Prediction of Heart Disease Using a Combination of Machine Learning and Spark

Introduction

The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time. In this paper a machine learning algorithm along with Apache Spark for processing large datasets is applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. The dataset consists of 12 main attributes used for performing the analysis. Various promising results are achieved and are validated using accuracy and confusion matrix. Using the Logistic Regression approach, 84.8% accuracy was obtained.

The proposed approach is implemented using ML-libs and Scala language on Apache Spark framework; MLlib is Apache Spark's scalable machine learning library. The key challenge in ECG classification is to handle the irregularities in the ECG signals which is very important to detect the patient status. Therefore, we have proposed an efficient approach to classify ECG signals with high accuracy. Each heartbeat is a combination of action impulse waveforms produced by different specialized cardiac heart tissues. Heartbeats classification faces some difficulties because these waveforms differ from person to person, they are described by some features. In our case, we have used a dataset with 918 records to evaluate the performance of our approach.

The model built by this project will be able to predict whether a patient is having any heart disease. With help of PySpark and Hadoop we can analyse the data of very large patients very efficiently. This will also enable the researchers to predict the overall health of a large population based on the results generated by this machine learning model.

Method

As we are processing the large amount of data, we used big data tools such as Hadoop from which we are accessing the dataset along with spark to statistical analysis on the dataset. Then we applied the machine learning model in jupyter notebook. We used binary logistic regression algorithm for project.

Method of approach for this is Logistic Regression. It is a statistical method used to analyse the relationship between a binary dependent variable and one or more independent variables. It is a type of regression analysis that is used when the dependent variable is binary, meaning it can only take two possible values, usually represented as 0 and 1.

The logistic regression model estimates the probability of the dependent variable taking the value of 1, given the values of the independent variables. It does this by fitting a sigmoidal curve to the data, which maps the independent variables to the probability of the dependent variable being 1. The curve is defined by the logistic function, which is a special type of sigmoid function that outputs values between 0 and 1.

This regression model can be trained using maximum likelihood estimation, which involves finding the parameters that maximize the likelihood of observing the data given the model. There are also various regularization techniques that can be used to prevent overfitting and improve the generalization performance of the model.

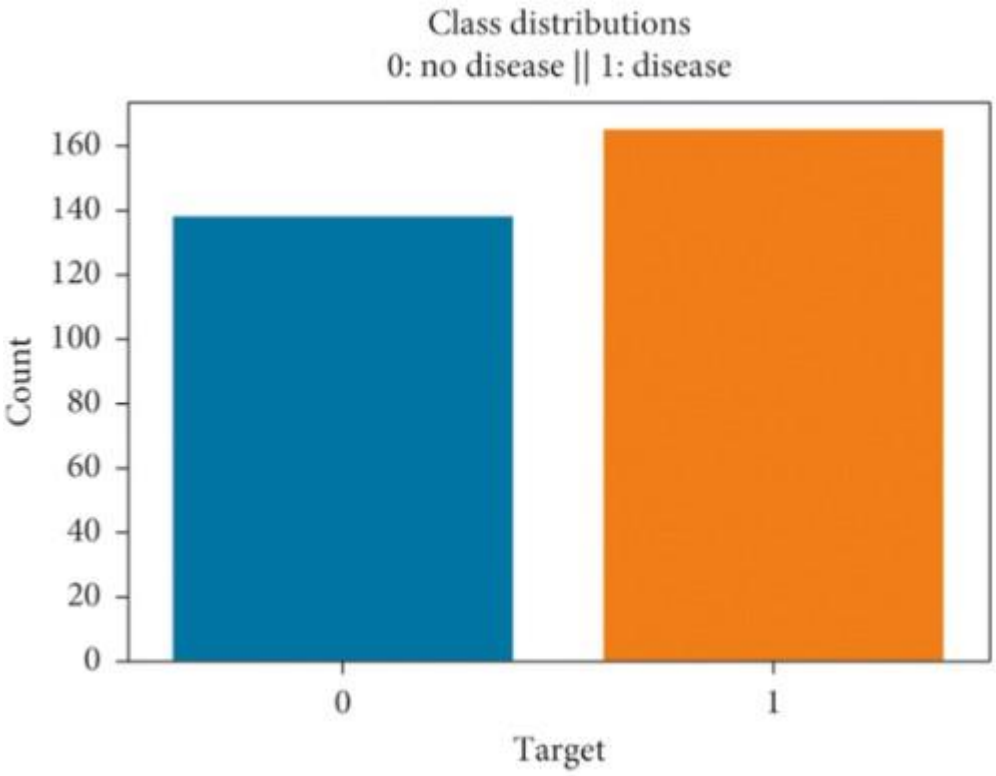
Using machine learning the model is trained on 75% of the given input dataset. The rest is used for the testing of the trained model. The overall accuracy of 84% is achieved.

Dataset Details

We have a dataset of details of patients who have heart disease or not based on the features in it. The dataset contains 12 attributes. The “target” attribute refers to the presence of heart disease in the patient. It is integer-valued 0 = no disease and 1 = disease. The first four rows and all the dataset features are shown in Table 1 without any pre-processing. Now the attributes which are used in this project purpose are described as follows and for what they are used or resemble:

- **Age**—age of patient in years, sex—(1 = male; 0 = female).
- **Cp**—chest pain type.
- **Trestbps**—resting blood pressure (in mm Hg on admission to the hospital). The normal range is 120/80 (if you have a normal blood pressure reading, it is fine, but if it is a little higher than it should be, you should try to lower it. Make healthy changes to your lifestyle).
- **Chol**—serum cholesterol shows the amount of triglycerides present. Triglycerides are another lipid that can be measured in the blood. It should be less than 170 mg/dL (may differ in different Labs).
- **Fbs**—fasting blood sugar larger than 120 mg/dl (1 true). Less than 100 mg/dL (5.6 mmol/L) is normal, and 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.
- **Restecg**—resting electrocardiographic results.
- **Thalach**—maximum heart rate achieved. The maximum heart rate is 220 minus your age.
- **Exang**—exercise-induced angina (1 yes). Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary artery disease.
- **Oldpeak**—ST depression induced by exercise relative to rest.
- **Slope**—the slope of the peak exercise ST segment.
- **Target** (T)—no disease = 0 and disease = 1, (angiographic disease status).

Some attributes contain string inputs so we have to convert them into integers as the machine learning model only takes integers or floats as parameters. The distribution of the data plays an important role when the prediction or classification of a problem is to be done. We see that heart disease occurred 54.46% of the time in the dataset, whilst 45.54% was no heart disease. So, we need to balance the dataset or otherwise it might get overfit. This will help the model to find a pattern in the dataset that contributes to heart disease and which does not as shown in Figure.



To increase the accuracy of the model we have given equal amount of both the results in the dataset so that the model can predict the logic behind accurately.

In the first phase the model is being trained under the given dataset and after that testing data can be passed through the trained model to predict whether the patient has heart disease or not.