**A picture containing text, sign

Description automatically generatedA picture containing text

Description automatically generated**

Πανεπιστήμιο Πειραιώς *Τμήμα Ψηφιακών Συστημάτων*

ΕΚΕΦΕ Δημόκριτος

Δι-ιδρυματικό Πρόγραμμα Μεταπτυχιακών Σπουδών στην

Τεχνητή Νοημοσύνη

**MACHINE LEARNING PROJECT**

Αντώνης Παπαχρήστου

Μεταπτυχιακός Φοιτητή

**Introduction**

A heart attack occurs when the flow of blood to the heart muscle suddenly becomes blocked. From WHO statistics every year 17.9 million dying from heart attack. The medical study says that human lifestyle is the main reason behind this heart problem. Apart from this there are many key factors which warns that the person may/may not getting chance of heart attack.

This project will focus on predicting heart disease using machine learning.

We have a dataset which classified if patients have heart disease or not according to features in it. I will try to use this data to create a model which tries predicting if a patient has this disease or not.

We will be using some common Python libraries, such as pandas, numpy, seaborn and matplotlib. Furthermore, for the machine learning side of this project, we will be using sklearn.

**The machine learning methods used in this project are:**

1. Logistic Regression
2. K-Nearest Neighbor (kNN)
3. Support Vector Machine (SVM)
4. Decision Tree
5. Random Forest

**About the Dataset**

I found this dataset (heart.csv) on Kaggle. It has 303 patients and contains 14 variables (features) for each one, 9 categorical and 5 continuous.

The structure of the dataset is given below:

1. age - age in years
2. sex - (male=1, female=0)
3. cp - chest pain type
   * 0: Typical angina: chest pain related decrease blood supply to the heart
   * 1: Atypical angina: chest pain not related to heart
   * 2: Non-anginal pain: typically esophageal spasms (non-heart related)
   * 3: Asymptomatic: chest pain not showing signs of disease
4. trestbps - Resting blood pressure (in mm Hg on admission to the hospital)
5. chol - serum cholesterol in mg/dl
6. fbs - (fasting blood sugar mg/ml)
   * 0: Less than 120 mg/ml
   * 1: Greater than 120 mg/ml
7. restecg - Resting electrocardiographic results
   * 0: Nothing to note
   * 1: ST-T Wave abnormality
     + can range from mild symptoms to severe problems
     + signals non-normal heart beat
   * 2: Possible or definite left ventricular hypertrophy
     + Enlarged heart's main pumping chamber
8. thalach - Maximum heart rate achieved
9. exang - exercise induced angina (yes=1, no=0)
10. oldpeak - ST depression induced by exercise relative to rest looks at stress of heart during exercise unhealthy heart will stress more
11. slope - the slope of the peak exercise ST segment
    * 0: Upsloping: better heart rate with exercise (uncommon)
    * 1: Flat sloping: minimal change (typical healthy heart)
    * 2: Downslopins: signs of unhealthy heart
12. ca - number of major vessels (0-3) colored by fluoroscopy
    * colored vessel means the doctor can see the blood passing through
    * the more blood movement the better (no clots)
13. thal – thalassemia types
    * 1: normal
    * 2: fixed defect: used to be defect but ok now
    * 3: reversable defect: no proper blood movement when exercising
14. target - have heart disease or not (1=yes, 0=no) (= the predicted attribute)

**Data Analysis**

**-Τhe original csv file is:**

**![Table

Description automatically generated]()**

**-Checking the data for missing values (no missing value in the dataset).**

**-Renaming the column headers for better understanding of visualizations.**

**-Checking the dataset for wrong values:**

* **Five patients having number of major vessels=4 which is incorrect (0-3)**
* **Two patients having thalassemia type=0, which is incorrect (1-3)**

**The patients with the wrong values dropped, so the length of dataset now is 296 instead of 303 (296,14).**

**-**Percentage of patients with presence of heart disease is 54.39% and percentage of patients with no presence of heart disease: 45.61%.

- Percentage of male patients is 68.58% and female patients: 31.42%.

**-Let’s see the heart disease frequency based on the categorical features:**

![Chart, line chart

Description automatically generated]()![Chart

Description automatically generated]()![Chart

Description automatically generated]()![Chart, line chart

Description automatically generated]()![Chart

Description automatically generated]()![Chart, line chart

Description automatically generated]()![Chart, line chart

Description automatically generated]()![Chart

Description automatically generated]()

**OBSERVATIONS**

* Chest Pain: People with chest pain value equal to 1, 2, 3 are more likely to have heart disease than people with chest pain value equal to 0.
* Resting electrocardiographic results: People with value 1 (signals non-normal heartbeat, can range from mild symptoms to severe problems) are more likely to have heart disease.
* Exercise induced angina: People with value 0 (No angina) have heart disease more than people with value 1 (angina)
* The slope of the peak exercise ST segment: People with slope value equal to 2 (Down sloping: signs of unhealthy heart) are more likely to have heart disease than people with slope value equal to 0 (Upsloping: better heart rate with exercise) or 1 (Flat sloping: minimal change (typical healthy heart)).
* Number of major vessels (0-3) colored by fluoroscopy}: the more blood movement the better so people value equal to 0 are more likely to have heart disease.
* Thalassemia: People with thalassemia value equal to 2 (fixed defect: used to be defect but ok now) are more likely to have heart disease.

**-Let’s see the heart disease frequency based on the continuous features:**

![Chart, histogram

Description automatically generated]()![Chart, bar chart

Description automatically generated]()![Chart, histogram

Description automatically generated]()![Chart, histogram

Description automatically generated]()![Chart

Description automatically generated]()

**OBSERVATIONS**

* Resting blood pressure: anything above 130-140 is typically cause for concern
* Cholesterol: above 200 is cause for concern.
* Maximum heart rate achieved: People achieved a maximum more than 140 are more likely to have heart disease.
* Oldpeak ST depression induced by exercise relative to rest looks at stress of heart during exercise unhealthy heart will stress more

![Timeline

Description automatically generated]()**Let’s see the Correlation Matrix:**

**In the matrix we can see that that the features “Fasting\_blood\_sugar” and “Cholesterol” are the lowest correlated with the Heart Disease. The other features are more correlated with Heart Disease.**

**Data Processing**

**-Transforming categorical variables using one hot encoding technique. So I have created dummy variables for categorical data “Chest\_pain”, “Resting\_ECG\_results”, “ST-slope”, “Major\_vessels” and “Thalassemia\_types” to increase the accuracy.**

**-Dropping the unnecessary categorical data.**

**-Using the 80% of the dataset for training.**

**-Using the 20% of the dataset for testing.**

**-Scaled the data for better result (MinMaxScaler()).**

**Model Building**

**Logistic Regression**

**From sklearn used GridSearchCV for hyperparameter tuning.**

**For parameter C, I found that C=0.5 is best.**

***Training Results***

**Logistic Regression Train Accuracy: 85.59 %**

**Confusion Matrix**

**[ 90,17 ]**

**[17,112]**

Τhis means that 17 people predicted with heart disease while healthy and 17 people predicted as healthy while they have heart disease.

**Classification Report**

**![Table, calendar

Description automatically generated]()**

***Testing Results***

**Logistic Regression Test Accuracy: 86.67 %**

**Confusion Matrix**

**[ 24,4 ]**

**[4,28]**

Τhis means that 4 people predicted with heart disease while healthy and 4 people predicted as healthy while they have heart disease.

**Classification Report**

**![Calendar

Description automatically generated]()**

**kNN**

**k:** Number of neighbors.

**I tried different values for k (from 1 to 20) and find that the best Test Accuracy was when k=3.**

***Training Results***

**kNN Train Accuracy: 88.56 %**

**Confusion Matrix**

**[ 90,17 ]**

**[10,119]**

Τhis means that 17 people predicted with heart disease while healthy and 10 people predicted as healthy while they have heart disease.

**Classification Report**

**![Calendar

Description automatically generated]()**

***Testing Results***

**kNN Test Accuracy: 88.33 %**

**Confusion Matrix**

**[ 25,3 ]**

**[4,28]**

Τhis means that 3 people predicted with heart disease while healthy and 4 people predicted as healthy while they have heart disease.

**Classification Report**

**![Calendar

Description automatically generated]()**

**SVM**

**From sklearn used GridSearchCV for hyperparameter tuning.**

**For regularization parameter C=0.1.**

For kernel type=polynomial.

5th degree on the polynomial kernel function.

***Training Results***

**SVM Train Accuracy: 91.53 %**

**Confusion Matrix**

**[ 94,13 ]**

**[7,122]**

Τhis means that 13 people predicted with heart disease while healthy and 7 people predicted as healthy while they have heart disease.

**Classification Report**

**![Table, calendar

Description automatically generated]()**

***Testing Results***

**SVM Test Accuracy: 86.67 %**

**Confusion Matrix**

**[ 24,4 ]**

**[4,28]**

Τhis means that 4 people predicted with heart disease while healthy and 4 people predicted as healthy while they have heart disease.

**Classification Report**

**![Table, calendar

Description automatically generated]()**

**Decision Tree**

**From sklearn used GridSearchCV for hyperparameter tuning.**

The minimum number of samples required to split an internal node=4.

The maximum depth of the tree=4.

The minimum number of samples required to be at a leaf node=2.

***Training Results***

**Decision Tree Train Accuracy: 87.71 %**

**Confusion Matrix**

**[ 94,13 ]**

**[16,113]**

Τhis means that 13 people predicted with heart disease while healthy and 16 people predicted as healthy while they have heart disease.

**Classification Report**

**![Calendar

Description automatically generated]()**

***Testing Results***

**Decision Tree Test Accuracy: 80.00 %**

**Confusion Matrix**

**[ 24,4 ]**

**[8,24]**

Τhis means that 4 people predicted with heart disease while healthy and 8 people predicted as healthy while they have heart disease.

**Classification Report**

**![Calendar

Description automatically generated]()**

**Random Forest**

**From sklearn used GridSearchCV for hyperparameter tuning.**

**For criterion=entropy**

**For n\_estimators=15.**

***Training Results***

**Random Forest Train Accuracy: 99.58 %**

**Confusion Matrix**

**[ 107,0 ]**

**[1,128]**

Τhis means that no one predicted with heart disease while healthy and 1 person predicted as healthy while they have heart disease.

**Classification Report**

![Table

Description automatically generated]()

***Testing Results***

**Random Forest Test Accuracy: 88.33 %**

**Confusion Matrix**

**[ 24,4 ]**

**[3,29]**

Τhis means that 4 people predicted with heart disease while healthy and 3 people predicted as healthy while they have heart disease.

**Classification Report**

![Calendar

Description automatically generated]()

**Model Evaluation**

-The accuracy of each model in testing is:

![Table

Description automatically generated with medium confidence]()

-I am using ROC curves and AUC scores to choose the best of the 5 machine learning models:

![Chart, scatter chart

Description automatically generated]()

**AUC Scores**

|  |  |
| --- | --- |
| *LogisticRegression* | 0.940 |
| *KNeighborsClassifier* | 0.895 |
| *SVC* | 0.919 |
| *DecisionTreeClassifier* | 0.874 |
| *RandomForestClassifier* | 0.935 |

AUC can be classed as follows,

* 0.90 - 1.00 = excellent
* 0.80 - 0.90 = good
* 0.70 - 0.80 = fair
* 0.60 - 0.70 = poor
* 0.50 - 0.60 = fail

The Logistic Regression has the best AUC score.