International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

Comparative Analysis of Supervised Machine Learning Algorithms for Diabetes Prediction

Sanmati Kumar Jain, Dr. Rajesh Keshavrao Deshmukh

Department of Computer Science & Engineering, Kalinga University, Naya Raipur, Chhattisgarh, India

Abstract – Because of its increasing prevalence and the implications that are associated with it, diabetes mellitus, which is a chronic metabolic disorder that is characterized by hyperglycemia, poses a significant threat to the health of people all over the world. Forecasting diabetes in a manner that is both accurate and timely is absolutely necessary for effective management and preventative approaches. Through the application of machine learning methods, this study comes up with a model that can accurately predict diabetes. The predictive model utilizes supervised machine learning techniques, specifically Decision Tree, Naïve Bayes, Artificial Neural Network, and Logistic Regression. These techniques are applied to provide accurate predictions. A number of performance criteria, like as accuracy, recall, precision, and F-score, have been utilized in order to carry out the comparison of different techniques.

Keywords: Supervised learning, Accuracy, Precision, Recall, Diabetes

I. INTRODUCTION

A chronic metabolic disease defined by persistently high blood glucose levels, diabetes mellitus has recently risen to the ranks of the world's most critical health issues. Factors including sedentary lifestyles, changes in food, and urbanization have contributed to the fast rise in the prevalence of diabetes. Worldwide, 451 million individuals are living with diabetes, and that number is expected to skyrocket in the next decades, according to the International Diabetes Federation (IDF). Cardiovascular disease, neuropathy, nephropathy, and retinopathy are some of the long-term consequences of diabetes that can be lessened with early diagnosis and treatment. Although they are efficient, traditional diagnostic methods can be difficult and expensive for patients to afford or schedule. One game-changing strategy for better diabetes prediction and diagnosis in this setting is the use of ML algorithms.

Machine learning is a branch of AI that uses various algorithms and approaches to teach computers to learn from examples, so they can make judgments or predictions without human intervention. There is a lot of hope that using ML algorithms in healthcare might increase diagnostic accuracy, decrease costs, and pave the way for customized therapy, especially in the area of illness prediction. In order to develop predictive models that can identify individuals at risk of developing diabetes before clinical symptoms appear, the field of diabetes prediction through ML utilizes a variety of data sources, such as medical imaging, genetic information, and electronic health records (EHRs).

Diabetes prediction has made use of many ML algorithms, each with its own set of benefits and drawbacks. Because of its simplicity and interpretability, logistic regression—a basic approach in statistical modeling—is frequently utilized as a baseline for comparison. When dealing with non-linear correlations and interactions between features, decision trees and random forests—which employ hierarchical decision-making processes—tend to be preferred. For high-dimensional data, strong frameworks for capturing complicated patterns are neural networks, support vector machines (SVMs), and deep learning models. A more accurate analysis of complex datasets, such medical pictures and time-series data, is now possible because to recent developments in deep learning, especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

The capacity of the model to generalize to previously unknown data, the amount and quality of the data, and the technique chosen are all crucial to the success of ML algorithms in diabetes prediction. The success or failure of machine learning models is heavily dependent on the quality of the data preparation and feature selection processes. To construct reliable prediction models, it is necessary to deal with missing values, standardize data, and choose pertinent characteristics. Further information about the model's performance and its possible therapeutic use may be gleaned from model assessment measures including recall, accuracy, precision, and the area under the receiver operating characteristic curve (ROC-AUC).

I. REVIEW OF LITERATURE

Febrian, Muhammad et al., (2023) Blindness, renal failure, heart attacks, and death are all possible outcomes of diabetes. In 2019, 463 million people were diagnosed with diabetes, as

reported by the International Diabetes Federation. This figure will reach 700 million by 2045, up 578 million by 2030, if forecasts pan out. An report published in 2020 by the Republic of Indonesia's Ministry of Health lists Indonesia among the top ten nations with the highest diabetes prevalence in 2019. Diagnosing the kind of diabetic condition requires the skill of specialists. Many people who are evaluated have a serious illness since they took so long to find out what it was. Severe complications can be avoided with the use of diabetes detection technologies. Modern medical professionals rely on it for precise and rapid illness diagnosis. In light of this, we may employ machine learning to put an end to this killer by creating an AI model that can foretell the onset of diabetes. To determine which algorithm is most suited for this task, we will compare KNN and Naive Bayes. Finally, the study compared the Naive Bayes method with two k-Nearest Neighbor algorithms to see which one was better at utilizing supervised machine learning to forecast diabetes based on various health characteristics in the dataset. Our studies and testing with the Confusion Matrix show that Naive Bayes is the superior algorithm than KNN.

Ganie, Shahid & Malik, Majid. (2022) Diabetes ranks among the top ten causes of mortality. Diabetes mellitus is a lethal condition that presents a distinct and considerable risk to millions worldwide. Notwithstanding the veracity of statistical data on diabetes from numerous sources such as the World Health Organization, International Diabetes Federation, and American Diabetes Association, there is an optimistic assertion that early detection coupled with suitable management can effectively control diabetes mellitus and avert its complications. Currently, machine learning techniques are becoming increasingly significant in the healthcare industry due to their analytical categorization skills. Researchers are utilizing machine learning methods to enhance diabetes prediction and preserve human lives. This research presents a comparison of various supervised machine learning classifiers, evaluating their performance through many measures for the early prediction of type-II diabetic mellitus (T2DM). The experimental study has been effectively conducted utilizing six machine learning classification techniques. Among all classifiers, the random forest (RF) has superior performance in predicting T2DM, achieving an accuracy rate of 93.75%. Furthermore, the ten-fold cross-validation approach has been employed to eliminate class bias in the dataset.

El Massari, Hakim et al., (2022) One of the chronic diseases that is steadily on the rise is diabetes. When diabetes is not identified correctly and in a timely manner, complications start to arise. An automated system that can detect diabetic patients has recently been developed by medical research using several machine learning approaches, including ontology-based ML techniques. This paper offers a comprehensive analysis and comparison of the most widely used machine learning methods, with a focus on ontology-based ML classification. We looked at several different categorization techniques, including Support Vector Machines, Kernel-Naive Bayes, Artificial Neural Networks, Decision Tree, and Logistic Regression. Performance measurements like as F-Measure, Accuracy, Precision, and Recall are used to assess the outcomes. These metrics are obtained from the confusion matrix. Ontology classifiers and support vector machines (SVM) achieved the highest levels of accuracy in the experiments.

Nahzat, Shamriz & Yaganoglu, Mete. (2021) The use of AI to healthcare systems has seen significant change during the past several years. There are several applications of machine learning (ML) in the field of medical diagnostics. Critical diseases such as cancer, diabetes, heart disease, thyroid disease, and many more are predicted or diagnosed with the use of machine learning algorithms. Simplifying and speeding up the diagnostic process would have a significant impact on treatment for chronic diabetes, one of the most common illnesses globally. The primary objective of this study is to develop and implement a diabetes prediction system utilizing many machine learning approaches, and then to analyze the results of these systems to determine which one produces the most accurate classifier. Using a variety of diabetes disease-related characteristics, this research looks into diabetes prediction. To forecast the occurrence of diabetes, we employed a variety of Machine Learning classification techniques on the Pima Indian Diabetes Dataset. These techniques included K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT). This analysis makes use of models with varying degrees of precision. A model that accurately predicts diabetes is demonstrated in this study. This study found that random forest outperformed other machine learning algorithms when it came to diabetes predicting.

Hasan, Md. Kamrul et al., (2020) Diabetes, classified as a chronic condition, comprises a collection of metabolic disorders characterized by prolonged elevated blood sugar levels. The risk factor and severity of diabetes can be substantially diminished with accurate early prognosis. The precise and reliable prediction of diabetes is significantly impeded by the scarcity of labeled data and the existence of outliers or missing values in diabetes datasets. This literature proposes a comprehensive framework for diabetes prediction that incorporates outlier rejection, imputation of missing values, data standardization, feature selection, K-fold cross-validation, and various Machine Learning classifiers, including k-nearest neighbors, Decision Trees, Random Forest, AdaBoost, Naive Bayes, XGBoost, and Multilayer Perceptron (MLP). The literature also proposes the weighted

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

ensembling of many ML models to enhance diabetes prediction, with weights determined by the respective Area Under the ROC Curve (AUC) of each model. The performance statistic used is AUC, which is subsequently optimized by hyperparameter optimization via the grid search method. All investigations in this literature were performed under identical experimental circumstances utilizing the Pima Indian Diabetes Dataset. Our proposed ensembling classifier, derived from extensive experiments, demonstrates superior performance with a sensitivity of 0.789, specificity of 0.934, false omission rate of 0.092, diagnostic odds ratio of 66.234, and an AUC of 0.950, surpassing state-of-the-art results by 2.00% in AUC. Our suggested framework for diabetes prediction surpasses the alternative methodologies examined in the study. It can yield superior outcomes on the identical dataset, perhaps enhancing performance in diabetes prediction. Our source code for diabetes prediction is freely accessible.

II. EXPERIMENTAL SETUP

A model for diabetes identification has been constructed in this research using several supervised learning classifiers. Specifically, this study makes use of the Pima Indian Diabetic data set, which is housed in the machine learning repository at UCI. Patients who are female and at least 21 years old are included in this dataset. Plasma glucose, diastolic blood pressure, triceps skin fold thickness, serum insulin, body mass index, age, diabetes pedigree, and diabetes class variable are some of the 9 features included in this labeled dataset that includes 678 occurrences. The simulations were conducted using the WEKA 3.8.4 simulator.

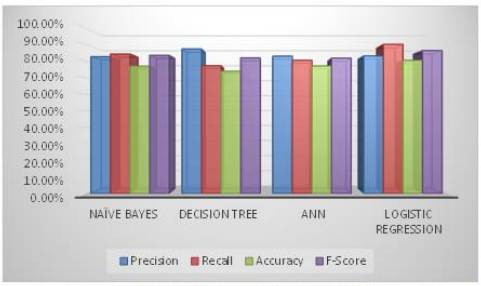

Figure 1: Comparison of various classifiers on diabetic dataset

Whether the data is organized or not, gathering raw data is the initial stage in building a predictive machine learning model. It is critical to clean the data after data collection to eliminate outliers and missing numbers. Step two involves selecting the most important and relevant characteristics from the provided data set. This process is known as feature selection. The following stage is to use a supervised learning classification technique to train the machine learning model. In this scenario, the training and testing portions are 70% and 30% of the total. After that, test data is used to evaluate the trained model. The supervised learning model uses its stored information about previous instances to make predictions about future outcomes. There are a number of performance metrics used to compare the efficiency of different classifiers. The accuracy value of each classifier is used to choose the best classifier for this model. However, F-score is also an important metric in this model because to the unbalanced diabetes dataset.

III. RESULTS AND DISCUSSION

Table 1: Comparison of various classifiers on diabetic dataset

Classification	**Naïve Bayes**	**Decision Tree**	**ANN**	**Logistic Regression**
Precision	82.59 %	87.27%	82.95 %	83.06%
Recall	84.21 %	77.17%	80.42 %	90.01%
Accuracy	76.86 %	73.85%	76.98 %	80.35%
F-Score	83.43 %	81.90%	81.71 %	86.28%

The table summarizes the performance of several classifiers on the diabetes dataset. It highlights their F-score, recall, accuracy, and precision. With an accuracy of 80.35 percent, Logistic Regression outperforms all of the classifiers when it comes to accurately categorizing cases. Though its accuracy of 83.06% is significantly lower than other models, it has the greatest recall rate of 90.01%, indicating that it is very good at detecting actual positive instances of diabetes.

Although Naïve Bayes is not as good at recall as Logistic Regression, it has competitive precision (82.59%) and F-score (83.43%), indicating a balanced performance between the two. However, its accuracy is the lowest at 76.86%. Among the classifiers, the Decision Tree model's accuracy of 87.27% stands out, suggesting it excels at reducing false positives. While it excels at precision, it fails to catch all positive instances due to its poorer recall (77.17%) and accuracy (73.5%). With a precision of 82.95% and an accuracy of 74.98%, Artificial Neural Networks (ANN) perform similarly to other models. From an accuracy and recall balancing perspective, it is positioned between Decision Tree and Naïve Bayes with an F-score of 81.71% and a recall rate of 80.42%.

IV. CONCLUSION

The comparative comparison of supervised machine learning algorithms for diabetes prediction reveals the considerable potential of sophisticated computational methods in improving early detection and diagnosis of diabetes. The results highlight the capacity of ensemble approaches to identify intricate patterns and interactions in clinical data, hence enhancing risk assessment. The study underscores the significance of model interpretability and the necessity for a balanced methodology that integrates predicted accuracy with therapeutic relevance. Subsequent research ought to concentrate on optimizing these models, rectifying data quality concerns, and amalgamating varied datasets to improve generalizability across diverse populations. Utilizing these data, healthcare systems may progress towards more individualized and proactive diabetic management, eventually enhancing patient outcomes and diminishing disease burden.

REFERENCES: -

[1] F. Febrian, M. Muhammad, F. Ferdinan, G. Sendani, K. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," *Procedia Computer Science*, vol. 216, pp. 21-30, 2023, doi: 10.1016/j.procs.2022.12.107.

[2] K. Kangra and J. Singh, "Comparative analysis of predictive machine learning algorithms for diabetes mellitus," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1728-1737, 2023, doi: 10.11591/eei.v12i3.4412.

[3] S. Ganie and M. Malik, "Comparative analysis of various supervised machine learning algorithms for the early prediction of type-II diabetes mellitus," *International Journal of Medical Engineering and Informatics*, vol. 1, no. 1, p. 1, 2022, doi: 10.1504/IJMEI.2021.10036078.

[4] H. El Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, "Diabetes prediction using machine learning algorithms and ontology," *Journal of ICT Standardization*, vol. 10, no. 1, pp. 319–338, 2022, doi: 10.13052/jicts2245-800X.10212.

[5] S. Firdous, G. A. Wagai, and K. Sharma, "A survey on diabetes risk prediction using machine learning approaches," *Journal of Family Medicine and Primary Care*, vol. 11, no. 11, p. 6929, 2022.

[6] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 90-100, 2022.

[7] I. Tasin, T. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 3, 2022, doi: 10.1049/htl2.12039.

[8] S. Nahzat and M. Yaganoglu, "Diabetes prediction using machine learning classification algorithms," *European Journal of Science and Technology*, vol. 24, pp. 53-59, 2021, doi: 10.31590/ejosat.899716.

[9] N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229-241, 2021.

[10] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Information Science and Systems*, vol. 8, no. 1, p. 1-14, 2020.

[11] 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. 10, pp. 1-1, 2020, doi: 10.1109/ACCESS.2020.2989857.

[12] K. M. Orabi, Y. M. Kamal, and T. M. Rabah, "Early predictive system for diabetes mellitus disease," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9728, pp. 420–427, 2016, doi: 10.1007/978-3-319-41561-1_31.

[13] A. Parashar, K. Burse, and K. Rawat, "A comparative approach for Pima Indians diabetes diagnosis using lda support vector machine and feed forward neural network," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 11, pp. 378-383, 2014.

[14] G. Visalatchi and S. J. Gnanasoundhari, "A survey on data mining methods and techniques for diabetes mellitus," *International Journal of Computer Science and Mobile Applications*, vol. 2, no. 2, pp. 100-105, 2014.