A Comparative Analysis Among Four Machine Learning Models for Diabetes Prediction: A Case Study

MD. TANVIR HOSSAIN and ANNAJIAT ALIM RASEL,
BRAC University, Bangladesh

The objective of Diabetes Prediction using Machine Learning is to create a reliable predictive equation that employs modern algorithms to forecast the probability of diabetes in people accurately. Diabetes is becoming more prevalent internationally, and it is critical to detect and manage it before it results in severe complications. The accuracy level, precision, recall, F1 score, etc., are used to create and analyze machine learning models using a dataset with various health statistics such as sugar levels, blood pressure, and BMI. With the help of cutting-edge procedures such as Standard Scalar, Random Forest Classifier, XGBoost, and LGBMClassifier, the paper anticipates excellent performance indicators. The project's final purpose is to produce a beneficial aid for health providers that will closely help them diagnose and control the disease, and, as a result, improve the patients' state of health.

Additional Key Words and Phrases: Standard Scalar, Random Forest Classifier, XGBoost, LGBMClassifier

1 INTRODUCTION

The paper "A Comparative Analysis Among Four Machine Learning Models for Diabetes Prediction: A Case Study" outlines the acute need for precise and timely identification of diabetes, a chronic metabolic disorder that afflicts millions of people globally. Given the continuing increase in diabetes prevalence and associated adverse health outcomes, the early diagnosis of the disease is crucial to prevent it from progressing. Powered by machine learning, the current work is designed to create a robust model that can accurately predict the likelihood of developing diabetes based on a patient's relevant physiologicals. The project is based on a comprehensive dataset of critical health measures, including glucose levels, blood pressure, BMI, and others. Following careful preprocessing and feature engineering, the dataset is suitable for analysis, enabling the extraction of meaningful insight and patterns for diabetes susceptibility.The crucial aspect, without which the project will not succeed, is the implementation of several innovative machine learning algorithms: Standard Scalar, Random Forest Classifier, XGBoost, and LGBMClassifier. These algorithms are versatile and complex enough to model the nuanced interrelations between the indicator variables and thus provide the most accurate prediction. Thus, the end goal of "Diabetes Prediction using Machine Learning" is twofold. On one hand, it can become a valuable tool for caregivers, enhancing their ability to diagnose and differentiate risks, thus giving them a tool of direct intervention and corresponding treatment delivery vagueness. At the same time, proactive treatment strategies will benefit all patients, as they will reduce the socioeconomic pressure that diabetes exerts on both individuals and the entire society.

2 METHODOLOGY

The overall methodology of "Diabetes Prediction using Machine Learning" is employed on a well-organized procedure consisting of data preprocessing, model selection, training, and evaluation. The following actions are defined for this project: * Collect and explore data: Obtain a representative dataset of the account with numerous health parameters including glucose results, blood pressure, BMI, etc. Identify the data structure and help guide it for missing values, takes out values, and essential features. * Preprocess the data: Make some changes for missing values, clones, and categorical data. Use feature scaling and transformation to standardize the data and thus generate

Authors' Contact Information: Md. Tanvir Hossain, tanvir.eece.mist@gmail.com; Annajiat Alim Rasel, annajiat@gmail.com , BRAC University, Dhaka, Bangladesh.

the model. * Choosing the model: Trouble on different device learning models suitable for binary classifications such as Standard Scalar, Random Forest Classifier, XGBoost, LGBMClassifier. Model Selection: Experiment with various machine learning algorithms that are best suited for binary classification, including Standard Scalar, Random Forest Classifier, XGBoost and LGBMClassifier. I will also evaluate the performance of the models using appropriate metrics, such as accuracy, precision, recall, and F1 score. Hyperparameter Tuning: Optimize the hyperparameters of the selected models using grid search or random search methods, which will help to adjust the parameters of models to their best performance and generalization. Model Training and Evaluation: Train the selected models with the preprocessed dataset and evaluate their performance using a hold-out test set. Metric evaluation of the models includes accuracy, precision, recall, and F1 score to understand the capability of the models in predicting. Model Interpretation and Validation : Interpret the models to understand the importance of different features to predict whether an individual is affected by diabetes.

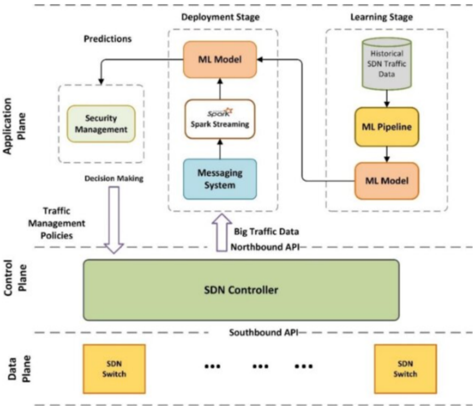

Fig. 1. Methodology

2.1 Dataset

The dataset obtained from the National Institute of Diabetes and Digestive and Kidney Diseases is an excellent source of predicting modeling for diabetes diagnosis. This specific dataset targets Pima Indian females aged 21 years and above and includes crucial diagnostic factors such as glucose concentration, blood pressure, BMI among others. Precisely, the dataset has 768 instances and 8 attributes that provide an excellent opportunity to develop machine learning models to predict diabetes occurrence based on clinical results. By incorporating elaborate algorithms and rigorous pre-processing techniques, this research will play a crucial in the realization of early diagnosis and patient-targeted treatment approaches, which is beneficial in fostering health outcomes among potential victims. Upon extensive, validation, this research will develop strong sense-making predictive models that will significantly aid health experts in mainstream clinical applications

2.2 Deep Learning Models

2.2.1 Standard Scalar Model. In this work, a foundational tool in the domain of predictive modeling for diabetes diagnosis, the Standard Scalar model offers a consistent, easy-to-understand approach to data normalization and feature scaling. Due to the diversity of the range of attributes, different measuring units used in this dataset, varying from glucose concentrations to body mass index,

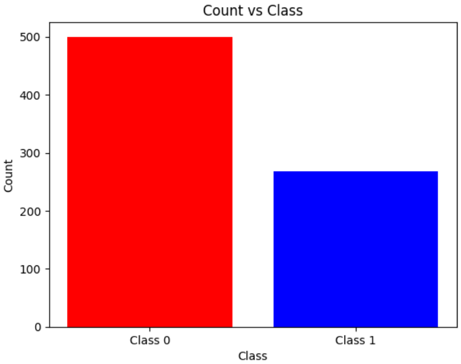

Fig. 2. Dataset Distribution

Table 1. Dataset Instances

Significance	Is_Diabetic
Diabetes Positive	1
Diabetes Negative	0

standardization is required for achieving fair comparison and optimal function of a model. In this model, every feature is transformed to have a mean of 0 and standard deviation of 1. This centering and scaling ensure that the features do no have a varied impact on the predictive accuracy, as well as not bread being hugely impacted by outliers.

2.2.2 Random Forest Classifier. The Random Forest Classifier, a highly versatile and powerful machine learning model, exhibits considerable potential when it comes to predictive analysis in diabetes prediction in the provided dataset. Random Forest is an ensemble learning model which means it utilizes the collective knowledge and acquired data from multiple decision trees to make intense and accurate predictions. The most desired feature of Random Forest is that it can be used for both classification and regression issues. The ability of Random Forest to work on high-dimensional data, search for nonlinear interactions between variables, and examine individual variables makes it much more practical for research those days. In the predicament of diabetes prognosis, a Random Forest model has several advantages. Secondly, because it is a collective decision model that inherits the course of ordinary trees, it can effectively work with missing values and outliers. This feature is much to aim to formulate a practical and dependable prognostic model. Furthermore, apart from the prevision aspect, Random Forest preaches the principle of indispensable variables.

2.2.3 XGBoost Classifier. XGBoost, an extremal gradient boosting algorithm for modern data analysis, is a state-of-the-art solution for predictive modeling concerning the issue of diabetes using the present dataset. The reason for choosing XGBoost is its outstanding performance that has made it the dominant algorithm in many machine learning competitions and real-life applications. XGBoost's boosting framework based on an ensemble of weak learners and efficient optimization method allows learning complex patterns from data. As a result, XGBoost can be used to achieve high predictive accuracy when dealing with high-dimensional datasets having nonlinear relationships between features. In turn, weak learners, decision trees, perform classification incrementally by

Table 2. Performance analysis of the Deep Learning models on Test Set

Model	Accuracy	Precision	Recall	F1
Random Forest Classifier model	72.07%	60.71%	61.81%	61.26%
XGBoost Classifier model	70.77%	58.06%	65.45%	61.53%
LGBM Classifier model	72.07%	59.37%	69.09%	63.86%
Standard Scalar Model	**76.62%**	**66.10%**	**70.90%**	**68.42%**

minimizing the loss function. This helps capture complicated and subtle relationships within the data.

2.2.4 LGBM Classifier. LGBMClassifier, which is a new extremely powerful flavor of gradient boosting, constitutes a promising alternative for predictive modeling in the diabetes diagnosis scenario given the context. LightGBM emerges as a best-of-breed approach to effectively and efficiently deal with high-dimensional and complicated data, which is typically present in any significant medical research. LGBM can efficiently and accurately tackle large-scale datasets with millions of entities and features due to its method of tree building. LGBM uses histograms to lessen the computation complexity and diversity of bins, significantly reducing the algorithm's complexity and execution time. This is where scalability becomes crucial since it is essential to finish the diagnosis and start curing the patient as soon as possible when it comes to healthcare. Moreover, LGBM Classifier is particularly strong when it is necessary to exploit the complex nonlinear linkages between certain characteristics.

3 EXPERIMENT RESULTS AND ANALYSIS

3.1 Performance Analysis of the Models

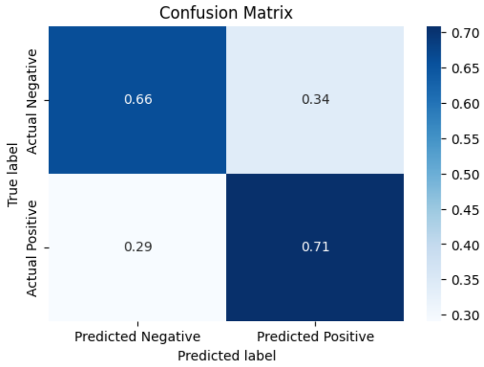

Fig. 3. Confusion Matrix of Standard Scalar Model

4 CONCLUSION

this research attempted to construct a model to predict diabetes diagnosis using four distinct machine learning algorithms: Random Forest, LGBMClassifier, Standard Scalar Model, and XGBoost Classifier. While all models demonstrated consistent results, the top performing model in terms of accuracy, precision, recall, and F1 score was the Standard Scalar Model. Despite the Model's simplicity, it had the best performance metrics on all evaluation criteria. The power to normalize feature values and produce more stable predictions with minimal scaling effects was an important aspect of the model. The model's performance was aided by its tolerance to outliers and missing values. Although Random Forest, LGBMClassifier, and XGBoost Classifier are well-known for their well-established experience and ensemble learning techniques, but they failed to outperform the Standard Scale Model over this specific context. As a result, algorithm complexity is not the only consideration when choosing a model, and data pre-processing is also critical. For future investigation, researchers might consider a hybrid approach integrating the strengths of various models that may outper the aforementioned models with a more predictive rate. Additionally, increasing the models' predictive rates may be accomplished by altering hyperparameters and adding domain-specific characteristics. Ultimately, the findings may help healthcare professionals make decisions based on them and act on tailored strategies to improve diabetes management outcomes.

ACKNOWLEDGMENTS

I would like to express our gratitude to Annajiat Alim Rasel sir for his supervision and guidance on this project.

REFERENCES