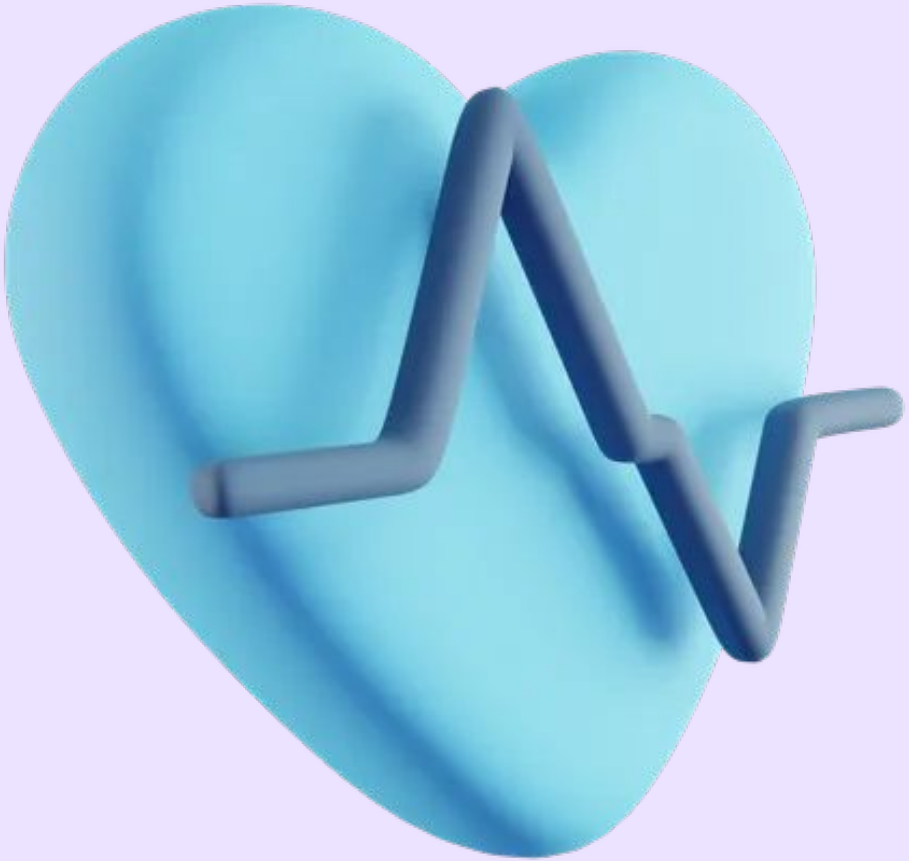
Gender-Based Risk Assessment of Cardiovascular Diseases

Team: Panda

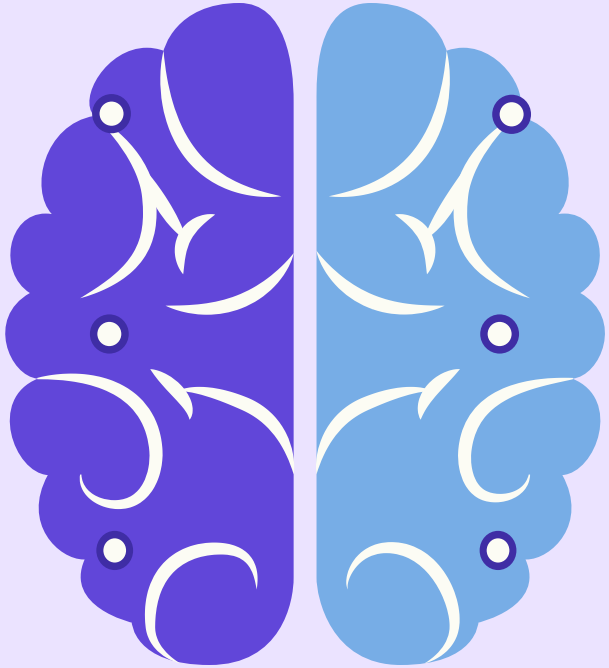
Karen Natalie (U2220586J)
Liu Yuheng (U2222313G)
Nichani Namya Ashok (U2223732C)
Su Rui (U2221036D)

Contents



1

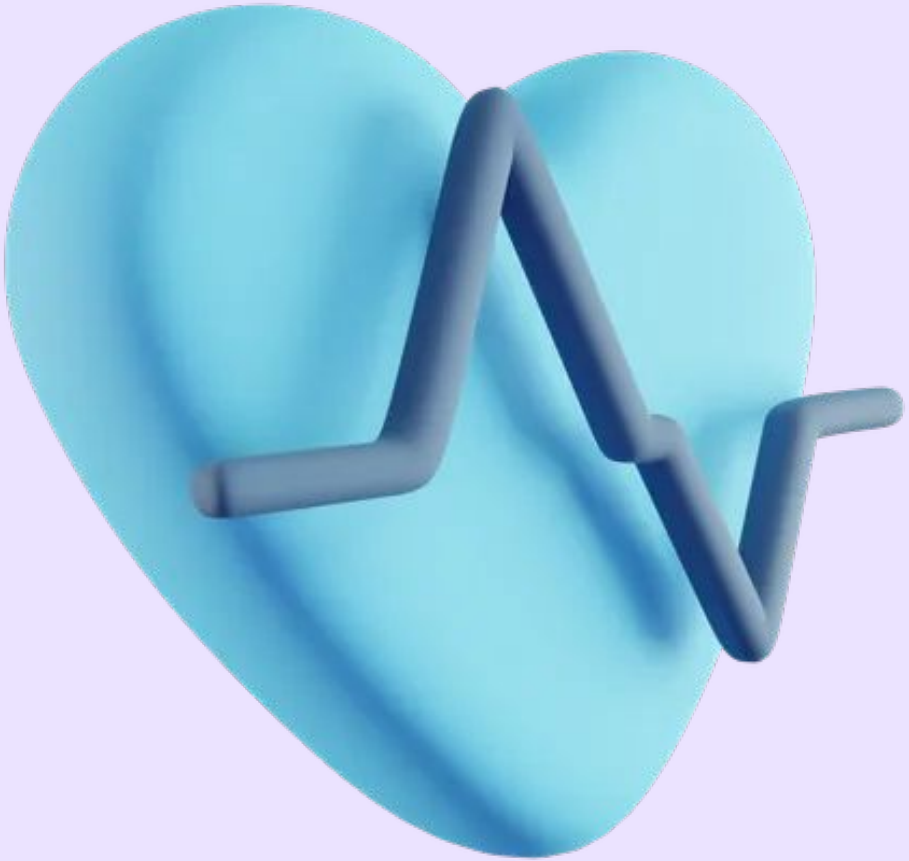
Introduction & Problem Statement



Problem Statement

How can we predict Cardiovascular Diseases based on gender?

Dataset: Cardiovascular Disease



2

Data Visualisation

Data Preparation

Before visualising the data, we cleaned the data by removing unrealistic values.

These included:

- ap_hi and ap_lo < 1
- ap_hi and ap_lo > 370
- ap_hi > ap_lo

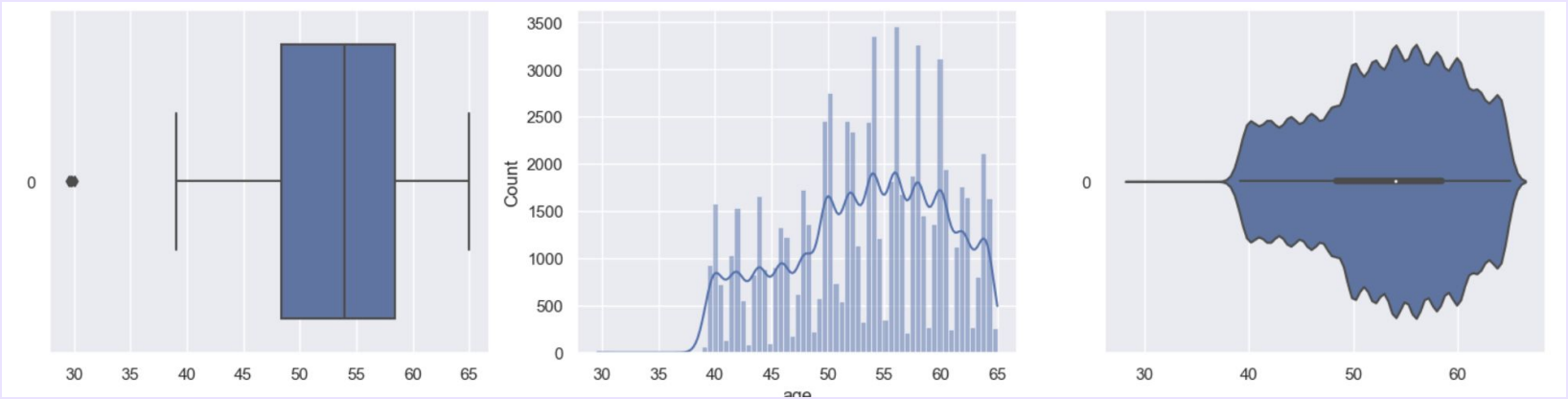
```
In [175]: # Filter
condition = ((cvd_df["ap_hi"]<370 )&(cvd_df["ap_hi"]>1) & (cvd_df["ap_lo"]<370)&(cvd_df["ap_lo"]>1) & (cvd_df["ap_hi"]>cvd_df["ap
unrealistic = cvd_df[~condition]["age"].count()
print(f"{unrealistic} unrealistic data removed")

cvd_df = cvd_df[condition]
cvd_df.describe().round(2)

1292 unrealistic data removed
```

Numerical Data Visualisation

Variables

1. Age
2. Height
3. Weight
4. Ap_hi (Systolic blood pressure)
5. Ap_lo (Diastolic blood pressure)

Numerical Data Visualisation

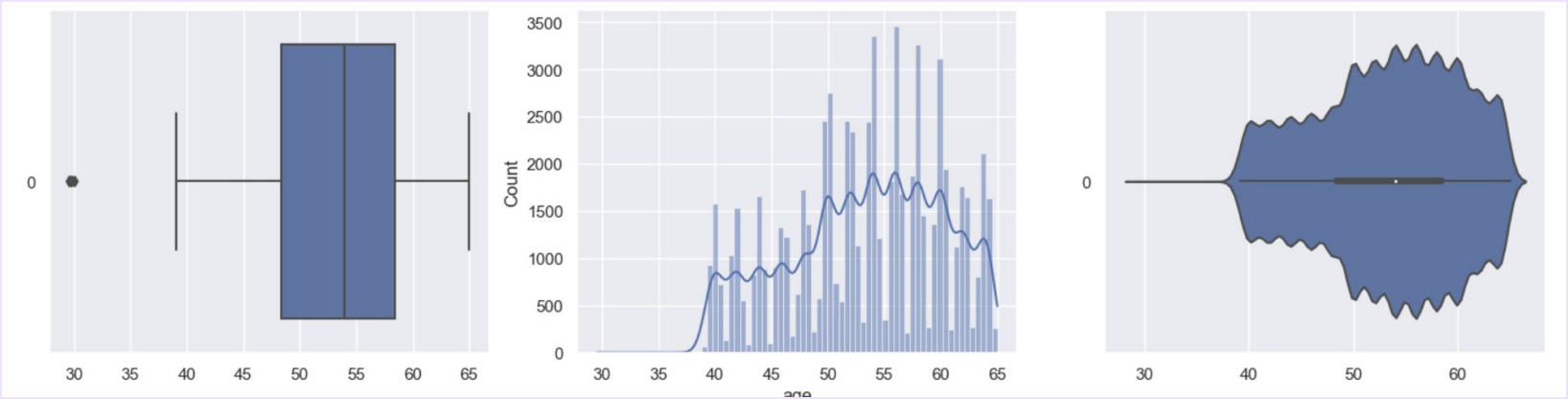


Box Plot

Histogram

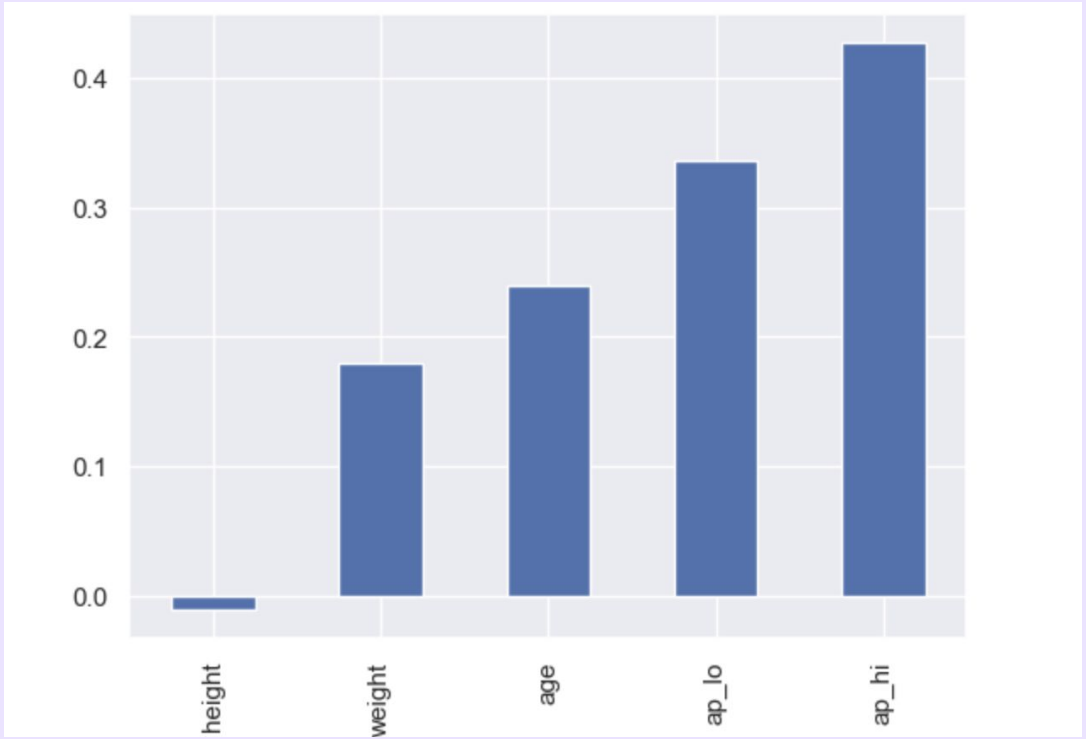
Violin Plot

Numerical Data Visualisation



Generally, the graphs showed that there were many outliers far from the mean for each variable. The trained model is expected to have better performance once the outliers are filtered out.

Numerical Data Visualisation



From the correlation plot:

Moderate positive correlation (0.3, 0.5] - ap_hi

Low positive correlation (0.1-0.3] - weight, age, ap_lo

No correlation (near zero) - height

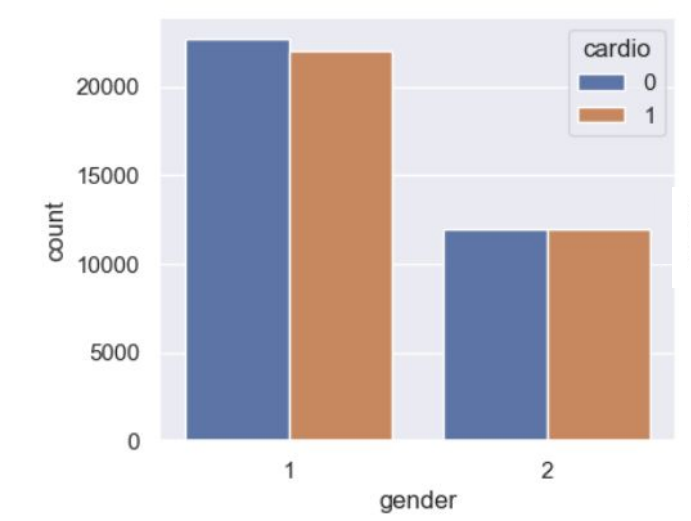
Categorical Data Visualisation

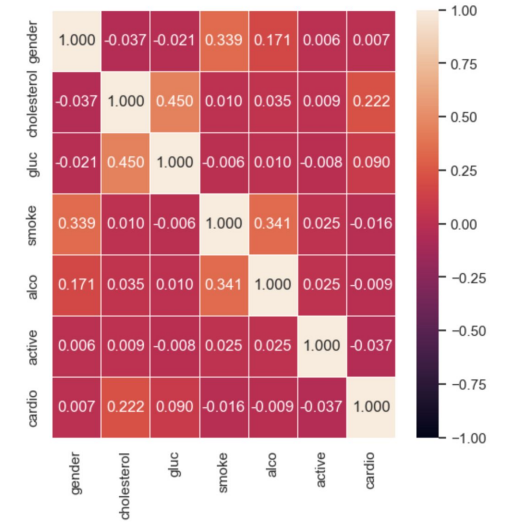
Variables

1. Gender
2. Cholesterol
3. Glucose
4. Smoke
5. Alcohol
6. Active

Categorical Data Visualisation

Subplot

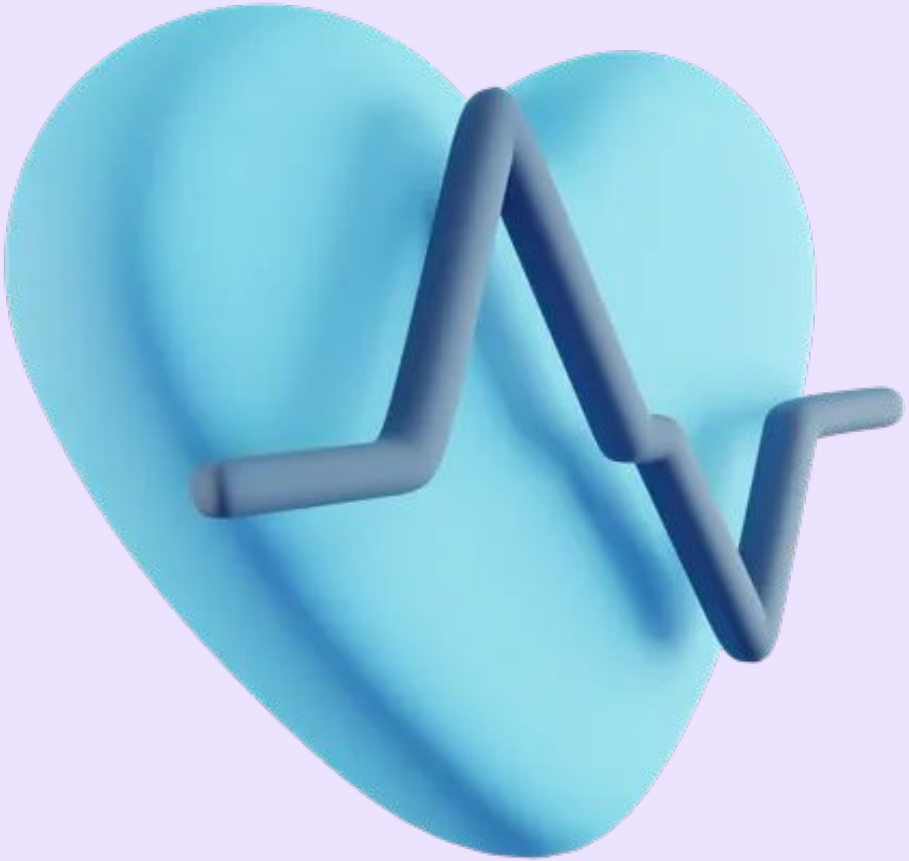




Heatmap

Data is balanced with regards to whether one has cardiovascular disease or not, but it is skewed towards the female gender.

Cholesterol has the highest correlation of 0.22 with cardiovascular disease. However, it is still a low positive correlation. The remaining variables have near zero correlation.



3

Data Manipulation

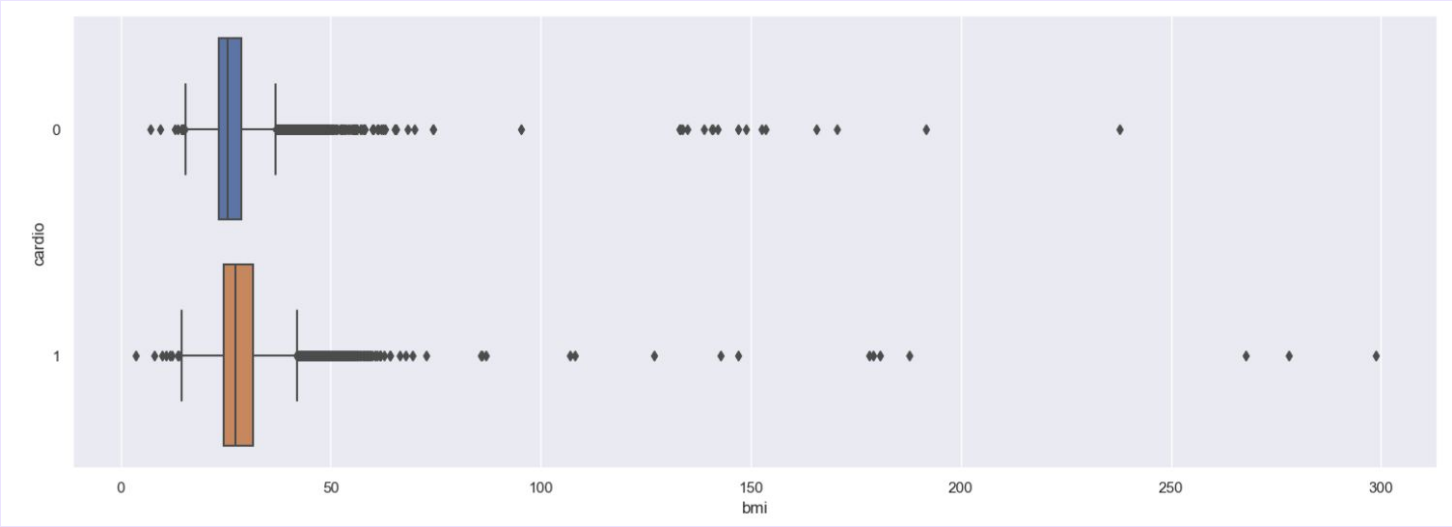
Body Mass Index (BMI)

```
# Calculate and include BMI for all rows
cvd_df["bmi"] = round((cvd_df["weight"] / (cvd_df['height']/100)**2), 1)
cvd_df.head(3)
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	bmi
0	50.4	2	168	62.0	110	80	1	1	0	0	1	0	22.0
1	55.4	1	156	85.0	140	90	3	1	0	0	1	1	34.9
2	51.7	1	165	64.0	130	70	3	1	0	0	0	1	23.5

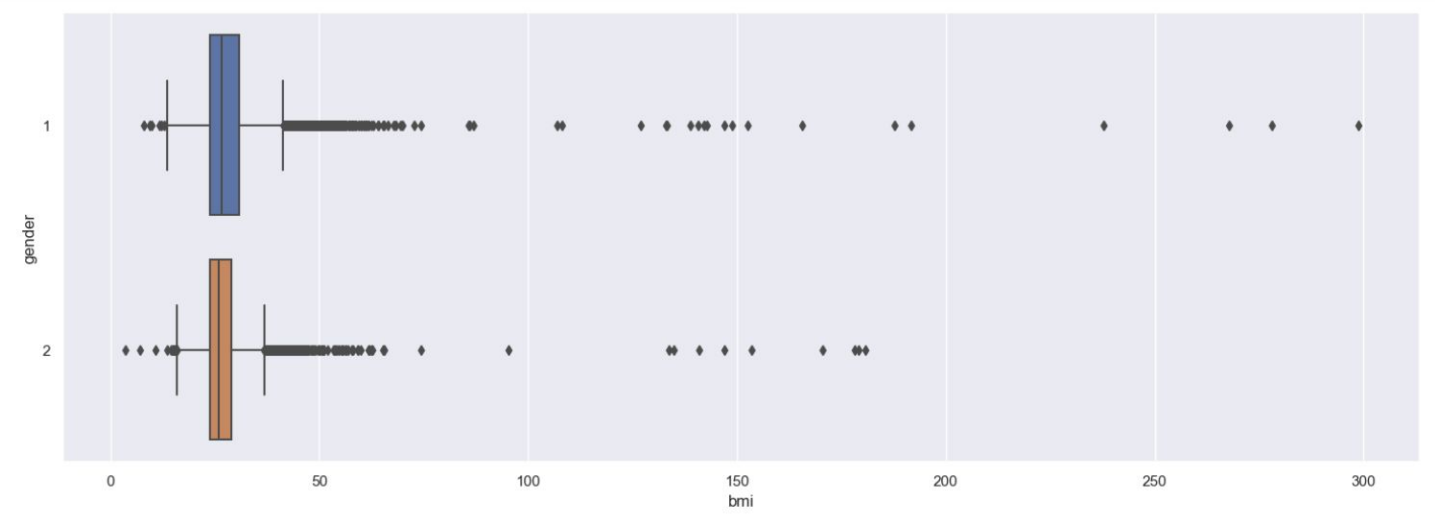
We calculated and added BMI into the dataset as it makes use of both height and weight for further analysis. This helps us to account for differences in body composition based on gender, and could be a potential predictor for cardiovascular disease.

BMI against Cardio



However, the boxplot visualisations of both variables did not yield any obvious relationships, as it appears that BMI seems to be similar for both cardiovascular patients and non-patients.

BMI across Genders



The boxplot visualisations of BMI and gender showed that males tend to have a lower BMI overall, while females have a much larger spread of BMI (likely due to greater number of data values) and a slightly higher mean.

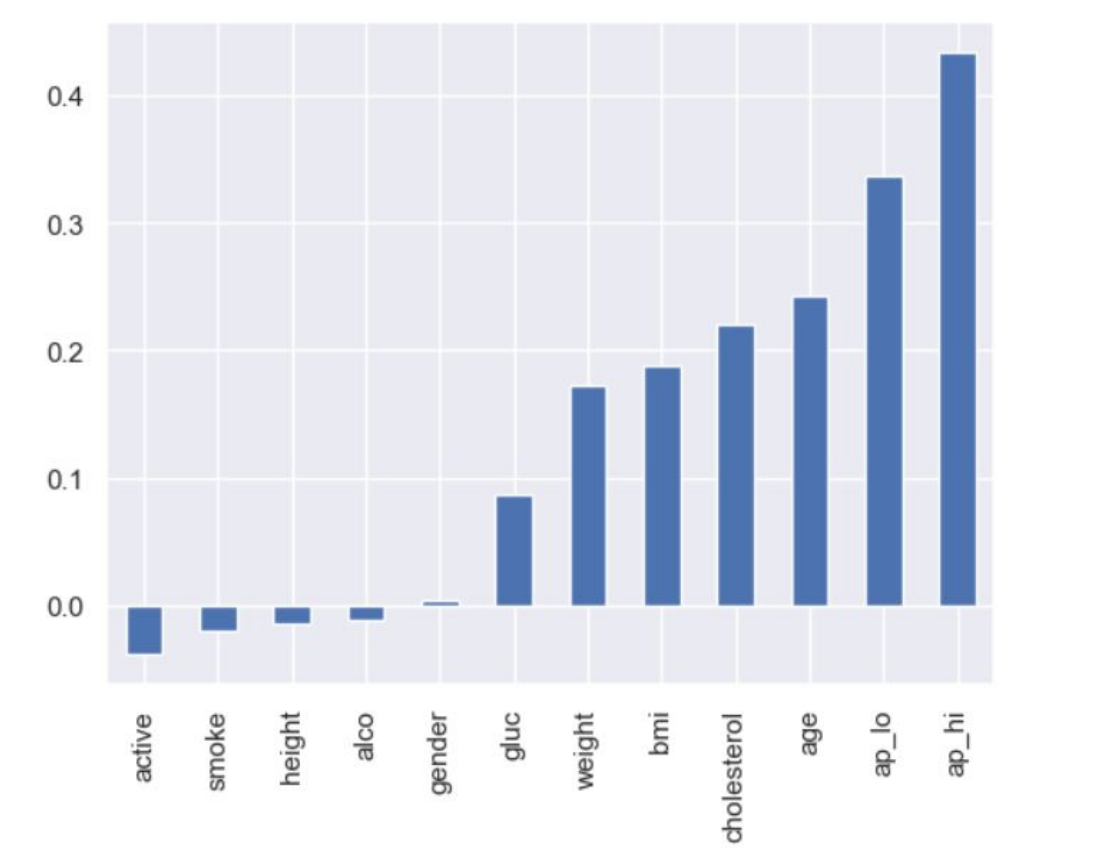
Removing Outliers

```
2901 outliers removed
<class 'pandas.core.frame.DataFrame'>
Index: 65807 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   age          65807 non-null  float64
 1   gender       65807 non-null  int64
 2   height       65807 non-null  int64
 3   weight       65807 non-null  float64
 4   ap_hi        65807 non-null  int64
 5   ap_lo        65807 non-null  int64
 6   cholesterol  65807 non-null  int64
 7   gluc         65807 non-null  int64
 8   smoke        65807 non-null  int64
 9   alco         65807 non-null  int64
 10  active       65807 non-null  int64
 11  cardio       65807 non-null  int64
 12  bmi          65807 non-null  float64
dtypes: float64(3), int64(10)
memory usage: 7.0 MB
```

To further balance the dataset, Z-Score method was used to remove outliers.

Only columns with numerical data (age, height, weight, ap_hi, ap_lo, bmi) were considered in this process.

Bar Plot for Correlation



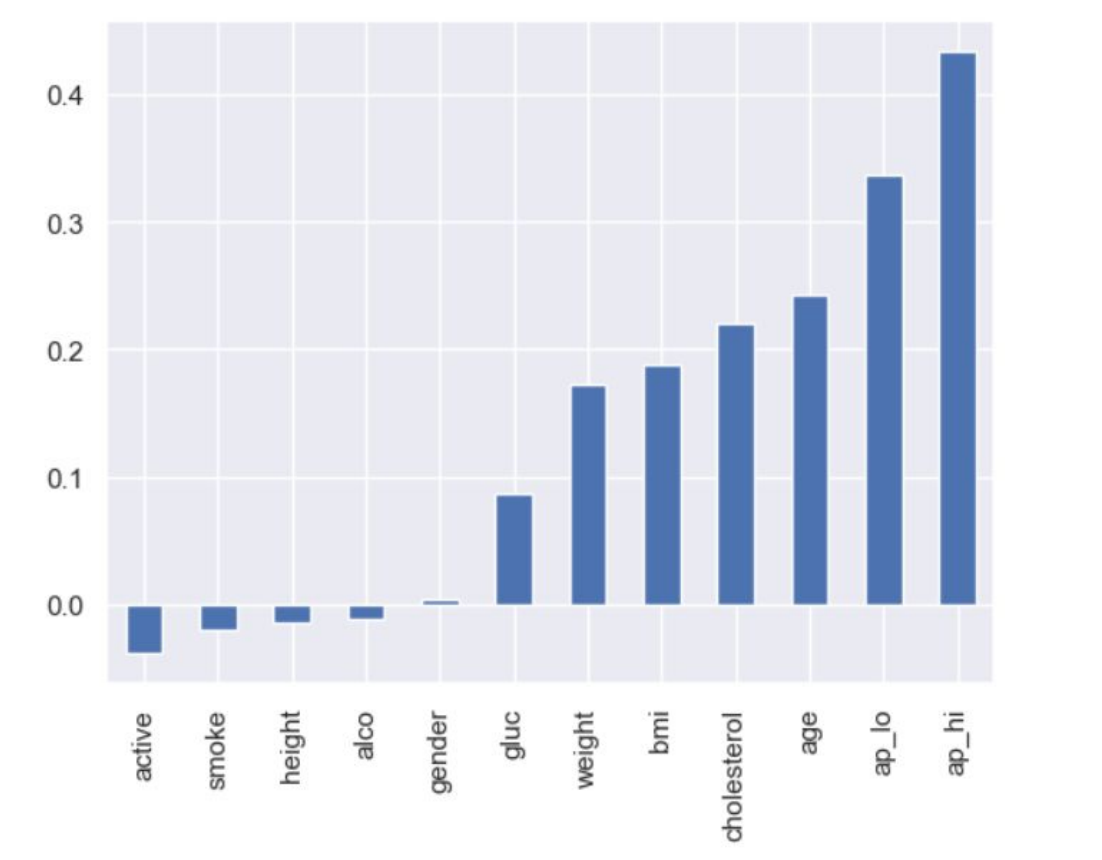
From the correlation plot:

Moderate positive correlation (0.3, 0.5] - ap_hi, ap_lo

Low positive correlation (0.1-0.3] - weight, bmi, cholesterol, age

No correlation (near zero) - active, smoke, height, alcohol, gender, glucose

Bar Plot for Correlation



As such, we can tell which variables have a higher positive correlation with the ‘cardio’ variable. It helps us select influential variables for further exploration of cardiovascular disease prediction.

Splitting the Dataset

Split into train and test

```
# Import train_test_split from sklearn
from sklearn.model_selection import train_test_split
```

For data with outliers:

```
# Split the Dataset into Train and Test
model_train, model_test = train_test_split(cvd_df, test_size = 0.25)
```

```
# Save test dataset
model_test.to_csv('model_with_outliers_test.csv', sep = ',', index = False)
```

For data without outliers:

```
# Split the Dataset into Train and Test
model2_train, model2_test = train_test_split(cvd_df_no_outliers, test_size = 0.25)
```

```
# Save test dataset
model2_test.to_csv('model_without_outliers_test.csv', sep = ',', index = False)
```

The dataset was split into train and test sets with 0.25 ratio, then each set was split further based on gender.

This process was repeated for the dataset with and without outliers to analyse whether the removal of outliers is effective.

Splitting the Dataset

Split according to gender

in dataset, male is represented by 1, female is 2, dtype int64

For data with outliers:

```
# Split model training dataset with outliers
female_df = model_train[model_train['gender'] == 1]
male_df = model_train[model_train['gender'] == 2]
```

```
# Save datasets
female_df.to_csv('female_with_outliers.csv', sep = ',', index = False)
male_df.to_csv('male_with_outliers.csv', sep = ',', index = False)
```

For data without outliers:

```
# Split model training dataset without outliers
female_df_no_outliers = model2_train[model2_train['gender'] == 1]
male_df_no_outliers = model2_train[model2_train['gender'] == 2]
```

```
# Save datasets
female_df_no_outliers.to_csv('female_without_outliers.csv', sep = ',', index = False)
male_df_no_outliers.to_csv('male_without_outliers.csv', sep = ',', index = False)
```

Classification

```
female_without_outliers = pd.read_csv('female_without_outliers.csv')
female_without_outliers.head()
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	bmi
0	56.2	1	165	72.0	100	80	1	2	0	0	1	0	26.4
1	53.6	1	155	72.0	120	80	1	1	0	0	1	0	30.0
2	47.8	1	150	59.0	140	90	2	3	0	0	1	1	26.2
3	64.2	1	169	67.0	120	80	1	1	0	0	0	0	23.5
4	63.6	1	156	83.0	150	100	3	3	0	0	1	1	34.1

```
female_with_outliers = pd.read_csv('female_with_outliers.csv')
female_with_outliers.head()
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	bmi
0	55.3	1	160	70.0	140	100	3	1	0	1	1	1	27.3
1	44.5	1	150	56.0	110	80	1	1	0	0	1	0	24.9
2	58.0	1	158	82.0	120	80	3	3	0	0	1	1	32.8
3	60.3	1	161	79.0	140	80	3	1	0	0	1	0	30.5
4	60.0	1	175	120.0	130	90	3	1	0	0	1	1	39.2

Female - with and without outliers

Classification

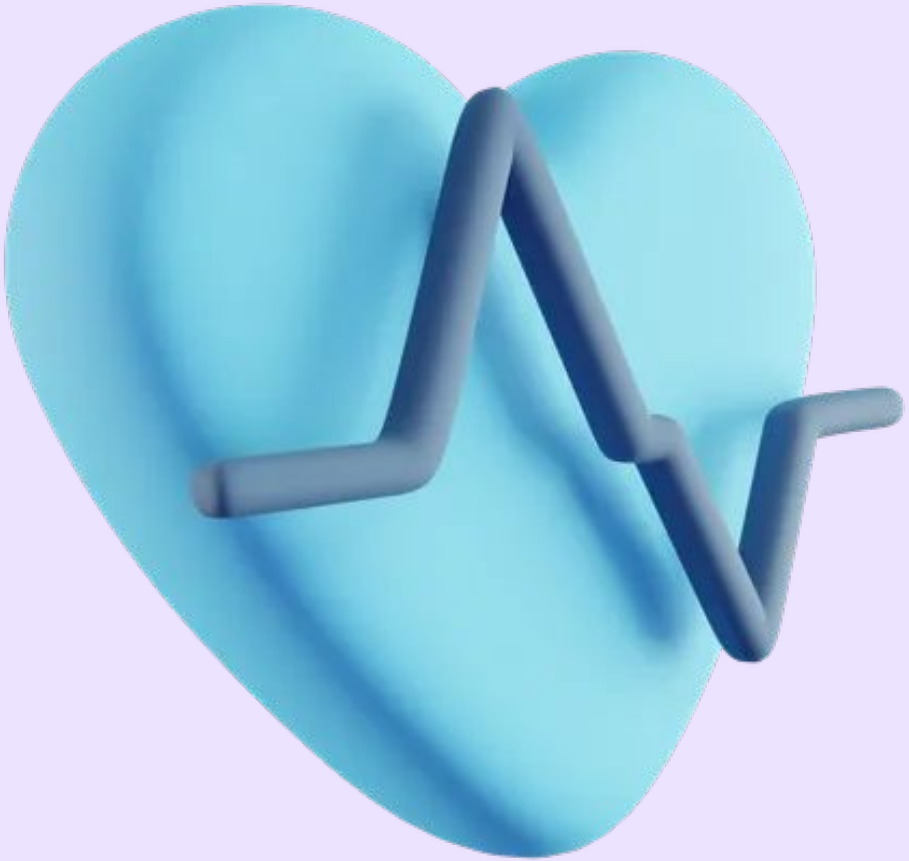
```
male_without_outliers = pd.read_csv('male_without_outliers.csv')
male_without_outliers.head()
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	bmi
0	46.2	2	164	72.0	80	60	1	1	1	0	0	0	26.8
1	64.4	2	163	50.0	140	90	2	1	0	0	1	1	18.8
2	60.1	2	170	85.0	120	60	1	1	0	0	1	0	29.4
3	53.9	2	168	83.0	130	80	1	1	0	0	1	0	29.4
4	54.7	2	170	75.0	120	80	1	1	1	0	1	0	26.0

```
male_with_outliers = pd.read_csv('male_with_outliers.csv')
male_with_outliers.head()
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	bmi
0	64.0	2	169	77.0	120	80	1	1	0	0	1	1	27.0
1	53.8	2	172	78.0	150	100	1	1	1	1	1	1	26.4
2	42.2	2	168	70.0	120	80	1	1	0	0	1	0	24.8
3	48.2	2	169	68.0	120	80	1	1	0	0	1	1	23.8
4	63.3	2	165	66.0	120	80	1	1	0	0	1	1	24.2

Male - with and without outliers



4

Analysing Data

Analysis

1. Female without Outliers (fwo)
2. Female with Outliers (fo)
3. Male without Outliers (mwo)
4. Male with Outliers (mo)

Analysis

We have performed the following on both male and female datasets, with and without outliers:

1. Simple Decision Tree
2. Multi-Variate Classification Tree
3. Chi Square Test of Independence

Model selection

Why **Decision Tree**?

Decision trees are easy to interpret and visualize.

It can easily capture non-linear patterns.

It requires fewer data preprocessing for example, there is no need to normalize columns.

The decision tree has no assumptions about distribution because of the non-parametric nature of the algorithm.

Model selection

What's the **trade offs**?

Sensitive to noisy data. It can overfit noisy data.

Decision trees are biased with imbalance dataset.

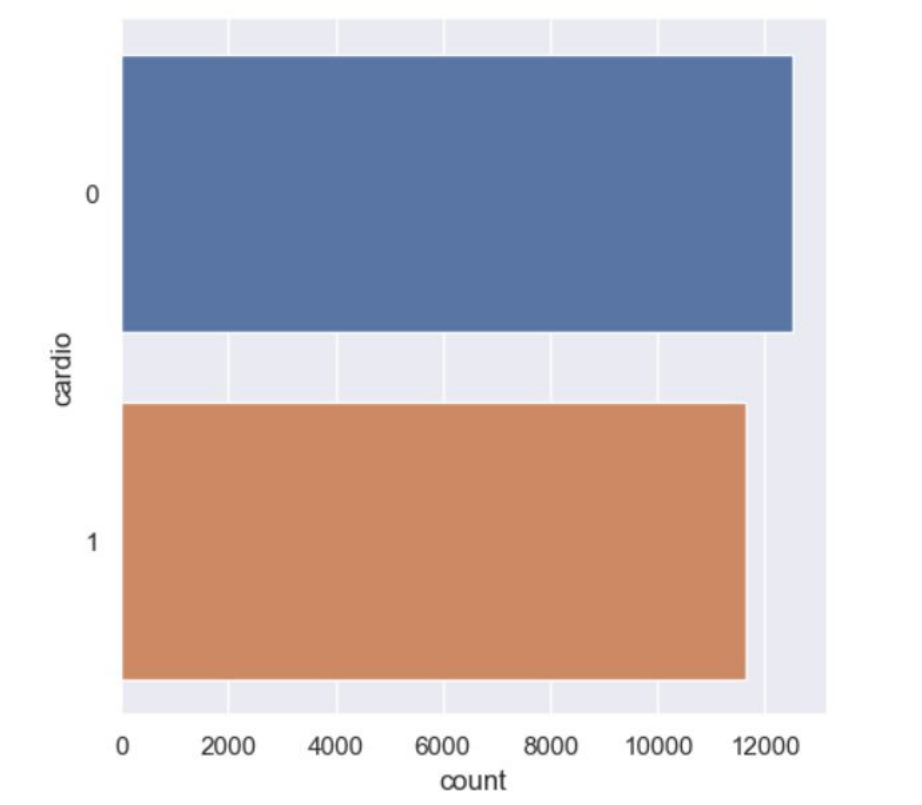
How we **minimise** it?

We cleaned the unrealistic data points and the outliers and trained separate models using data with and without outliers.

Analysis was done on each gender set separately, ensuring that gender did not imbalance the dataset as it was skewed towards females.

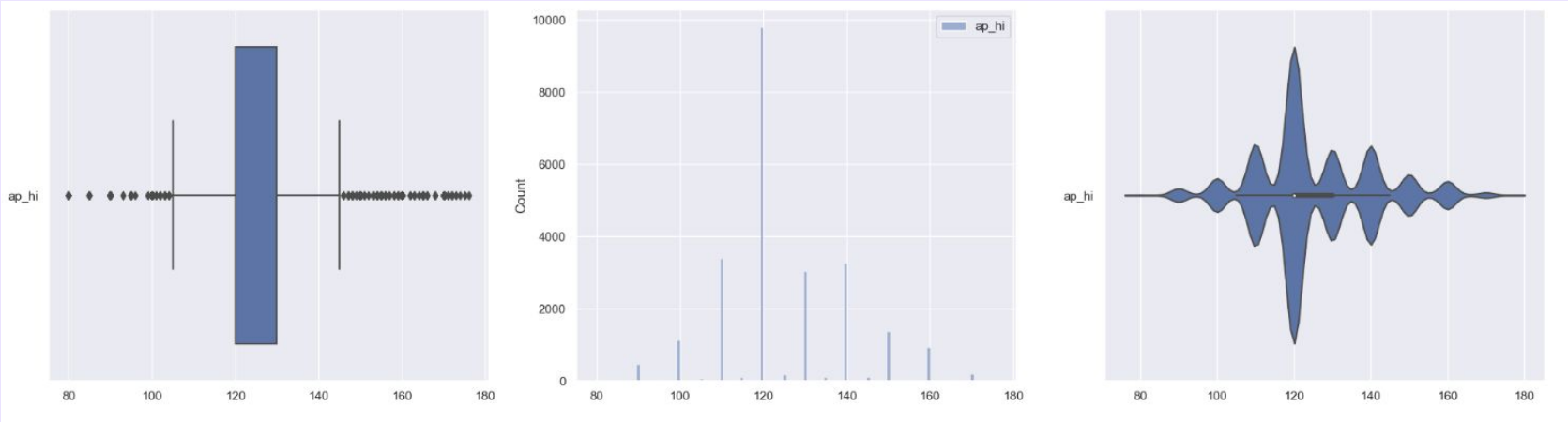
Tests Conducted on Each Gender Specific Dataset (Female without Outliers)

Basic Exploration: Count Plot



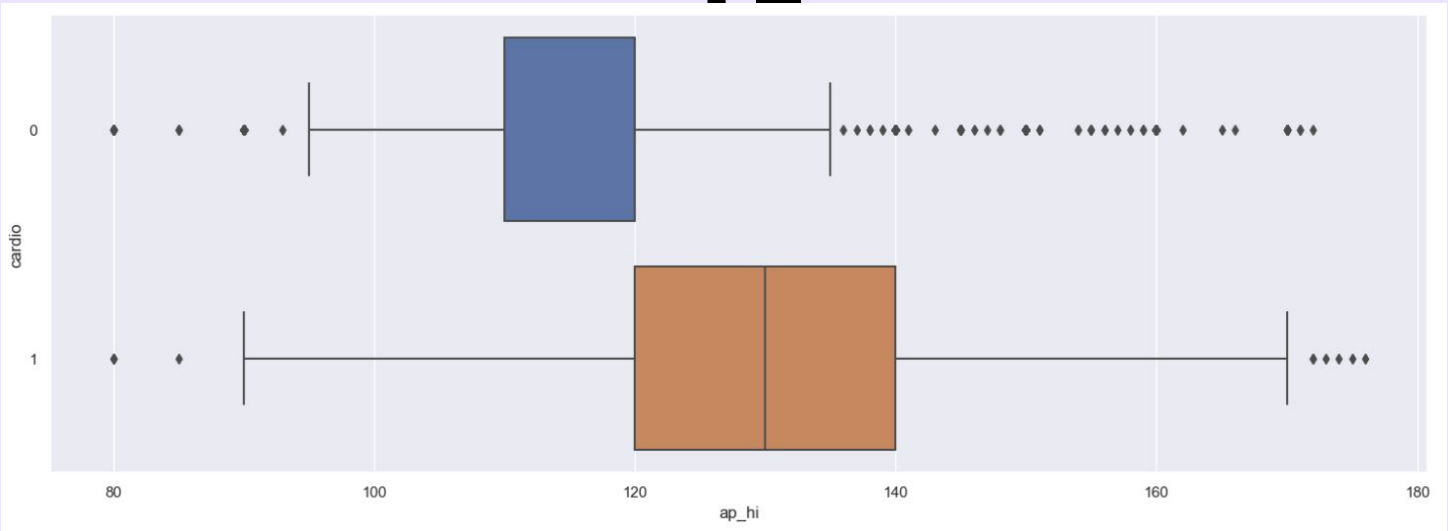
The count plot for cardio_fwo Train presents the distribution of the variable and shows a balanced dataset.

Basic Exploration: Distributions



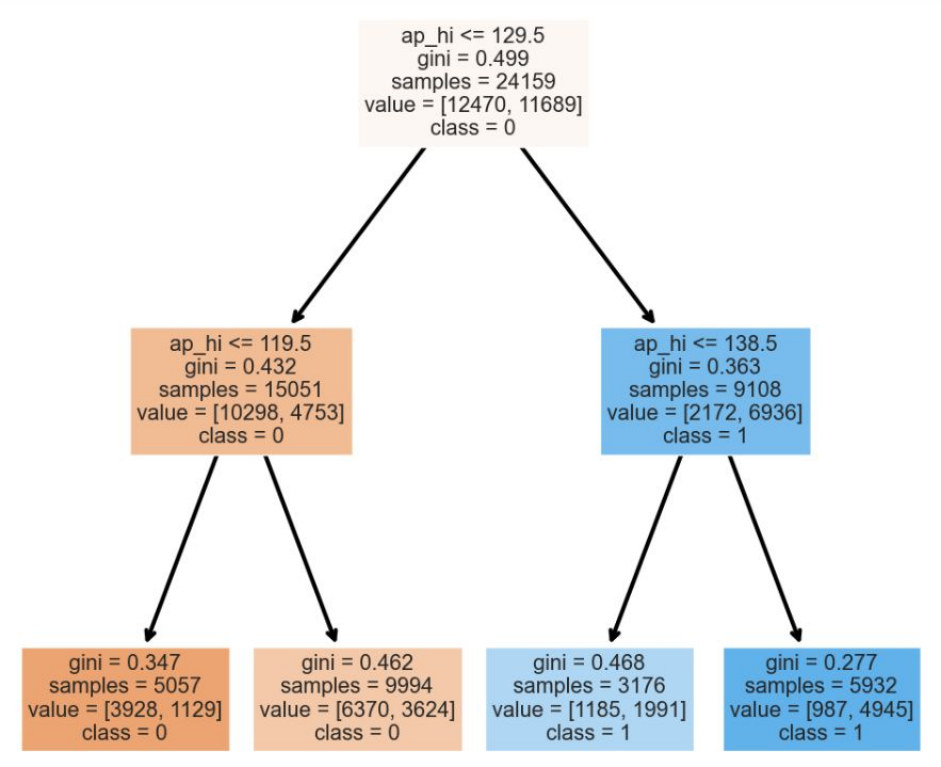
Box plot, histogram plot and violin plot were utilised to visualise the spread of data for ap_hi.

Basic Exploration: Cardio against Ap_hi



Since ap_hi has the highest correlation for predicting cardiovascular disease, both variables were visualised in box plots.

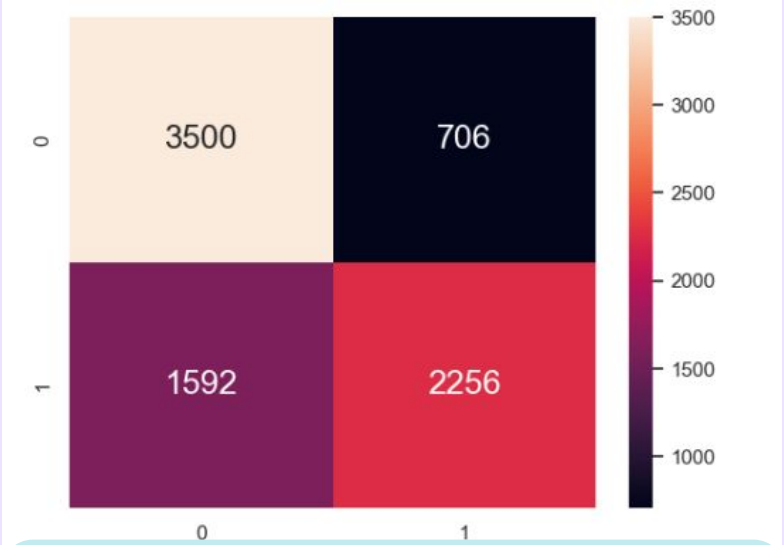
Simple Decision Tree



Ap_hi was also similarly selected for the simple decision tree to predict the risk of cardiovascular disease in patients within certain ranges of ap_hi.

Goodness of Fit of the Model

Confusion matrix



$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = 0.715$$

TP: Correct prediction on one having cvd
TN: Correct prediction on one not having cvd
FP: One with no cvd but predicted to have
FN: One with cvd but predicted to not have

It is important to not miss any potential cvd patient. Thus we calculate the False Negative Rate (FNR) and the Recall (True Positive Rate - TPR).

$$FNR = \frac{FN}{TP + FN} = \frac{1592}{2256 + 1592} = 0.413$$

$$TPR = \frac{TP}{TP + FN} = \frac{2256}{2256 + 1592} = 0.586$$

Chi Square Test of Independence

In [100]:
```
crosstab_fwo, test_results_fwo, expected_fwo = rp.crosstab(female_without_outliers["cholesterol"], female_without_outliers["cardi
                                               test= "chi-square",
                                               expected_freqs= True,
                                               prop= "cell")

crosstab_fwo
```

Out[100]:

	cardio		
cardio	0	1	All
cholesterol			
1	43.10	31.45	74.55
2	5.91	7.82	13.73
3	2.71	9.01	11.72
All	51.73	48.27	100.00

The Chi Square Test was chosen as another means to investigate the relationship between the categorical variables and the risk of cardiovascular disease, by analysing the difference between expected values and observed values.
Firstly, researchpy was installed and a crosstab of the categorical variable was created.

Chi Square Test of Independence

	Chi-square test	results
0	Pearson Chi-square (2.0) =	1755.1172
1	p-value =	0.0000
2	Cramer's V =	0.2334

Cholesterol

	Chi-square test	results
0	Pearson Chi-square (2.0) =	324.5630
1	p-value =	0.0000
2	Cramer's V =	0.1004

Glucose

The Chi Square Test was run using researchpy and results were tabulated containing the test statistic and p-value.

Chi Square Test of Independence

	Chi-square test	results
0	Pearson Chi-square (1.0) =	0.3103
1	p-value =	0.5775
2	Cramer's phi =	0.0031

Smoke

	Chi-square test	results
0	Pearson Chi-square (1.0) =	29.2962
1	p-value =	0.0000
2	Cramer's phi =	0.0302

Active

	Chi-square test	results
0	Pearson Chi-square (1.0) =	0.0090
1	p-value =	0.9245
2	Cramer's phi =	0.0005

Alcohol

Each categorical variable was investigated individually and results are as shown.

Chi Square Test of Independence

Variable	p-value
Cholesterol	0.0000
Glucose	0.0000
Smoke	0.5775
Active	0.0000
Alcohol	0.9245

At 5% significance level, the significant variables were cholesterol, glucose and active as their p-values were <0.05.

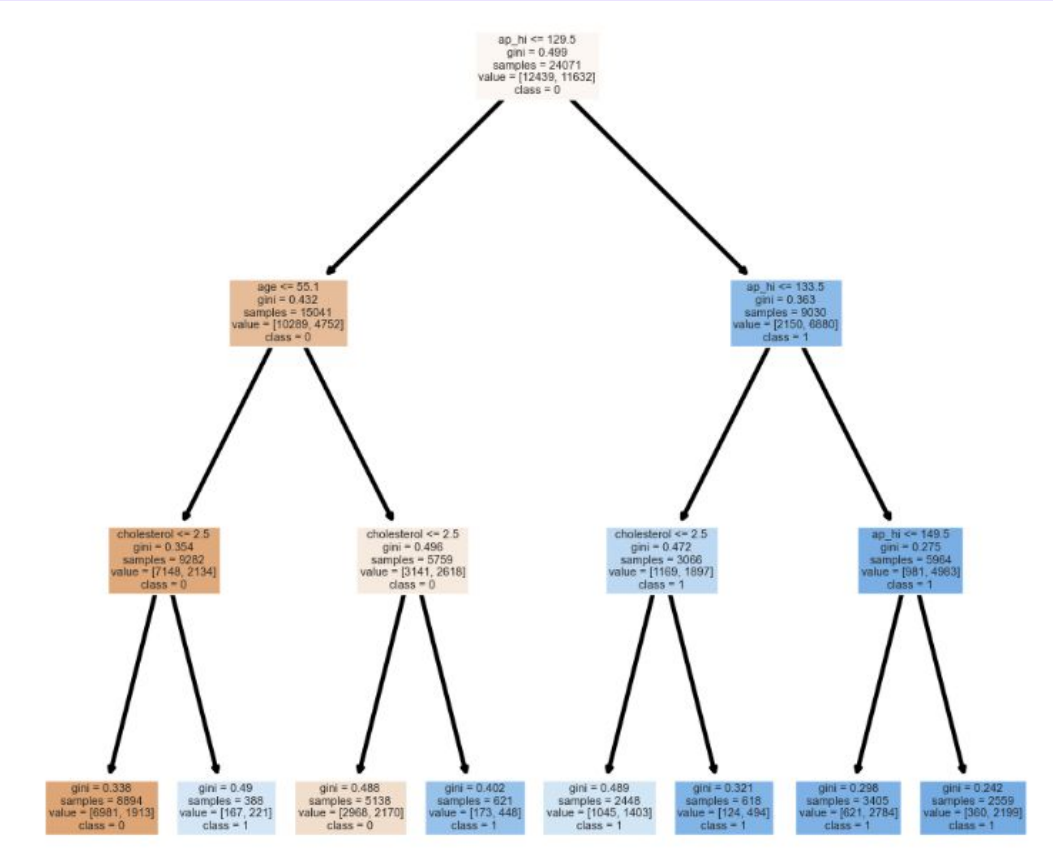
Chi Square Test of Independence

Phi and Cramer's V	Interpretation
>0.25	Very strong
>0.15	Strong
>0.10	Moderate
>0.05	Weak
>0	No or very weak

Variable	Cramer's V
Cholesterol	0.2334
Glucose	0.1004
Active	0.0302

Cramer's V was also calculated to investigate the strength of the relationship with cardiovascular disease. From our results, cholesterol had a strong relationship, glucose had a moderately strong relationship, active had a very weak relationship. These significant variables were subsequently used in the multi-variate decision tree.

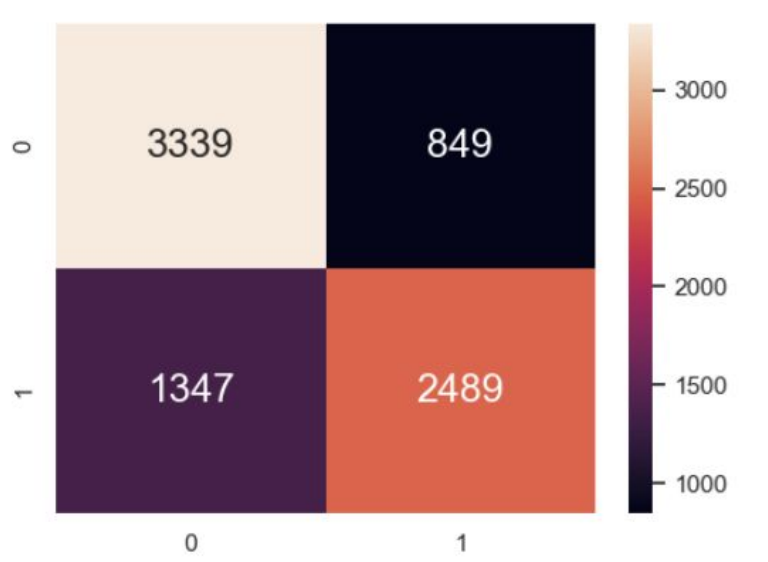
Multi-Variate Classification Tree



The most important variables are Ap_hi, Age and Cholesterol.

Goodness of Fit of the Model

Confusion matrix



$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = 0.726$$

It is important to not miss any potential cardiovascular disease patient. Thus we calculate the FNR and Recall (TPR).

$$\text{FNR} = \frac{FN}{TP + FN} = \frac{1347}{2489 + 1347} = 0.351$$

$$\text{TPR} = \frac{TP}{TP + FN} = \frac{2489}{2489 + 1347} = 0.649$$

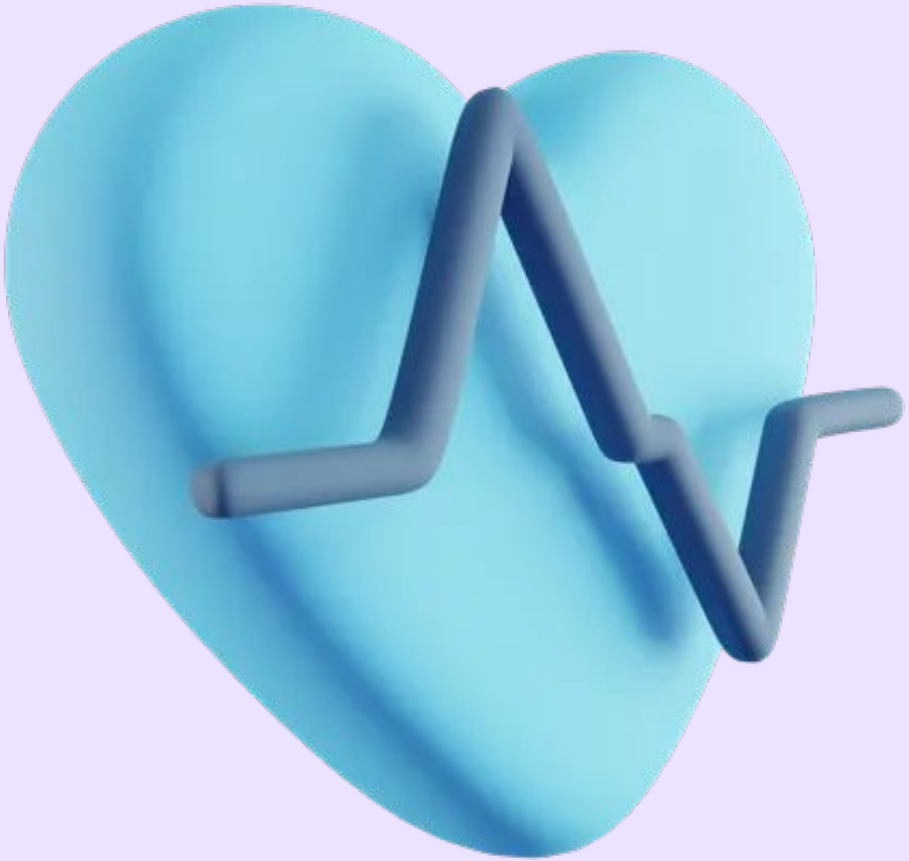
Multivariate model has both greater Accuracy and Recall, and lower FNR.

Male without Outliers (mwo)

By repeating the same steps above, we found that the significant factors related to CVD in males are:
Ap_hi, Age, Ap_lo, cholesterol

Analysis on Test Set

We analysed FWO and MWO test sets using chi square test and multivariate decision tree. We found the results of the significant variables consistent with the train set results.



5

Conclusion

Train Set: Comparison between models

Gender	Variables used in Decision Tree	With/without Outliers	Accuracy	Recall (TPR)
Female	Uni: Ap_hi	Without	0.715	0.687
Female	Multi: All significant	Without	0.726	0.649
Female	Uni: Ap_hi	With	0.710	0.606
Female	Multi: All significant	With	0.722	0.659
Male	Uni: Ap_hi	Without	0.712	0.619
Male	Multi: All significant	Without	0.711	0.526
Male	Uni: Ap_hi	With	0.709	0.641
Male	Multi: All significant	With	0.715	0.646

- Removal of outliers increases the consistency of model (results of significant variables were more consistent after repeated runs), while accuracy and tpr/fnr are similar
- Female model is slightly more accurate than male in general
- Multi-variate Decision Tree is more accurate than Uni-variate as it considers more than one feature

Test Set: Chi Square Test of Independence

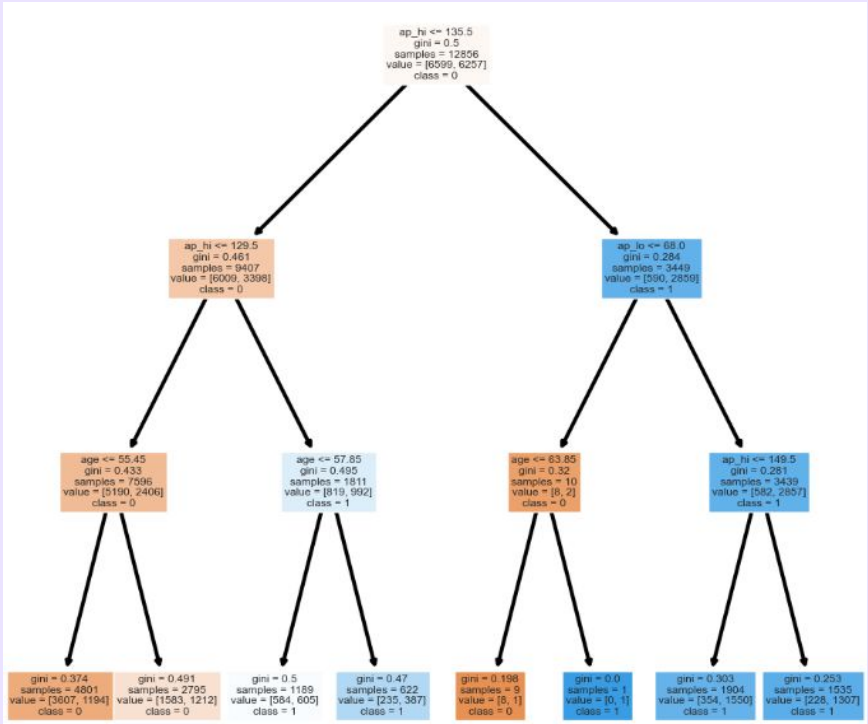
	Chi-square test	results
0	Pearson Chi-square (2.0) =	1755.1172
1	p-value =	0.0000
2	Cramer's V =	0.2334

	Chi-square test	results
0	Pearson Chi-square (2.0) =	682.0624
1	p-value =	0.0000
2	Cramer's V =	0.1995

Comparing the chi-square tests for both genders, at 5% significance level:

1. **Females: Cholesterol, Glucose, Active**
2. **Males: Cholesterol, Active, Smoke, Alcohol**

Test Set: Comparison Between Genders



From the analysis of Female and Male multi-variate decision trees, the major factors that affect CVD in different genders are:

1. Females: Age, Ap_hi, Cholesterol
2. Males: Age, Ap_hi and Ap_lo, Cholesterol

Conclusion: We proved our assumption that gender causes a difference in cardiovascular disease prediction.

Additional Research

1. Modify or add criteria to improve classification accuracy
2. Since there are more female than male data, future research should include more balanced data between the 2 genders for more accuracy

Than k You

Contributions

Karen Natalie (U2220586J) - Code: Analysis on Test Set, Conclusion, Troubleshooting & Editing of other sections; Slides: Section 1, 2, 5

Liu Yuheng (U2222313G) - Code: Data Preparation, Numerical and Categorical Data Visualization, Data Manipulation; Slides: Section 1, 4, 5

Nichani Namya Ashok (U2223732C) - Code: Female with and without outliers, Male with and without outliers; Slides: Section 3 & 4, and editing of all sections

Su Rui (U2221036D) - Code: Female with and without outliers, Male with and without outliers, Editing of other sections; Slides: Section 2, 3, 4