heart disease detection

Shreshtha Kumar Gupta
20HCS4159

Introduction

I would like to present a project that utilizes machine learning algorithms to predict heart disease. The project uses a dataset containing information about various factors that contribute to heart disease such as age, sex, cholesterol levels, and more.

The project imports necessary libraries such as NumPy and Pandas to manipulate and analyze the dataset. Then, the dataset is divided into training and testing sets using the train_test_split method from scikit-learn.

After this, the Logistic Regression algorithm is applied to the training set to create a model for heart disease prediction. Finally, the accuracy of the model is measured by comparing the predictions made by the model to the actual outcomes in the testing set.

Dataset

The Kaggle Heart Disease Prediction dataset is a dataset containing patient data from the Cleveland Clinic Foundation. The dataset consists of 14 features, including age, sex, cholesterol level, and presence of various heart disease risk factors such as hypertension and diabetes. The target variable is a binary classification of whether or not the patient has heart disease. The dataset contains 1025 instances and has been preprocessed to remove missing values and outliers. The data has been normalized and scaled to ensure that each feature has equal importance in the prediction model. This dataset has been used in many machine learning studies to develop models that can accurately predict the presence of heart disease in patients based on their demographic and clinical features. The dataset provides an excellent opportunity for researchers to develop predictive models for heart disease that can assist doctors in diagnosing the disease and providing appropriate treatment to patients.

Model selection

Logistic Regression is a statistical method used to analyze and classify data when the dependent variable is categorical. In other words, it is used for classification problems where the outcome variable is binary (0/1), such as predicting if a person will buy a product or not, if an email is spam or not, or in this case, if a patient has a heart disease or not.

In our project, the objective is to build a model that can predict whether a patient has heart disease or not based on several features such as age, sex, cholesterol level, and so on. Therefore, Logistic Regression is a suitable model for this problem as it can classify patients into two groups (those with heart disease and those without) based on the values of the input features.

Model Training

In our project, the logistic regression model is trained using the heart disease dataset. The dataset is first split into training and testing sets, with 80% of the data used for training and 20% used for testing. The training data is then used to fit the logistic regression model, which learns the relationships between the independent variables (features) and the dependent variable (presence of heart disease). Once the model is trained, it is evaluated on the testing data to assess its performance and accuracy.

Model Testing

For testing our model we use the accuracy score metric and on feeding the x_train_prediction and y_train data we get to see the accuracy score as 85.24% while on feeding the x_test_prediction and y_test data we get the accuracy as 80.48% as we can see that there is not huge difference between both the scores so we can say that our model is decent.

Accuracy score is a metric used to evaluate the performance of a classification model. It represents the percentage of correct predictions made by the model out of all the predictions made.

Conclusion

At last we have finally trained and tested our model so now it is ready to make the prediction and show us the result, one thing to keep in mind that 0 means that the particular individual is not having any disease and 1 means he/she is having some heart disease.