Diabetes Prediction in Healthcare

Apurva Deshpande, Rushikesh Gholap, Sushmitha Rajeswari Muppa and Veda Varshita

Abstract

As the prevalence of diabetes continues to rise globally, early prediction and intervention have become imperative for effective healthcare management. This research explores the application of machine learning classifiers to predict diabetes, aiming to identify the most accurate and robust approach. The study evaluates the performance of various individual classifiers, including Decision Trees, Linear discriminant analysis (LDA), Naive Bayes, and Logistic Regression, on a comprehensive dataset of health parameters. The research involves the analysis of a diverse range of features, such as age, body mass index, blood pressure, and cholesterol levels, to train and test the classifiers. Results indicate that individual classifiers exhibit varying degrees of accuracy, sensitivity, and specificity. To enhance predictive capabilities, an ensemble classifier is proposed, leveraging the strengths of multiple algorithms. **The ensemble classifier is constructed through a systematic integration of individual classifiers, employing techniques such as bagging or boosting.** The research evaluates the ensemble classifier's performance against standalone classifiers, considering metrics such as precision, recall, F1-score, and **area under the receiver operating characteristic curve. The findings demonstrate that the ensemble classifier outperforms individual classifiers in terms of predictive accuracy and robustness.** Furthermore, the study investigates the interpretability of the ensemble model, providing insights into the critical features contributing to diabetes prediction. This contributes to the understanding of the underlying factors influencing the disease and facilitates personalized healthcare strategies. In conclusion, this research highlights the importance of employing ensemble classifiers in diabetes prediction, showcasing their ability to enhance accuracy and reliability. The findings contribute valuable insights to the field of healthcare analytics, emphasizing the potential for machine learning to revolutionize early detection and intervention in diabetes, ultimately improving patient outcomes.

1 Background and Related Work

Diabetes mellitus has emerged as a major global health concern, with an escalating prevalence that poses significant challenges to healthcare systems worldwide. The chronic nature of diabetes and its associated complications necessitate proactive and personalized healthcare strategies. Early prediction of diabetes is crucial for timely intervention, enabling healthcare providers to implement preventive measures and improve patient outcomes. Traditional methods of diabetes prediction often rely on clinical risk factors and demographic information. However, with the advent of machine learning techniques, there is an opportunity to harness the power of computational models to analyze complex datasets and extract valuable insights. Machine learning classifiers, such as Decision Trees, Linear discriminant analysis (LDA), Naive Bayes, and Logistic Regression, have demonstrated promise in predicting diabetes based on diverse sets of health parameters.

The study on *Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine*[1] explores applying Gradient Boosting for diabetes prediction. It combines weak models, like decision trees,

iteratively correcting errors, contributing to the effectiveness of ensemble methods in diabetes diagnosis. In *Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method,*[2] the focus is on optimizing K-Nearest Neighbor (KNN) for diabetes prediction. The study explores finding the optimum K value and incorporates feature selection for accurate diabetes prognosis, providing insights into optimizing KNN for improved classification accuracy. The study *Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach*[3] applies Random Forest Classifier for diabetes prognosis. This ensemble approach, using multiple decision trees, enhances predictive accuracy, emphasizing the potential of Random Forest Classifier as a robust forecasting model for diabetes. *Nadeem et al.*[4] introduce a fusion-based ML method for diabetes onset prediction, combining Support Vector Machines (SVM) and Artificial Neural Networks (ANN). Achieving a classification accuracy of 94.67%, it surpasses other models, showcasing the effectiveness of combining SVM and ANN for diabetes prediction.

In summary, these studies collectively highlight the versatility of machine learning techniques in improving predictive models for diabetes mellitus, contributing to the ongoing advancement of predictive analytics in healthcare.

2 Machine Learning Models

2.1 Decision Trees

Decision Trees are hierarchical structures systematically navigate health parameters, such as age, BMI, and blood pressure, to classify individuals into diabetic or non-diabetic categories. Decision Trees inherently capture complex relationships, offering transparency and interpretability crucial for medical contexts. Their ability to handle non-linear interactions makes them effective in discerning patterns within diverse datasets. This research leverages Decision Trees to unravel key features influencing diabetes, contributing valuable insights for early prediction and tailored healthcare strategies. Hence, they are utilized to sift through extensive datasets employing decision rules, efficiently categorizing data. This research explores diverse decision-tree-based techniques for data classification, as depicted in Figure 1. For instance, as indicated the decision tree checks blood glucose level, age, history of heart disease, body mass index and then decides if the sample is diabetes positive or not.

3 Evaluation of results

Preceding model input, the dataset undergoes cleaning, and data values are normalized. Employing SMOTE for oversampling and Random under-sampling for the majority class achieves a balanced dataset exceeding 20,000 rows. A 2:1 ratio splits the dataset into training and validation sets. Post-training, classifiers undergo evaluation through various metrics using the confusion matrix (Fig. 2). TP (true positive) signifies correctly predicted positive instances, TN (true negative) for negatives, FP (false positive) for incorrect positive predictions, and FN (false negative) for inaccuracies in negative predictions. Evaluation metrics encompass accuracy, **sensitivity, specificity, ROC-AUC curve, and Precision–Recall curve. Ensemble outperforms other diabetes models, attaining the highest accuracy (82.26%), while NB demonstrates the lowest accuracy at 70.56%. Nonetheless, accuracy alone is insufficient for comprehensive**

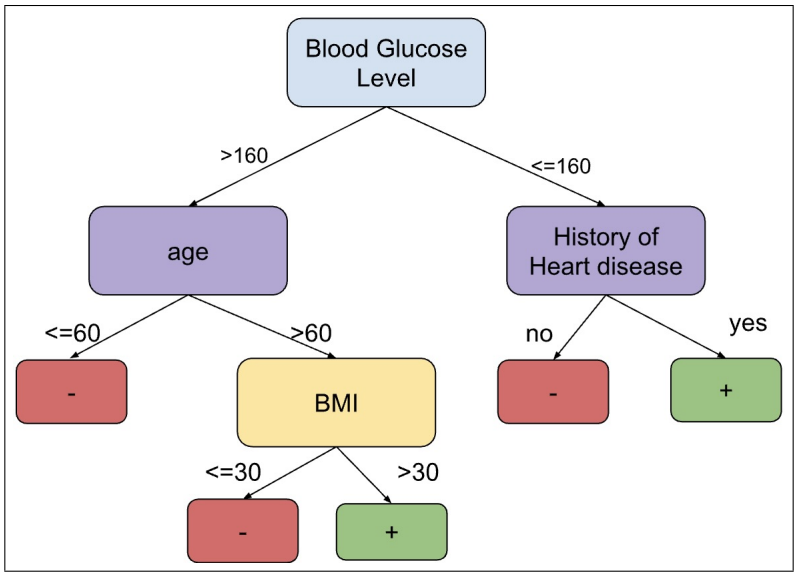
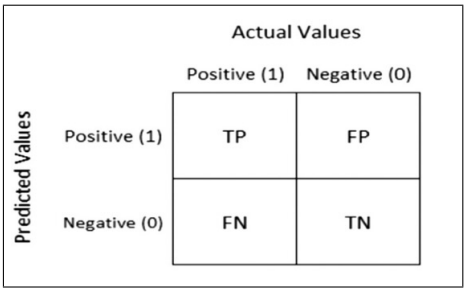


Figure 1: Decision Tree flowchart

model assessment and prediction.



	precision	recall	f1-score	support
0	0.81	0.84	0.83	43671
1	0.83	0.80	0.82	43663
accuracy			0.82	87334
macro avg	0.82	0.82	0.82	87334
weighted avg	0.82	0.82	0.82	87334

Figure 2: Confusion matrix and best classifier metrics

$$\textbf{Precision} = \textbf{TP} \;/\; (\textbf{TP} + \textbf{FP})$$

In the realm of disease detection, the prediction of false positives can result in misdiagnosis and the inefficient utilization of healthcare resources. Enhancing the precision of diagnostic models serves as a crucial solution to mitigate this issue. Precision specifically measures the accuracy of positive predictions, achieved by tallying correctly predicted positive samples (TP) and dividing by the total number of positive predictions (TP + FP), whether correct or incorrect. According to **Table 1**, among the evaluated classifiers, the **RF model exhibited the highest precision at 83.47%, followed by DT at 83.02%, indicating accurate predictions for over 83% of cases identified as diabetic. LR showed the least precision, registering a rate of 70.56%.**

Recall = TP / (TP + FN)

Recall, also known as sensitivity, akin to precision, seeks to determine the proportion of accurately detected true positives. It achieves this by dividing the correctly predicted positive samples (TP) by the total number of positives, whether correctly or incorrectly predicted as positive (TP, FN). Recall assesses the number of accurately predicted positive instances out of all potential positive predictions. The highest sensitivity, or recall, is observed at **80.45% for RF, indicating that 80.45% of diabetes cases in the test set were correctly identified by the RF classifier. NB exhibits the poorest performance in this regard, with a recall rate of 67.07%.** Avoiding false negatives is crucial, as overlooking the presence of the disease can lead to delayed treatment and real harm.

F-measure = 2(Precision x Recall) / (Precision + Recall)

A robust diabetes detection system needs to minimize both missed diagnoses and misdiagnoses. However, accuracy and specificity often present conflicting performance metrics. The F-measure takes into account both precision and recall, ranging from 0 to 1. A maximum value of 1 indicates perfect precision and recall, while a minimum value of 0 suggests that either precision or recall is zero. **RF attains the highest F1 score at 82.26%, showcasing superior performance across all evaluation metrics. This substantiates RF as the most effective classifier for diabetes in this study.**

4 Analysis and Conclusions

Our research's primary contribution lies in creating machine learning predictive models for early diabetes detection. **We investigated five classifiers (DT, RF, KNN, LR, and NB) to predict diabetes likelihood, with the RF classifier achieving the highest accuracy at 82.26%. The study focused on evaluating diabetes prediction using crucial features. Leveraging the advanced classification capabilities of machine learning algorithms, our model holds substantial potential to assist medical practitioners significantly in the diagnosis process.**

add accuracy table
add dataset
say something about ensemble

5 Future Work

Incorporating temporal analysis into the study involves a comprehensive exploration of how the predictive performance of the models evolves over time. This approach recognizes the dynamic nature of healthcare data, especially in the context of diabetes prediction. By scrutinizing temporal trends, one can discern patterns, fluctuations, and potential variations in predictive accuracy. Investigating whether the models adapt effectively to evolving trends becomes pivotal in understanding their long-term efficacy. This temporal consideration not only enhances the models' ability to capture nuanced changes in health parameters but also provides insights into their robustness for continuous and real-world applications. This approach aligns with the evolving nature of healthcare data,

ensuring that predictive models remain relevant and effective in addressing the dynamic landscape of diabetes prevalence and risk factors over extended periods. Ultimately, incorporating temporal aspects enriches the models' adaptability, contributing to their sustained accuracy and utility in healthcare decision-making.

References

[1] Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine https://www.mdpi.com/2075-4418/11/9/1714

[2] Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method https://ieeexplore.ieee.org/abstract/document/9214129

[3] Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach https://link.springer.com/chapter/10.1007/978-981-16-2164-2_19

[4] An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators https://www.sciencedirect.com/science/article/pii/S2772442522000582#b7