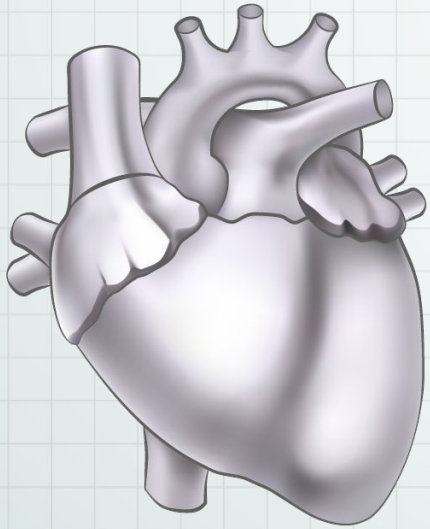


IE0005 Mini-Project

Dataset: Cardiovascular Disease Prediction

Team Macbook: Chen Mei Ling
Vernis Aw Ning Min
Tan Tse Teng
Kester Toh En Le

INTRODUCTION

Singapore Statistics

In Singapore, 21 people die from cardiovascular disease (heart diseases and stroke) every day. Cardiovascular disease accounted for 32% of all deaths in 2021. This means that almost 1 out of 3 deaths in Singapore is due to heart diseases or stroke.

DEATHS FROM CARDIOVASCULAR DISEASE

	2021	2020	2019
Total No. of Deaths	24,292	22,054	21,446
Ischaemic Heart Diseases	20.1%	20.5%	18.8%
Cerebrovascular Diseases (including stroke)	6.1%	6.0%	5.8%
Hypertensive Diseases (including hypertensive heart disease)	3.4%	2.9%	2.6%
Other Heart Diseases	2.3%	2.1%	2.0%
Atherosclerosis	0.2%	0.2%	0.1%
Total % of Deaths from Cardiovascular Disease	32.0%	31.7%	29.3%
Total No. of Deaths from Cardiovascular Disease	7,762	6,990	6,291

Source: Singapore Heart Foundation

Dataset and Objective

Dataset

Cardiovascular Disease Prediction

Objective

To build a prediction model to determine the likelihood of cardiovascular disease

Exploratory Data Analysis and Observation

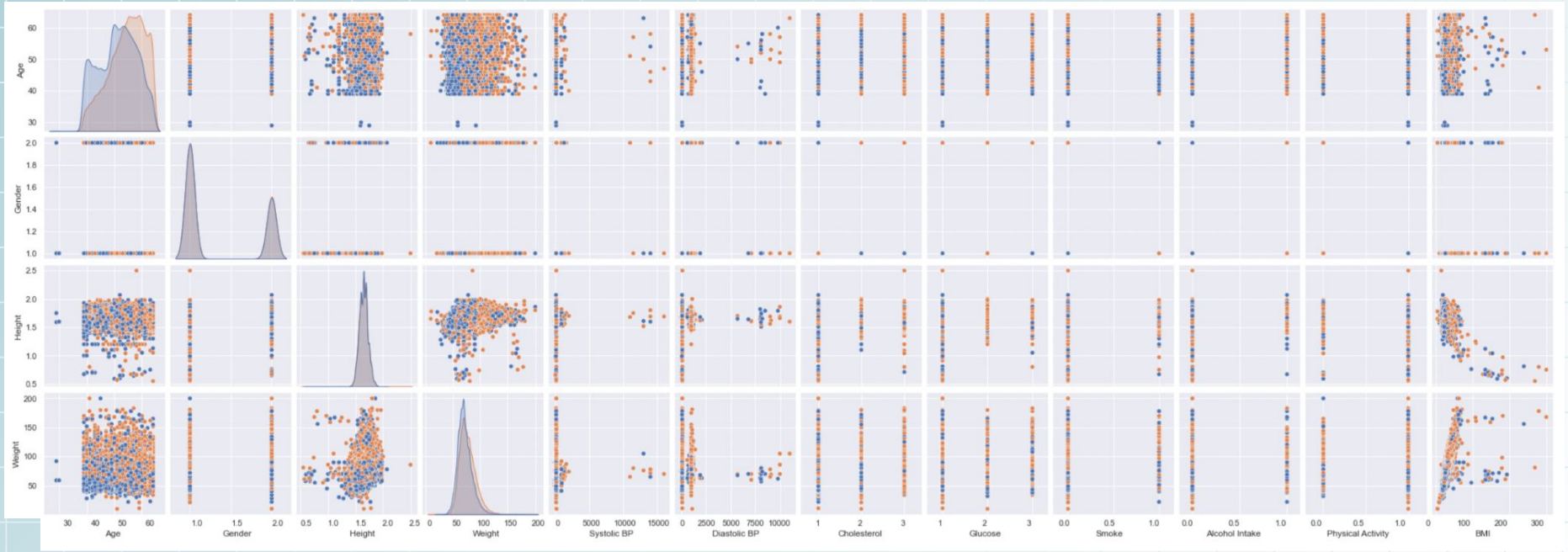
01

Exploratory Analysis and Observations

01

Pairplot

We chose to visualise the relation of pairs of variables using Seaborn pairplot.

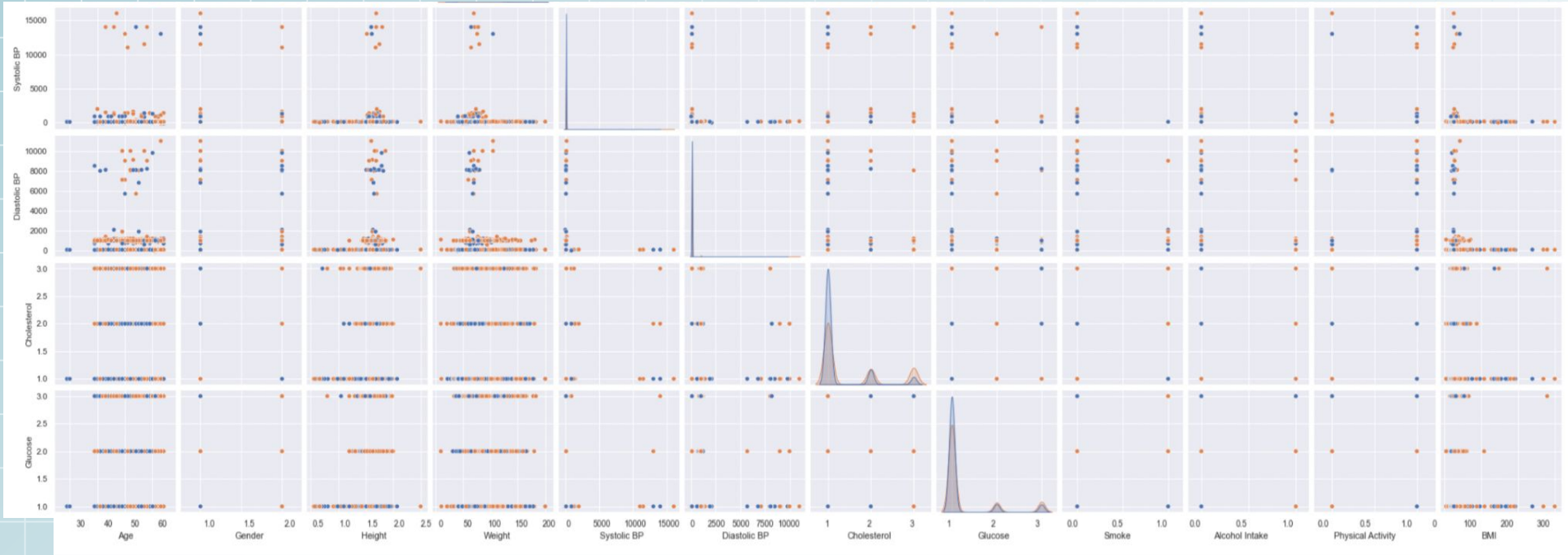


Exploratory Analysis and Observations

01

Pairplot

We chose to visualise the relation of pairs of variables using Seaborn pairplot.

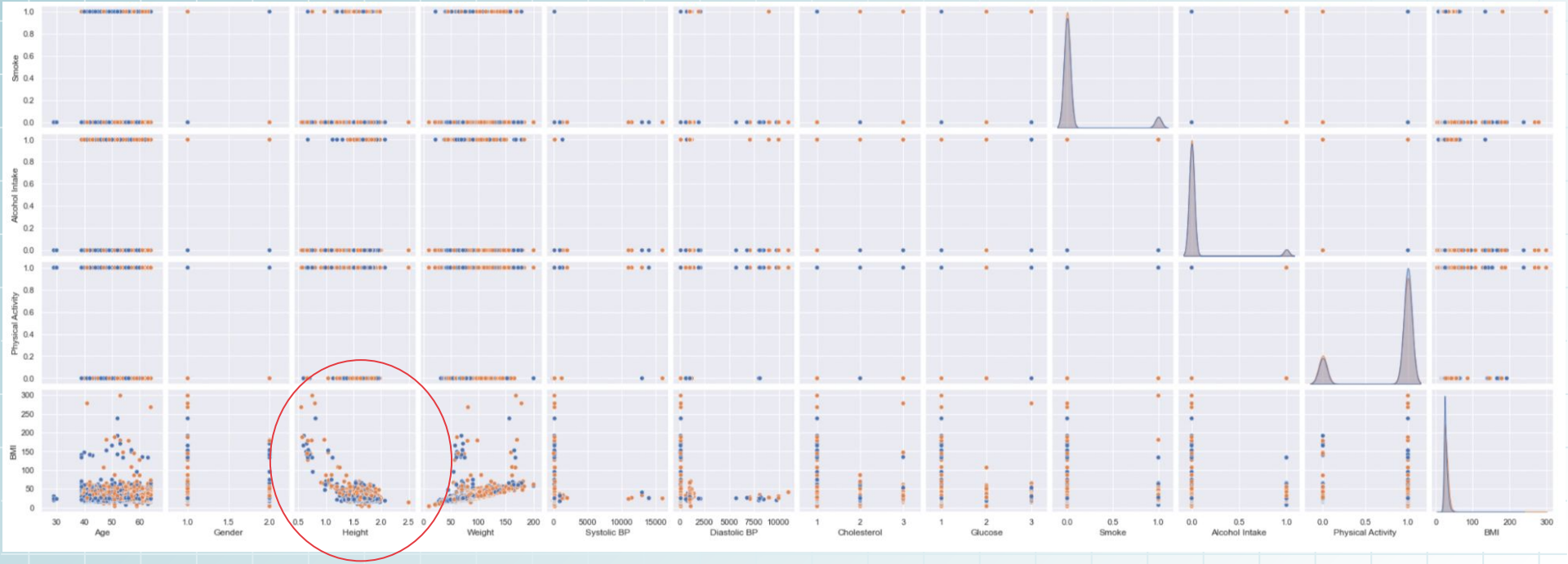


Exploratory Analysis and Observations

01

Pairplot

We chose to visualise the relation of pairs of variables using Seaborn pairplot.



Data Preparation

02

Cleaning of the data

Data Preparation

Purpose:
To modify the format of the given dataset to achieve high quality data to make good decisions upon.

Steps done to achieve this:

- Renamed the headers of each data to have clearer view of what each columns stands for
- Modification of data:
    - Changed of data type (e.g int to float)
    - Changed of age from days to years
    - Added a “BMI” column
- Removed irrelevant columns (e.g. id)
- Removed outliers (e.g. Data above maximum and below minimum are removed)

Data Preparation

1. Cleaning of Data ✓

2. Add column 'BMI' ✓

	Age	Gender	Height	Weight	BMI	Systolic BP	Diastolic BP	Cholesterol	Glucose	Smoke	Alcohol Intake	Physical Activity	Cardiovascular Disease
0	50	2	1.68	62.0	21.97	110	80	1	1	0	0	1	0
1	55	1	1.56	85.0	34.93	140	90	3	1	0	0	1	1
2	51	1	1.65	64.0	23.51	130	70	3	1	0	0	0	1
3	48	2	1.69	82.0	28.71	150	100	1	1	0	0	1	1
4	47	1	1.56	56.0	23.01	100	60	1	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	52	2	1.68	76.0	26.93	120	80	1	1	1	0	1	0
69996	61	1	1.58	126.0	50.47	140	90	2	2	0	0	1	1
69997	52	2	1.83	105.0	31.35	180	90	3	1	0	1	0	1
69998	61	1	1.63	72.0	27.10	135	80	1	2	0	0	0	1
69999	56	1	1.70	72.0	24.91	120	80	2	1	0	0	1	0

70000 rows × 13 columns

Data Preparation

3. Remove Outliers

	Age	Gender	Height	Weight	BMI	Systolic BP	Diastolic BP	Cholesterol	Glucose	Smoke	Alcohol Intake	Physical Activity	Cardiovascular Disease
0	50.0	2	1.68	62.0	21.97	110.0	80.0	1	1	0	0	1	0
1	55.0	1	1.56	85.0	34.93	140.0	90.0	3	1	0	0	1	1
2	51.0	1	1.65	64.0	23.51	130.0	70.0	3	1	0	0	0	1
3	48.0	2	1.69	82.0	28.71	150.0	100.0	1	1	0	0	1	1
5	60.0	1	1.51	67.0	29.38	120.0	80.0	2	2	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69993	53.0	1	1.72	70.0	23.66	130.0	90.0	1	1	0	0	1	1
69994	57.0	1	1.65	80.0	29.38	150.0	80.0	1	1	0	0	1	1
69995	52.0	2	1.68	76.0	26.93	120.0	80.0	1	1	1	0	1	0
69998	61.0	1	1.63	72.0	27.10	135.0	80.0	1	2	0	0	0	1
69999	56.0	1	1.70	72.0	24.91	120.0	80.0	2	1	0	0	1	0

61784 rows × 13 columns

Clean Dataset

	Age	Gender	Height	Weight	BMI	Systolic BP	Diastolic BP	Cholesterol	Glucose	Smoke	Alcohol Intake	Physical Activity	Cardiovascular Disease
0	50.0	Men	1.68	62.0	21.97	110.0	80.0	Normal	Normal	No	No	Yes	No
1	55.0	Women	1.56	85.0	34.93	140.0	90.0	Well Above Normal	Normal	No	No	Yes	Yes
2	51.0	Women	1.65	64.0	23.51	130.0	70.0	Well Above Normal	Normal	No	No	No	Yes
3	48.0	Men	1.69	82.0	28.71	150.0	100.0	Normal	Normal	No	No	Yes	Yes
5	60.0	Women	1.51	67.0	29.38	120.0	80.0	Above Normal	Above Normal	No	No	No	No
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69993	53.0	Women	1.72	70.0	23.66	130.0	90.0	Normal	Normal	No	No	Yes	Yes
69994	57.0	Women	1.65	80.0	29.38	150.0	80.0	Normal	Normal	No	No	Yes	Yes
69995	52.0	Men	1.68	76.0	26.93	120.0	80.0	Normal	Normal	Yes	No	Yes	No
69998	61.0	Women	1.63	72.0	27.10	135.0	80.0	Normal	Above Normal	No	No	No	Yes
69999	56.0	Women	1.70	72.0	24.91	120.0	80.0	Above Normal	Normal	No	No	Yes	No

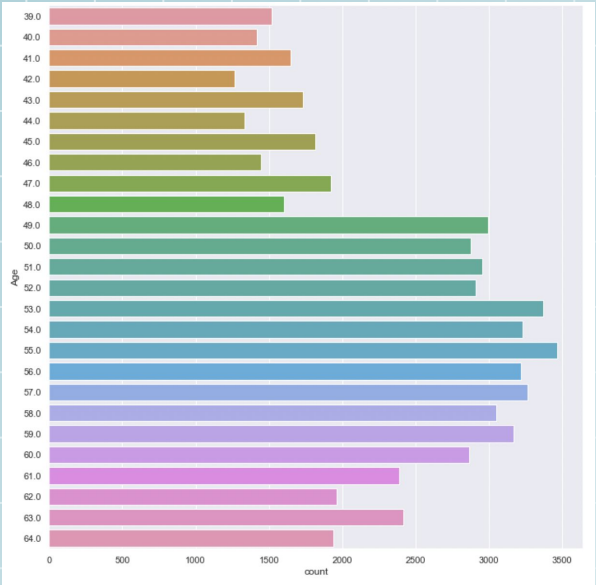
61784 rows × 13 columns

Exploratory Analysis and Observations

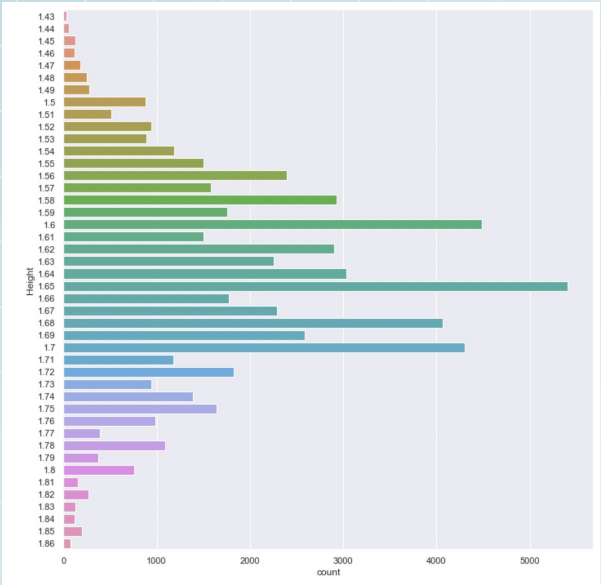
02

Catplot

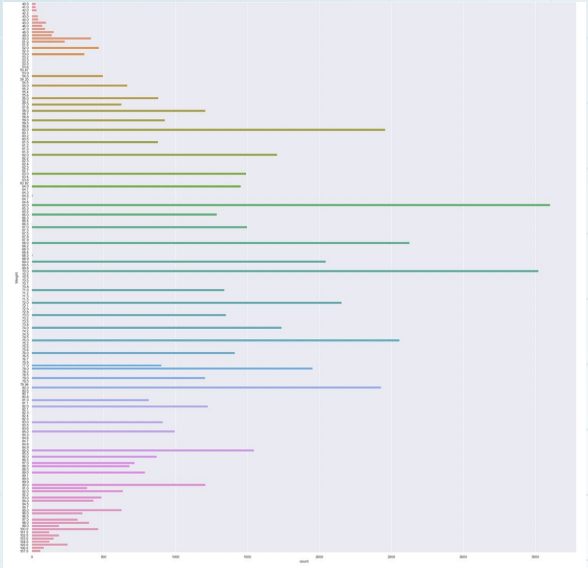
We chose to visualise the individual factors using Seaborn catplot.



Age



Height



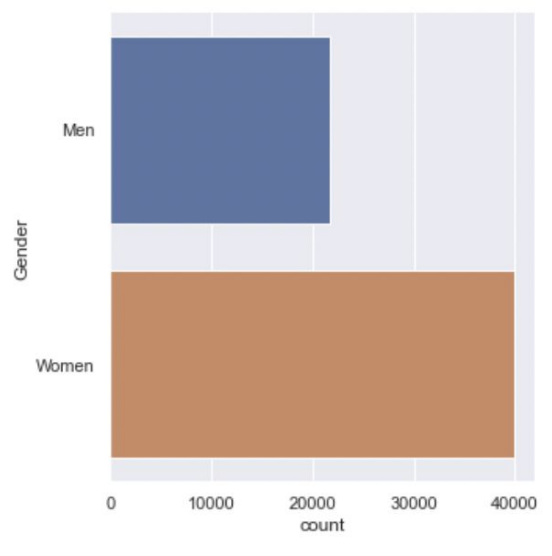
Weight

Exploratory Analysis and Observations

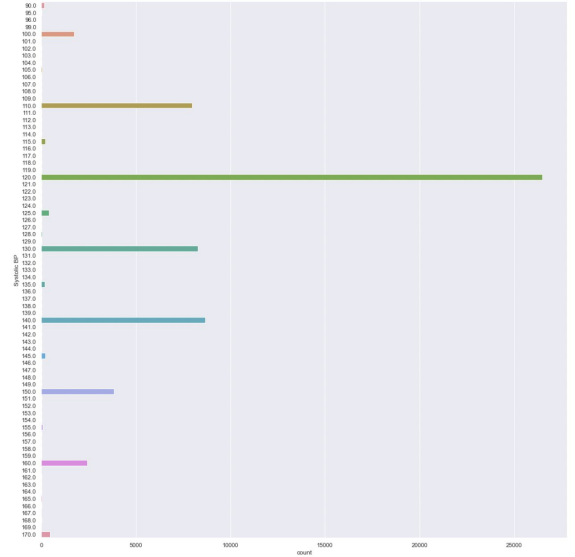
02

Catplot

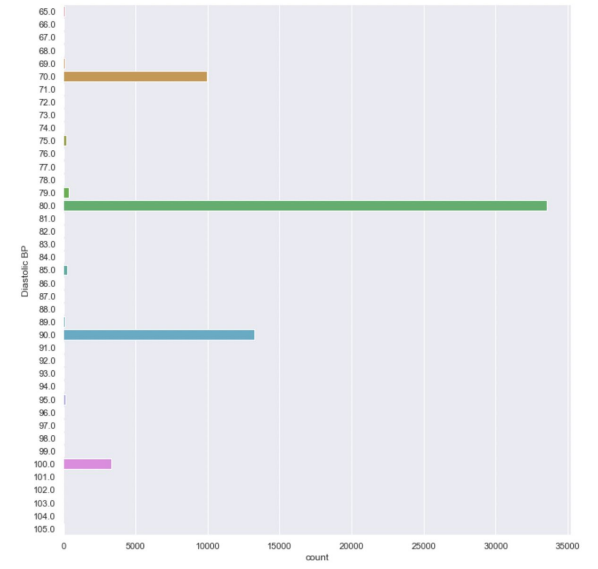
We chose to visualise the individual factors using Seaborn catplot.



Gender



Systolic BP



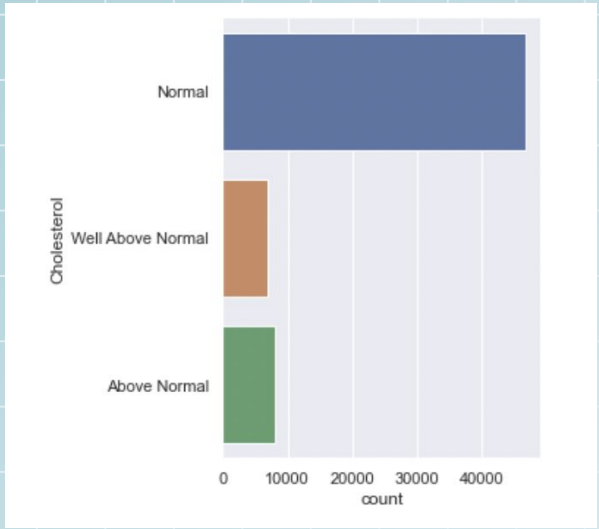
Diastolic BP

Exploratory Analysis and Observations

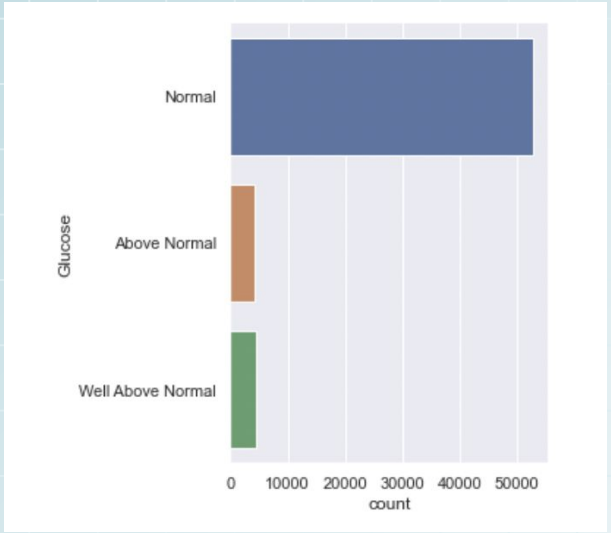
02

Catplot

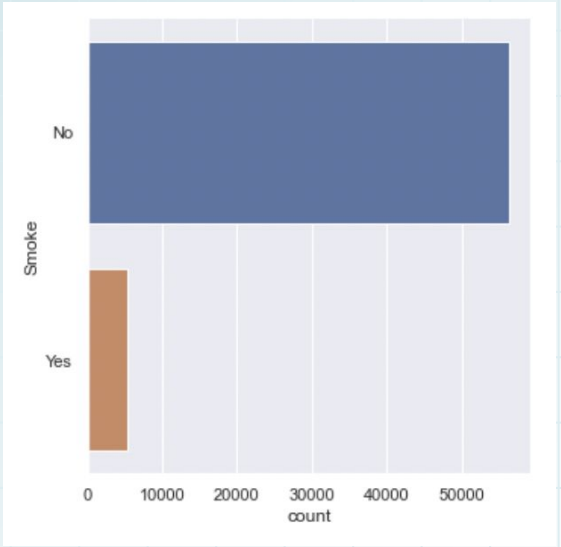
We chose to visualise the individual factors using Seaborn catplot.



Cholesterol



Glucose



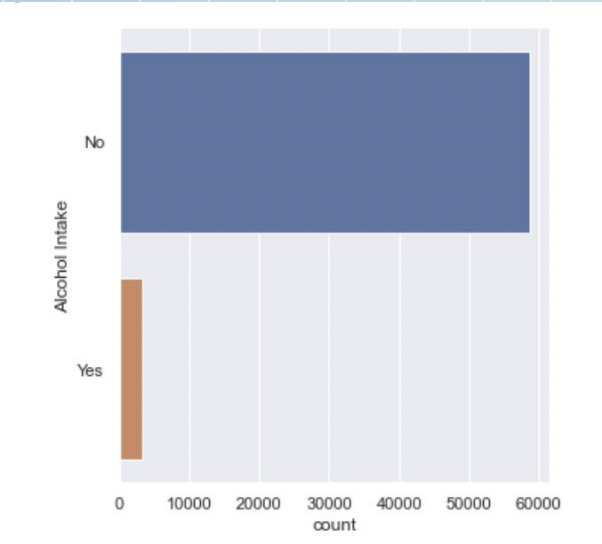
Smoke

Exploratory Analysis and Observations

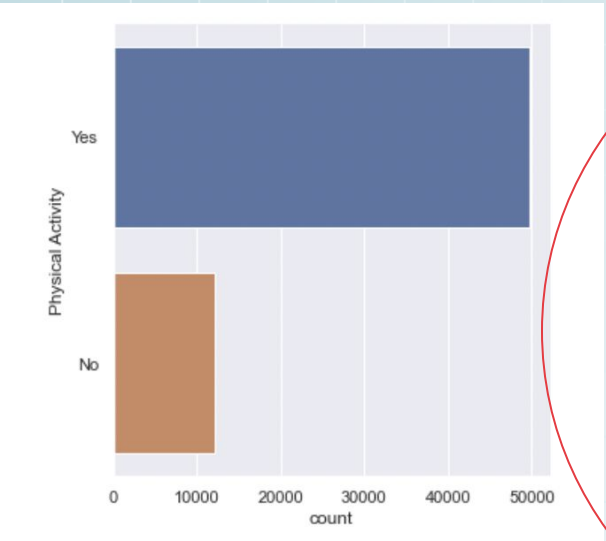
02

Catplot

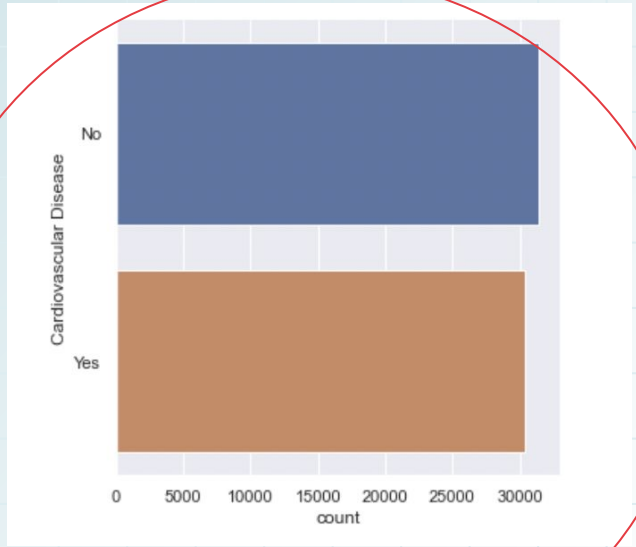
We chose to visualise the individual factors using Seaborn catplot.



Alcohol Intake



Physical Activity

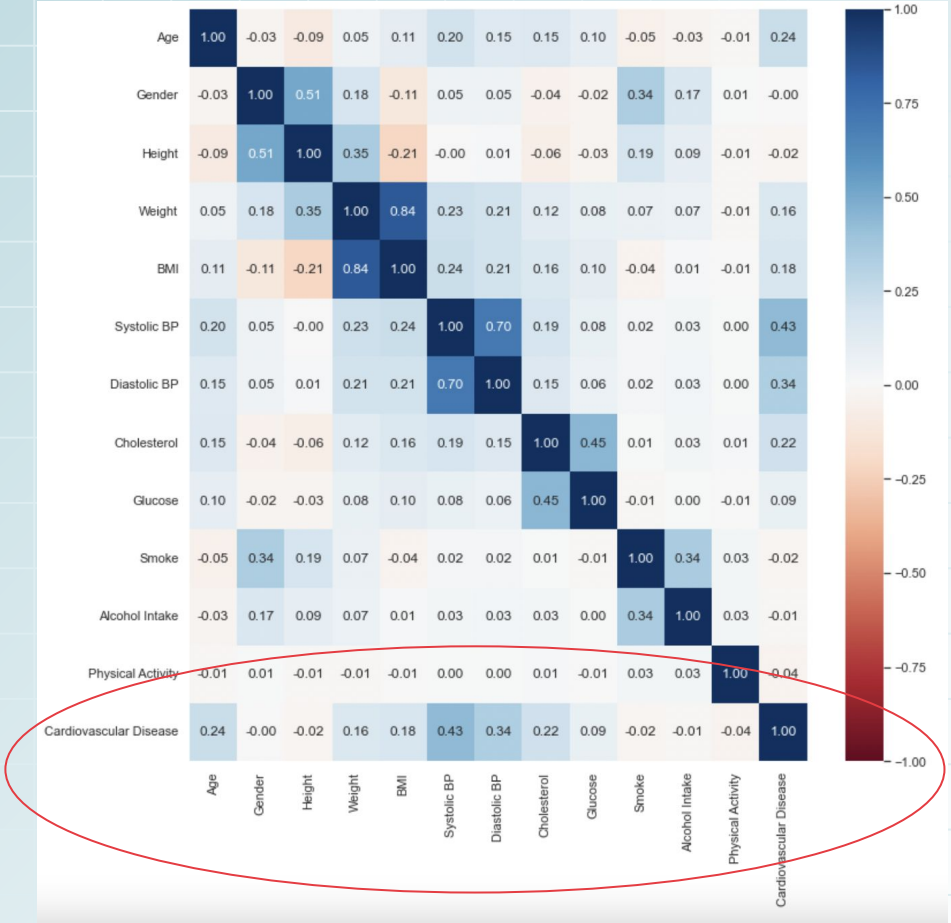


Presence of Cardiovascular Disease

Exploratory Analysis and Observations

03 Heatmap

We chose to visualise relation of factors using Seaborn heatmap.



Machine Learning

03

ML Tool Chosen: **Clustering, Anomaly detection, Classification**

How does it help to achieve our objective?

- It detects anomalous datas and removes them, then predicts the likelihood of having cardiovascular disease based on its different factors, and classifies them into positive and negative classes

Machine Learning Techniques and Models

01	02	03	04
One-Hot Encoding (OHE)	K-Means Clustering	Local Outlier Factor	Random Forest Classifier

Machine Learning Techniques and Models

1. One-Hot Encoding (OHE)

Description:
Representation of categorical data in binary values as machine learning algorithms cannot work with categorical data directly

Steps done to achieve this:

1. Identified and picked out the categorical columns in our data
2. Used manual OHE technique to load the data

Obtained:

- 3 Columns for Cholesterol (Normal, Above Normal, Well above Normal)
- 2 Columns for Cardiovascular disease (Yes, No)

Machine Learning Techniques and Models

1. One-Hot Encoding (OHE)

	Age	Height	Weight	BMI	Systolic BP	Diastolic BP	Cholesterol_Above Normal	Cholesterol_Normal	Cholesterol_Well Above Normal	Cardiovascular Disease_No	Cardiovascular Disease_Yes
0	50.0	1.68	62.0	21.97	110.0	80.0	0	1	0	1	0
1	55.0	1.56	85.0	34.93	140.0	90.0	0	0	1	0	1
2	51.0	1.65	64.0	23.51	130.0	70.0	0	0	1	0	1
3	48.0	1.69	82.0	28.71	150.0	100.0	0	1	0	0	1
5	60.0	1.51	67.0	29.38	120.0	80.0	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...
69993	53.0	1.72	70.0	23.66	130.0	90.0	0	1	0	0	1
69994	57.0	1.65	80.0	29.38	150.0	80.0	0	1	0	0	1
69995	52.0	1.68	76.0	26.93	120.0	80.0	0	1	0	1	0
69998	61.0	1.63	72.0	27.10	135.0	80.0	0	1	0	0	1
69999	56.0	1.70	72.0	24.91	120.0	80.0	1	0	0	1	0

61784 rows × 11 columns

Machine Learning Techniques and Models

2. K-Means Clustering

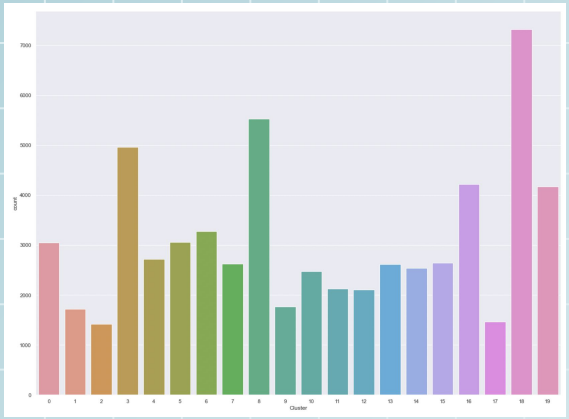
Description:
To group similar data points together and discover underlying patterns

Steps done to achieve this:

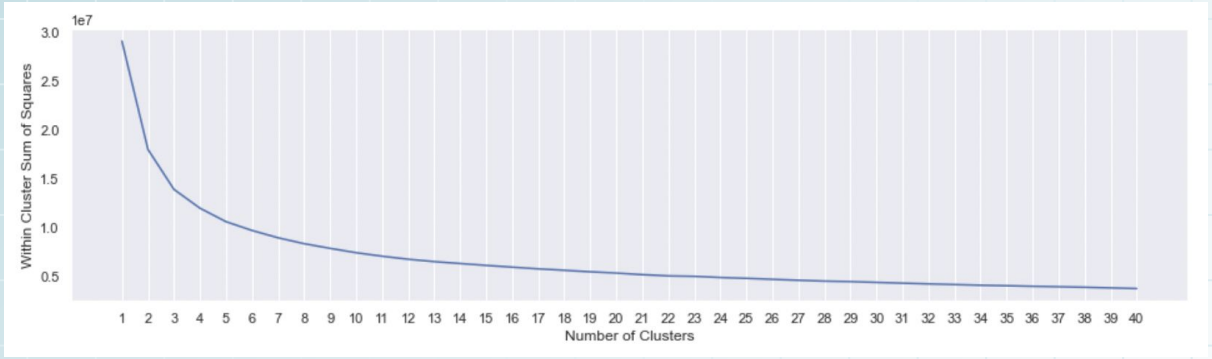
1. Define a fixed number (k) of clusters in the dataset
2. Using the clustering model, allocate each data point to each of the clusters through reducing the within cluster sum of squares

Machine Learning Techniques and Models

2. K-Means Clustering



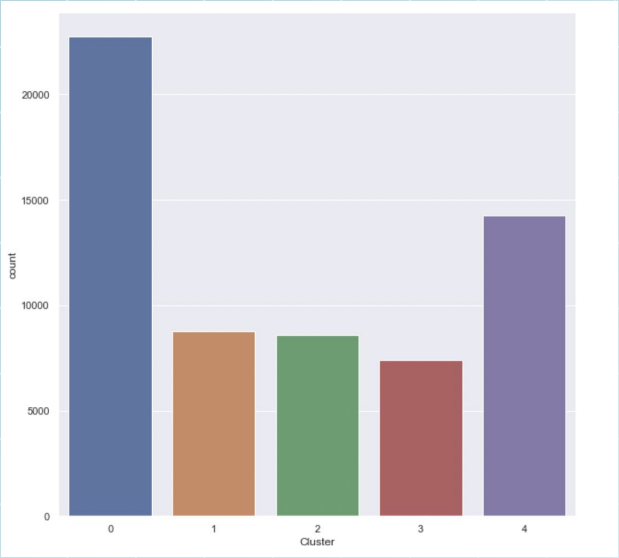
Clustering



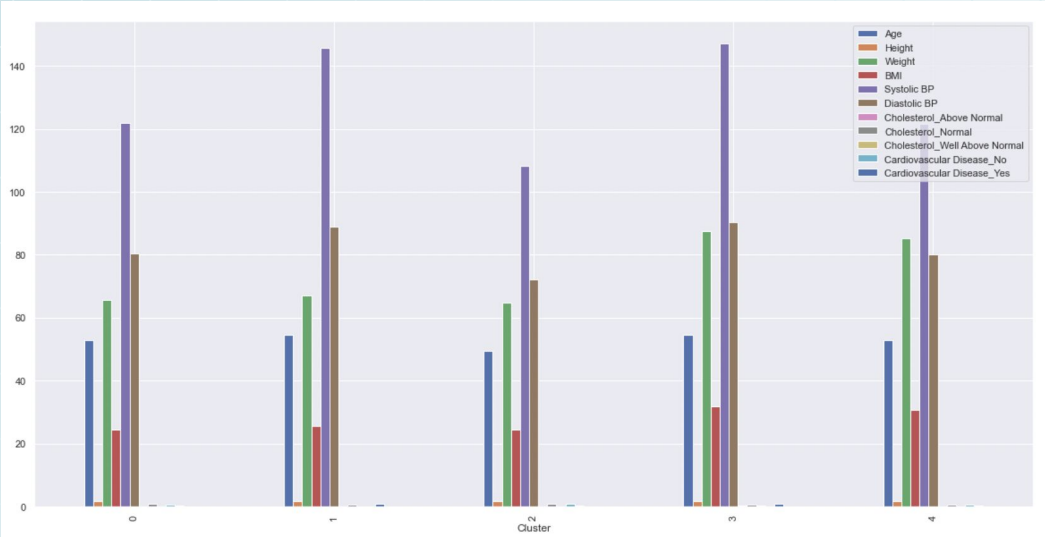
Number of Clusters against Within Cluster Sum of Squares

Machine Learning Techniques and Models

2. K-Means Clustering



Cluster Sum of Squares



Average behaviour of each cluster

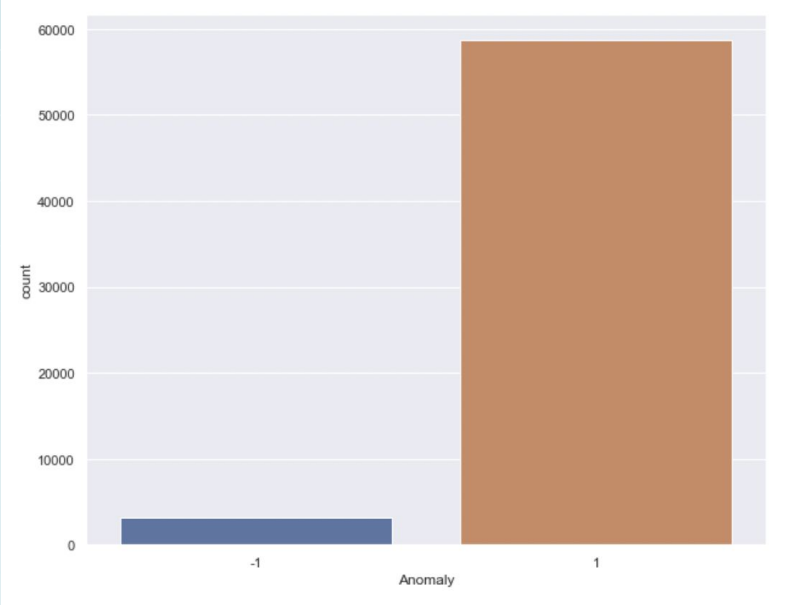
Machine Learning Techniques and Models

3. Local Outlier Factor (LOF)

Description:
Computes the local density deviation of a given data point with respect to its neighbors

Steps done to achieve this:

1. Define a fixed number (k) of neighbours for consideration
2. Using LOF, find out density of a certain point and compare if with density of other points
3. Predict Anomalies

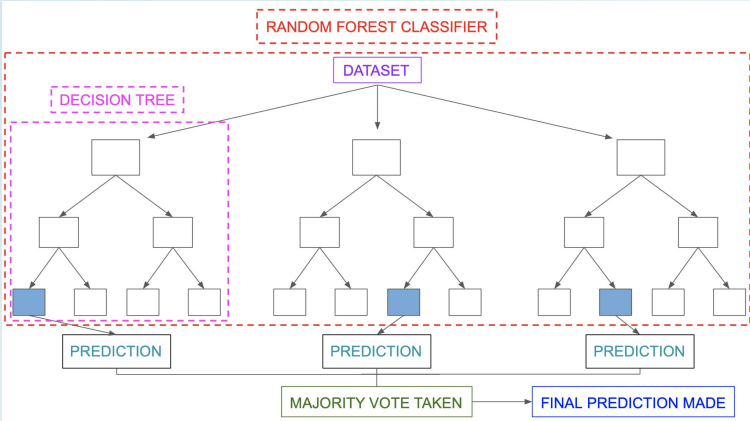


Machine Learning Techniques and Models

4. Random Forest Classifier

Description: Uses Ensemble Learning: Multiple decision trees are used to give a class prediction based on the factors in relation to Cardiovascular Disease

Purpose of Random Forest in the context of our project: Classifies the factors used, measures the effectiveness for predicting the likelihood of cardiovascular disease.



How Random Forest works:

1. Select random samples from a given dataset and split into Train and Test sets
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.

Evaluation

04

ML Tool Chosen: **Confusion Matrix, ROC Curve and AUC**

Evaluation

01	02
Confusion Matrix	ROC Curve and AUC

Evaluation

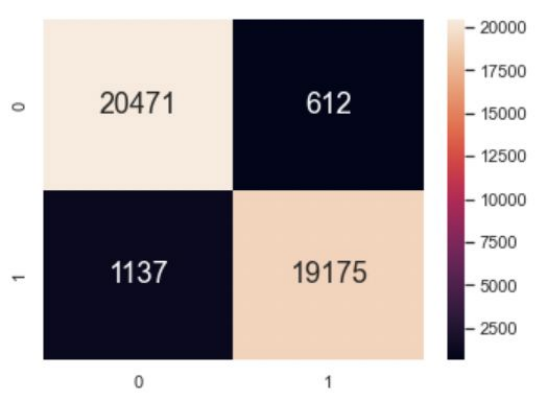
1. Confusion Matrix

Description: A confusion matrix is a summary of prediction results on a classification problem.The number of correct and incorrect predictions are summarized with count values and broken down by each class.

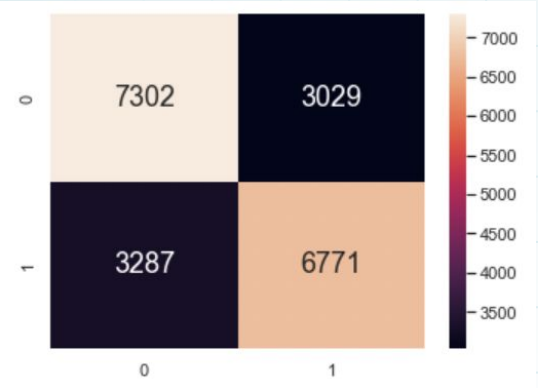
Purpose of Confusion Matrix in the context of our project:
To determine the accuracy of the likelihood of cardiovascular disease. It also gives us an idea on the proportion of data that are wrongly predicted and which category they fall under (false positive or false negative).

How Confusion Matrix works:

1. Apply classification model to the testing data (75% train, 25% test)
2. Construct confusion matrix using results (TP, TN, FP, FN)
3. Determine the accuracy rate of the classification model using TPR and TNR



Train dataset



Test dataset

Evaluation

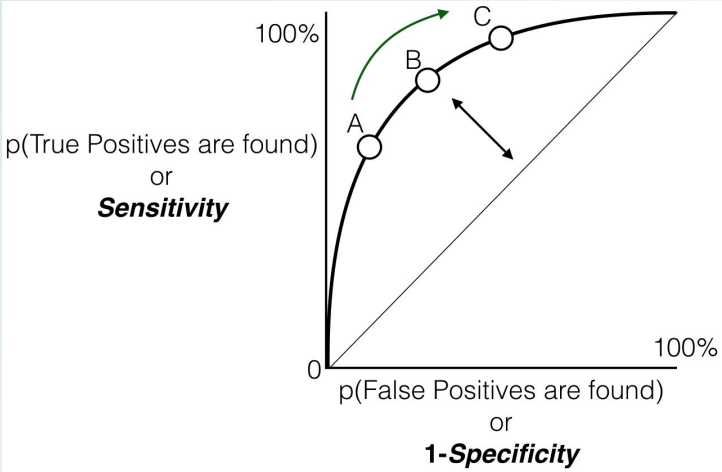
2. ROC Curve and AUC

Description: An ROC (Receiver Operating Characteristic) curve evaluate the performance of a binary classifier and AUC(area under curve) is used as the measurement for evaluation.

Purpose of ROC curve and AUC in the context of our project: To evaluate the performance of the random forest classifier

How ROC curve and AUC works:

1. Apply classification model to the testing data
2. Calculate the FPR and TPR
3. Plot a ROC curve using FPR and TPR
4. Calculate the AUC

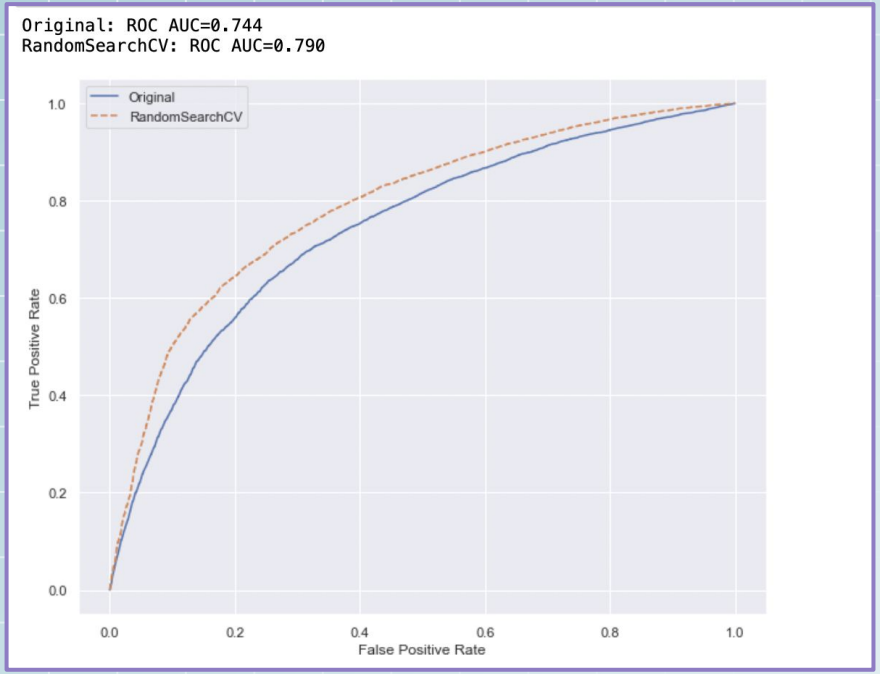


Evaluation

ROC curve

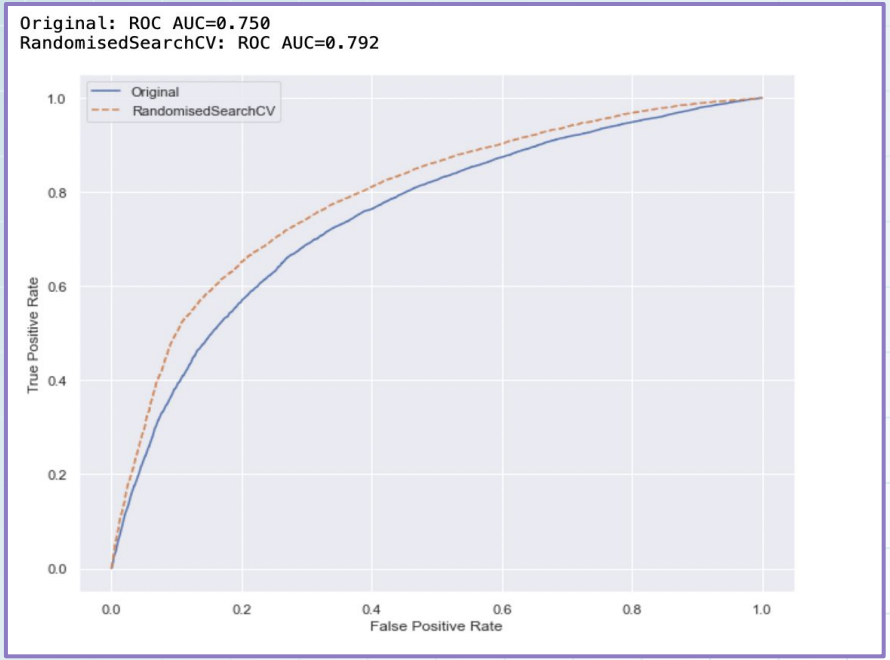
With Anomalies

Original: ROC AUC=0.744
RandomSearchCV: ROC AUC=0.790



Without Anomalies

Original: ROC AUC=0.750
RandomisedSearchCV: ROC AUC=0.792



Optimising Hyperparameters

05

Tool: RandomizedSearchCV

Optimising Hyperparameters

1. RandomizedSearchCV

Description: RandomizedSearchCV finds the best parameters to be applied on our model within a parameter grid, with stratified k-fold cross validation

Purpose of RandomizedSearchCV:
To explore if there exist a better set of parameters for our model

```
{'n_estimators': 500,
 'min_samples_split': 2,
 'min_samples_leaf': 1,
 'max_depth': 8}
```

How it works:

1. We split our train dataset into 5 parts randomly and applied our model with different parameters onto 4 of the subsets of data, using the 5th set as the test set.
2. The process is repeated with each of the 5 sets as the test set.
3. RandomizedSearchCV then picks out the most optimal parameters

Summary of Results

Hyperparameter	Original Value (Without removing Anomalies)	RandomizedSearchCV (Without removing Anomalies)
max_depth	None	15
min_samples_leaf	None	1
min_samples_split	None	2
n_estimators	None	500
Classification Accuracy	0.6902251214	0.7237726225
AUC	0.743606411	0.7926799353

Hyperparameter	Original Value (Removed Anomalies)	RandomizedSearchCV (Removed Anomalies)
max_depth	None	15
min_samples_leaf	None	1
min_samples_split	None	2
n_estimators	None	500
Classification Accuracy	0.6923076923	0.7293236964
AUC	0.7472283082	0.7944624545

Outcome

Through this project, we analyzed the data set, trained a classification model with classification, clustering and anomaly predication and evaluated the data.

We have built a relatively effective model of 0.794 AUC score to predict the likelihood of cardiovascular disease.

Job Distribution

Name	Chen Mei Ling	Vernis Aw Ning Min	Tan Tse Teng	Kester Toh
Exploratory Data and Analysis (Plots)	✓	✓	✓	x
Data Preparation	✓	supporting	supporting	x
Machine Learning	supporting	✓	✓	x
Evaluation + Optimising Hyperparameters	✓	✓	✓	supporting

Thank You!

References:

https://www.myheart.org.sg/health/heart-disease-statistics/
https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/