A machine learning approach to predict diabetes and heart disease

Abu Kowcher, Md. Nurul Alam, Md Rezaul Karim
Rangamati Science and Technology University

Abstract

Nowadays, Diabetes and Heart Diseases are among the two most common diseases of humans. People having diabetes have a high risk of other diseases such as heart disease, kidney disease, stroke etc. Therefore, Early-stage prediction of diabetes and heart disease have a significant importance in health sector. The aim of this study is to create a clinical support system that will help to diagnosis diabetes and heart disease at primary stage. In this paper, we have proposed a model to predict diabetes and heart disease more accurately using logistic regression.

Materials and Methods

We have collected both datasets from Kaggle. The diabetes dataset [1] contains 768 instances. Each instance has 9 features. There are 1025 instances and 14 features in the heart dataset [2]. After the collection of the dataset, we applied the following steps to build machine learning models and predict accuracy.

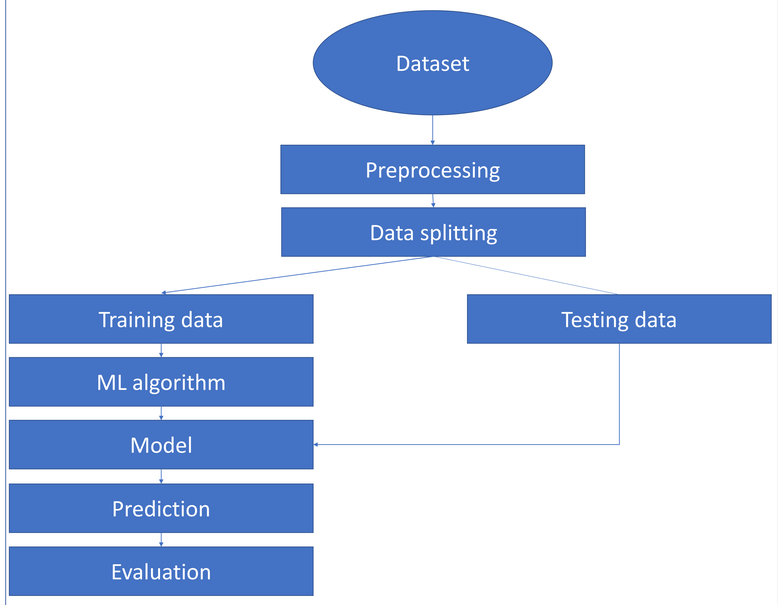

Figure 1. Flow diagram of working method

We have evaluated several machine learning models trained with algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Multinomial Naïve Bayes, K Nearest Neighbors, and Stochastic Gradient Descent algorithm to predict diabetes and heart diseases.

The outcomes of the experiment are shown in the result section.

Results

We have applied multiple machine learning algorithms to predict diabetes and heart disease. For evaluation of the model, we have considered cross validation result, test score, precision, recall and F1 score. According to these parameters, we have found that Logistic regression model is the winner. It provides highest accuracy (best cv score and test score) for both diabetes and heart disease prediction.

For diabetes prediction, Logistic Regression model gives 78% accuracy in cross validation (10-Fold CV) and 77% accuracy in test result.

For heart disease prediction, same model gives 82% and 87% accuracy respectively.

Model Comparison

Diabetes Prediction	Cross validation score	Test score	Precision	Recall	F1 score
Logistic Regression	**78**	**77**	**74**	**74**	**74**
Decision Tree	69	66	63	64	63
Random Forest	73	65	63	64	63
SVM	77	68	65	65	65
MNB	60	58	54	54	54
KNN	78	76	74	73	73
SGD	62	64	56	54	52

Heart Disease Prediction	Cross validation score	Test score	Precision	Recall	F1 score
Logistic Regression	**82**	**87**	**90**	**86**	**86**
Decision Tree	75	75	76	76	75
Random Forest	82	84	83	83	83
SVM	80	87	89	86	86
MNB	74	75	75	75	75
KNN	**61**	**52**	**53**	**53**	**52**
SGD	**61**	**54**	**27**	**50**	**35**

Histogram of cross validation result and test result

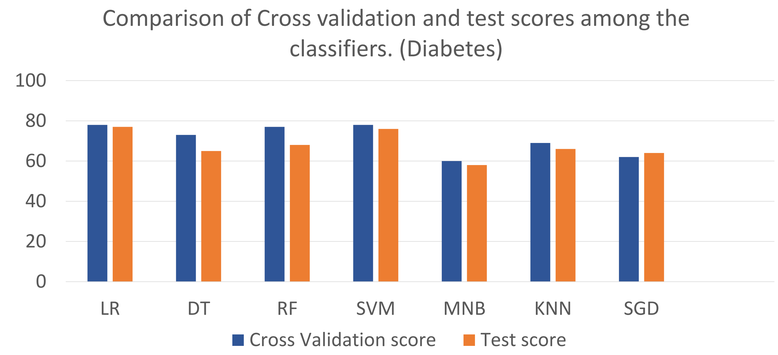

Figure – Diabetes Prediction

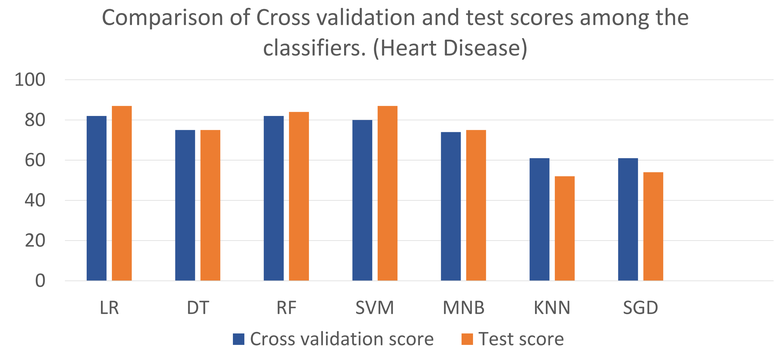

Figure – Heart Disease Prediction

Conclusions

In this study, various machine learning algorithms are applied to the datasets and the classification has been done using various algorithms of which Logistic Regression gives the highest accuracy of 77% and 87% respectively for diabetes and heart disease prediction.

Clearly, our machine-learning models can predict diabetes and heart disease at the primary stage. It will definitely help in the diagnosis of diseases early and will increase the chance of recovery. Further, this study can be extended to find how likely non-diabetic and people with no heart disease at present can have the risk of diabetes and heart disease in the next few years.

Contact

Abu Kowcher (contact.abukowcher@gmial.com)
Md Rezaul Karim (rkarim1506@gmail.com)
Md. Nurul Alam (mdnurulalam6005@gmail.com)
Rangamati Science and Technology University.

References

1. https://www.kaggle.com/datasets/kandij/diabetes-dataset
2. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset