**A picture containing logo

Description automatically generated**

**Project: Using Exploratory Data Analysis and Visualization to predict heart diseases.**

**Analytical report**

**Course: Data visualization**

**Instructor: Dr. Osama Abdulhay**

**Group Members:**

**Reema Ma’en - 20190150**

**Osama Rakan - 20190341**

**Amr Amr - 20190633**

**Saif Abu Al Raisi - 20190465**

**Abdulrahman Shabaneh - 20190546**

**Rama Al Alawneh – 20190153**

# **Brief Overview**

The aim of this project was to provide a way to classify whether the patient has CHD or not, in order to do so our team followed the following protocol

* Basic data exploration
* Data cleaning
* Detected and dealt with data inconsistencies
* Checked the distribution and correlations between variables
* Added more columns that would be of greater benefit to the outcome.

# **Main statistical features of the variables in the dataset:**

The dataset consisted of 310 rows and 15 columns, 5 of which are numerical. Here are their main statistical features after cleaning: Figure 1

* Age: most of the recorded ages lie in the range between 29 and 77, their average is 54, in other words most of the recorded instances are for patients around that age.
* Restbps: stands for the resting blood pressure, was recorded a maximum of 200 and a minimum of 94 the average was around 131.
* Chol: cholesterol, maximum is 564 the minimum is 126 and the average is 247.
* Thalack: maximum heart rate achieved, the range was between 202 and 71 the average was 149.
* Oldpeak: which was the ST depression induced by exercise relative to rest, the range was between 6.2 and 0 meaning that the patient could have no ST depression at all or any value up to 6.2 (according to the data that was present) the average was 1.
* Max\_heart\_rate\_difference : is a new column that we added to the data as it might give us more insight into the problem. The range is between -29 and 82 the average is 15.8

As for the remaining columns which were mainly categorical, here are their main statistical features: Figure 2

* Sex: around 65% of the patients were males while the rest were female
* Cp: which stands for the chest pain type in which numbers represented the chest pain type. Around 50% were “4”, 30% were “3”, 20% were “2” and 10% were “1”.
* Fbs: fasting blood sugar, around 80% were below the threshold (120) and 20% were above it.
* Restecg: resting ECG results in which numbers represented different results. 0 and 2 contributed the most with 50% and 48% while 1 was around 2%
* Exang: exercise induced angina, 67% did not while the other 33% do
* Slope: the slope of the peak exercise ST segment, results in which numbers represented different kinds of slope. 2 and 1 contributed the most 46% each while 3 was 7%
* Ca: number of major vessels colored by flourosopy, 58% were ‘0’ , 22% were ‘1’ , 13% were ‘2’ and 7% were ‘3’.
* Thal: thalassemia, were numbers represent different defects, 3 made up 55% of the data ,7 made up 39% of the data while 6 made up 6% of it.
* Outcome: presence or absence of CHD, 54% of the patients had CHD while 46% didn’t.

# **Central statistical relationships and correlations**

In order to better understand the given data, first of all we used a correlation heatmap which is a 2D visualization that represents how strongly correlated each column is using colored cells to represent how strong the correlation is usually a monochromatic scale, the color range is from red to blue, red color indicates strong relation between columns while blue indicates weak relations. It was clear to us that there was a strong relation between the desired outcome and ‘thal’, ‘ca’, ’oldpeak’ and ‘cp’ columns as shown in Figure 4. After wards, we took each of the most relevant features and saw how did it interact with the ‘outcome’ column as shown in Figure 3

We could safely conclude the following :

# **Data visualizations**

Chart, bar chart, box and whisker chart

Description automatically generatedChart, bar chart, box and whisker chart

Description automatically generatedChart, bar chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, bar chart, box and whisker chart

Description automatically generated

Figure 1

Chart, bar chart

Description automatically generatedChart

Description automatically generatedChart, waterfall chart

Description automatically generatedChart, bar chart, waterfall chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart

Description automatically generated

Figure 2

oldpeak

thal

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure

cp

ca

Chart

Description automatically generated

Figure

# **Discussion of the explanatory analysis about the potential factors that influenced the outcome of the interest.**

* As the ST depression induced by exercise (oldpeak) increases there is a higher chance that the patient has CHD.
* As the number of major vessels colored by florosopy (ca) increases there is a lower chance that the patient will have CHD
* When thalassemia (thal) is a fixed or reversable defect (3 or 7) there is a high chance the person has CHD
* When the chest pain (cp) type is asymptomatic (4) there is a very high chance the person has CHD

# **Recommendations**

We recommend collecting more data regarding the patient’s medical background since it aids in the process of reaching the desired outcome. As well as examining thalassemia, chest pain type, number of vessels colored by florosopy and ST depression induced by exercise all before anything else since they contribute the most to the problem at hand.