Analysis and Prediction of Heart Disease

Bertan Güneş / July 2019

1. Introduction :

The aim of this project is to explore the Heart Disease UCI data set and, the construction of machine learning and deep learning model that predicts whether or not the patient has heart disease through the variables in this data set.

1.1. Business Problem

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects you're born with (congenital heart defects), among others.Heart disease is usually based on signs, symptoms and physical examination of the patient. There are several factors that increase the risk of heart disease. If we look at the risk factors related to heart disease ;

Main factors that cannot be changed :

- Age
- Gender
- Heredity

Main factors that can be changed :

- Smoking
- High cholesterol
- Hypertension
- Physical inactivity
- Overweight
- Diabetes

Other factors such as stress, alcohol, and diet-nutrition.

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. In this study, an effective heart disease prediction system (EHDPS) is developed using neural network for predicting the risk level of heart disease. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The EHDPS predicts the likelihood of patients getting heart disease. It enables significant knowledge, relationships between medical factors related to heart disease and patterns, to be established.

2. Data :

This database contains 76 attributes, but all published experiments refer to using a subset of 1 4 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to t his date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

1. age: The person's age in years
2. sex: The person's sex (1 = male, 0 = female)
3. cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
4. trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
5. chol: The person's cholesterol measurement in mg/dl
6. fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
7. restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave ab normality, 2showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. thalach: The person's maximum heart rate achieved
9. exang: Exercise induced angina (1 = yes; 0 = no)
10. oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on th e ECG plot. See more here)
11. slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
12. ca: The number of major vessels (0-3)
13. thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defe ct)
14. target: Heart disease (0 = no, 1 = yes)

Dataset has information about 303 people. There are 13 features which are made available an d there is one target variable.

2.1. Data Requirements :

Exploration of the data set will help data scientists to transform data and , historical data abo ut hearth disease can be used as the basis to build and train prediction algorithms.What kind of data do we need to make prediction?Preprocessing of the data set is necessary to predict target variable and , so me health related information such as people’s behavior, habits and genetic factors as well.

2.2. Data Cleansing :

After the exploration of data set, data cleansing is a major factor to predict correctly.Heart dis ease data set has not null values and, some records were not categorized.Our data set has not missing fe atures data and, missing classification data.As a result , we have a good quality data set to continue our d ata science process.

3. Methodology :

To explain and identify the problem and resolve medical objectives, different data science tec hnique, which interpret the medical goals have been implemented to predict the heart disease and to im prove the success standards of the algorithms for prediction. Suitable machine learning algorithms, like: R andom Forest, SVM (Support Vector Machine), Gradient Boosted Tree, Naive Bayes, Decision tree and M ultilayer Perceptron were preferred for the training and implementation in python for developing and ev olving the predictive model. These algorithms executed on the model will help medical experts to predict and diagnose heart attacks in the patient dataset. The main goal is to identify which machine-learning alg orithm has the best accuracy for the prediction of heart disease from the patient dataset.

3.1. Feature Engineering :

After the data cleansing is done, the next step in our process is feature creation.In feature eng ineering process :

- One Hot Encoding to the categorical integer features.These are sex, cp, fbs, restecg, exang, slope, ca ve thal.
- Merging all features and one hot encoded features into the single feature vector
- Creating a Feature Engineering ML pipeline.
- Creating the normalized features and drop all the other columns from our dataset and features vector except target and features.

3.2. Model Definition :

Selecting the model which is most appropriate for our usecase the model that is most appropriate for accurately predicting the possibility of a prediction of heart disease data set.Target variable or label gives us a binary value of 0 or 1 which a patient does not have heart disease and have heart disease.Selection of Machine learning and deep learning algorithms can be changed according to data set features and train and test accuracy.

Our model is binary classification .Hence, supervised machine learning algorithm will be best suited for binary classification.The list for our classification use case are listed:

- Gradient Boosted Tree
- Support Vector Machine(SVM)
- Random Forest
- Naive Bayes
- Decision Tree

Based on heart disease data set, feed forward neural network(multi-layer perceptron) is a binary linear classifier and suitable deep learning algorithm to predict label.

3.3. Building prediction model :

Before starting the process of building model , splitting the data into training and test data set is 80% training data and 20% test data. The selected framework is Apache Spark machine learning framework.After that, classification algorithms are implemented and, the algorithm accuracy comparison is created.

3.4. Model Training :

In this step, training machine learning and deep learning models on SparkML using the training data set is processed.After training of the model, we have measured the performance for classification model using the data set.The training accuracies of our models :

- Gradient Boosted Trees
  Training Accuracy : %97

- Random Forest Classification
  Training Accuracy : %93

- Naive Bayes Algorithm
  Training Accuracy : %84

- Decision Tree Algorithm
  Training Accuracy : %92

- Linear Support Vector Machine
  Training Accuracy : %87

- Feed Forward Neural Network
  Training Accuracy : %84

Prediction models are giving us a good accuracy.After that, our models will be evaluated.

3.5. Model Evaluation :

Evaluation of the model is processed by using test data set.After that , measuring the accuracy of our models against the test data set.The validation accuracy measured for these models are listed :

Gradient Boosted Trees
Validation Accuracy : %73

Random Forest Classification
Validation Accuracy : %84

Naive Bayes Algorithm
Validation Accuracy : %89

Decision Tree Algorithm
Validation Accuracy : %73

Linear Support Vector Machine
Validation Accuracy : %82

Feed Forward Neural Network
Validation Accuracy : %84

So Naive Bayes is the best here and Random Forest is the 2nd best and, Decision tree and GBT are the worst algorithms.

4. Results and Discussion :

After data exploration, cleansing, etl, modeling, training and evaluating process , the most efficient supervised machine learning algorithm is Naive Bayes (%89 validation accuracy).Feed forward neural network's (Multi Layer perceptron) validation accuracy is %84.Based on the results we can confirm that the performance of prediction models is good.For future processes, cross validation and model deployment will be implemented in this project.