BVRIT HYDERABAD College of Engineering for Women

Authors - Manaswini, Deepika, Sravani, Bhavya, Harshitha, Srinika

Mentored by - Dr Geetika S Pandey, Assistant Prof, AIML

Diabetes Prediction using Machine Learning

Abstract

Getting a rapid understanding of the context of a patient's overall health has been particularly important during the COVID-19 pandemic as healthcare workers around the world struggle with hospitals overloaded by patients in critical condition. Intensive Care Units (ICUs) often lack verified medical histories for incoming patients. A patient in distress or a patient who is brought in confused or unresponsive may not be able to provide information about chronic conditions such as heart disease, injuries, or diabetes. Medical records may take days to transfer, especially for a patient from another medical provider or system.Knowledge about chronic conditions such as diabetes can inform clinical decisions about patient care and ultimately improve patient outcomes.

To achieve this goal,in this project we do an early prediction of Diabetes in the human body through applying various Machine Learning techniques. Machine learning techniques provide better results for prediction by constructing models from datasets collected from patients. In this project we used Machine Learning Classification and ensemble techniques on a dataset to predict diabetes, which

are Logistic Regression (LR), Decision Tree (DT) ,Random Forest (RF) and XGBoost(XGB). The accuracy is different for each model.The model with higher accuracy is capable of predicting diabetes effectively.

Our result shows that XGBoost achieved higher accuracy compared to other machine learning techniques.

Methodology

In this section we shall learn about the various classifiers used in machine learning to predict diabetes. We shall also explain our proposed methodology to improve the accuracy. Five different methods were used in this paper. The different methods used are defined below. The output is the accuracy metrics of the machine learning models. Then, the model can be used in prediction

Dataset Description

The data set used is https://www.kaggle.com/competitions/widsdatathon2021/data. The objective is to predict if the patient is diabetic or not.

The dataset contains 130,157 rows and 181 columns. Diabetes Mellitus is the feature we are going to predict.

1. **Logistic Regression**

Training Accuracy 70-30	81.42
Training Accuracy 80-20	81.23

2. **Decision Tree**

Training Accuracy 70-30	74.78
Training Accuracy 80-20	74.63

3. **Random Forest**

Training Accuracy 70-30	80.56
Training Accuracy 80-20	80.12

4. **XGBoost**

Training Accuracy 70-30	82.61
Training Accuracy 80-20	82.38

The above data shows that XG Boost gives the highest accuracy of 82.61%

Conclusion

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, four machine learning classification algorithms are studied and evaluated on various measures.In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.