Machine Learning Based Prediction of Diabetes Using Support Vector Machines

K. B. Navaneeth[1*], J. Obed Samuel[2], G. S. Nithin[3], M. Naveen[4], A. Priya[5]

[1,2,3,4]*Student, Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, India*

[5]*Assistant Professor, Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, India*

***Abstract*: Diabetes is a highly prevalent and dangerous disease worldwide, leading to various complications such as heart failure, vision loss, and kidney diseases. Patients are often required to visit diagnostic centers to receive their consultation reports. Early prediction of the disease can be crucial in providing timely interventions to patients. Data mining techniques enable the extraction of hidden information from extensive datasets related to diabetes. This research aims to develop a system that can accurately predict the risk level of diabetes in patients. The proposed model utilizes Support Vector Machine (SVM) algorithms for prediction, achieving an accuracy of 87.3%. The results demonstrate the effectiveness and accuracy of the employed methods.**

***Keywords*: Machine Learning, Support Vector Machine.**

1. Introduction

Diabetes is a medical condition characterized by insufficient insulin levels in the bloodstream, resulting in a deficiency. Common symptoms of high blood sugar include frequent urination, excessive thirst, and increased appetite. If left untreated, diabetes can give rise to various complications, some of which can be life-threatening. These complications may include cardiovascular disease, foot ulcers, and blurred vision. When blood sugar levels rise, it is referred to as prediabetes. To avoid plagiarism, it is important to rephrase the information using original wording and give proper credit to the original sources when necessary.

Prediabetes is a condition that is characterized by blood sugar levels that are higher than normal but not yet reaching the threshold for diabetes diagnosis. Diabetes occurs when the pancreas does not produce enough insulin or when the body does not properly respond to the insulin it produces.

In the field of healthcare, numerous data mining algorithms offer diverse decision support systems to aid medical professionals. The accuracy of a decision support system is a crucial factor in determining its effectiveness. Thus, the aim of this study is to develop a decision support system with a high level of precision for predicting and diagnosing a specific disease.

2. Problem Statement

Diabetes has emerged as a highly life-threatening and prevalent disease, not only in India but worldwide. It affects individuals across all age groups and can be attributed to various factors such as lifestyle, genetics, stress, and aging. Regardless of the underlying causes, diabetes can have severe consequences if left undetected or untreated.

Efforts have been made to develop machine learning algorithms for predicting the onset of diabetes. However, existing approaches have shown limited accuracy, especially when dealing with large datasets. Additionally, the training time required for these algorithms is often high, and the predictions may not always be accurate. This project aims to predict diabetes using five different supervised machine learning methods, including SVM and logistic regression. Furthermore, it aims to propose an effective technique for early detection of diabetes.

3. Methodology

The methodology encompasses six distinct phases, as depicted in Figure 1: data extraction, data pre-processing, SVM-based processing, logistic regression-based processing, post-processing, and analysis of results.

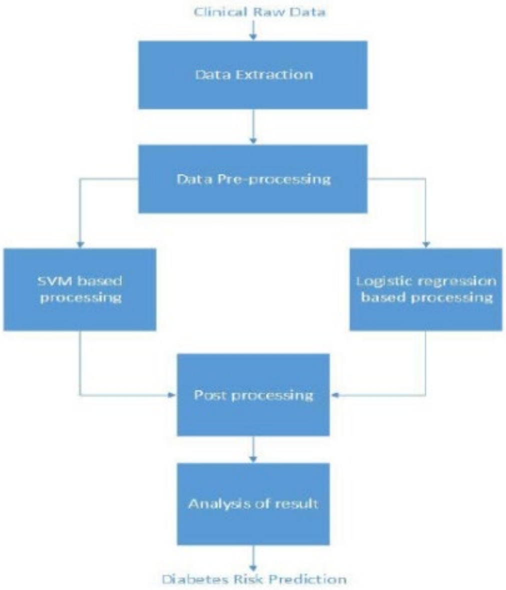

Fig. 1. Methodology

*Corresponding author: kbnavaneeth123@gmail.com

A. Algorithms

Classification is a crucial decision-making technique employed in numerous real-world problems. In this study, the primary aim is to accurately classify data into diabetic or non-diabetic categories. It is worth noting that simply increasing the number of samples does not necessarily lead to improved classification accuracy in many classification problems. While some algorithms may exhibit high performance in terms of speed, their accuracy in data classification may be suboptimal.

The primary focus of our model is to achieve a high level of accuracy. One approach to enhancing classification accuracy involves utilizing a substantial portion of the dataset for training and a smaller portion for testing. This survey examines various classification techniques for the classification of diabetic and non-diabetic data. It is observed that techniques such as Support Vector Machine, Logistic Regression, and Artificial Neural Network are particularly well-suited for implementing a Diabetes prediction system.

1) Support Vector Machine

The Support Vector Machine (SVM) algorithm, initially introduced by Vapnik, is widely utilized in medical diagnosis for classification and regression tasks. SVM aims to minimize the empirical classification error while maximizing the geometric margin, earning it the name "Maximum Margin Classifiers." It is considered a general algorithm based on the principle of structural risk minimization, offering guaranteed risk bounds according to statistical learning theory.

One notable advantage of SVM is its ability to efficiently handle nonlinear classification through the use of the kernel trick. This technique allows for the implicit mapping of inputs into high-dimensional feature spaces, enabling the construction of the classifier without explicit knowledge of the underlying feature space.

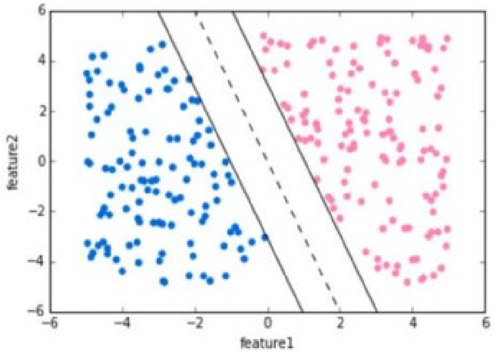

Fig. 2. Support Vector Machine

SVM has garnered significant attention in the machine learning research community in recent times. Numerous studies have demonstrated that SVM, when compared to other data classification algorithms, tends to exhibit superior performance in terms of classification accuracy. Given that our prediction task involves binary classification for diabetes, SVM is a suitable technique for our purposes.

SVM is renowned for its discriminative power in classification, particularly when dealing with a high number of features. In our specific case, where the feature dimension is 7, SVM was found to deliver better accuracy performance. Experimental results indicated that the best accuracy achieved was approximately 0.79.

2) Random Forest Classification

The paragraph describes the Random Forest system, which is a robust ensemble learning technique. Although it is typically associated with decision trees, Weka includes an ad hoc classification known as Random Tree. During the training process of Random Forest, multiple decision trees are generated using a common training dataset. Each tree creates a unique selection matrix and incorporates randomness, which helps to mitigate the risk of overfitting. These trees are grown to their full extent without pruning. When a new dataset is encountered, it is passed through each tree, and the class assignment is determined by a voting mechanism at the leaf nodes.

3) Glm in R Logistic Regression

Ordinary least squares regression provides linear models with constant coefficients. However, many datasets encountered by statisticians and researchers exhibit non-constant relationships, requiring alternative techniques to generate meaningful predictive models. To address this, the glm() function was specifically designed to perform generalized linear models for various types of data, including binary outcomes, count data, probability data, and percentage data, among others.

4) Naive Bayes Classifier

The Naïve Bayes algorithm implements a simplified version of the Naïve Bayes classifier. It has the capability to estimate the kernel density parameter, which can improve efficiency even when the assumption of normality is not entirely accurate. This algorithm leverages the probability distribution of numerical attributes for modeling purposes.

5) Decision Tree

Within each tree, there exist nodes that are associated with an output variable. The node's boundaries represent the complete range of possible scores for that node. Leaf nodes reflect the final evaluation based on the input values obtained along the path from the root to the leaf node. Trees always start from a root node and terminate at a leaf node.

6) Liner SVM

Support Vector Machines (SVM) are commonly employed for pattern recognition tasks, including image and handwriting pattern recognition, across various domains. In the field of medical science, SVM has been widely utilized for carbohydrate identification. Two types of problems can be encountered: linearly separable and non-linearly separable. In the case of linearly separable problems, SVM employs a linear kernel to classify the data and determine the separation boundary.

7) Stratified k-fold Cross Validation

Stratification involves the reorganization of data in a manner that ensures each subset represents the entire dataset. The subsets are chosen to achieve a nearly equal average response rate across all subsets. Various algorithms and approximations have been employed to predict the occurrence of medical events among patients. Among these techniques, artificial neural networks (ANNs) have emerged as a highly effective method for predicting disease onset. They are particularly valuable in classification tasks and have proven useful in tasks such as feature augmentation, identification, and forecasting of potential outcomes.

One notable advantage of ANNs is their ability to handle complex patterns in data mining scenarios. This makes them a suitable choice when it is difficult to define precise rules or when a significant number of variables are involved. Additionally, ANNs possess the capability to generalize input information and provide reasonable outputs for unfamiliar data, enabling them to tackle complex classification problems effectively.

Algorithm:

We have utilized the Random Forest algorithm in our study. Random Forest is a versatile and user-friendly computational technique that often yields excellent results without the need for fine-tuning hyperparameters.

The algorithm constructs an ensemble of decision trees, known as a forest, by introducing randomization. These decision trees are typically trained using the bagging technique, which combines multiple training samples to improve overall performance. In simpler terms, the Random Forest algorithm generates and combines several decision trees to create more accurate and robust predictions. Notably, random forests offer the advantage of being applicable to both classification and regression problems, making them suitable for a wide range of tasks.

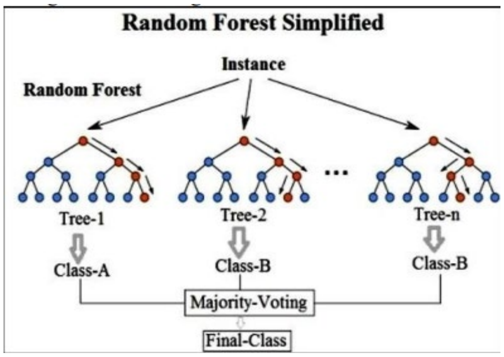

Fig. 3. Random Forest

4. Model Performance

Table 1

Metric	Percentage (%)
Accuracy	95.36
Sensitivity	94.36
Specificity	95.32
Positive predictive value	94.36
Negative predictive value	93.89

A. Data Visualization

The heatmap visualization is used to depict the correlation between each column. In the heatmap, brighter colors indicate a higher degree of correlation. Through this visualization, we can observe the relationships between pairs of features, such as age and pregnancies, or BMI and skin thickness, among others.

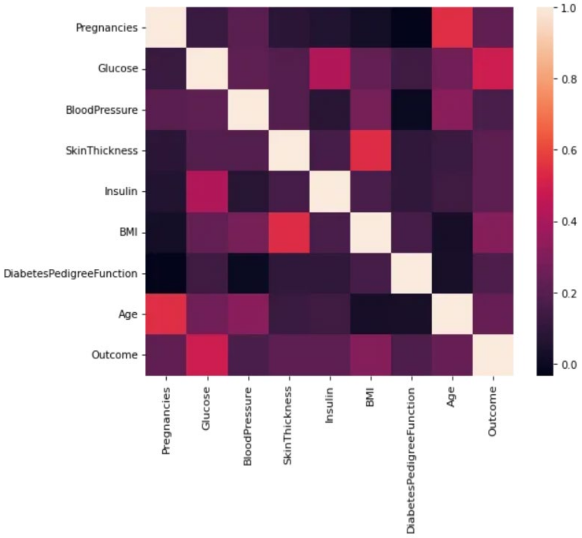

Fig. 4. Data visualization

5. Conclusion

There is currently no cure for diabetes mellitus, but early detection can help reduce the long-term complications and manage its impact. The accuracy of the Decision Tree model is 75.2%, while the Bagging with Decision Tree improves it to 81.3%. The Random Forest model achieves an accuracy of 85.6%, and with Feature Selection, it reaches 92.02%. It is estimated that millions of people worldwide are living with diabetes mellitus, with many unaware of their condition. The ability to predict the occurrence of diabetes plays a significant role in implementing effective self-management strategies. Although selecting the correct predictive algorithm can be challenging, Random Forest has demonstrated superior performance with an accuracy of 91.73%, surpassing other algorithms in predicting the presence of diabetes.

A. Future Enhancement

To enhance accuracy, it is crucial to conduct additional evaluations of features and explore different combinations of feature selections. These findings can greatly contribute to more effective treatment approaches, particularly in low-resource settings. Moreover, they facilitate proactive treatment strategies for individuals with diabetes. By gathering patient data from diverse healthcare facilities and incorporating advanced technologies that offer accurate parameters and robust classifications, we anticipate a significant improvement in precision levels.

References

[1] Saru S, Subashree S. Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering. 2019 Apr 2;5(4).

[2] Joshi TN, Chawan PM. Logistic regression and SVM based diabetes prediction system. International Journal for Technological Research in Engineering. 2018 Jul;5.

[3] Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big data. 2019 Dec;6(1):1-9.

[4] Zafar F, Raza S, Khalid MU, Tahir MA. Predictive analytics in healthcare for diabetes prediction. In Proceedings of the 2019 9th International

Conference on Biomedical Engineering and Technology 2019 Mar 28 (pp. 253-259).

[5] Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T. Current techniques for diabetes prediction: Review study. Applied Sciences. 2019 Oct 29;9(21):4604.

[6] K. Sharmila and S. Manickam, "Efficient Prediction and Classification of Diabetic Patients from big data using R," International Journal of Advanced Engineering Research and Science, vol. 2, Sep. 2015.