Data Science Techniques and Applications
Coursework II

Baran Buluttekin
13153116

March 23, 2019

We can start of our analysis by getting the dataset. Corresponding Heart Diseases dataset[2][1] can be downloaded from kaggle by clicking download link in the website. When the download finished we received compressed version of file named *heart-disese-uci.zip*. Fallowing the un-compression of the file we will obtain csv file *heart.csv* which is in final format for analysis.

Next step is to loading python libraries to start the analysis. In this coursework fallowing open source libraries used:

1. **Pandas:** Tabular data manipulation library that will be use to load the data and cleaninig/reshaping.[4]
2. **Matplotlib:** Will be used as main plotting library that has many pre build high level plotting tools.[3]
3. **Seaborn:** Python library for special plotting styles such as pairplots.[6]
4. **Scikit-Learn:** Open source machine learning library that includes various machine learning algorithms.[5]

As a first step libraries above loaded in python in the fallowing format.

```
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from sklearn.decomposition import PCA
import seaborn as sns
```

With the help of pandas library we can load the dataset and view first 5 rows to check if the data frame is in expected format.

```
df = pd.read_csv("heart.csv")
df.head()
```

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63.0	1.0	3.0	145.0	233.0	1.0	0.0	150.0	0.0	2.3	0.0	0.0	1.0	1.0
37.0	1.0	2.0	130.0	250.0	0.0	1.0	187.0	0.0	3.5	0.0	0.0	2.0	1.0
41.0	0.0	1.0	130.0	204.0	0.0	0.0	172.0	0.0	1.4	2.0	0.0	2.0	1.0
56.0	1.0	1.0	120.0	236.0	0.0	1.0	178.0	0.0	0.8	2.0	0.0	2.0	1.0
57.0	0.0	0.0	120.0	354.0	0.0	1.0	163.0	1.0	0.6	2.0	0.0	2.0	1.0

Fallowing the guide of coursework, I choose "age", "chol" and "trestbps" as important dimensions. Description of the columns from the coursework I is given below as a reminder.

1. **age**: Age in years (max:77, min:29)
2. **chol**: cholesterol level (mg/dl)
3. **trestbps**: Blood pressure recording for resting

In order to get a better grasp of the data distribution with these three variables, I will use pairplot. Pairplot will help us see different variation of these three variable with each other and with relation to the target outcome of the dataset.

```
selected_dims = ["age", "chol", "trestbps", "target"]
sns.pairplot(df[selected_dims], hue="target")
```

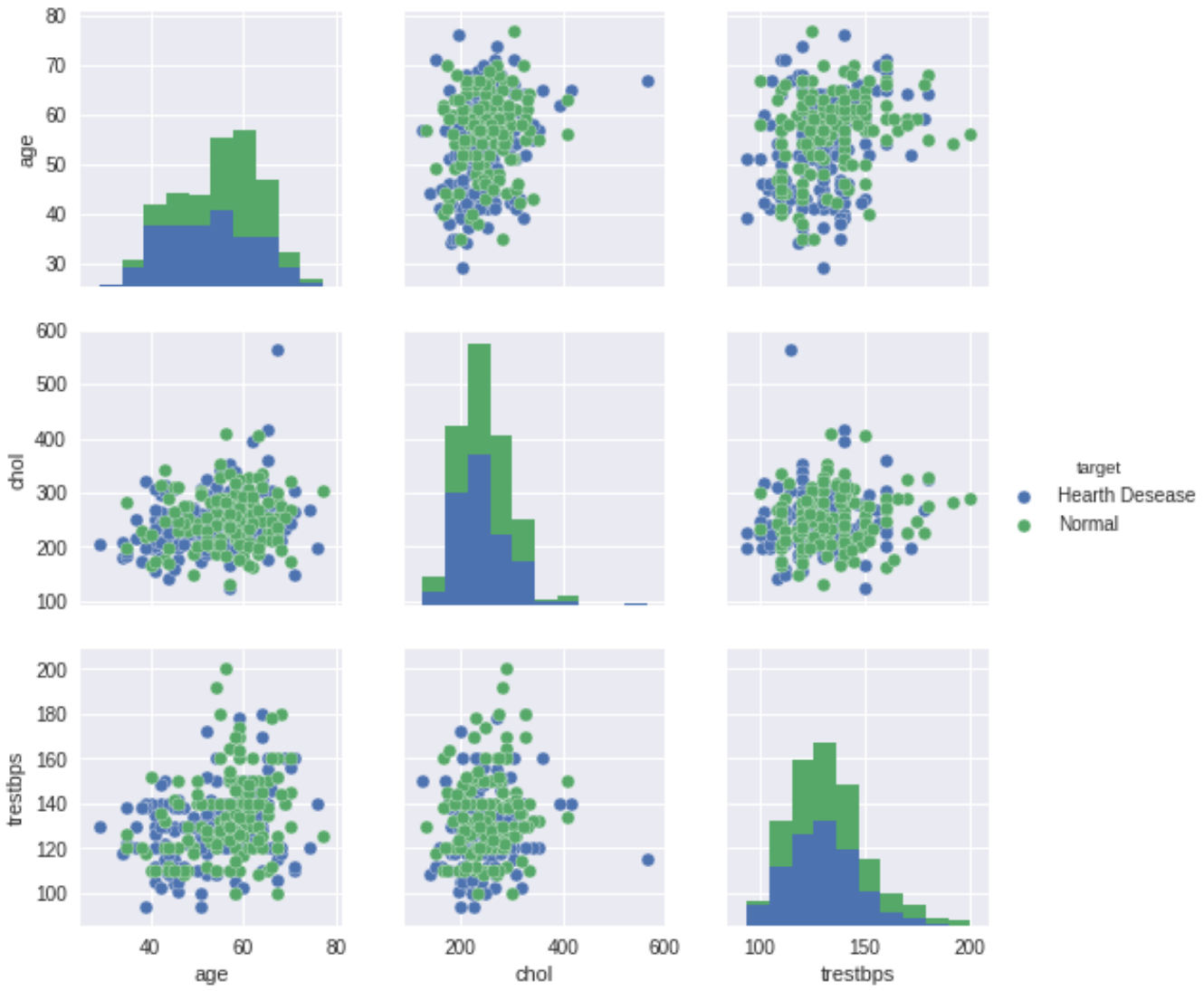

Figure 1: Pairplot with chosen dimensions.

We can also plot these 3 dimensions in 3D plot to see their relation to target variable. I created helper function quickly plot 3D plots given dimensions in dataset. Code for the function follows.

```
def plot_3d(dataframe, x_label, y_label, z_label, target_label, title):
    labelTups = [("Normal", 0), ("Hearth Desease", 1)]
    fig = plt.figure(1, figsize=(8, 6))
    ax = Axes3D(fig, elev=-150, azim=110)
    ax.scatter(dataframe[x_label], dataframe[y_label], dataframe[z_label], c
                                          =dataframe[target_label],
               cmap=plt.cm.Set1, edgecolor='k', s=40)
    ax.set_title(title)
    ax.set_xlabel(x_label)
```

```
    ax.w_xaxis.set_ticklabels([])
    ax.set_ylabel(y_label)
    ax.w_yaxis.set_ticklabels([])
    ax.set_zlabel(z_label)
    ax.w_zaxis.set_ticklabels([])

    sc = ax.scatter(dataframe[x_label], dataframe[y_label], dataframe[
                                            z_label], c=dataframe[
                                            target_label], cmap="Spectral",
                                            edgecolor='k')

    colors = [sc.cmap(sc.norm(i)) for i in [1, 0]]
    custom_lines = [plt.Line2D([],[], ls="", marker='.',
                    mec='k', mfc=c, mew=.1, ms=20) for c in colors]
    lgd = ax.legend(custom_lines, [lt[0] for lt in labelTups],
                loc='center left', bbox_to_anchor=(1.0, .5))
    plt.show()
```

We can observe their relationships by calling the function with the *"age"*, *"chol"* and *"trestbps"*.

```
plot_3d(df, "age", "chol", "trestbps", "target", "Heart Desease clustering
                                        by 3 dimensions")
```

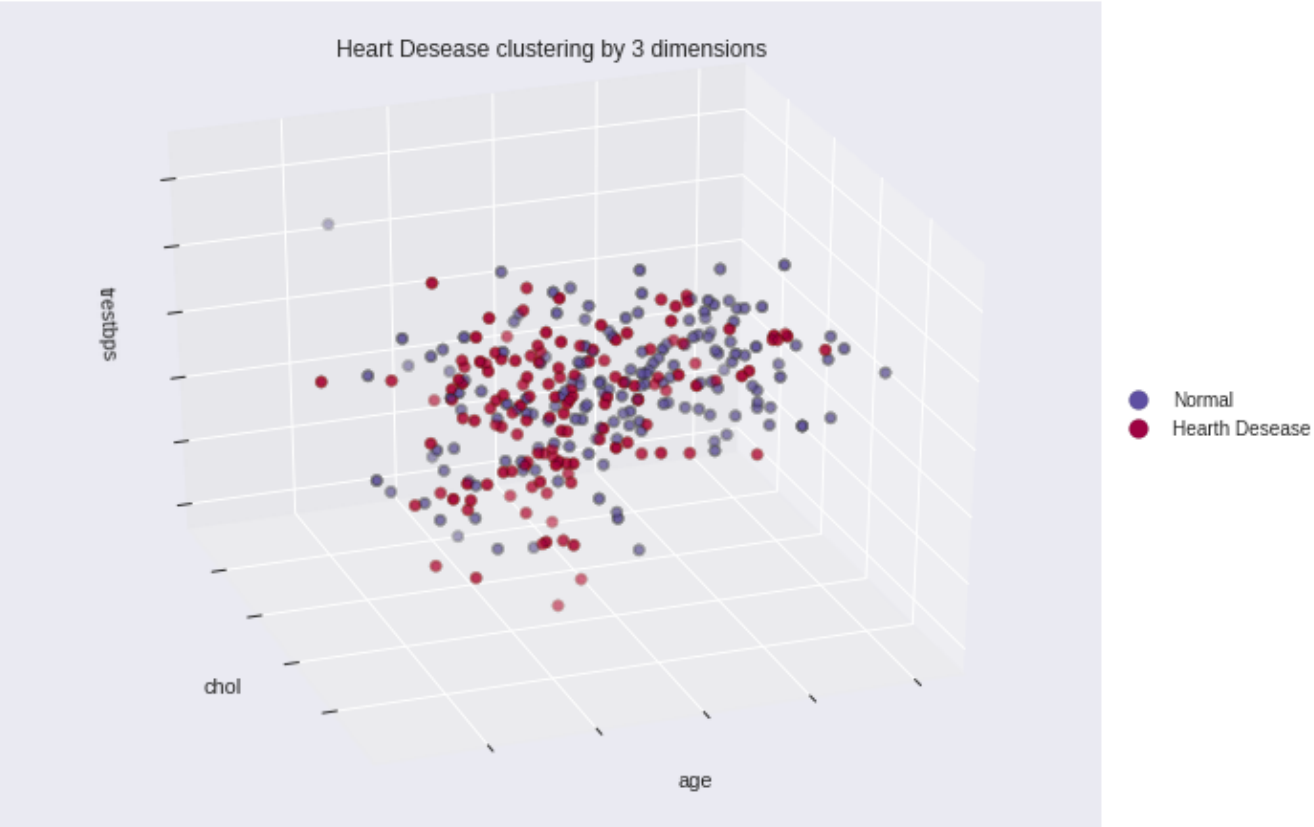

Figure 2: 3D graph of the selected dimensions.

Applying PCA On Dataset

References

[1] Janosi et al. *UCI Machine Learning Repository*. 1988. URL: `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`.

[2] *Heart Disease UCI*. URL: `https://www.kaggle.com/ronitf/heart-disease-uci`.

[3] J. D. Hunter. "Matplotlib: A 2D Graphics Environment". In: *Computing in Science Engineering* 9.3 (May 2007), pp. 90–95. ISSN: 1521-9615. DOI: `10.1109/MCSE.2007.55`.

[4] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 51–56.

[5] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[6] Michael Waskom et al. *mwaskom/seaborn: v0.8.1 (September 2017)*. Sept. 2017. DOI: 10.5281/zenodo.883859. URL: https://doi.org/10.5281/zenodo.883859.