

A SYNOPISIS ON

**HEART STROKE PREDICTION**

Under the Guidance of (Mentor)

**Ms. Aishwarya Saxena**

By: Mr. Sushant Jadhav

Machine Learning - Python

College: Terna Engineering College, Nerul

Summer Internship: June 2020 - July 2020



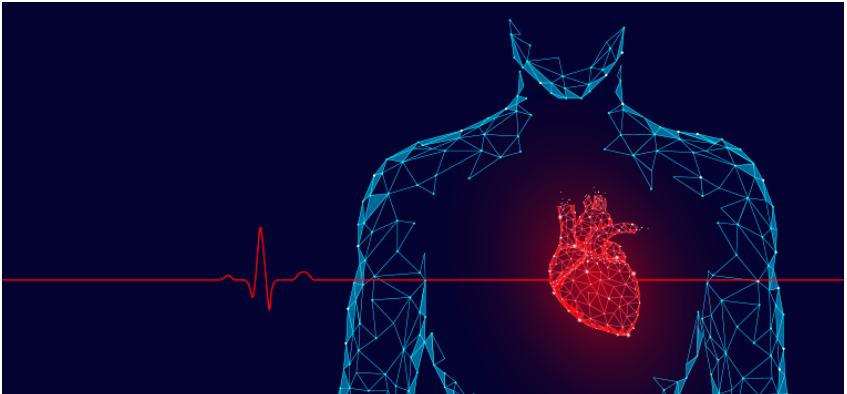
INDEX

|  |  |  |
| --- | --- | --- |
| S no. | CONTENT | PAGE NO. |
| 1 | Introduction | 1 |
| 2 | Objective | 1 |
| 3 | Background | 2-3 |
| 4 | Hardware software requirements | 4 |
| 5 | Coding | 5-7 |
| 6 | Output-Screenshots | 8-10 |
| 7 | Future scope | 11 |
| 8 | conclusion | 11 |
| 9 | Bibliography and references | 11 |

**HEART STROKE PREDICTION**

**INTRODUCTION**

The main topic is prediction using machine learning technics. Machine learning is widely used now a days in many business applications like e-commerce and many more. Prediction is one of area where this machine learning used, our topic is about prediction of heart stroke by processing patient’s dataset and a data of patients to whom we need to predict the chance of occurrence of a heart stroke disease. The aim is to achieve better accuracy and to make the system more efficient so that it can predict the chances of heart stroke.



**OBJECTIVE**

1. Heart stroke prediction by machine learning will reduce the time by predicting the value risk of heart stroke using dataset which include normal information of individual.

2. In this, we are predicting risk of heart stroke without novel calcium scan of human heart but by putting input and calculating the same.

3. Hence, overall analysing time will be reduced, cross validation of feature selection.

4. In the end, Patient who has high risk of heart stroke can be easily diagnosed and will be put under treatment procedure instantly.

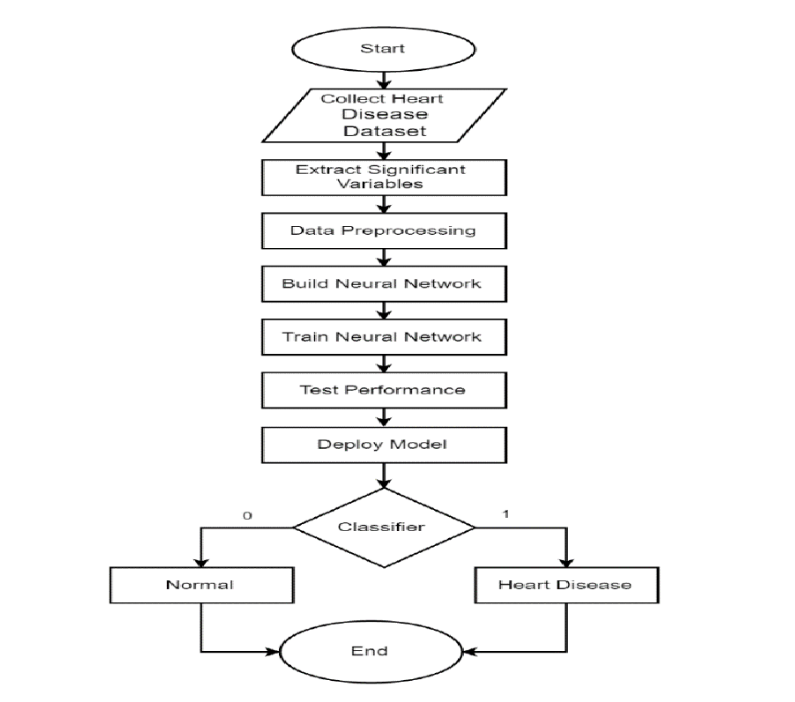
5. Overall test cost will also get reduced as man power gets reduced.

**BACKGROUND**

Data analysis proves to be crucial in the medical field. It provides a meaningful base to critical decisions. It helps to create a complete study proposal. One of the most important uses of data analysis is that it helps in keeping human bias away from medical conclusion with the help of proper statistical treatment. By use of data mining for exploratory analysis because of nontrivial information in large volumes of data.

Heart predictor system will use the data mining knowledge to give a user-oriented approach to new and hidden patterns in the data. The knowledge which is implemented can be used by the healthcare experts to get better quality of service and to reduce the extent of adverse medicine effect.

Data mining has been used in a variety of applications such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile computing. Of late, data mining has been applied successfully in healthcare fraud and detecting abuse cases.



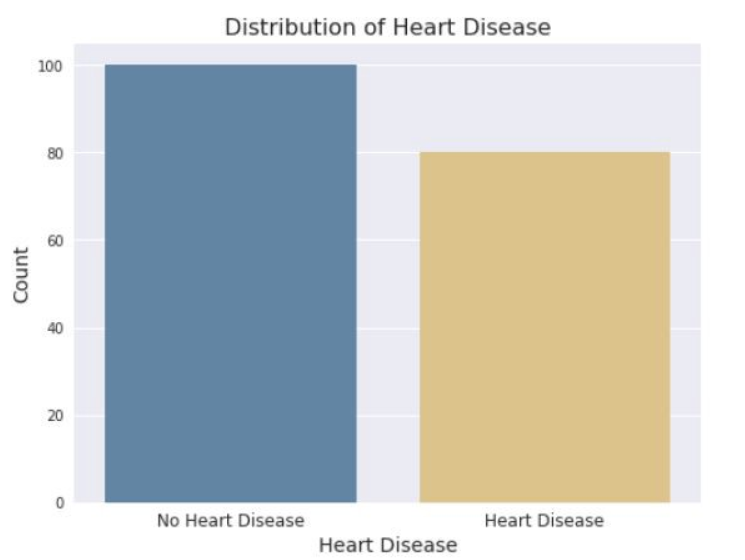
**MODEL TERMINOLOGY**

**K-Neighbours Classifier:**

This classifier looks for the classes of K nearest neighbours of a given data point and based on the majority class, it assigns a class to this data point. However, the number of neighbours can be varied.

**Random Forest Classifier:**

This classifier takes the concept of decision trees to the next level. It creates a forest of trees where each tree is formed by a random selection of features from the total features.



**Data Processing:**

After exploring the dataset, I observed that I need to convert some categorical variables into dummy variables and scale all the values before training the Machine Learning models. First, I'll use the get\_dummies method to create dummy columns for categorical variables.

dataset = pd.get\_dummies(df, columns = ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'])

**HARDWARE AND SOFTWARE REQUIREMENTS**

**Hardware Requirements:**

|  |  |
| --- | --- |
| Hardware Tools | Minimum Requirements |
| Processor | I5 or above |
| Hard Disk | 10 GB |
| RAM | 8 GB |
| Monitor | 17'' Coloured |
| Mouse | Optical |
| Keyboard | 122 Keys |

**Software Requirements:**

|  |  |
| --- | --- |
| Software Tools | Minimum Requirements |
| Platform | Windows, Linux and Mac OS |
| Operating System | Windows, Linux and Mac OS |
| Technology | Machine Learning - Python |
| Scripting Language | Python |
| IDE | Pycharm/ jupyter (notebook) |

**CODE: (Jupyter Notebook)-**

#Importing Modules-

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from matplotlib import rcParams

from matplotlib.cm import rainbow

%matplotlib inline

import warnings

warnings.filterwarnings('ignore')

#Here we are using 2 Algorithms-

#1. KNeighborsClassifier

#2.RandomForestClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

df = pd.read\_csv('datasetheart.csv')

df.info()

df.describe()

#Feature Selection-

import seaborn as sns

#get correlations of each features in dataset

corrmat = df.corr()

top\_corr\_features = corrmat.index

plt.figure(figsize=(20,20))

#plot heat map

g=sns.heatmap(df[top\_corr\_features].corr(),annot=True,cmap="PuBu")

#Histogram-

df.hist(color="orange",edgecolor="red")

plt.rcParams['figure.figsize']=(12,10)

sns.set\_style('whitegrid')

sns.countplot(x='target',data=df,palette='RdPu',edgecolor="darkblue")

plt.rcParams['figure.figsize']=(5,3)

#Data Processing-

dataset = pd.get\_dummies(df, columns = ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'])

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

standardScaler = StandardScaler()

columns\_to\_scale = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']

dataset[columns\_to\_scale] = standardScaler.fit\_transform(dataset[columns\_to\_scale])

dataset.head()

y = dataset['target']

X = dataset.drop(['target'], axis = 1)

from sklearn.model\_selection import cross\_val\_score

knn\_scores = []

for k in range(1,21):

knn\_classifier = KNeighborsClassifier(n\_neighbors = k)

score=cross\_val\_score(knn\_classifier,X,y,cv=10)

knn\_scores.append(score.mean())

plt.plot([k for k in range(1,21)], knn\_scores, color = 'darkblue')

for i in range(1,21):

plt.text(i, knn\_scores[i-1], (i, knn\_scores[i-1]))

plt.xticks([i for i in range(1,21)])

plt.xlabel('Number of Neighbors (K)')

plt.ylabel('Scores')

plt.title('K Neighbors Classifier scores for different K values')

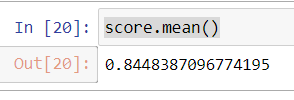
plt.rcParams['figure.figsize']=(21,14)

knn\_classifier = KNeighborsClassifier(n\_neighbors = 12)

score=cross\_val\_score(knn\_classifier,X,y,cv=10)

score.mean()

#OUTPUT-

#Random Forest Classifier

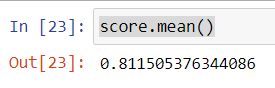
from sklearn.ensemble import RandomForestClassifier

randomforest\_classifier= RandomForestClassifier(n\_estimators=10)

score=cross\_val\_score(randomforest\_classifier,X,y,cv=10)

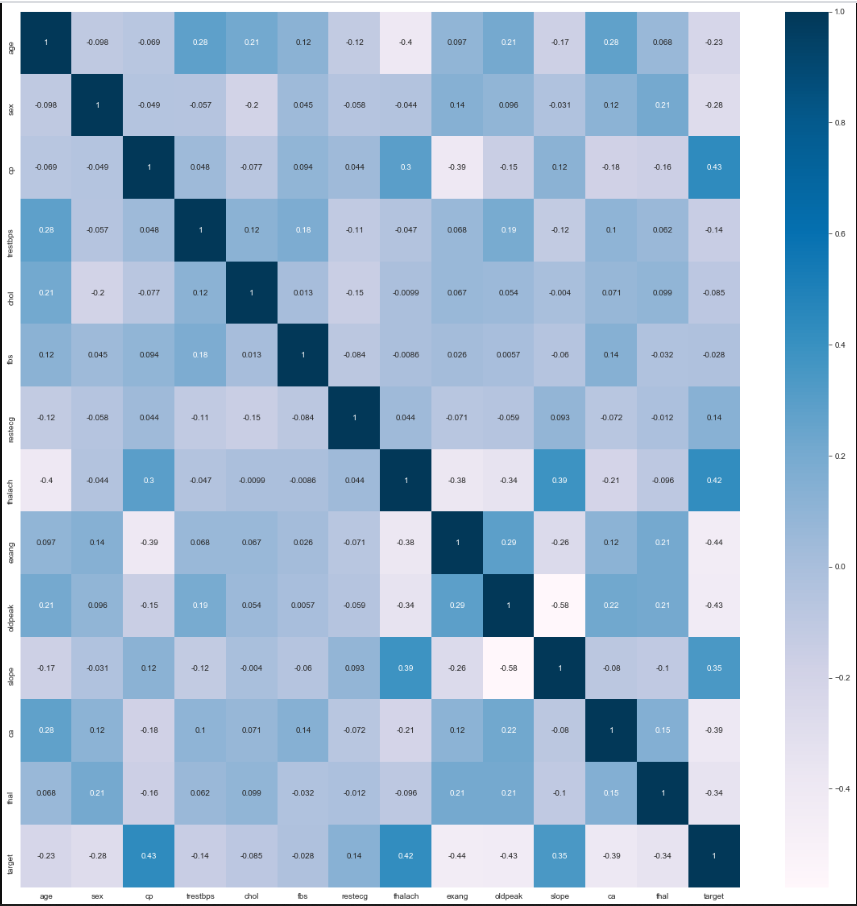
score.mean()

#OUTPUT-

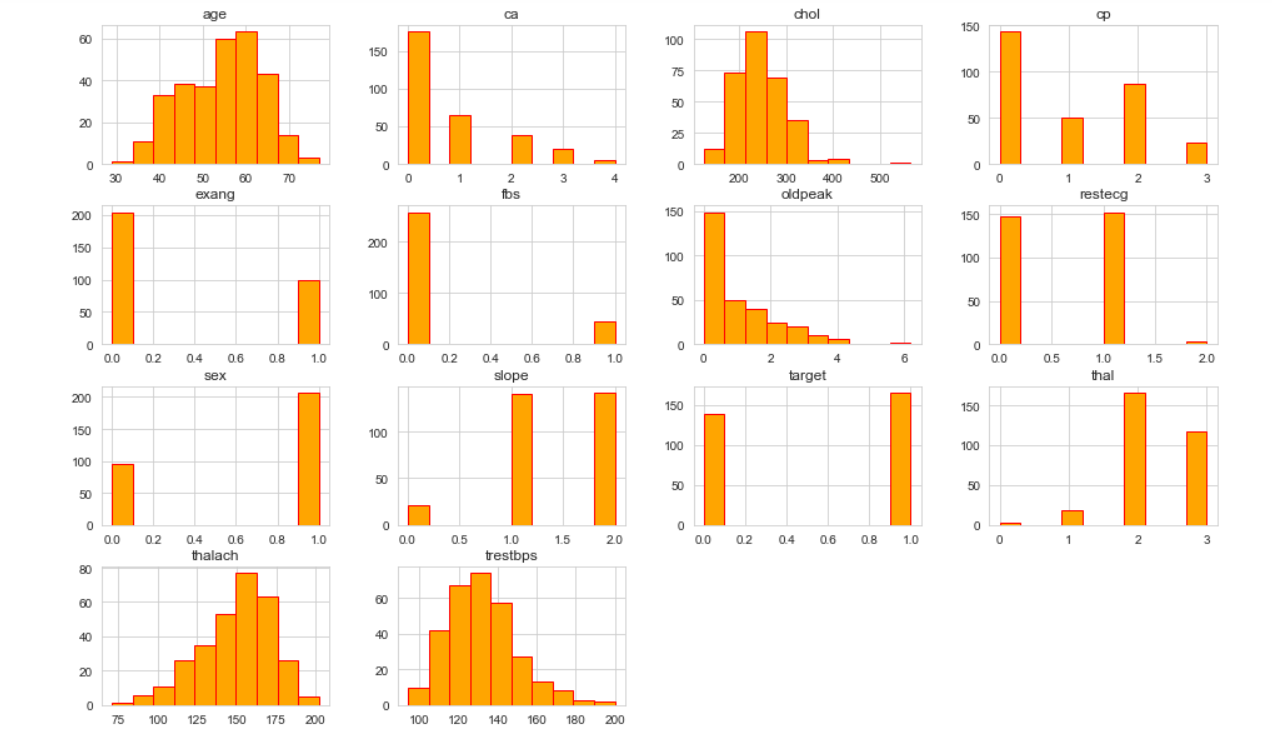


**Output-Screenshots:**

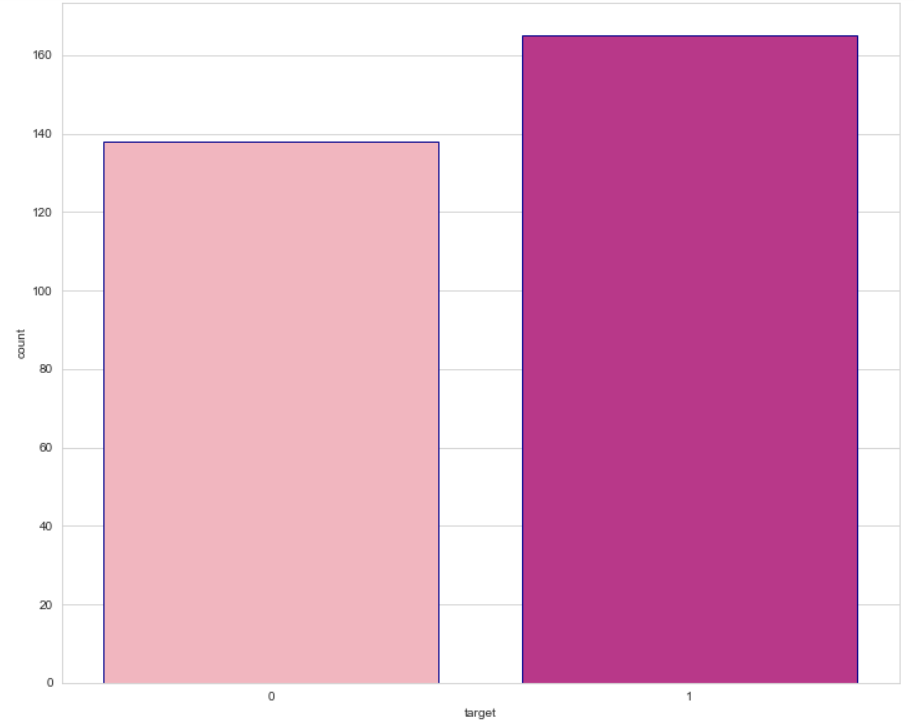
Featuring Sections of attributes in Dataset-



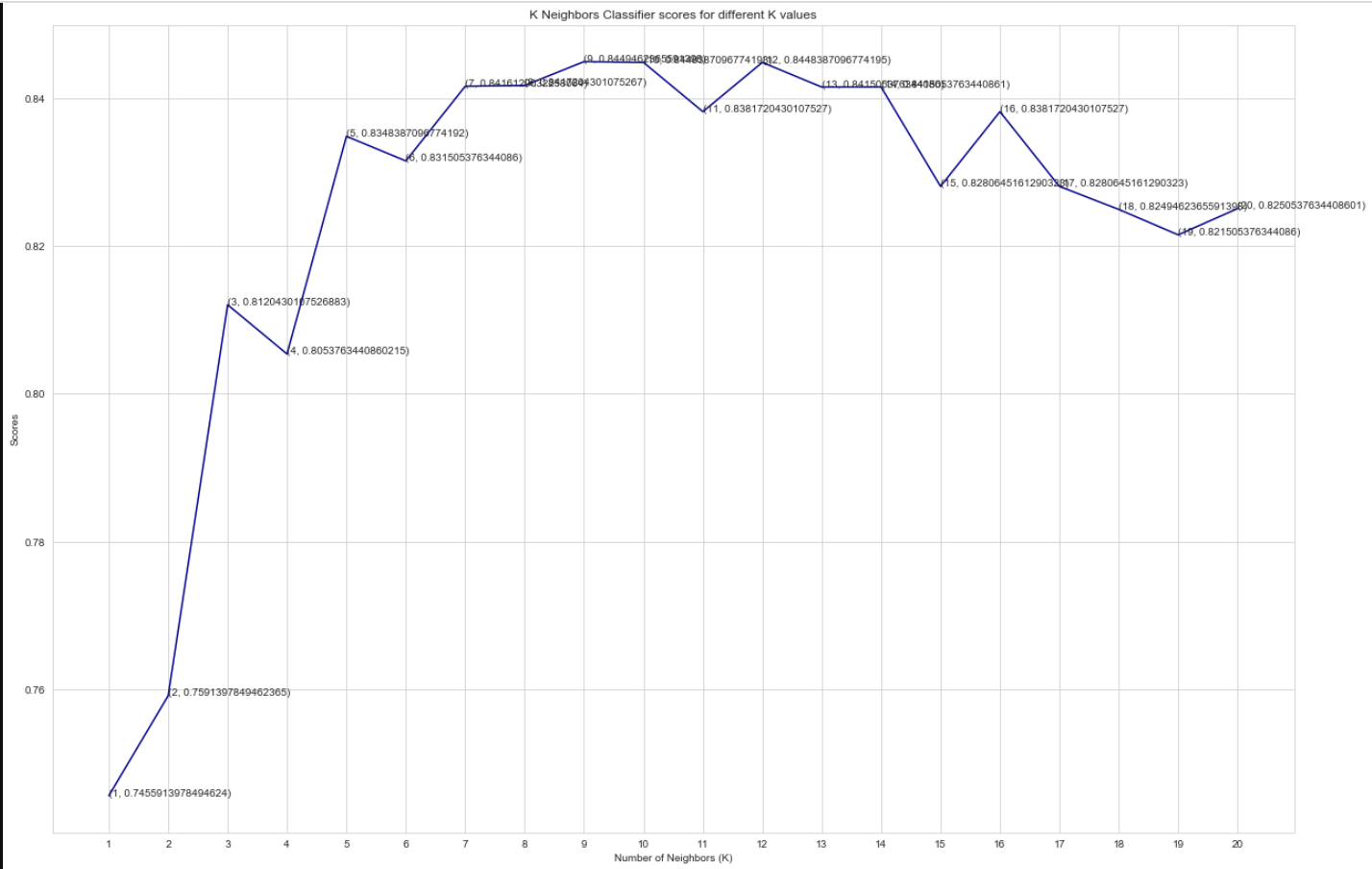
#Histogram Graph-



#countplot-



#K Neighbors Classifier scores for different K values-



Out- 0.8448387096774195

#Random Forest Classifier-

Out- 0.811505376344086

**FUTURE SCOPE**

The heart stroke prediction can be done using other machine learning algorithms. Advance Logistic regression can also perform well in case of binary classification problems such as heart stroke prediction. Random forests can perform well than decision trees. Also, the ensemble methods and artificial neural networks can be applied to the dataset. The results can be compared and improvised.

So when a patient is predicted as positive for heart stroke, then the medical data for the patient can be closely analysed by the doctors. An example would be - suppose the patient has diabetes which may be the cause for heart disease/stroke in future and then the patient can be given treatment to have diabetes in control which in turn may prevent the heart stroke. Hence, In future our model’s architecture and purpose will get modified as well as add new attributes.

* We will be able to easily recognise condition of individual with its risk percentage.
* We can easily know exact accuracy and implement it in medical case papers.
* With help of this prediction system we can lead to much healthier lifestyle.

**CONCLUSION**

Heart Stroke is one of the major concerns for society today. It is difficult to manually determine the odds of getting heart disease based on risk factors. However, machine learning techniques are useful to predict the output from existing data. The performance of the health’s diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

**BIBLIOGRAPHY AND REFRENCES**

[1] Datasets ([https://www.kaggle.com/ronitf/heart-disease-uci/data#](https://www.kaggle.com/ronitf/heart-disease-uci/data))

[2] Sklearn libraries (<https://scikit-learn.org/>)

[3] Algorithms Reference(<https://towardsdatascience.com/>)

[4] Matplotlib Reference(<https://matplotlib.org/tutorials/index.html>)

[5] SVM Algorithm Study Case(<https://ml-cheatsheet.readthedocs.io/en/latest/>)