HEART DISEASE PREDICTION

Subject Name	Big Data Fundamentals - Data Storage Networking
Subject Teacher	Parisa Naraei

Submitted Date	2 November 2019
Submission By	
Jinal Shah	763513
Tom Joseph	760915
Pankaj Narang	761237
Dolly M Shah	764495
Gurpreet Kaur	763505

Lambton College In Toronto

CBD 2214 -1

Table of Contents

1. Abstract 2
2. Introduction 2
3. Data Preparation 2
3.1. Load/Query data 2
3.2. Clean data 3
3.2.1. Replacing '?' with NaN (Not a number) 3
3.2.2. Filled missing values with median 3
3.2.3. Checking for outliners 3
4. Feature Engineering 3
5. Feature Construction 5
6. Dimension Reduction 5
6.1. Feature Selection 5
7. Data Modelling 5
7.1. Supervised Learning Classification 5
8. Performance Measure 6
8.1. Predicting your dataset models 6
8.1.1. Training set and testing set 6
8.2. Choosing the right performance indicator 6
8.2.1. Accuracy 6
8.2.2. Confusion matrix 6
8.2.3. ROC Curve 7
9. Result 8
10. References 8

1. Abstract

It can be said that cardiovascular disease is one of the most vital human illnesses in the world and influences human existence very badly. In this project, we are going to develop a machine-learning-based diagnosis system for heart disorder prediction using Cleveland heart sickness dataset. The proposed system can effortlessly discover and classify people with heart disorder from healthy people and is helping doctors in analyzing heart sufferers efficiently.

2. Introduction

The heart sickness has been regarded as one of the most complicated and life threatening human diseases in the world. In this disease, commonly the heart is unable to push the required amount of blood to other components of the body to fulfill the normal functionalities of the body, and due to this, eventually the coronary heart failure occurs. The correct and acceptable diagnosis of the heart sickness risk in patients is integral for decreasing their associated dangers of extreme heart issues and enhancing safety of heart.

In our project we are developing the machine learning model using the random forest algorithm for classification into heart disease or not. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

3. Data Preparation

Data preparation is the procedure of cleaning and re-modelling raw data prior to processing and analysis. It is a vital step prior to processing and often includes reformatting data, making corrections to data and the combining of records sets to enrich data.

3.1.Load/Query data

In this step we load the dataset file (processed.cleveland) and then start by slicing it into different portions for better understanding of data. Slicing of data in our project is omitted as the dataset is very small and limited.

We have used Pandas python library to read the data and add column headers “age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","oldpeak","slope","ca","thal","num".

3.2. Clean data

Cleaning your data capability filtering out the parts you don’t want or want so that you don’t need to appear at or manner them and modifying the parts, you do need in the structure so that you can right use them. It also includes dealing with lacking values and outliers.

3.2.1. Replacing ‘?’ with NaN (Not a number)

```
print("> Replace empty values with NaN and meaningful value")
dataframe_re = dataframe.replace(to_replace ="?",value ="NaN")
```

3.2.2. Filled missing values with median

Missing values were filled using the imputer function in python. It replaced all NaN vales with the median value.

```
from sklearn.preprocessing import Imputer
imp = Imputer(missing_values='NaN', strategy="median",axis=0)
imp = imp.fit(dataframe_re)
imp_df = imp.transform(dataframe_re)
```

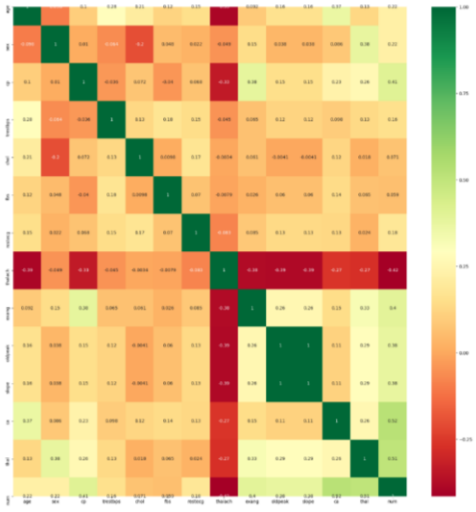
3.2.3. Checking for outliners

We have checked for outliners but have not removed any data.

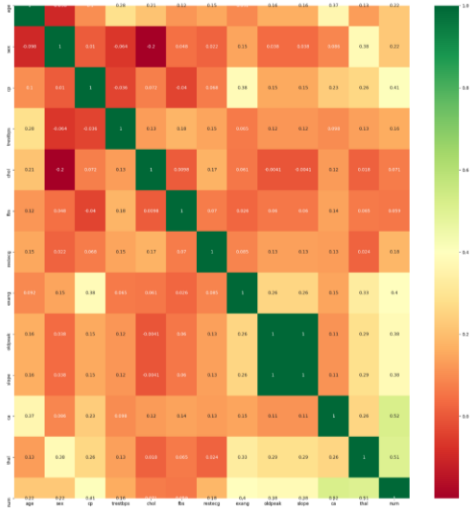
4. Feature Engineering

Feature engineering is the process of using domain understanding of the information to create features that make machine learning algorithms work. A feature is an attribute that might help to find solution to fix the problem.We have done the feature engineering using the heat map correlation matrix of attribute and removed the “thalach”.

Output of heat map correlation matrix



Output after removing “thalach” column



5. Feature Construction

Feature construction is the application of a set of constructive operators to a set of existing features resulting in construction of new features.

We have added a column named “heatdisease” having values 0 or 1 based on “num” column.

```
print("Convert and categorise num into 0 or 1 and remove num")
heartdisease_map = {0:0,1:1,2:1,3:1,4:1}
df['heartdisease'] = df['num'].map(heartdisease_map)
```

6. Dimension Reduction

6.1. Feature Selection

Feature decision improves the classification accuracy and reduces the model execution time. For characteristic resolution in our system, we have used correlation heat map algorithm to remove “thalach” column.

7. Data Modelling

7.1. Supervised Learning Classification

Classification problem is when your output is a category. For example, disease or no disease. We used classification method in machine learning algorithm.

We have used a **random forest classifier**, which is a Meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [5].

8. Performance Measure

8.1. Predicting your dataset models

8.1.1. Training set and testing set

In the predicted model we have divided the dataset into 80 – 20% where 80% of the dataset is given as training data and 20% as test data since we have a small data set.

8.2. Choosing the right performance indicator

8.2.1. Accuracy

Accuracy is measured by calculating number of correctly predicted labels/total number of labels in test set.

```
accuracy = rf.score(x_test,y_test)*100
print("Random Forest Algorithm Accuracy Score : {:.2f}%".format(accuracy))
```

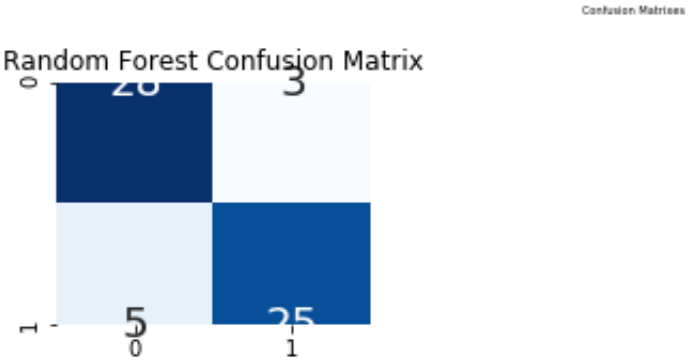
```
Training Random Forest Classification
Random Forest Algorithm Accuracy Score : 86.89%
```

In our model we have 86.89% accuracy

8.2.2. Confusion matrix

A confusion matrix is table is used to demonstrate performance of classification model on a set of test data for which values are known. It allows the visualization of the performance.

```
Plotting the confusion matrix
[[28  3]
 [ 5 25]]
```



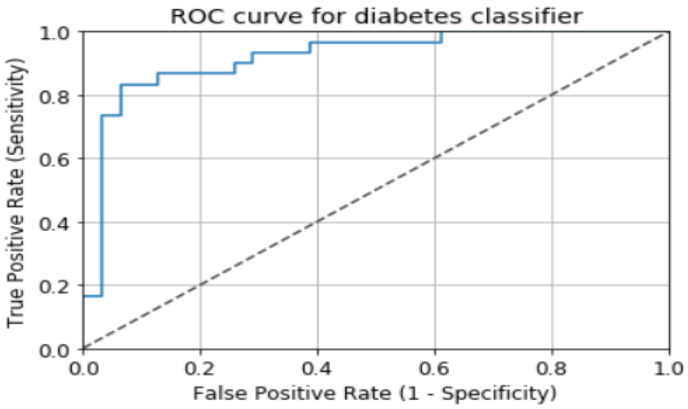

```
Calculating the sensitivity and Specificity
Sensitivity :  0.8484848484848485
Specificity :  0.8928571428571429
```

```
print("Calculating the sensitivity and Specificity")
total=sum(sum(cm_rf))
sensitivity = cm_rf[0,0]/(cm_rf[0,0]+cm_rf[1,0])
print('Sensitivity : ', sensitivity )
specificity = cm_rf[1,1]/(cm_rf[1,1]+cm_rf[0,1])
print('Specificity : ', specificity)
```

8.2.3. ROC Curve

In a **ROC curve** the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter.

A ROC curve demonstrates several things such as it shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.



9. Result

We were able to create a machine learning model with 86.89% accuracy. The machine-learning-based diagnosis system for heart disorder can effortlessly discover and classify people with heart disorder from healthy people and will help doctors in analyzing heart sufferers efficiently.

10. References

1. https://www.hindawi.com/journals/misy/2018/3860146/#B12
2. https://www.talend.com/resources/what-is-data-preparation/
3. https://en.wikipedia.org/wiki/Feature_engineering
4. https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_303
5. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html