determine the optimal machine-learning approach for predicting diabetes.

REFERENCES

[1] W. H. Organization, “Diabetes,” World Health Organisation, 04 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, “Prediction of diabetes using machine learning algorithms in healthcare,” in *2018 24th international conference on automation and computing (ICAC)*. IEEE, 2018, pp. 1–6.

[3] A. Mujumdar and V. Vaidehi, “Diabetes prediction using machine learning algorithms,” *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[4] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, “Diabetes prediction using ensembling of different machine learning classifiers,” *IEEE Access*, vol. 8, pp. 76 516–76 531, 2020.

[5] N. Yuvaraj and K. SriPreethaa, “Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster,” *Cluster Computing*, vol. 22, no. Suppl 1, pp. 1–9, 2019.

[6] D. H. Sulistyawati and A. Murtadho, “Performance accuration method of machine learning for diabetes prediction: Performance accuration method of machine learning for diabetes prediction,” *Jurnal Mantik*, vol. 4, no. 1, pp. 164–171, 2020.

[7] C. C. Olisah, L. Smith, and M. Smith, “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective,” *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106773, 2022.

[8] S. M. Ganie and M. B. Malik, “An ensemble machine learning approach for predicting type-ii diabetes mellitus based on lifestyle indicators,” *Healthcare Analytics*, vol. 2, p. 100092, 2022.

[9] L. Ismail, A. Hennebelle, H. Materwala, J. A. Kaabi, P. Ranjan, and R. Janardhanan, “Secure and privacy-preserving automated end-to-end integrated iot-edge-artificial intelligence-blockchain monitoring system for diabetes mellitus prediction,” *arXiv preprint arXiv:2211.07643*, 2022.

[10] D. Dutta, D. Paul, and P. Ghosh, “Analysing feature importances for diabetes prediction using machine learning,” in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2018, pp. 924–928.

[11] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting diabetes mellitus with machine learning techniques,” *Frontiers in genetics*, vol. 9, p. 515, 2018.

[12] S. I. Ayon and M. M. Islam, “Diabetes prediction: a deep learning approach,” *International Journal of Information Engineering and Electronic Business*, vol. 12, no. 2, p. 21, 2019.

[13] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, “Predictive models for diabetes mellitus using machine learning techniques,” *BMC endocrine disorders*, vol. 19, no. 1, pp. 1–9, 2019.

[14] A. Choudhury and D. Gupta, “A survey on medical diagnosis of diabetes using machine learning techniques,” in *Recent Developments in Machine Learning and Data Analytics: IC3 2018*. Springer, 2019, pp. 67–78.

[15] T. N. Joshi, P. Chawan *et al.*, “Diabetes prediction using machine learning techniques,” *Ijera*, vol. 8, no. 1, pp. 9–13, 2018.

[16] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021.

[17] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, “Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction,” in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1. IEEE, 2021, pp. 141–146.

[18] A. TIWARI, “Diabetes prediction-eda + 10 models,” kaggle.com, 2021. [Online]. Available: https://www.kaggle.com/code/aryantiwari123/diabetes-prediction-eda-10-models/input