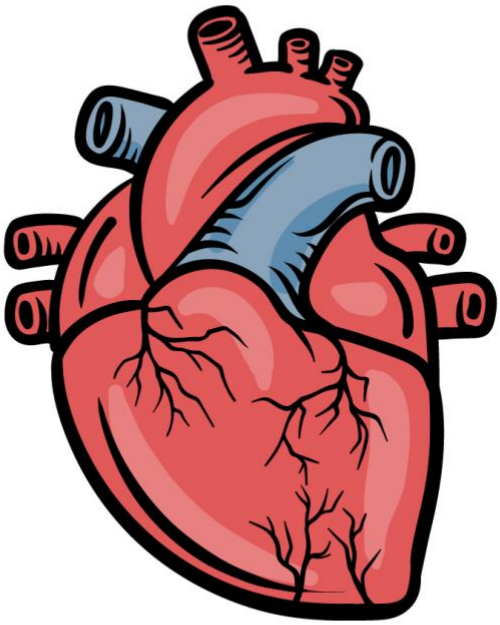

COMP5310 –
PROJECT STAGE 1

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, 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

```
heart_df = pd.read_csv('heart.csv')
heart_df
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	Exercis
0	40	M	ATA	140	289	0	Normal	172	
1	49	F	NAP	160	180	0	Normal	156	
2	37	M	ATA	130	283	0	ST	98	
3	48	F	ASY	138	214	0	Normal	108	
4	54	M	NAP	150	195	0	Normal	122	
...	...	...	...	...	...	...	...	...	
913	45	M	TA	110	264	0	Normal	132	
914	68	M	ASY	144	193	1	Normal	141	
915	57	M	ASY	130	131	0	Normal	115	
916	57	F	ATA	130	236	0	LVH	174	
917	38	M	NAP	138	175	0	Normal	173	

1.2: Numerical and Catagorical predictors in heart dataset

```
# Numerical Variables
numerical = heart_df.select_dtypes(exclude = object).columns
numerical
```

```
Index(['Age', 'RestingBP', 'Cholesterol', 'FastingBS', 'MaxHR', 'Old
peak',
       'HeartDisease'],
      dtype='object')
```

```
# Categorical Variables
categorical = heart_df.select_dtypes(include = object).columns
categorical
```

```
Index(['Sex', 'ChestPainType', 'RestingECG', 'ExerciseAngina', 'ST_S
lope'], dtype='object')
```

1.3: Correlation matrix of all the numerical variables

```
# Descriptive statistics of dataset - mean, std, min, max etc
heart_df.describe().style.background_gradient(cmap = 'Purples')
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

1.4: Checking for null values

```
# To check null values
heart_df.isna().sum()
```

```
Age               0
Sex               0
ChestPainType     0
RestingBP         0
Cholesterol       0
FastingBS         0
RestingECG        0
MaxHR             0
ExerciseAngina    0
Oldpeak           0
ST_Slope          0
HeartDisease      0
dtype: int64
```

1.5: Checking for duplicate values

```
# Check duplicate values
duplicates = heart_df.duplicated()
print(duplicates)
```

```
0      False
1      False
2      False
3      False
4      False
       ...
913    False
914    False
915    False
916    False
917    False
Length: 918, dtype: bool
```

```
heart_df.drop_duplicates(inplace = True)
heart_df.shape
```

```
(918, 12)
```

1.6: Encoding categorical variables to numerical variables

```
# Select categorical variables
categ = heart_df.select_dtypes(include=object).columns

# One hot encoding
heart_df = pd.get_dummies(heart_df, columns=categ, drop_first=True)
heart_df.head()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease	Sex_M	ChestPainTy
0	40	140	289	0	172	0.0	0	1	
1	49	160	180	0	156	1.0	1	0	
2	37	130	283	0	98	0.0	0	1	
3	48	138	214	0	108	1.5	1	0	
4	54	150	195	0	122	0.0	0	1	

```
heart_df.tail()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease	Sex_M	ChestPain
913	45	110	264	0	132	1.2	1	1	
914	68	144	193	1	141	3.4	1	1	
915	57	130	131	0	115	1.2	1	1	
916	57	130	236	0	174	0.0	1	0	
917	38	138	175	0	173	0.0	0	1	