SC1015 Mini Project

Kaggle: Cardiovascular Disease Dataset

Christian Asher Widjaja (U2320188K), David Cheong (U2322508J), Elbert Gunawan (U2323822J)

Cardiovascular Health...

Did You Know?

According to the World Health Organization, cardiovascular diseases (CVDs) are the number one cause of death globally, claiming approximately 17.9 million lives each year.

Doing These Can Help?

No Smoking/Drinking

Exercising Regularly

Balanced Diets

Problem Definition

Are we able to predict BP based on various lifestyle variables (active, alcohol, cardio, and smoke) ?

Not likely to contract CVDs?

More likely to contract CVDs?

Preliminary Feature Selection

Lifestyle

smoke, alco, active, cardio

Health

cholesterol, gluc, bmi, bp_category

Data Cleaning: BMI

Removing Outliers

<10:
extremely underweight

>40:
super-morbid obese

The range we took

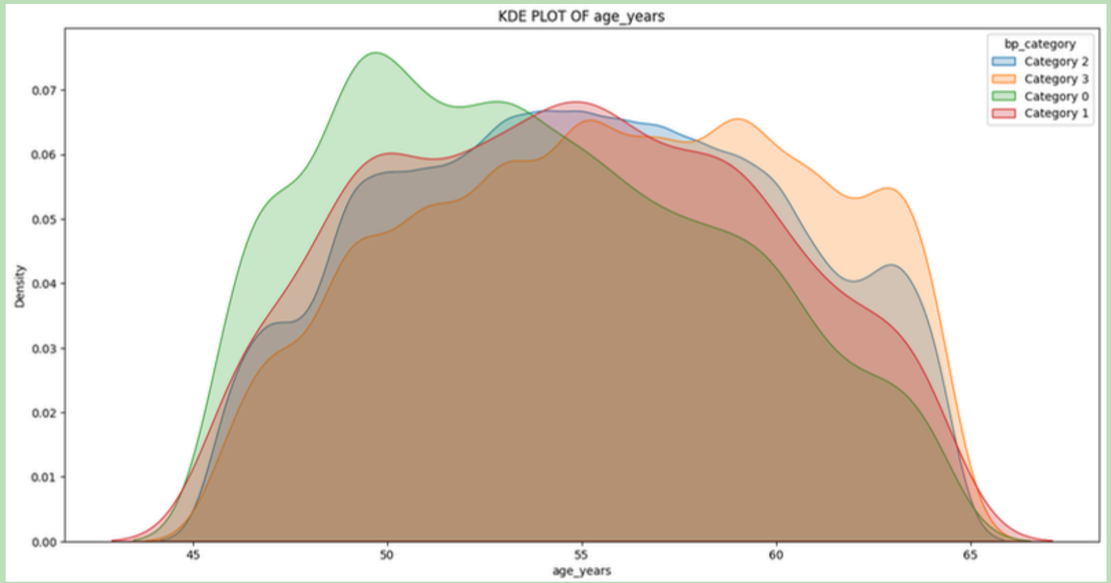
Encoding ‘bp_category’ to easily identify levels of blood pressure and simplify data preprocessing

- ‘0’ - Normal Heart Rate
- ‘1’ - Elevated Heart Rate
- ‘2’ - Stage 1 Hypertension
- ‘3’ - Stage 2 Hypertension

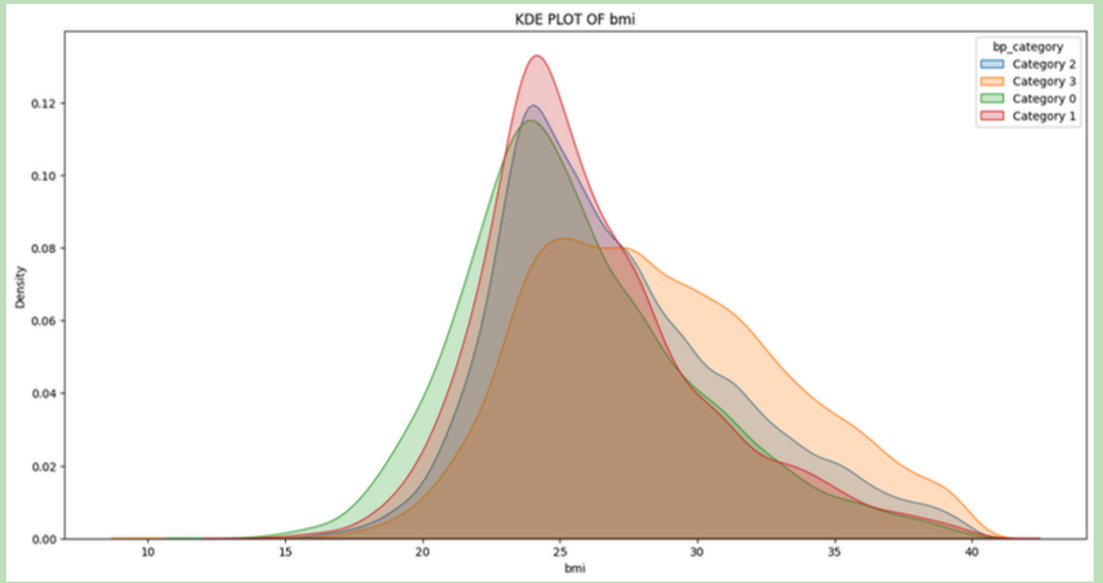
Also ensured there were no null values in our dataset

Exploratory Data Analysis

Age vs bp_category



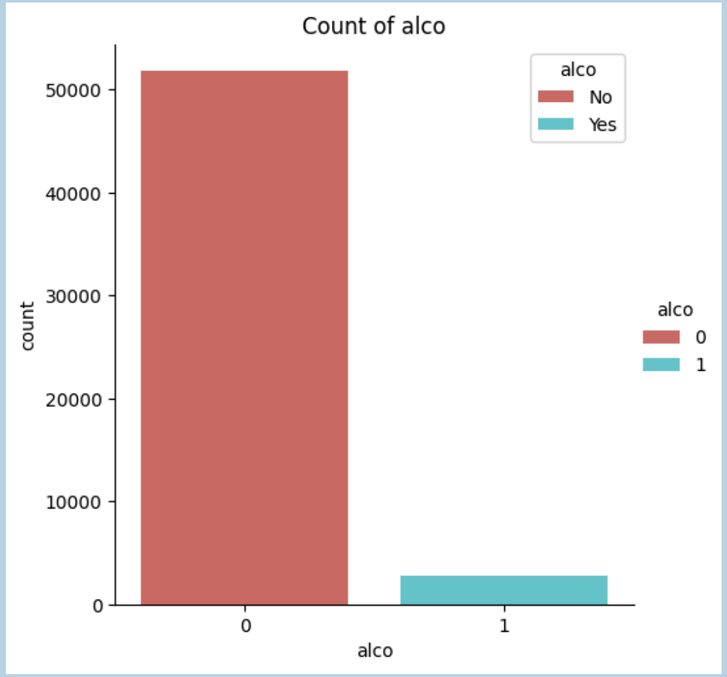
BMI vs bp_category

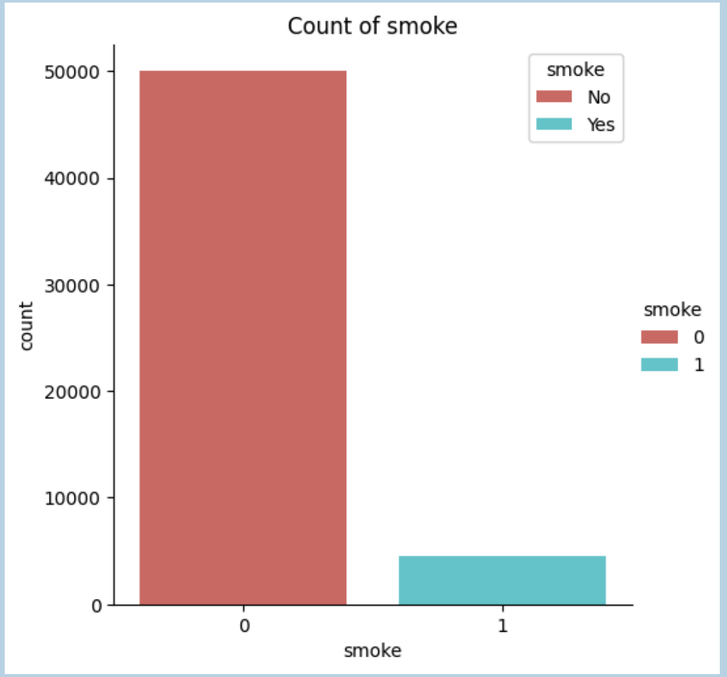


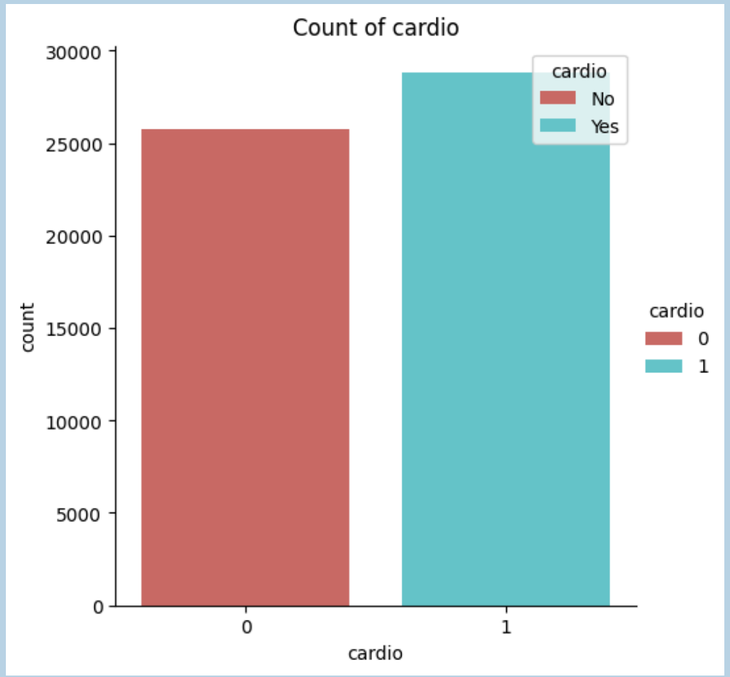
Participants with...

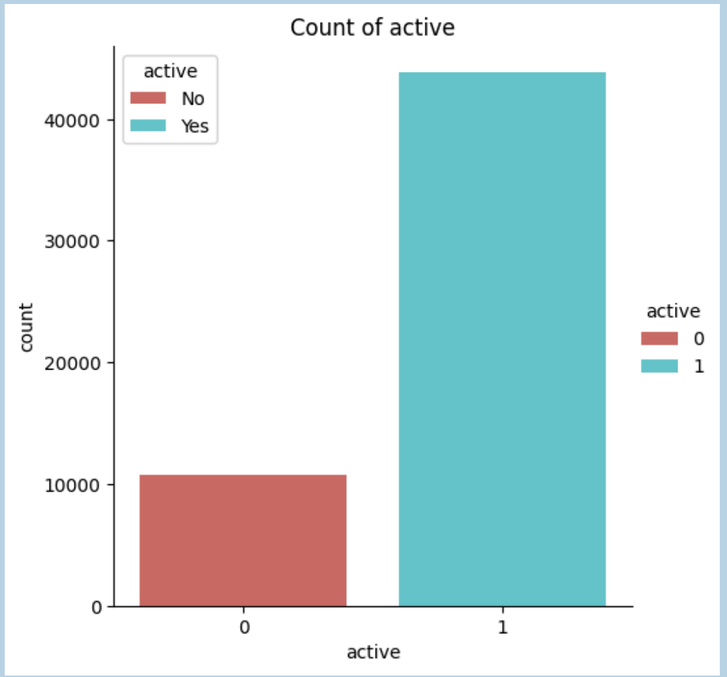
Normal blood pressure tend to be **younger** and have **lower** BMI

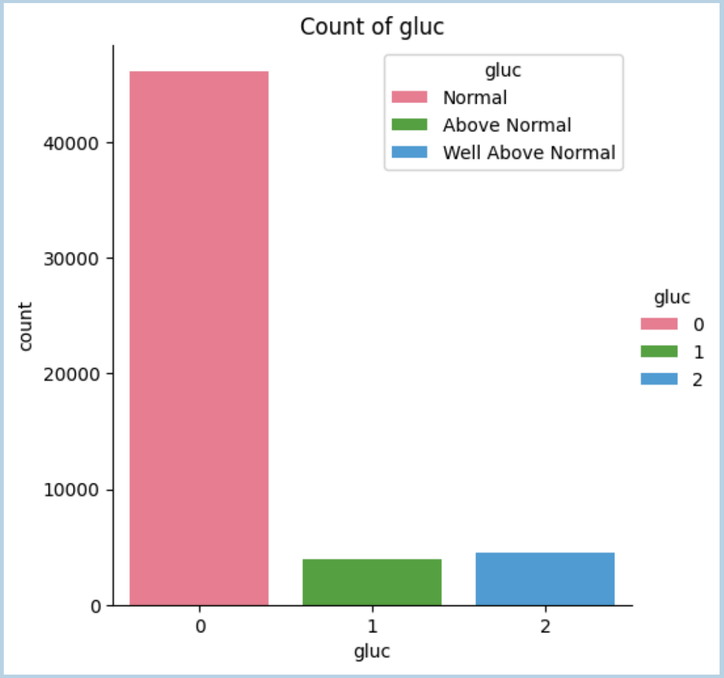
Hypertension Stage 2 tend to be **older** and have **higher** BMI.

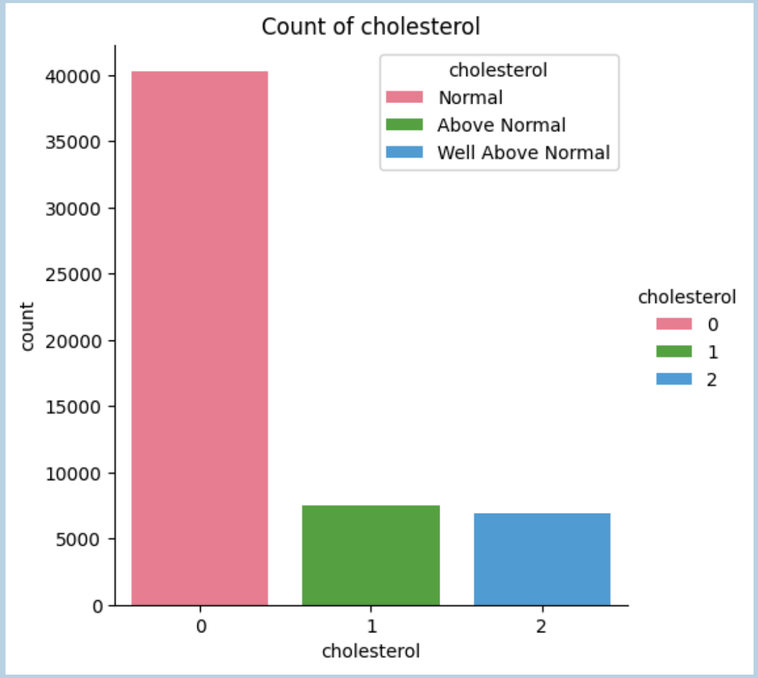

Count of alco
alco
No
Yes
50000
40000
30000
20000
10000
0
count
alco
0
1
0
1
alco

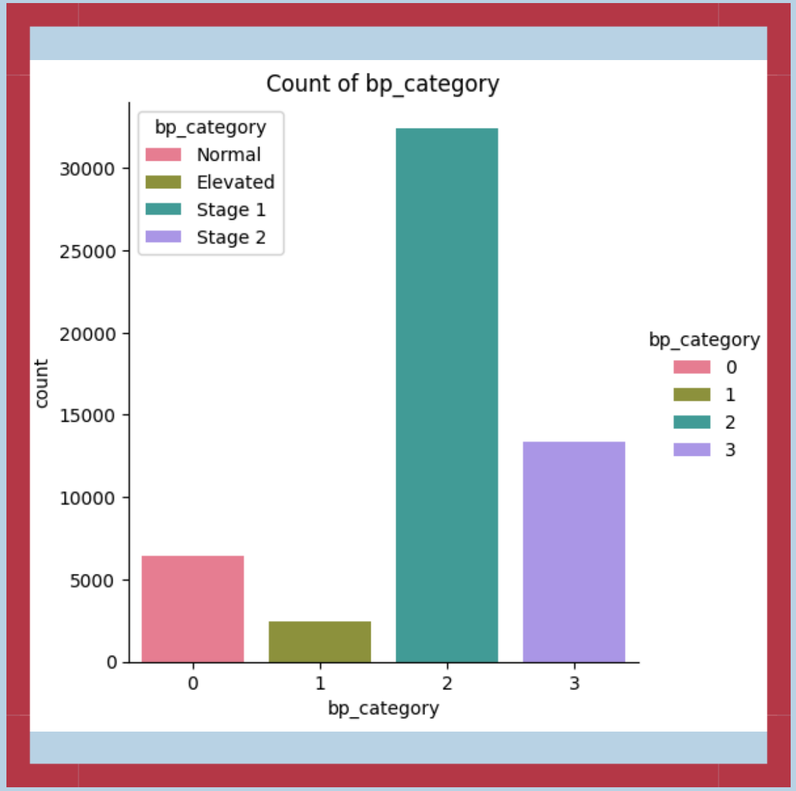

Count of smoke
smoke
No
Yes
50000
40000
30000
20000
10000
0
count
smoke
0
1
0
1
smoke

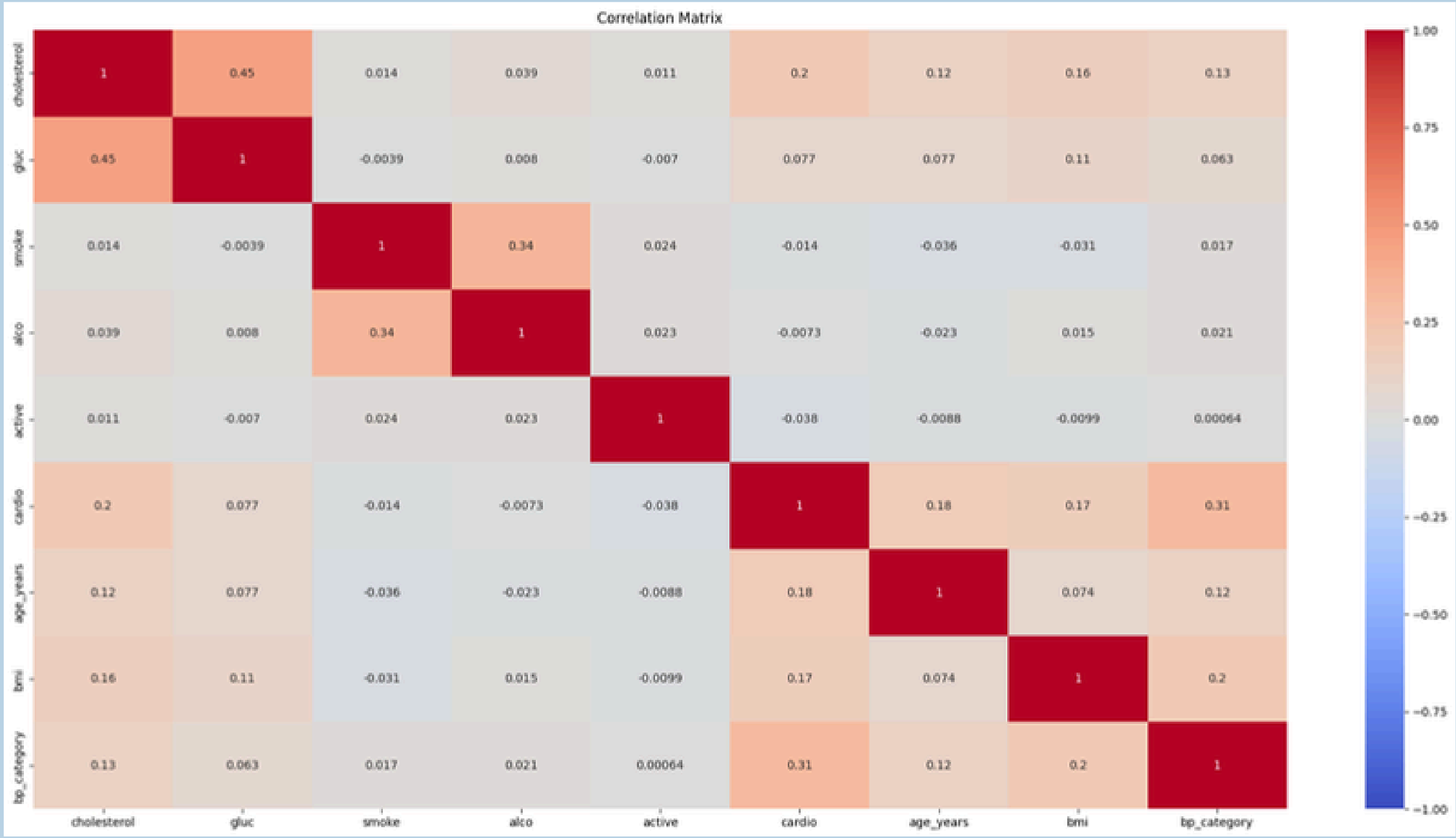

Count of cardio
cardio
No
Yes
30000
25000
20000
15000
10000
5000
0
count
cardio
0
1
0
1
cardio


Count of active
active
No
Yes
40000
30000
20000
10000
0
count
active
0
1
0
1
active


Count of gluc
gluc
Normal
Above Normal
Well Above Normal
count
40000
30000
20000
10000
0
0
1
2
gluc
gluc
0
1
2

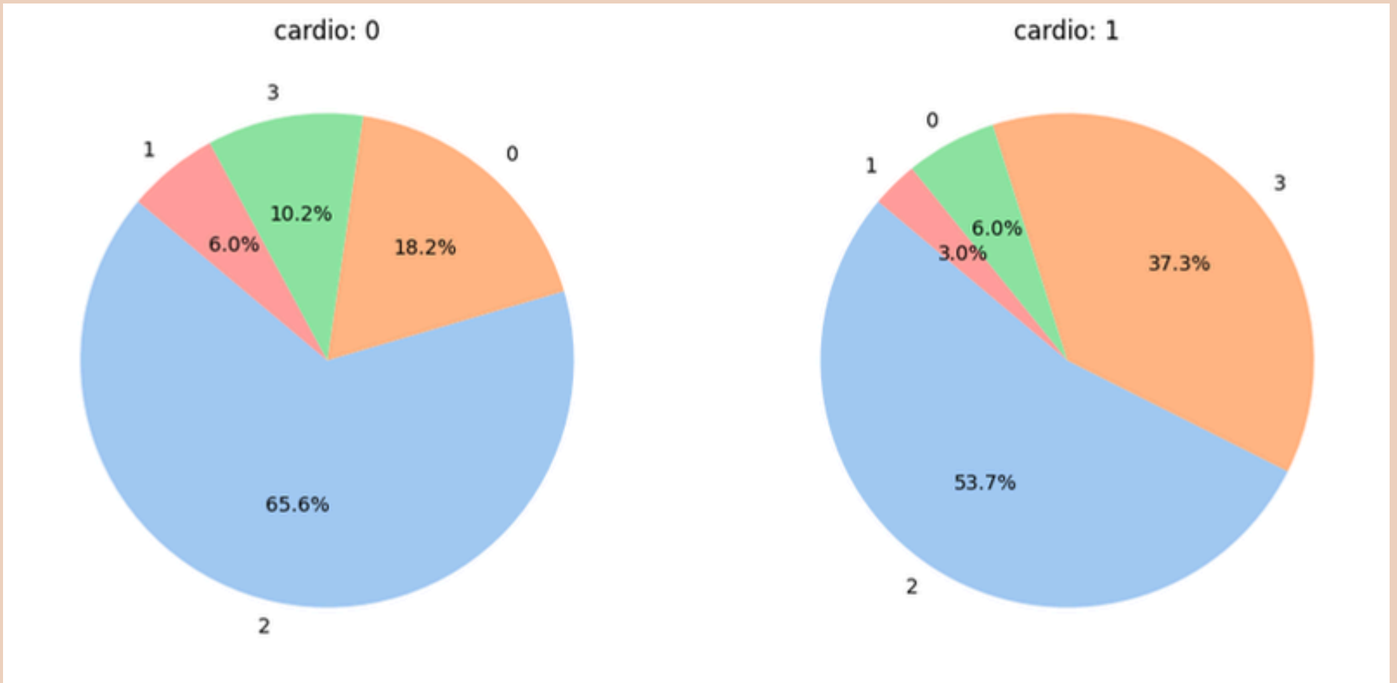

Count of cholesterol
cholesterol
Normal
Above Normal
Well Above Normal
count
40000
35000
30000
25000
20000
15000
10000
5000
0
0
1
2
cholesterol
cholesterol
0
1
2


Count of bp_category
bp_category
Normal
Elevated
Stage 1
Stage 2
count
30000
25000
20000
15000
10000
5000
0
0
1
2
3
bp_category
bp_category
0
1
2
3



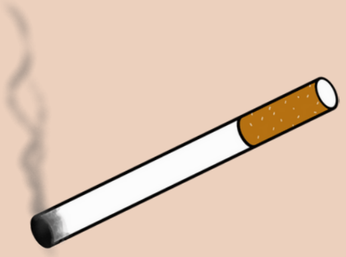
No Strong Correlations!

	cardio	bp_category	probability
0	0	0	0.181511
1	0	1	0.059947
2	0	2	0.656313
3	0	3	0.102229
4	1	0	0.059525
5	1	1	0.030422
6	1	2	0.536978
7	1	3	0.373075



Lifestyle Variables

Positive	Negative
Active	Alcohol, Cardio, Smoke
Weaker Correlation	Stronger Correlation
Weaker and more long-term impact	Stronger and more immediate impact

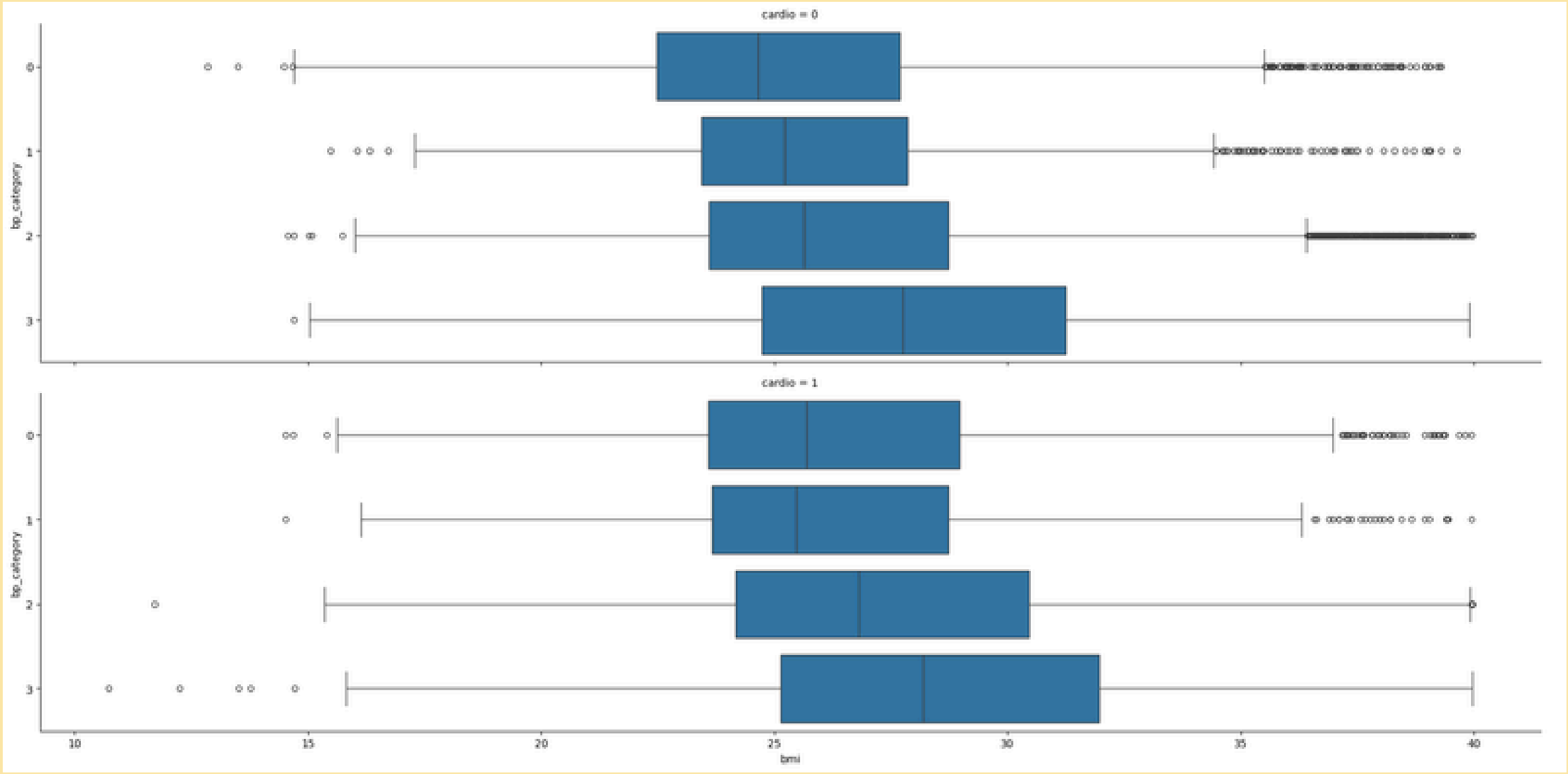


Health Variables

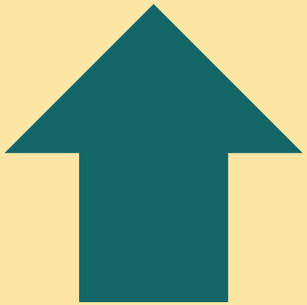
BMI	Cholesterol	Glucose
0.2	0.13	0.063

What is the relationship between BMI and blood pressure?

BMI and bp_catergory



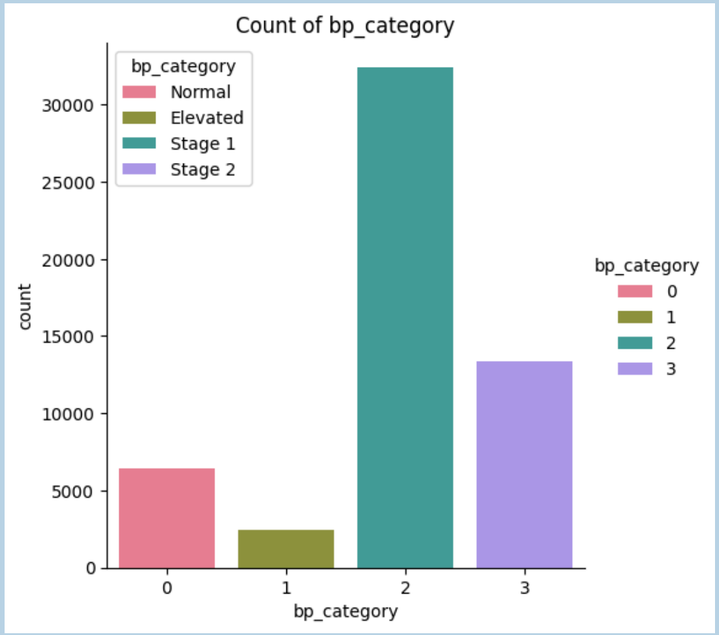
BMI

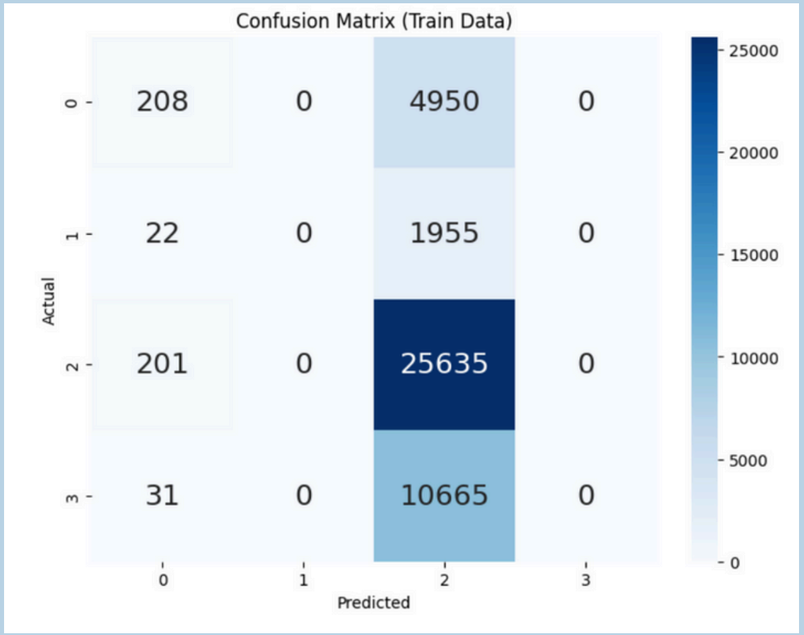


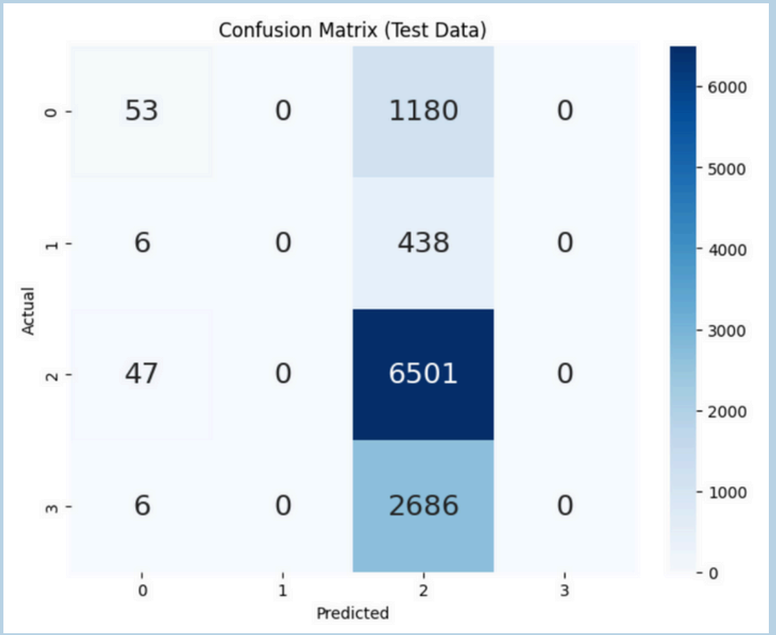
Blood Pressure

Most evident for Hypertension Stage 2

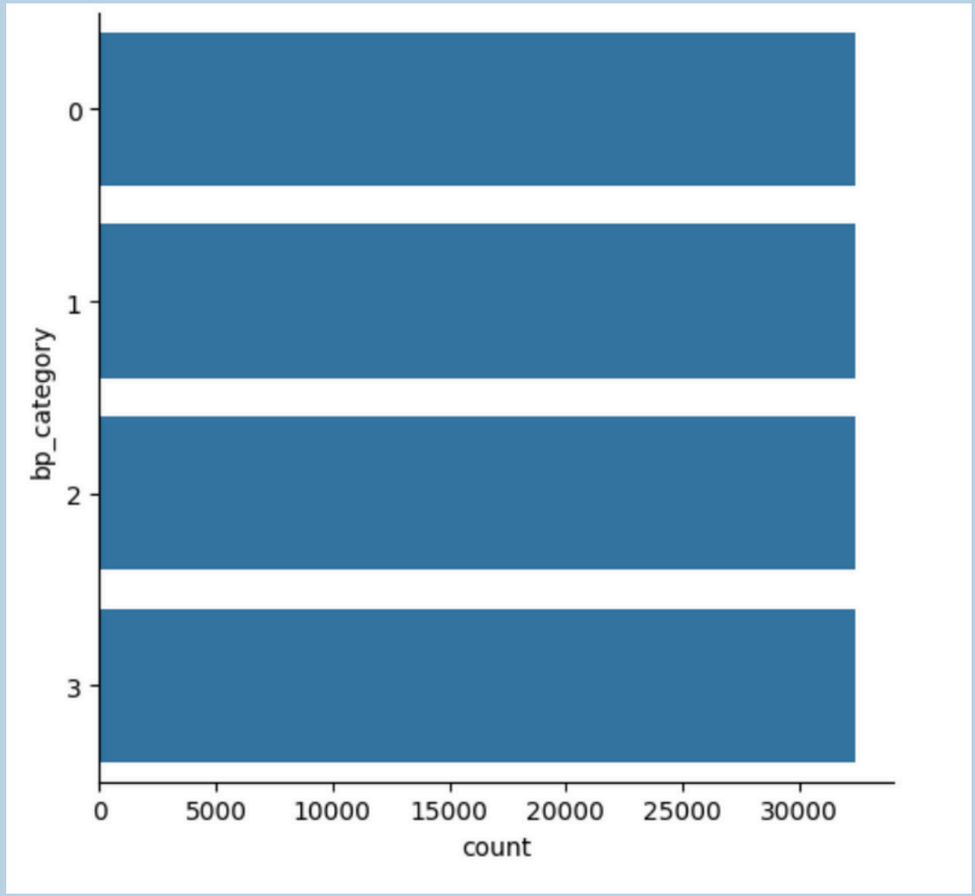
Trend continues from earlier analysis



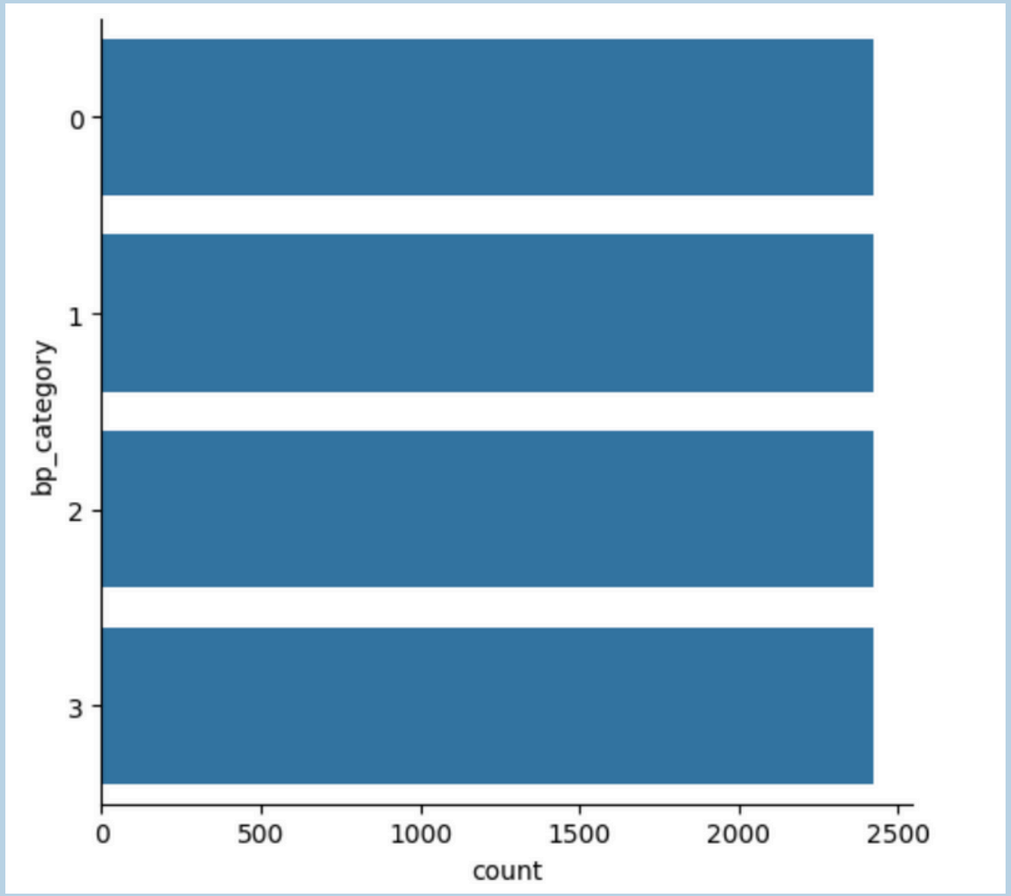




Highly Biased toward Class 2!



Upsampled All Classes to Equal Value of Class 2!

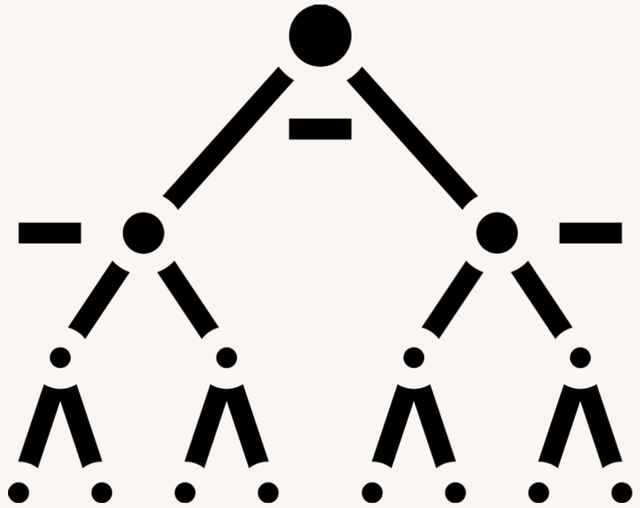


Downsampling All Classes to Equal Value of Class 1!

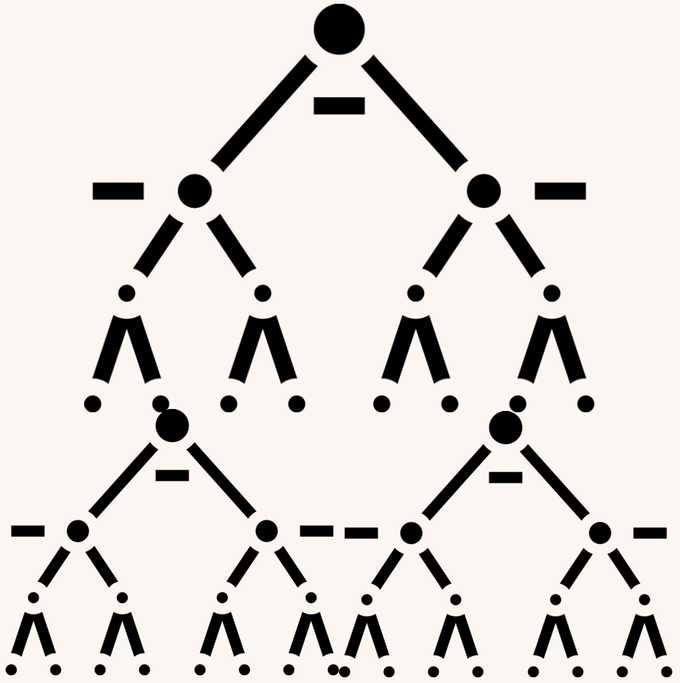
Machine Learning

Classification Models using Sckit-Learn (Upsampled)

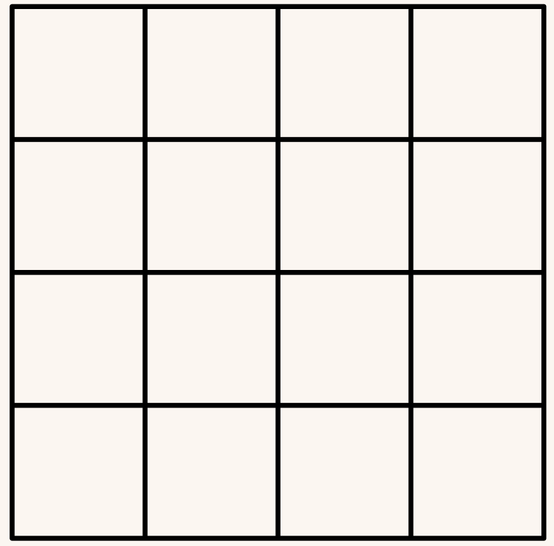
What did we use?



Decision Tree



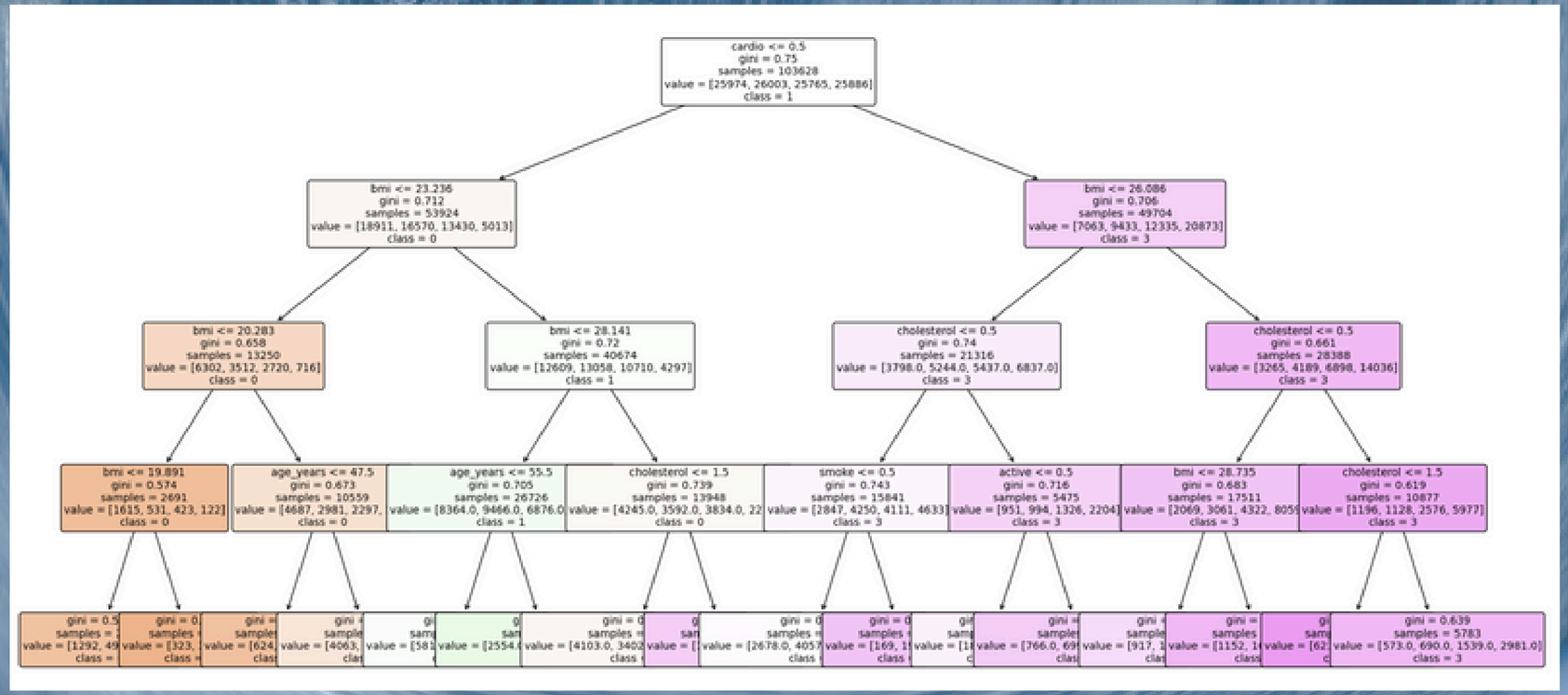
Random Forest



GridSearch

Decision Tree (Upsampled)

- Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression
- Decision Tree Model using a Depth of 4 for classifying response variable
- No Class 2 values seen in the tree, Why?

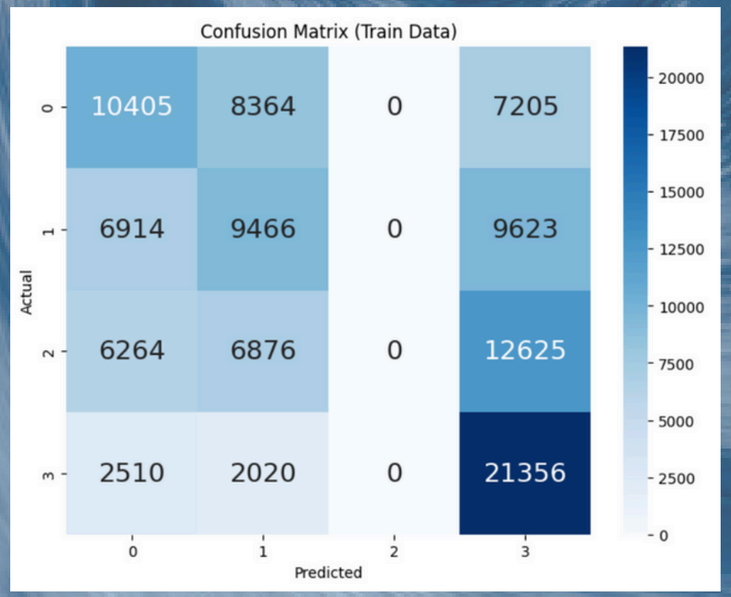


Confusion Matrix (Upsampled)

Train Set

Accuracy: 39.78%

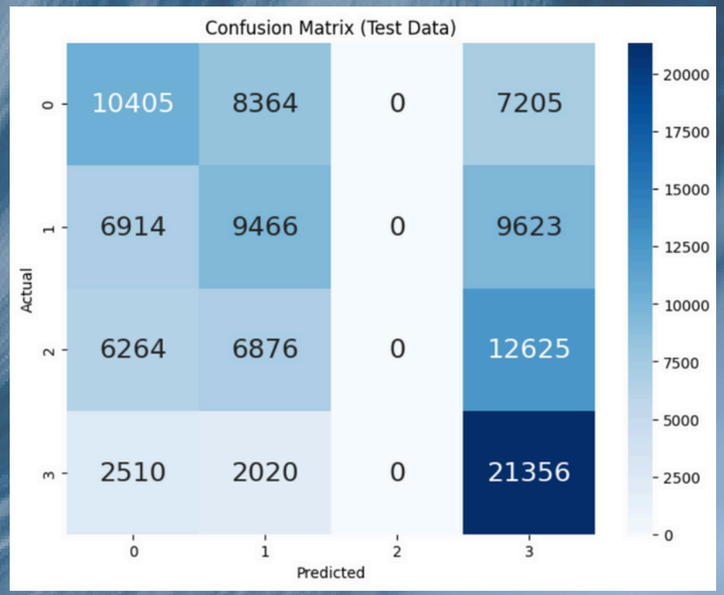
- Class 0 :
  - TPR : 40.06%
  - FPR : 20.20%
- Class 1:
  - TPR : 36.4%
  - FPR : 22.49%
- Class 2:
  - TPR : 0
  - FPR : 0
- Class 3:
  - TPR : 82.5%
  - FPR : 37.89%



Test Set

Accuracy: 39.15%

- Class 0 :
  - TPR : 38.88%
  - FPR : 20.20%
- Class 1:
  - TPR : 36.04%
  - FPR : 22.86%
- Class 2:
  - TPR : 0
  - FPR : 0
- Class 3:
  - TPR : 82.35%
  - FPR : 37.91%



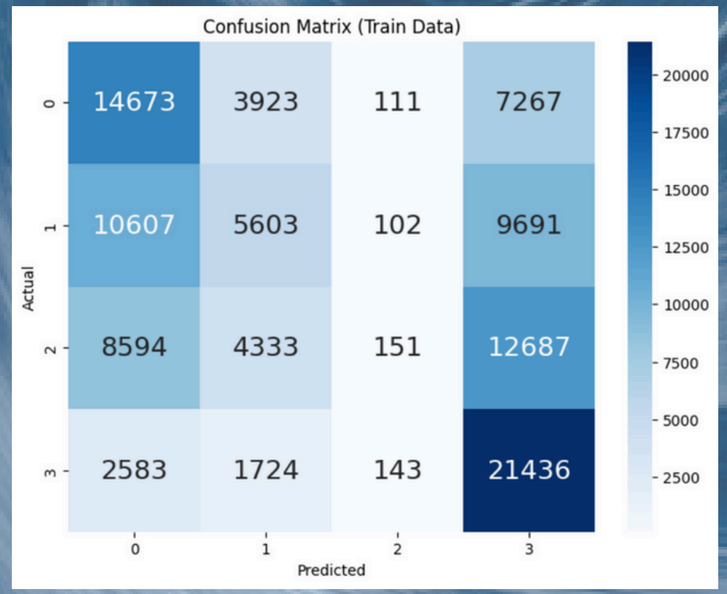
Random Forest (Upsampled)

- Random forest is a commonly used machine learning algorithm that combines the output of multiple decision trees to produce a single result.
- Improvements on Train and Test set :
  - The goodness of fit model increased slightly
  - Class 2 values finally appear, and better classification accuracy
  - Class 0 True Positive Rate increased

Train Set

Accuracy: 40.04%

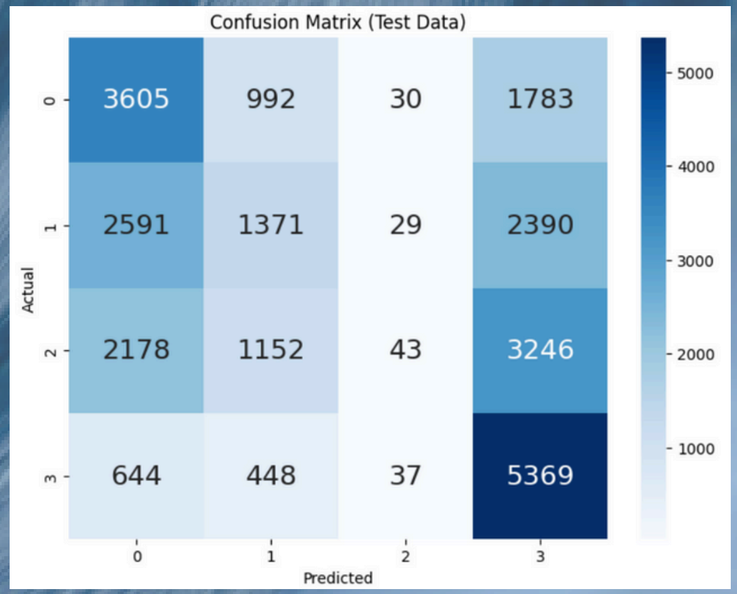
- Class 0 :
  - TPR: 56.49%
  - FPR : 28.05%
- Class 1:
  - TPR: 21.55%
  - FPR: 12.86 %
- Class 2:
  - TPR: 0.6%
  - FPR: 99.41%
- Class 3:
  - TPR: 82.81%
  - FPR: 38.13%



Test Set

Accuracy: 39.15%

- Class 0 :
  - TPR: 38.88%
  - FPR : 20.20%
- Class 1:
  - TPR: 36.04%
  - FPR: 22.86%
- Class 2:
  - TPR: 0.6%
  - FPR: 0.5%
- Class 3:
  - TPR: 82.35%
  - FPR: 37.91%



Random Forest After GridSearch

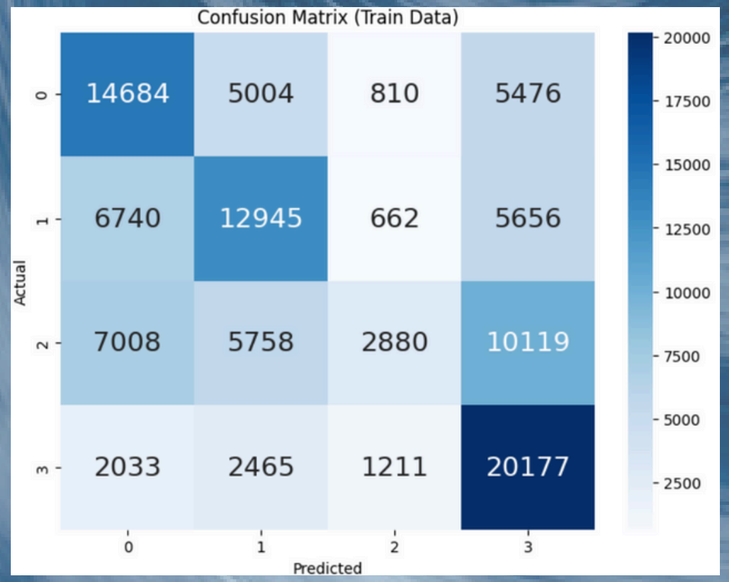
- It will focus on optimizing the performance of the random forest model, and find the best optimal parameters so it can produce maximum accuracy in the goodness of fit of the model itself.
- Improvements on Train and Test set :
  - The goodness of fit model increased significantly
  - Class 2 True Positive Rate Increased in both sets
  - Class 0 True Positive Rate increased in the Test Set
  - Class 1 True Positive Rate Increased in both sets
  - Class 3 False Positive Rate Decreased in both sets

```
Fitting 3 folds for each of 90 candidates, totalling 270 fits
Best Parameters: {'max_depth': 10, 'n_estimators': 1000}
Best Score: 0.47043270251215213
```

Train Set

Accuracy: 48.91%

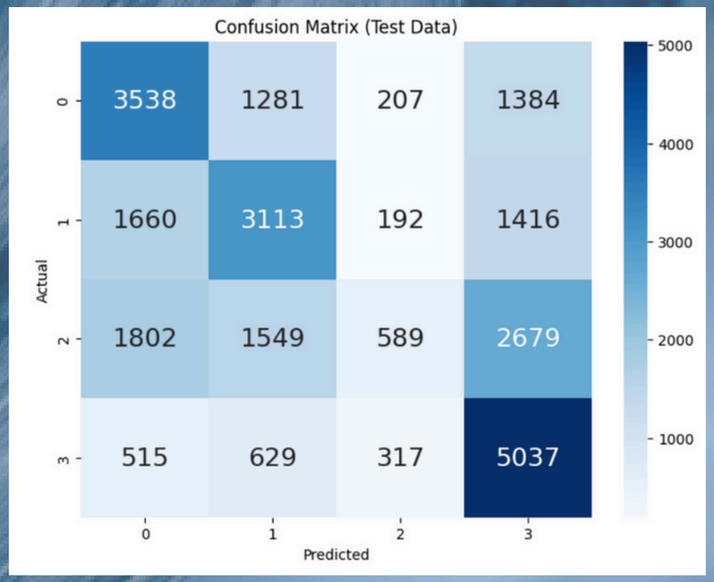
- Class 0 :
  - TPR: 56.53%
  - FPR : 20.32%
- Class 1:
  - TPR: 49.78%
  - FPR: 17.04%
- Class 2:
  - TPR: 11.18%
  - FPR: 3.45%
- Class 3:
  - TPR: 77.95%
  - FPR: 27.34%



Test Set

Accuracy: 47.39%

- Class 0 :
  - TPR: 55.2%
  - FPR : 20.40%
- Class 1:
  - TPR: 48.79%
  - FPR: 17.71%
- Class 2:
  - TPR: 8.9%
  - FPR: 3.71%
- Class 3:
  - TPR: 77.52%
  - FPR: 28.23%

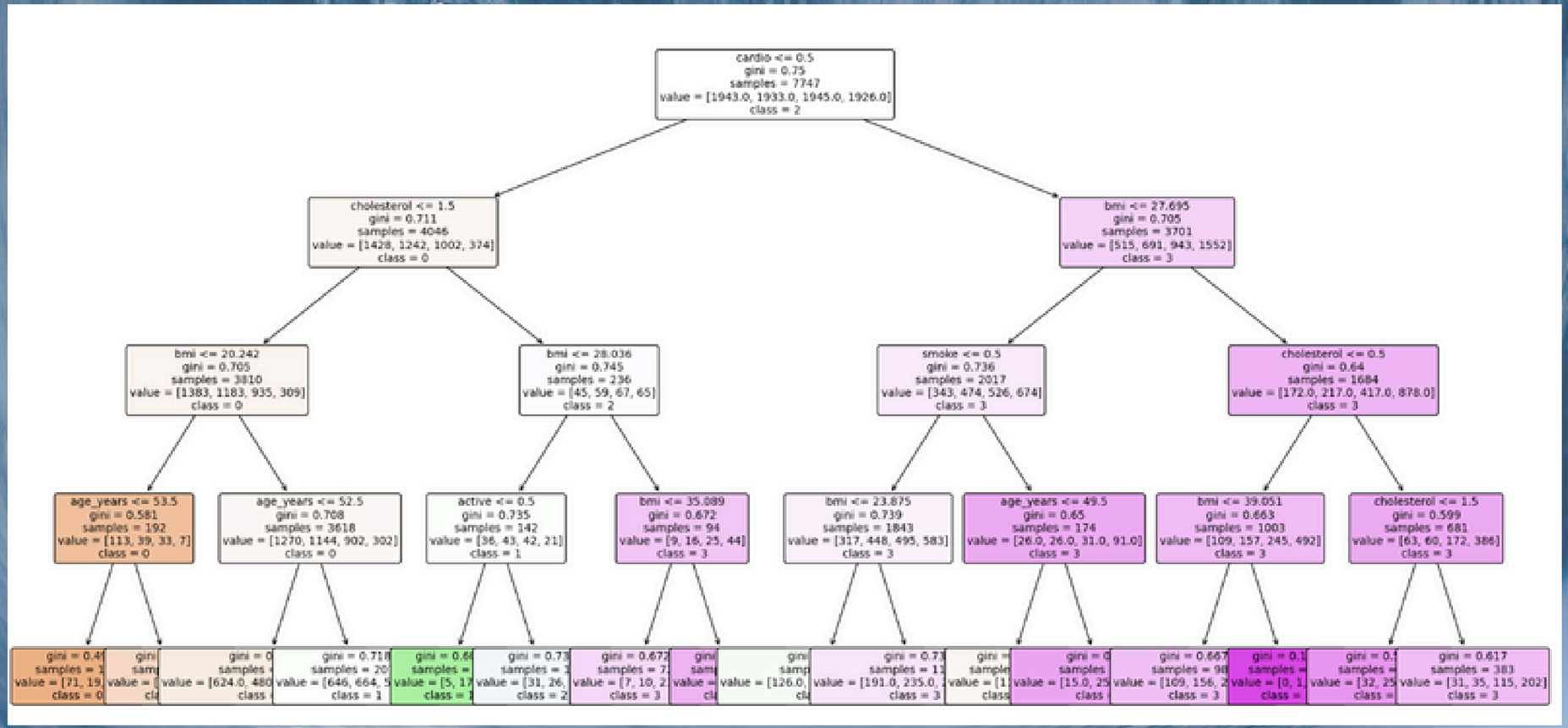


Machine Learning

Classification Models using Sckit-Learn (Downsampled)

Decision Tree (Downsampled)

- Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression
- Decision Tree Model using a Depth of 4 for classifying response variable
- Class 1 values not really seen in the tree, Why?

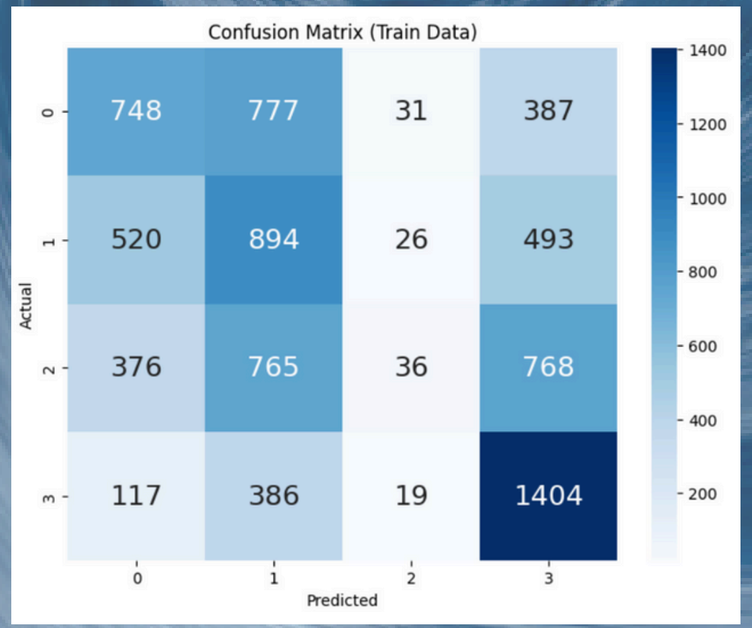


Confusion Matrix (Downsampled)

Train Set

Accuracy: 39.78%

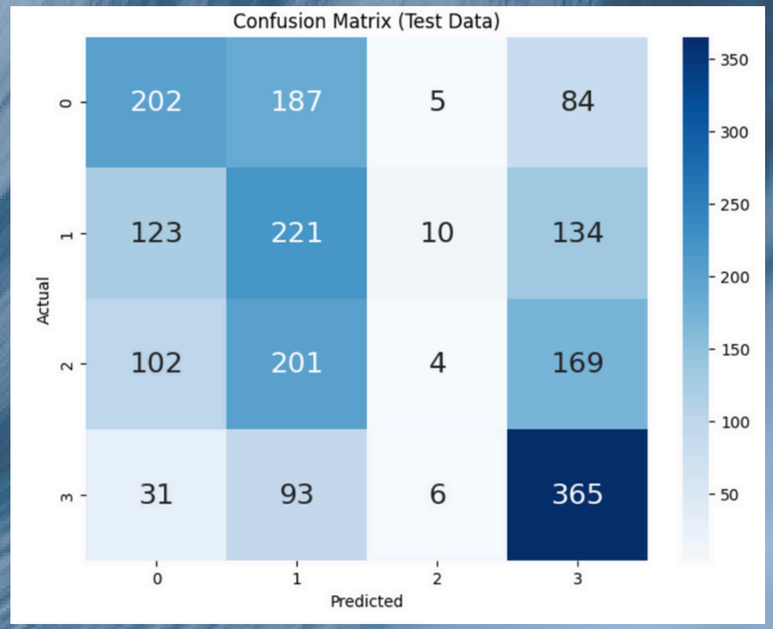
- Class 0 :
  - TPR: 38.50%
  - FPR : 17.45%%
- Class 1:
  - TPR: 46.25%
  - FPR: 33.16%
- Class 2:
  - TPR: 1.85%
  - FPR: 1.31%
- Class 3:
  - TPR: 72.90%
  - FPR: 28.31%



Test Set

Accuracy: 40.89%

- Class 0 :
  - TPR: 42.26%
  - FPR : 17.55%
- Class 1:
  - TPR: 45.29%
  - FPR: 33.20%
- Class 2:
  - TPR: 0.84%
  - FPR: 1.44%
- Class 3:
  - TPR: 73.74%
  - FPR: 26.84%



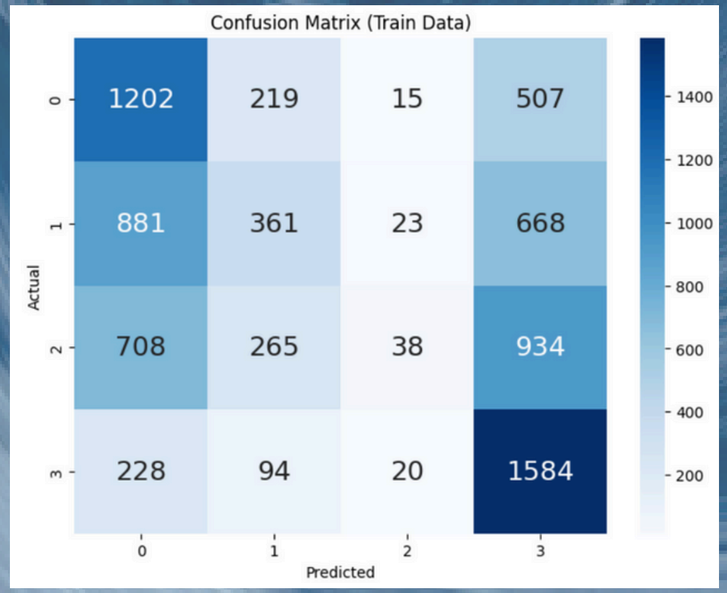
Random Forest (Downsampled)

- Random forest is a commonly used machine learning algorithm that combines the output of multiple decision trees to produce a single result.
- Improvements on Train and Test set :
  - The goodness of fit model increased slightly
  - Class 0 True Positive Rate increased significantly in both sets
  - Class 3 True Positive Rate increased significantly in both sets

Train Set

Accuracy: 41.11%

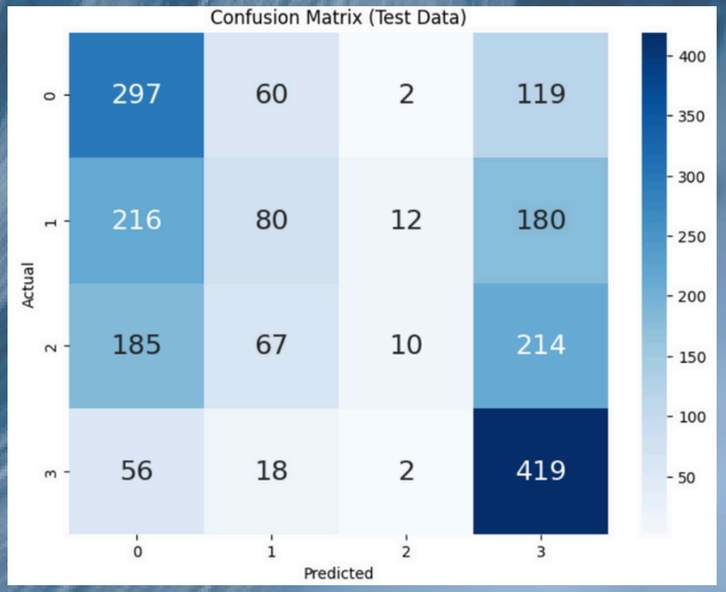
- Class 0 :
  - TPR : 61.86%
  - FPR : 31.31%
- Class 1:
  - TPR : 18.68%
  - FPR : 9.94%
- Class 2:
  - TPR : 1.95%
  - FPR : 1%
- Class 3:
  - TPR : 82.24%
  - FPR : 36.23%



Test Set

Accuracy: 41.61%

- Class 0 :
  - TPR : 62.13%
  - FPR : 31.32%
- Class 1:
  - TPR : 16.39%
  - FPR : 10.01%
- Class 2:
  - TPR : 2.10%
  - FPR : 1.10%
- Class 3:
  - TPR : 84.65%
  - FPR : 35.58%



GridSearch (Downsampled)

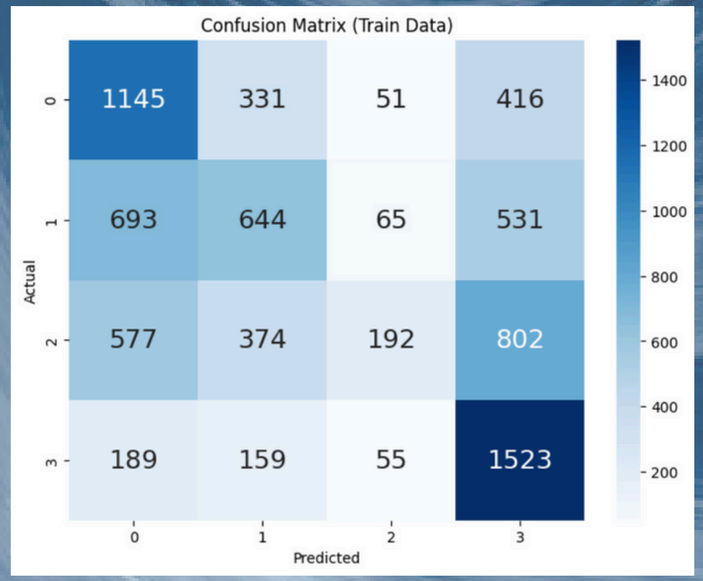
- It will focus on optimizing the performance of the random forest model, and find the best optimal parameters so it can produce maximum accuracy in the goodness of fit of the model itself.
- Improvements on Train and Test set :
  - The goodness of fit model increased significantly
  - Class 2 True Positive Rate Increased in both sets
  - Class 0 False Positive Rate Decreased in both sets
  - Class 1 True Positive Rate Increased in both sets
  - Class 3 False Positive Rate Decreased in both sets

```
Fitting 3 folds for each of 90 candidates, totalling 270 fits
Best Parameters: {'max_depth': 7, 'n_estimators': 100}
Best Score: 0.40493118774277265
```

Train Set

Accuracy: 45.23%

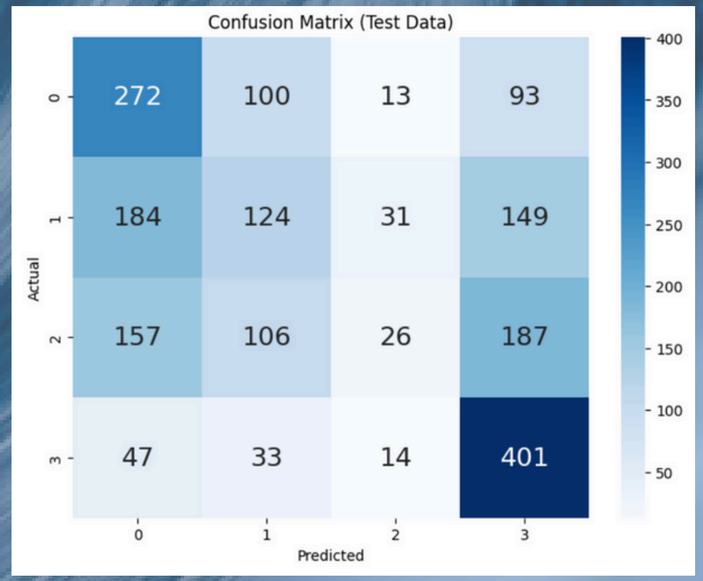
- Class 0 :
  - TPR: 58.93%
  - FPR : 25.14%
- Class 1:
  - TPR: 33.32%
  - FPR: 14.86%
- Class 2:
  - TPR: 9.87%
  - FPR: 2.95%
- Class 3:
  - TPR: 79.08%
  - FPR: 30.05%



Test Set

Accuracy: 42.49%

- Class 0 :
  - TPR: 56.90%
  - FPR : 26.59%
- Class 1:
  - TPR: 25.41%
  - FPR: 16.49%
- Class 2:
  - TPR: 5.46%
  - FPR: 3.97%
- Class 3:
  - TPR: 81.01%
  - FPR: 29.75%



Data Analysis of ML Results

What Does this Data mean?

Analysis of the Results

01.

Hypertuning and Random Forest Helped with balancing the distribution of data,

- Better Goodness of fit model accuracy

02.

The Upsampled Dataset is a slightly better alternative to the Downsampled Dataset. Why?

- Better Goodness of Fit model
- Better balance of TPR and FPR throughout all Classes.

03.

Our Model is better used for predicting Class 0 and Class 3 (Normal and Hypertension Stage 2).

- Classification Accuracy seemed to be the best
- The accuracy between TPR and FPR seemed to have the best correlation

Outcomes and Lesson learnt

Lessons Learned :

- Random Forest
- GridSearch
- Resampling

Outcomes

- Better for predicting different types of elevated heart rate
- Helps with early detection of cardiovascular health based on lifestyle choices.

Interesting Observations :

- Class 2 (Hypertension Stage 1) has the lowest values in Upsampled and Downsampled dataset
- Goodness of fit Model Accuracy can't get above 50%

Recommendations

01. Obtain a better dataset to use as trying to predict more than 4 response variables may not lead to best accuracy. Gathering more nuanced data can help a lot.

02. Try to work with medical professionals to help better understand the correlations between lifestyle choices and cardiovascular health - ensures clinical significance and more accurate predictive models

03. Employing GridSearch with increased cross-validation folds can refine model parameters. Exploring XGBoost or deep learning may also enhance results, and nested cross-validation could offer a more reliable performance assessment.

Larana, Inc.