Diabetes Prediction

By Athrwa Deshmukh (19BCE7381)

Colab Link

PROBLEM STATEMENT

Diabetes is a chronic disease or group of metabolic disease where a person suffers from an extended level of blood glucose in the body, which is either the The objective of this project is to make use of significant features, design a prediction algorithm using Machine learning and find the optimal classifier to give the closest result comparing to clinical outcomes. This project aims to focus on selecting the attributes that ail in early detection of Diabetes. The result shows the ID3 and the Random forest has the highest accuracy of 97.88% and 97.58%, respectively holds best for the analysis of diabetic data.

ABOUT THE DATASET

- The dataset was downloaded from Kaggle
- Training samples-2000
- Features-9
- Training data-80%
- Testing data-20%

ALGORITHMS USED

1. Logistic regression.
2. KNN.
3. SVM.
4. Naive Bayes.
5. ID3
6. Random forest

FEATURES IN THE DATASET

1. Pregnacies
2. Glucose
3. Blood pressure
4. Skin thickness
5. Insulin
6. BMI
7. Diabetes pedigree function
8. Age

CODE DESCRIPTION

1. Describing dataset.
2. Checking for null values.
3. Outliers detection and removal.
4. Extracting features and target.
5. Splitting data for training and testing.
6. Applying all the 6 algorithms and displaying confusion matrix, accuracy, precision and recall for each.
7. Plotting all accuracy for comparison.

RESULT

The performance evaluation of the classification techniques is done through the various performance measure such as accuracy, sensitivity, specificity, and recall, precision. Our research paper focus on the five classification techniques such as support vector machine, Random forest, Naïve Bayesian, decision tree and K-nearest neighbour. Table 2 shows the results of the classification technique

- Support vector machine: The accuracy of SVM is 80.06%, precision 76.39% and recall 52.88%.
- Random forest: The accuracy of random forest is 97.58%, precision 95.28% and recall 97.11%.
- Naive Bayesian Classification: The accuracy is 75.83%, precision 61.32% and recall 62.5%
- K-Nearest Neighbour: The accuracy is 85.80%, precision 74.78% and recall 82.69%.
- ID3 decision tree: The accuracy is 97.88%, precision 98.02% and recall 95.19%
- Logistic Regression: The accuracy is 79.75%,precision 74.67% and recall 53.84%.

Thank You