CMPE 407

Heart Disease Prediction

BY:
MERVE BAYER 11575008

OZAN ER 117200105

SUPERVISED BY:
ÖZGÜR ÖZDEMİR

ISTANBUL BILGI UNIVERSITY

Contents

List of Figures

List of Tables

Heart Disease

Merve Bayer
Ozan Er

May 1, 2020

1 Introduction

In this project, it is investigated a possible data-science application to predict the presence of heart disease in the patient. It is important to early diagnosis for successful treatment. Also earlier detection increases the chance of surviving. The aim of the project is exploring the person has a heart disease or not. In that way, this model can be used in hospital to gain time and prevent wasting money during diagnosis.

In this dataset, classification method which is a supervised learning algorithm is used. To find the best model, Random Forest Classifier, Decision Tree, Support Vector Machine, Logistic Regression, Naive Bayes and K-Nearest Neighbors are experienced.

2 Presentation of The Data

Heart disease dataset contains 14 attributes concerning heart disease diagnosis. The ”target” field refers to the presence of heart disease in the patient. The data was collected from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation and the creator is Robert Detrano, M.D., Ph.D..

2.1 Data Source

The dataset has been discovered on Kaggle web site:
”https://www.kaggle.com/ronitf/heart-disease-uci” [8]

The description indicates that this database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by Machine Learning researchers to this date.The source claims that more details on their current and past projects on related topics are available on:
”https://archive.ics.uci.edu/ml/datasets/Heart+Disease”. [9]

Dataset is provided as a single file with size 11.328 bytes in CSV format.

2.2 Data Structure

Dataset contains both numerical and categorical input variables. It consists of 303 rows and 14 columns including the target column. Numerical values are age, trestbps, chol, thalach and oldpeak.

Feature *age* is age in years.

Feature *thalach* is maximum heart rate achieved.

Feature *chol* is serum cholestoral in mg/dl.

Feature *trestbps* is resting blood pressure (in mm Hg on admission to the hospital).

Feature *oldpeak* is ST depression induced by exercise relative to rest.
Categorical variables are sex, cp, fbs, exang, slope, thal, restecg and target.

Feature *sex* is a nominal variable (dichotomous), (1 = male; 0 = female).

Feature *exang* is a nominal variable, exercise induced angina (1 = yes; 0 = no).

Feature *slope* is an nominal variable, the slope of the peak exercise ST segment; value 1: upsloping, value 2: flat, value 3: downsloping.

Feature *restecg* is a nominal variable, resting electrocardiograph results (values 0,1,2). 0 = normal; 1 = ST-T wave abnormality; 2 = left ventricular hypertrophy.

Feature *ca* is an ordinal variable number of major vessels (0-3) colored by flourosopy.

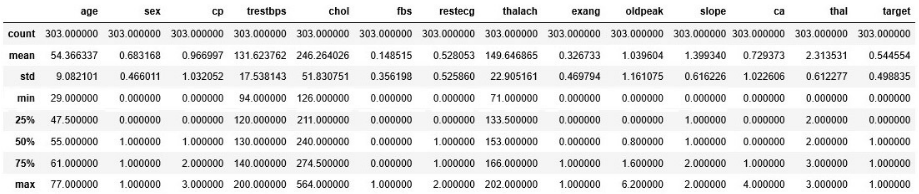
Feature *cp* is a nominal variable, experienced chest pain type as typical angina, atypical angina, non-anginal pain, asymptomatic.

Feature *thal* is a nominal variable 3 = normal; 6 = fixed defect; 7 = reversible defect.

Feature *fbs* is a nominal variable (dichotomous), (fasting blood sugar >120 mg/dl) (1 = true; 0 = false).

Feature *target* is the target variable for the prediction; a value of 1 represents the presence of heart disease, otherwise it takes a value of 0. It is dichotomous variable.

According to heatmap (see in Figure 2), thalach, cp and slope are the most correlated features with target. However, these are only individual correlation.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figure 1: Features Description

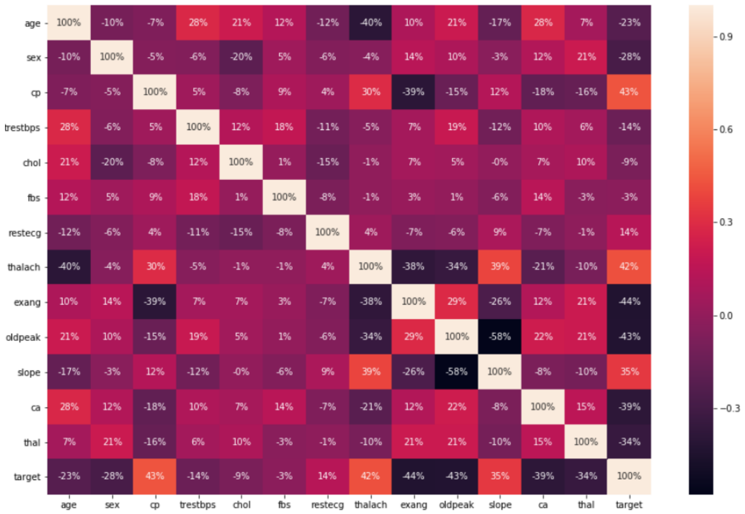
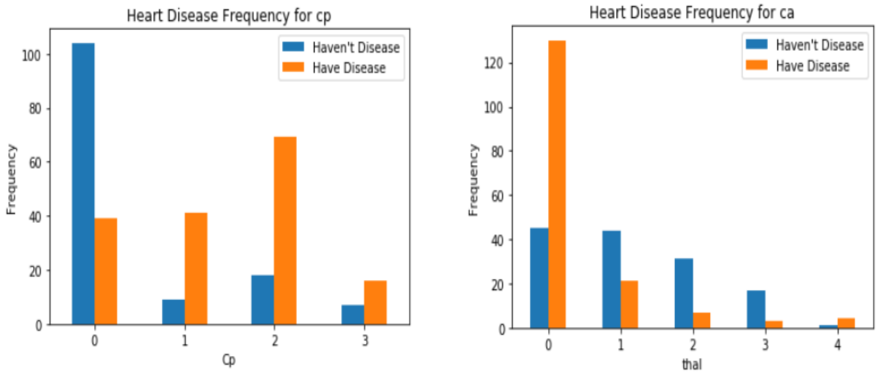


Figure 2: Heatmap

2.3 Visualization

In dataset, there are 4 chest pain type and distribution of types are 143 typical angina values, 50 atypical angina values, 87 non-anginal pain values and 23 asymptomatic values. As seen in the Fig. 3a, chance of having disease is less if sample has typical angina. Chance is getting higher when the sample has non-anginal pain. However, since there are 87 values, it cannot be claimed that possibility of having disease on non-anginal has much bigger than atypical angina nor asymptomatic.



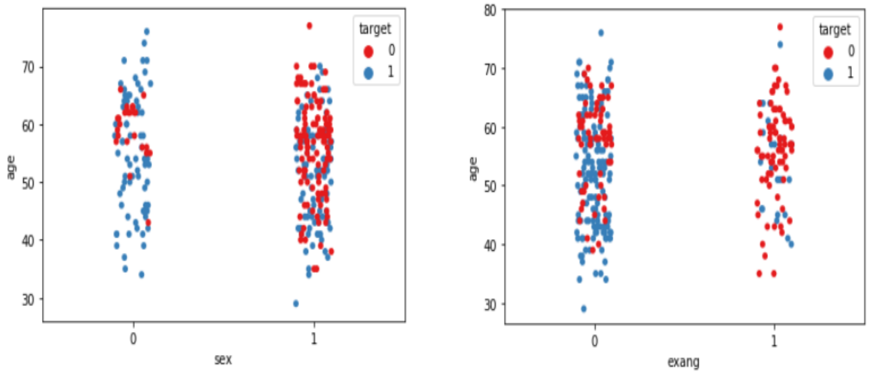
(a) cp Distribution

(b) ca Distibution

Figure 3: ca and cp Distribution

Ca is number of major vessels (0-3) colored by flourosopy. There are 175 zero values, 65 one values, 38 two values, 20 three values and 5 four values in dataset. According to Fig. 3a, it can be claimed that if number of major vessel is 0, chances of having disease is much bigger.

As shown in Fig. 4a, number of males are much bigger than number of females. However, number of samples that have heart disease is too much on female class. The ratio is bigger in female class, 75% of females are sick. In addition, age is not effective. It seems like chance of sickness is less between 55 and 70, but it is probably just because of sample selections.



(a) Sex and Age Relation On Target

(b) Exang and Age Relation On Target

Figure 4: Age Relations on Target

As seen in Fig. 4b, distribution of exang is not proportional. There are 204 zero values and 99 one values in dataset. Again age does not have a big effect on exang-target relation. However, if exang is zero, chance of having disease is much bigger. Easily can say exang and target is correlated.

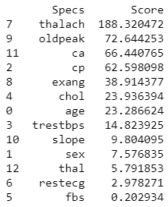
3 Objective

The objective in this project is to create a model which predicts whether the patient has a heart disease or not. The model will make this prediction based on the learned relationships between the feature “target” and the rest of the features which are best for all models individually present in this dataset. If an acceptable accuracy is achieved this model can be used in hospitals to perform early diagnosis before it is too late.

4 Preprocessing

Heart disease dataset is well prepared. It does not need any label encoding and also data does not have any NaN values.

4.1 Feature Selection



```
      Specs       Score
7    thalach  188.320472
9   oldpeak   72.644253
11       ca   66.440765
2        cp   62.598098
8     exang   38.914377
4      chol   23.936394
0       age   23.286624
3  trestbps   14.823925
10    slope    9.804095
1       sex    7.576835
12     thal    5.791853
6   restecg    2.978271
5       fbs    0.202934
```

Figure 5: Chi-squared

In this dataset, to see the correlation between features chi-squared test is done and get top 10 best features. As seen in the Figure 5, thalach has the most correlation. However, for all classifiers to train model, SelectFromModel function is used. In that way, all models use the best features based on their own algorithms.

For decision tree, random forest classifier, logistic regression and support vector machine classifier sklearn.feature_selection [5] module is used.

For K-Nearest Neighbors and Naive Bayes classifiers the best features of logistic regression are used, because performance of models are calculated as higher compared to selecting manually.

5 Experiments

Supervised learning method is used in this project. The aim is predicting the sample has a heart disease or not, so there are two classes. First one is "no" and shown as "0", and second one is "yes" and shown as "1". Therefore, this is a classification problem which includes two classes. Decision tree, Random Forest, SVM, Logistic Regression, KNN and Naive Bayes algorithms are experienced to find the best model.

5.1 Decision Tree

Decision tree chooses the best features and divides the inputs into smaller decisions (nodes) and repeats until find the leaf nodes in all branches. In other words, "it formulates some set of rules. These rules can be used to perform predictions." [7]

Decision tree is one of the choices, because it helps to understand the logic behind the data.

Inbuilt class feature_importances is used to see the graphic of feature's importance. Then using SelectFromModel function selects 5 features which are best for the model. These are cp, chol, oldpeak, ca and thal. After grid search, best parameters are selected as 'max_depth': 3, 'min_samples_leaf': 2, 'min_samples_split': 2. For criterion 'gini' is used. In the end, after 10-Fold Cross Validation, mean of experiment scores is 0.8217.

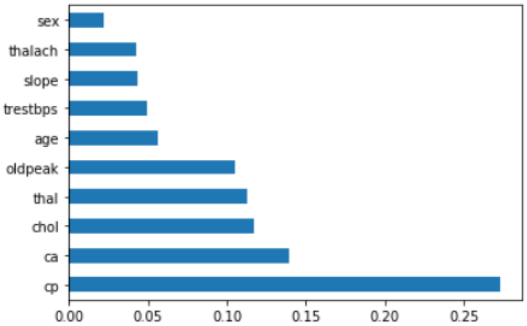


Figure 6: Decision Tree Features

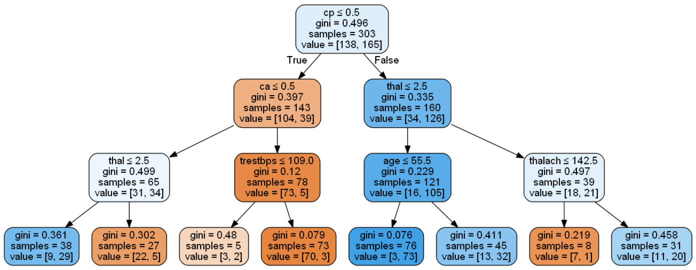

cp ≤ 0.5 gini = 0.496 samples = 303 value = [138, 165]
True
False
ca ≤ 0.5 gini = 0.397 samples = 143 value = [104, 39]
thal ≤ 2.5 gini = 0.335 samples = 160 value = [34, 126]
thal ≤ 2.5 gini = 0.499 samples = 65 value = [31, 34]
trestbps ≤ 109.0 gini = 0.12 samples = 78 value = [73, 5]
age ≤ 55.5 gini = 0.229 samples = 121 value = [16, 105]
thalach ≤ 142.5 gini = 0.497 samples = 39 value = [18, 21]
gini = 0.361 samples = 38 value = [9, 29]
gini = 0.302 samples = 27 value = [22, 5]
gini = 0.48 samples = 5 value = [3, 2]
gini = 0.079 samples = 73 value = [70, 3]
gini = 0.076 samples = 76 value = [3, 73]
gini = 0.411 samples = 45 value = [13, 32]
gini = 0.219 samples = 8 value = [7, 1]
gini = 0.458 samples = 31 value = [11, 20]

Figure 7: Decision Tree

This is the decision tree example without feature selection which maximum depth is 3. Cp is the most important feature, that's why it is a root node.

5.2 Random Forest Classifier

Random Forest algorithm "creates a forest by some way and makes it random." [7]

Random Forest is the other choice because it is wanted to see the performance of model without rules, unlike decision tree.

Inbuilt class feature_importances is used to see the graphic of feature's importance. Then using SelectFromModel function selects 6 features which are best for the model.

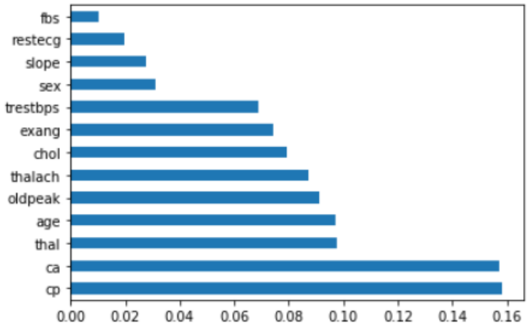


Figure 8: Random Forest Features

criterion='entropy', n_estimators=1000, max_depth=4 are the best parameters for Random Forest Classifier. After feature selection, grid search and cross validation mean of experiment scores is 0.8152.

5.3 Support Vector Machine

Support Vector Machine aim is finding a separating hyperplane between two classes."Best hyperplane is the one that represent the largest separation between two classes." [6]

The model is chosen to discover whether the data can be separable such points.

After automatically selecting, SVM uses 6 features and these are sex, cp, exang, oldpeak, ca and thal and it is done 10-Fold cross validation, mean of experiment scores is 0.81505.

5.4 Logistic Regression

"Logistic Regression is used to predict the probability of a categorical dependent variable" [1]. In this project, the target is dichotomous (0=no, 1=yes), and logistic regression uses logit function. That's why logistic regression is the best possible choice for this dataset.

By automatically selecting, model uses 7 features and these are sex, cp, exang, slope, oldpeak, ca and thal. After 10-Fold cross validation, mean of experiment scores is 0.8349. Logistic Regression gets the best accuracy among 6 models and the best parameters are the default ones.

5.5 K-Nearest Neighbors

"KNN is a model that classifies data points based on the points that are most similar to it." [2] To find distance Euclidean Distance formula is used. Unlike k-means, KNN classifies based on similarities and some labels known beforehand. This algorithm is chosen to see accuracy on this dataset and it is easy to use.

For model, 7 best features of logistic regression are used: sex, cp, exang, slope, oldpeak, ca and thal.
Model is experimented with different values of k ranging from 1 to 16. It is picked the best k value as 10 based on accuracy over data while employing a 4 fold cross validation. Algorithm hyperparameter is chosen as "auto". Mean of experiment scores is calculated as 0.8317. KNN is the second best model among these models.

5.6 Naive Bayes

"Naive Bayes works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. By, using the conditional probability, it can be calculate the probability of an event using its prior knowledge." [4].

Fig. 9 is the formula of conditional probability. P(A l B) is the probability of A given B. P(B l A) is the probability of B given A. P(A) is probability of A and P(B) is the probability of B.(Figure 9)

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Figure 9: Conditional Probability Formula

Naive Bayes is another choice, because the goal is finding possibility.

Default hyperparamters are used, since it gives the best accuracy and mean of experiment scores is 0.8186. Accuracy is not low, so it can be said that the data is balanced.

6 Methodology and Result

Model	Accuracy
Random Forest	0.8151612903225807
Decision Tree	0.821720430107527
Support Vector Machine	0.8150537634408602
Logistic Regression	0.8349462365591398
K-Nearest Neighbors	0.8317204301075268
Naive Bayes	0.8186021505376344

Table 1: Results

In this project, 6 classifier models are used to find which one is the best for predicting the having disease by using sklearn library with Python programming language.

These steps are followed: Importing libraries, reading data, checking for missing values, checking for categorical data, visualising, preprocessing feature selection, creating models, doing grid search and using best hyperparameters and features, evaluting model with cross validation.

Classifiers are Random Forest, Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine and k-Nearest Neighbors. The accuracies are 81.5%, 82.2%, 81.9%, 83.5%, 81.5% and 83.2%, respectively.

For Logistic Regression, it is utilized default parameters.

For Decision Tree, maximum depth is chosen as 3.

For KNN, model is experimented with k ranging from 1 to 16. 10 is the best choice. For KNN and Naive Bayes, best features of Logistic Regression are used.

It can be said Logistic Regression is the best model, because the target is dichotomous and Logistic Regression gets better results in binary variable, due to logit function. However, it cannot be said Random Forest or SVM is the worst one. Since in every cross validation, Random Forest creates new forests randomly, therefore accuracy changes between 0.0% and 0.02% in every try.

Cp, thal and oldpeak are most selected features. That means that these 3 features have a big impact on predicting whether having a disease or not.

7 Conclusion

As a conclusion, it is used supervised learning methods. Data is well labeled to create the presence of heart disease model. It is preferred to use Decision Tree, SVM, Random Forest, Logistic Regression and KNN models to find the best accuracy.

In this project for assessing the performance of machine learning models K-Fold Cross Validation is used and ten fold cross validation is performed.
As a result, Logistic Regression has the best accuracy (83.5%).

For improving the result, it can be used 76 features from real data and selects the best features for model among them. For KNN, it can be used other features and k to improve result, it is not certain that Logistic Regression features are best for KNN. After some manual and random selections, Logistic Regression's features were found the best for KNN. For other models, maybe it can be experimented with better hyperparameters.
In addition, in dataset thal is Thallium Stress Test, it is not mentioned in dataset page. More information can be found in healthline website[3].

References

[1] Rajesh S. Brid. https://medium.com/greyatom/logistic-regression-89e496433063.

[2] Rajesh S. Brid. https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26.

[3] healthline. https://www.healthline.com/health/thallium-stress-test.

[4] Adipta Martulandi. https://medium.com/datadriveninvestor/a-gentle-introduction-to-naive-bayes-classifier-9d7c4256c999.

[5] scikit-learn developers. https://scikit-learn.org/stable/modules/feature$_selection.html$.

[6] Morgan Shields. *Research Methodology and Statistical Methods*.

[7] Synced. https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674.

[8] UCI. https://www.kaggle.com/ronitf/heart-disease-uci.

[9] UCI. https://archive.ics.uci.edu/ml/datasets/Heart+Disease.