Abstract

Diabetes is a persistent medical condition which effect people of all age groups. It is caused by the body's excessive glucose levels. The indicators of this elevated glucose level include high urinary frequency, increased appetite, and increased thirst. Diabetes shouldn't be ignored because, if untreated, it can have major side effects on the body's organs, including damage to the eyes, kidneys, heart, and other systems. If it is anticipated earlier, it can be managed. It's crucial to identify diabetes earlier and lessen its symptoms. In this paper, we intend to create a model that predicts diabetes more precisely using a variety of machine learning (ML) classification techniques. Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), Decision Tree (DT), and Naive Bayes are the algorithms that are examined to determine which model is most effective for predicting diabetes.

Keywords

Diabetes, Classification, Machine learning

Introduction

Diabetes mellitus is a chronic disease described by hyperglycemia [1]. Diabetes might bring numerous entanglements. As per the developing morbidity in coming years, it implies that ten of every hundred people will be suffering from diabetes [1]. Diabetes represents a great danger to human health. It causes serious complications. The trait of diabetes is that the abnormal level of blood glucose, which is brought about by imperfect insulin secretion or impaired biological effects, or both [2]. High appetite, thirst and urinary frequency are few of the features. Certain risk factors such as age of a person, body mass index, glucose levels, blood pressure, etc., play crucial role in identifying the disease. Diabetes has the potential to permanently harm and impair a variety of tissues, especially the eyes, kidneys, heart. Type 1 diabetes and type 2 diabetes are two categories that can be used to categorize diabetes [1].

Predicting diabetes using ML in conjunction with nanotechnology involves leveraging nanoscale materials and devices to enhance the accuracy, efficiency, and sensitivity of diagnostic tools and methods for detecting diabetes. The convergence of ML and nanotechnology presents a transformative approach in predicting diabetes, revolutionizing healthcare. The synergy between ML and nanotechnology significantly enhances diabetes prediction, enabling precise, personalized, and proactive healthcare strategies, ultimately improving patient outcomes and revolutionizing disease management.

With the advancement of expectations for everyday comforts, diabetes is progressively normal in people's daily life. In this way, it is good that research is being done on how to rapidly and precisely analyze diabetes. ML can assist
people make primer discernment about diabetes mellitus as per their everyday checkup data, and it can act as a source of perspective for specialists. In order to determine which ML algorithm would accurately predict diabetes, we are comparing multiple algorithms of ML in this study. We employ RF, LDA, KNN and few other algorithms as our ML techniques. In addition, the best model is deployed into the GUI application.

**Experimentation**

**Existing work**

The already existing systems are predicting the possibility of type 2 diabetes mellitus using artificial neural networks (ANN) [3], prediction of diabetes using ANN [4], prediction of diabetes using back propagation algorithm [4], prediction of diabetes by Ada Boost algorithm with an accuracy of 80% [5], anticipate the chances of getting coronary heart disease for a diabetic patient through Naive Bayes algorithm [6].

**Proposed model**

The proposed system uses ML to predict diabetes. To determine which model can predict diabetes more accurately, different ML approaches, including RF, KNN, SVM, LDA, Naive Bayes and DT, are compared in the system. The most precise model is also deployed in the GUI application, which enables users to anticipate diabetes with ease (Figure 1).

**Proposed model advantages**

- The system is more effective due to fitting datasets for different ML models by applying ML algorithms.
- By analyzing the characteristics of a person, the early anticipation of a disease can be made possible in the proposed system using ML.
- Diabetes is predicted more accurately.
- An intuitive GUI tool that enables users to predict diabetes.

**Experimental procedure**

**Data collection**

The data used is diabetes dataset which is collected from online sources. The attributes in dataset are age, skin thickness, number of pregnancies, diabetes pedigree function, glucose, insulin, body mass index, blood pressure, and outcome.

**Pre-processing**

Real world data generally includes missing data, noisy values making us difficult to develop ML models. Pre-processing is to clean the unwanted data and prepare it for the ML model. It increases the effectiveness of model and also model’s accuracy.

**Training dataset**

The training data, which is the largest portion of the actual dataset, is used to train the ML model.

**Testing dataset**

Once the model has been trained, then we have to test the developed model using the testing dataset. This dataset evaluates the model’s performance and guarantees that it can works well to new datasets.

**ML models**

The supervised ML algorithms like RF, LDA, KNN, SVM, DT, and Naive Bayes are used to develop a predictive model.

**Choosing the best model**

To choose the best model for more accurately predicting diabetes, all ML models are compared based on precision, recall, f1-score, receiver operating characteristics (ROC-AUC) value, accuracy.

**Results and Discussion**

Table 1 displays the comparison results of the different ML models. RF has the highest accuracy out of the six models with around 95-96%. The second highest accuracy is for DT model with 93% accuracy. The Naive Bayes has the least accuracy of 76%.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.78</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.76</td>
</tr>
<tr>
<td>KNN</td>
<td>0.78</td>
</tr>
<tr>
<td>DT</td>
<td>0.93</td>
</tr>
<tr>
<td>RF</td>
<td>0.95</td>
</tr>
<tr>
<td>LDA</td>
<td>0.79</td>
</tr>
</tbody>
</table>

**ROC curve for six models and comparison chart**

The figure 2 acutely tells that the ROC-AUC value is higher than the accuracy. The highest ROC curve value is for RF model is 0.994. The figure 3 compares the accuracies of six models. After comparing both the accuracies and ROC curve values for all the six models, RF is selected as the best model for predicting the diabetes.
Future Scope

In Future, in order to improve user interaction with administrators, we can add more features to the GUI application. These features include a query module that allows users to post queries and receive answers from administrators, a prescriptive module that can suggest treatments for diabetic patients, and a diet module that outlines healthy eating practices for people to follow.

Acknowledgements
None.

Conflict of Interest
None.

References

Conclusions

In this paper, we compared different ML algorithms for anticipating the diabetes and the best model is deployed into the GUI application. The dataset used in this paper is the information about the characteristics of diabetic patient. In order to find a predictive model, six ML methods including SVM, Naive Bayes, KNN, DT, RF, and LDA were evaluated. Accuracy and ROC curves were computed for each model and compared to others to determine the model’s robustness. The outcomes showed that the best model, with the highest ROC curve value and accuracy, is RF. As a result, the RF algorithm was chosen to predict diabetes more effectively.