|  |
| --- |
| Cartoon Heart Drawing - How To Draw A Cartoon Heart Step By Step  comp5310 – project Stage 1  Data Acquisition and Cleaning |

­

**Section 1: Problem**

**1.1: Introduction**

Heart diseases encompasses a range of conditions that impact the functioning of the heart, including coronary artery disease, arrhythmia, and heart failure. Making healthy lifestyle choices can serve as a preventative measure against many of these conditions, and in cases where they do occur, medication can assist in their management (Smith, Y. (2019, March 15)). According to the World Health Organisation (WHO), Cardiovascular diseases (CVDs) are the leading cause of death globally, with an estimated of 17.9 million deaths each year (World Health Organisation. (n.d.)). Cardiovascular diseases. CVDs are concertedly contributed by overweight, hypertension and unhealthy lifestyles (Jain, A. (2019, April 25)). Although the cessation of alcohol, smoking and poor diet can help assist in avoiding this multifaceted clinical disease, oftentimes the indication of heart problems may not be detectable until the patient encounters a heart attack. Therefore, it is necessary that healthcare industries generate data to accommodate knowledge or pattern for decision making (Asgari, S., Ghaemmaghami, Z., & Mohammadzadeh, N. (2020)).

**1.2: Problem Definition**

What is more beneficial than merely detecting the presence or absence of heart diseases is to classify these diseases and understand the relationship between patients from datasets on a molecular level. By performing this analysis, misdiagnosis of these diseases can be reduced through these vastly accurate methods using several machine learning and data analysis techniques. Initially, our analysis will require us to define a few key research questions:

1. What are the most important factors in predicting the presence of heart disease, and how can these factors be incorporated into a predictive model?
2. How can the use of ensemble methods improve the accuracy of heart disease detection models?
3. Which supervised learning model will be the most effective in determining the factors in contributing towards cardiovascular diseases?
4. Which clinical and demographic factors are the strongest predictors of heart disease in patients, and how can this information be used to improve early detection and prevention efforts?

**Section 2: Problem Approach**

Any analysis that is completed on a dataset would requirement exploratory data analysis (EDA) which involves understanding the general structure and content of the data in its csv form and then producing some basic statistics and visualisations to gather an initial understanding. This is my preliminary approach towards this project, followed by machine learning techniques which will involve classification algorithms. This is due to the scope of the research questions aforementioned in conjunction with the structure of the data provided. Using a range of different classification models, collating and processing these models and comparing them through evaluation metrics will be ideal in providing a deterministic view of understanding the factors which may affect chances of patients in getting heart disease. Through this approach, clinicians will find useful information during treatment.

**Section 3: Data**

**3.1: Acquisition of relevant Dataset**

The data was acquired from the link provided [here](https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction?datasetId=1582403&sortBy=voteCount), which was collated from different datasets in the UCI Machine Learning Repository from the following link: <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>. The usage of various datasets across different regions and observations can improve the validity of our assessment of these variables when determining their effect on heart diseases. The final dataset contains 918 observations alongside attribute information which includes numerical and categorical variables, some of them include the age, sex, ChestPainType, RestingBP (resting blood pressure), Cholesterol and more. We can establish a prediction model using various classification algorithms that will assist in research purposes as this dataset contains a diverse and large amount of information useful that is crucial in early detection of heart diseases.

**3.2: Data Cleaning and Transformation**

After data acquisition, we have a dataset in csv format containing 12 columns and 918 rows of both numerical and categorical data. We then ingest the dataset into Jupyter notebook for pre-processing by using the pandas library as shown in appendix 1.1. Observing the variables, we see that HeartDisease is the outcome variable which is binary, meanwhile the other predictors comprise of quantitative and qualitative variables. The numerical and categorical variables area shown in appendix 1.2. To check the basic descriptive statistics, which will be useful in gathering our initial understanding of the data, we see that FastingBS and HeartDisease are binary variables, indicated from their respective minimum and maximum values of 0 and 1, while the other predictors displayed are numerical and the predictors which are not shown, such as Sex and ExerciseAngina are categorical. Cholesterol has the highest mean amongst the predictors, whereas FastingBS has the lowest.

A correlation analysis has been performed in appendix 1.3, where relatively all predictors are not considered highly correlated, indicated from the various R score values which do not exceed 0.7 hence no multicollinearity issues will be encountered. However, we can scrutinise some observations such as the negative correlation between MaxHR and HeartDisease with an R score of -0.4. Further analysis through linear regression will be required to help quantify parameter estimations and relationship between predictors and outcome. Furthermore, we can check the null values using the .isna().sum() command, as we see in appendix 1.4 that there are no null values. Checking duplicate values as seen in appendix 1.5, we can first use the .drop\_duplicates() command then confirm if there was any duplicate values by checking the shape of the dataset. We confirm that there are no duplicate values here.

Data transformation will be necessary in some of the features we are working with as models can only work with numerical values. It is necessary to convert the categorical values of the features into numerical values. When building and testing our models in stage 2 and 3 of the project, incorporating all the variables will be fruitful in obtaining a well performing model. In appendix 1.6, we observe how encoding these categorical features is performed. This process is done by first identifying the categorical features in the dataset, then encoding the variables as dummies. This means that the variables take on values of 0 or 1 to represent the presence or absence of a particular characteristic or attribute. The dataset presents more variables such as the different types of ChestPaintType and ST\_Slope types with either 0 or 1 as the values. We may also need to transform RestingBP, Cholesterol and MaxHR to 1 or 2 decimal places so we can ensure consistency and accuracy in the analysis of the data. Omitting this can cause skewness or inaccuracies due to rounding errors especially when working with coefficients during modelling using regression.

**Reference**

Ale, T. B. M. (2020). Cardiovascular diseases EDA + Modeling. Kaggle. Retrieved May 5, 2023, from <https://www.kaggle.com/code/aletbm/cardiovascular-diseases-eda-modeling>

Asgari, S., Ghaemmaghami, Z., & Mohammadzadeh, N. (2020). Machine learning-based prediction of heart failure readmission or death: Implications of choosing the right algorithm and feature selection. Journal of Cardiac Failure, 26(12), 1016-1019. doi: 10.1016/j.cardfail.2020.07.011

David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the  
Cleveland database."

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., &  
Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of  
coronary artery disease. American Journal of Cardiology, 64,304--310.

Gennari, J.H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. Artificial  
Intelligence, 40, 11--61.

Jain, A. (2019, April 25). Exploratory Data Analysis on Heart Disease UCI Data Set. Towards Data Science. <https://towardsdatascience.com/exploratory-data-analysis-on-heart-disease-uci-data-set-ae129e47b323>

Miguel, F. Z. (2021). Heart Failure Prediction: Classification Models. Kaggle. Retrieved May 5, 2023, from <https://www.kaggle.com/code/miguelfzzz/heart-failure-prediction-classification-models>

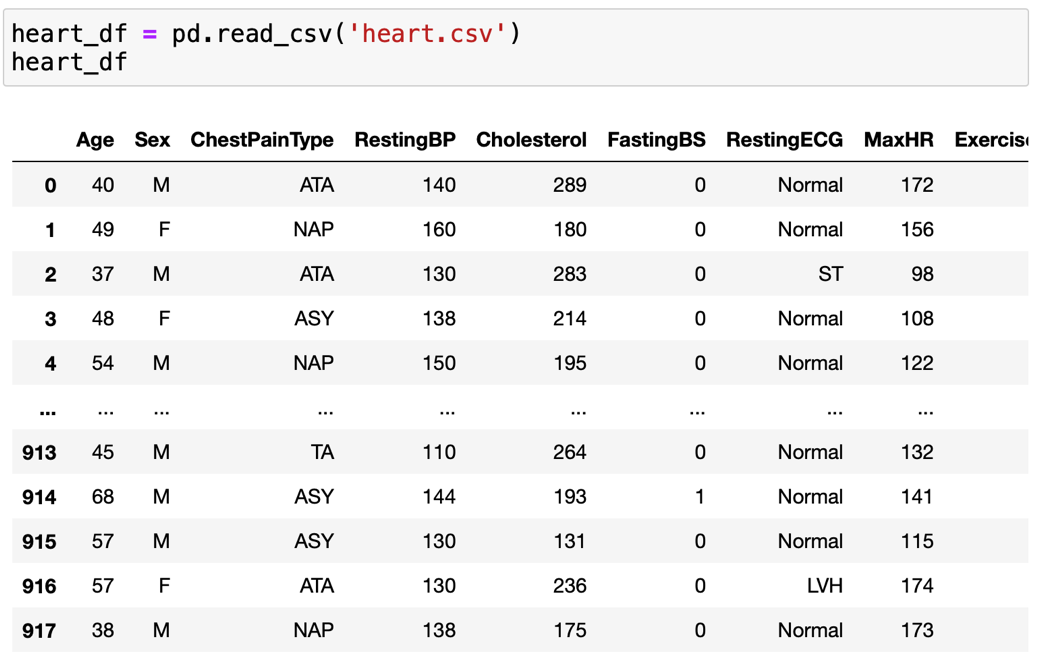
Smith, Y. (2019, March 15). What are the symptoms of heart disease in men? Medical News Today. <https://www.medicalnewstoday.com/articles/237191>

Wasef, M. (2021). Heart Disease EDA with ML. Kaggle. Retrieved May 5, 2023, from <https://www.kaggle.com/code/mohamedwasef/heart-disease-eda-with-ml>

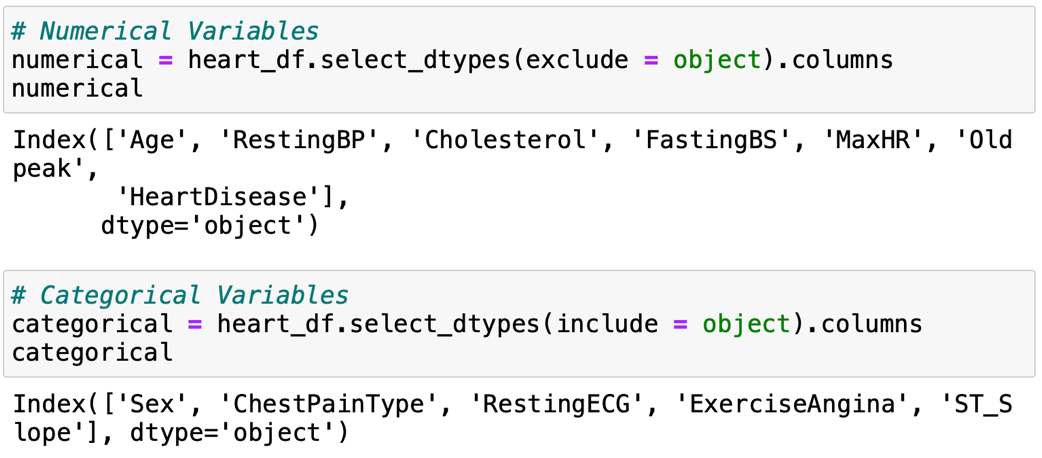
World Health Organization. (n.d.). Cardiovascular diseases. Retrieved May 3, 2023, from <https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1>

**Appendix**

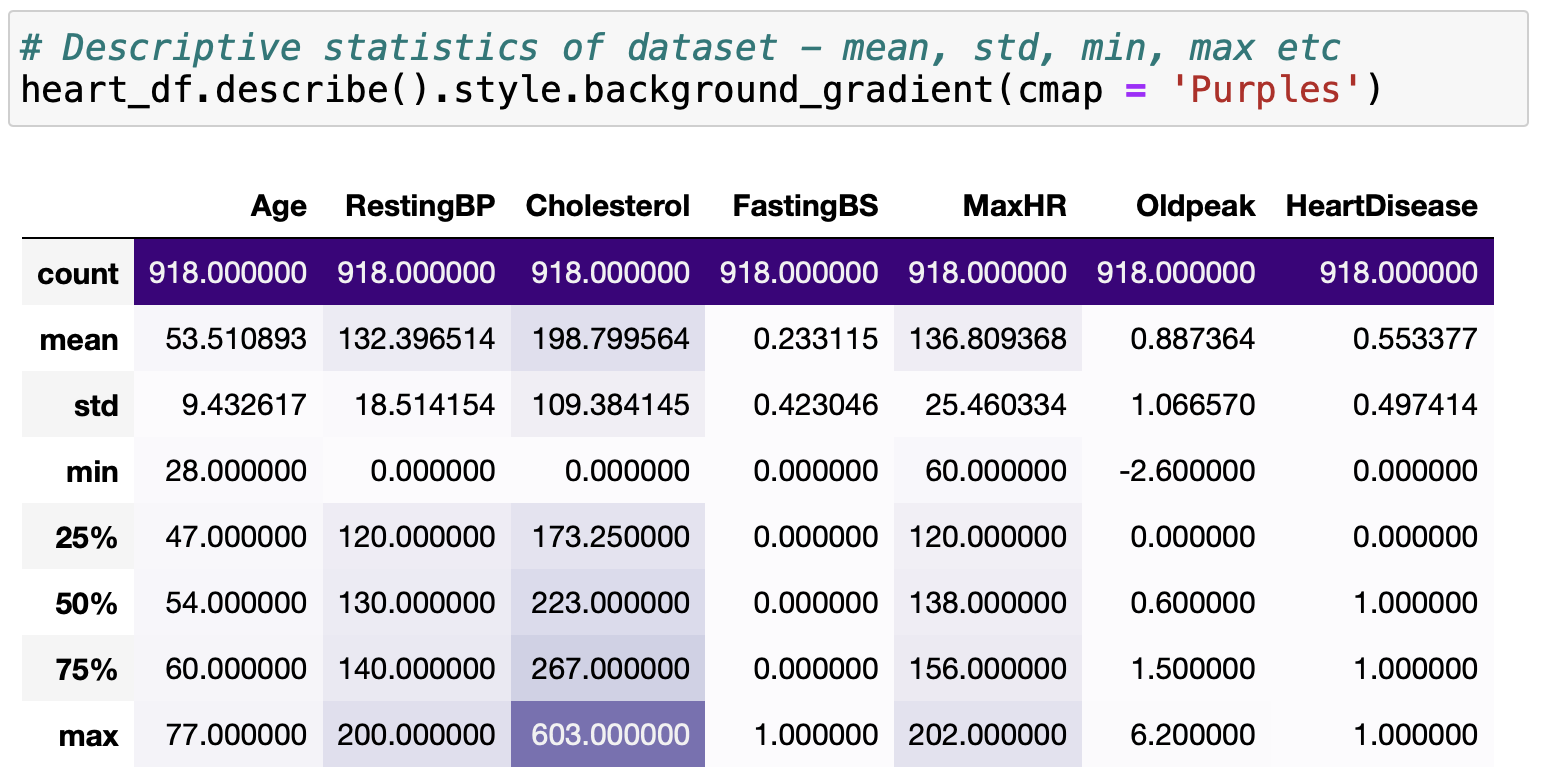
**1.1: Reading the data in Jupyter Notebook**



**1.2: Numerical and Catagorical predictors in heart dataset**



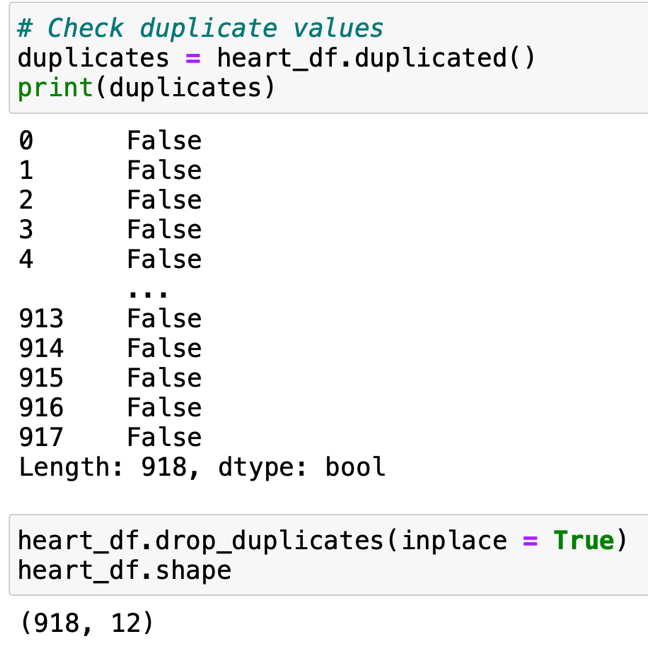
**1.3: Correlation matrix of all the numerical variables**



**1.4: Checking for null values**



**1.5: Checking for duplicate values**



**1.6: Encoding categorical variables to numerical variables** 