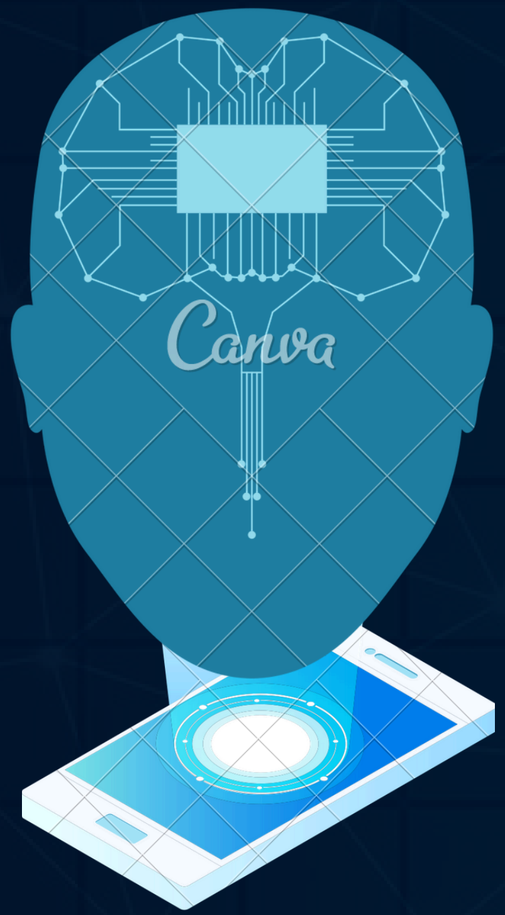


SC1015 MINI-PROJECT : CARDIOVASCULAR DISEASE

FDAA GROUP 5

NADEN JAREL

ANTHONY KOH,

KHAW BOON KIAT,

GUNASEKARAN

ABINESHKUMAR

PRACTICAL MOTIVATION

Cardiovascular Disease(CVD) is the leading cause of death globally, representing 32% of global deaths.(World Health Organization: WHO, 2021)

Managing risk factors such as smoking and physical inactivity can significantly reduce the likelihood of developing CVD.

PROBLEM DEFINITION

We aim to predict CVD risks using a dataset of health records of 70,000 patients. The goal is to develop a machine learning model that is not only accurate but also practical. A health checkup in Singapore costs about 1.5k on average, which can be expensive for some individuals, hence with this model, we will help reduce the need for unnecessary health checkups, saving costs.

SAMPLE COLLECTION

kaggle

Description of the dataset, as available on Kaggle, is as follows. Learn more: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Features:
Numerical Features:
Data | Description | DataType
age : age | int (days) |
height : height | int (cm) |
weight : weight | float (kg) |
ap_hi : Systolic blood pressure | int |
ap_lo : Diastolic blood pressure | int |

Categorical Features
gender : Gender | 1: women, 2: men |
cholesterol : Cholesterol | 1: normal, 2: above normal, 3: well above normal |
gluc : Glucose | 1: normal, 2: above normal, 3: well above normal |
smoke : Smoking | binary |
alco : Alcohol intake | binary |
active : Physical activity | binary |
cardio : Presence or absence of cardiovascular disease | binary |

There is a total of 13 Features and 70000 Entries

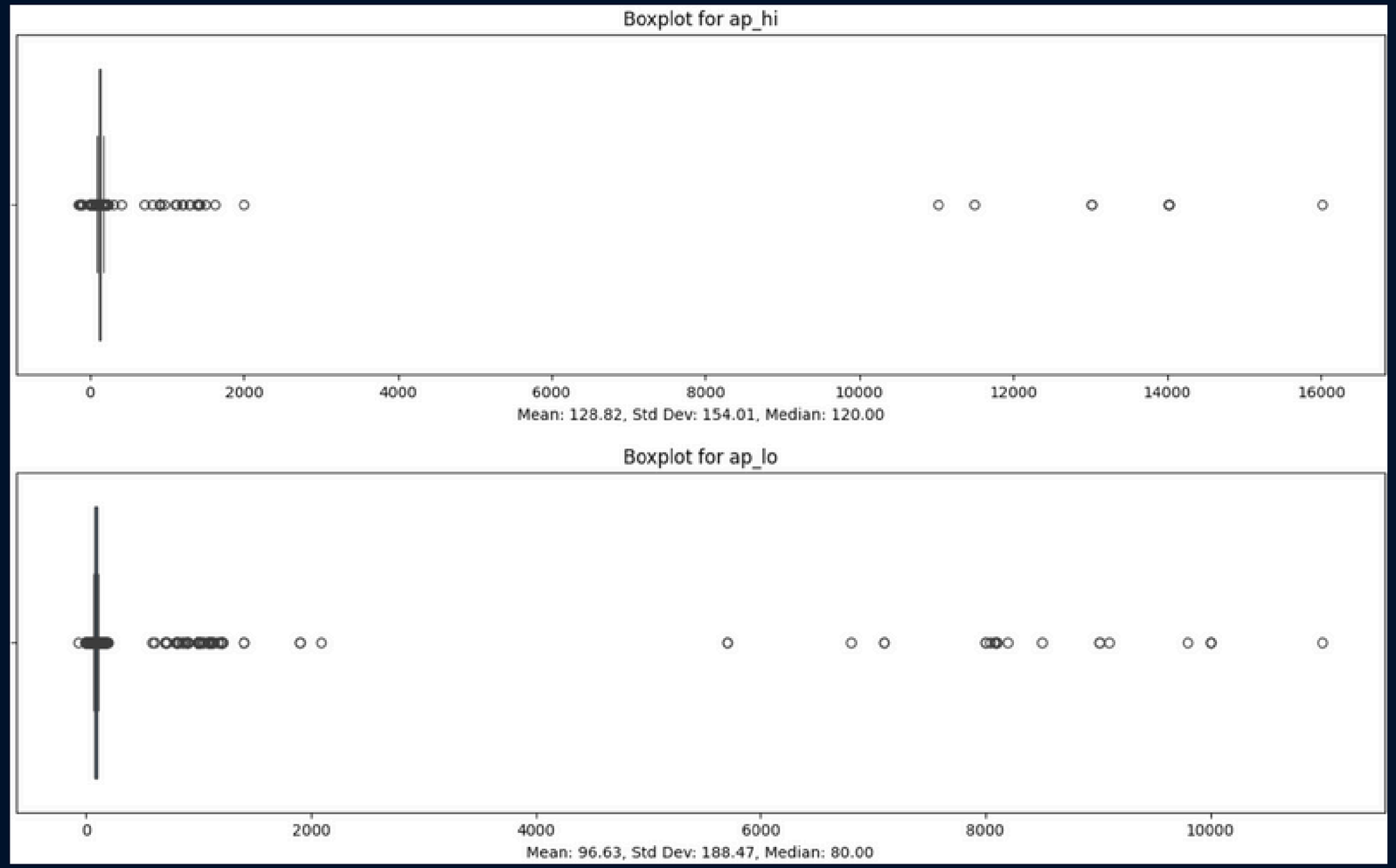
DATA PREPARATION

DATA CLEANING:

- REMOVED "ID" COLUMN
- CHECKED FOR NULL DUPLICATE VALUES
- CONVERTED AGE FROM DAYS TO YEARS
- REMOVED OUTLIERS

DATA PREPARATION

BEFORE REMOVING OUTLIERS



DATA PREPARATION

REMOVING OUTLIERS

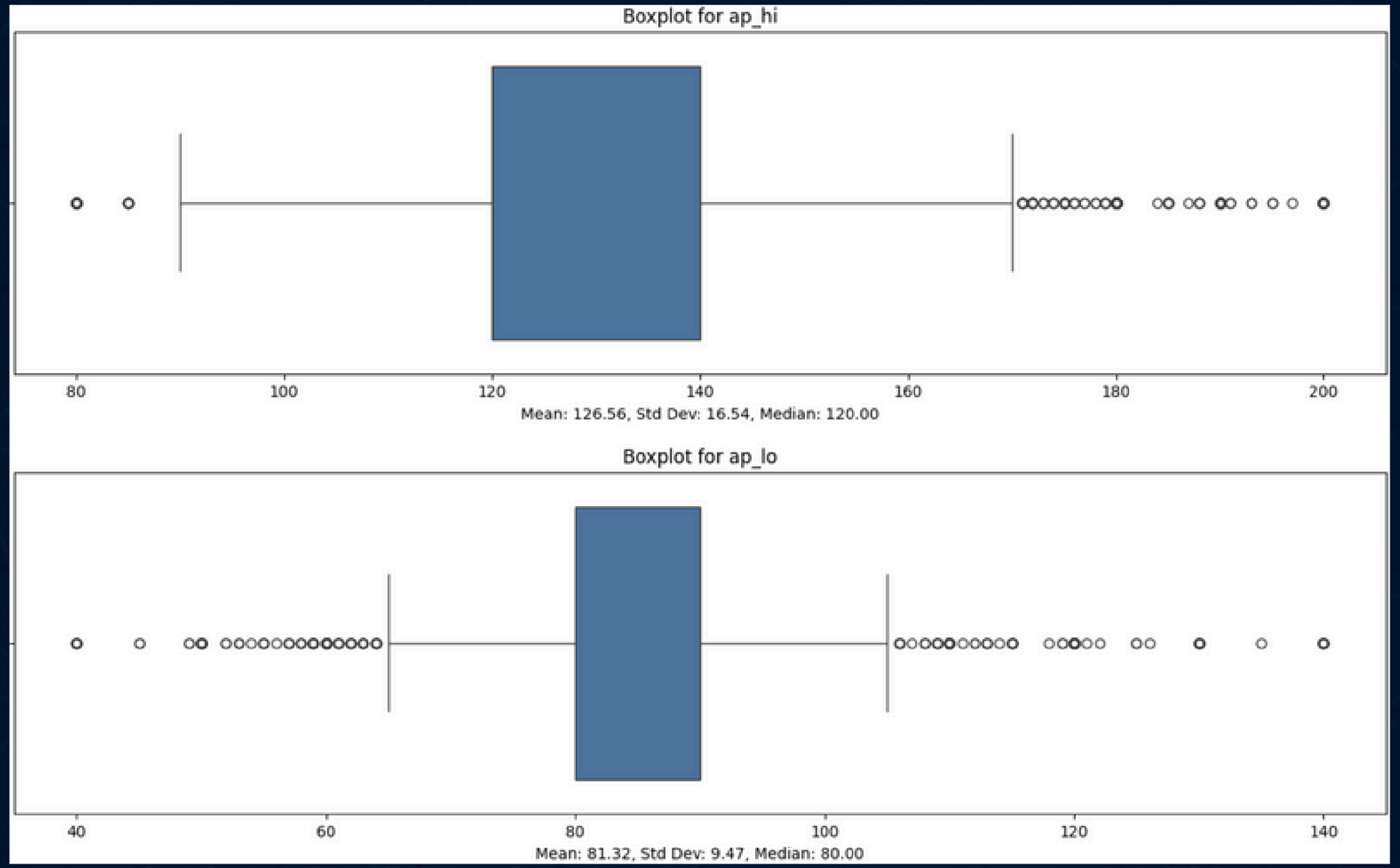
```
ap_hi = data['ap_hi']
ap_lo = data['ap_lo']

# Print number of unusual training examples
print("number of outliers:")
data["cardio"].loc[(ap_hi < 80) | (ap_hi > 200) | (ap_lo < 40) | (ap_lo > 140)].count()
```

```
number of outliers:
1320
```

Since there are only 1320 outliers, we will go ahead and remove them

DATA PREPARATION

AFTER REMOVING OUTLIERS



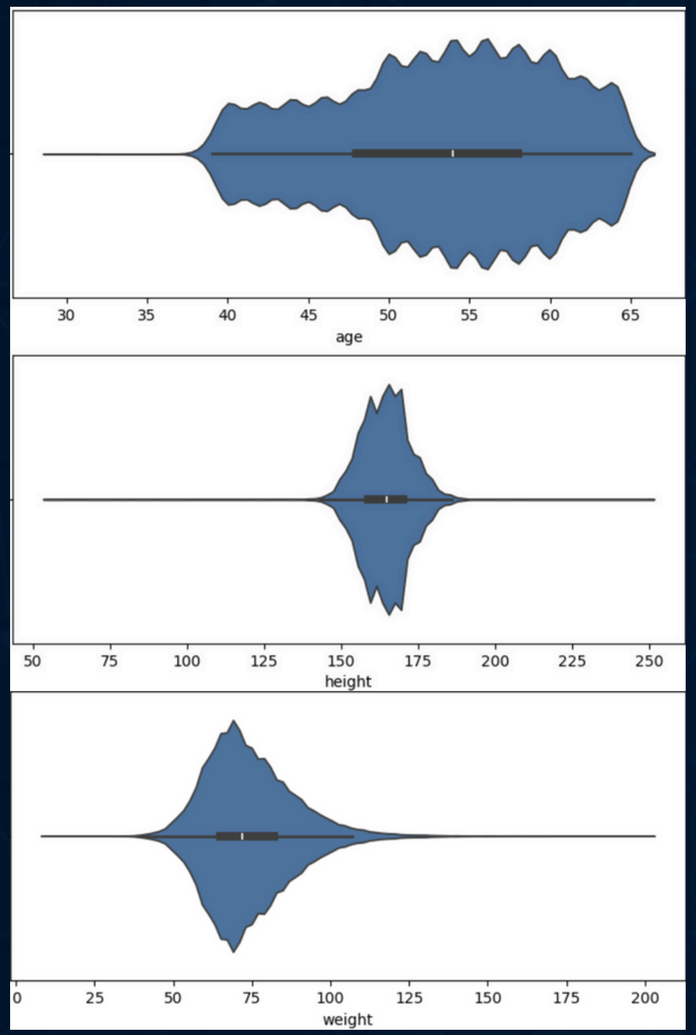
EXPLORATORY ANALYSIS

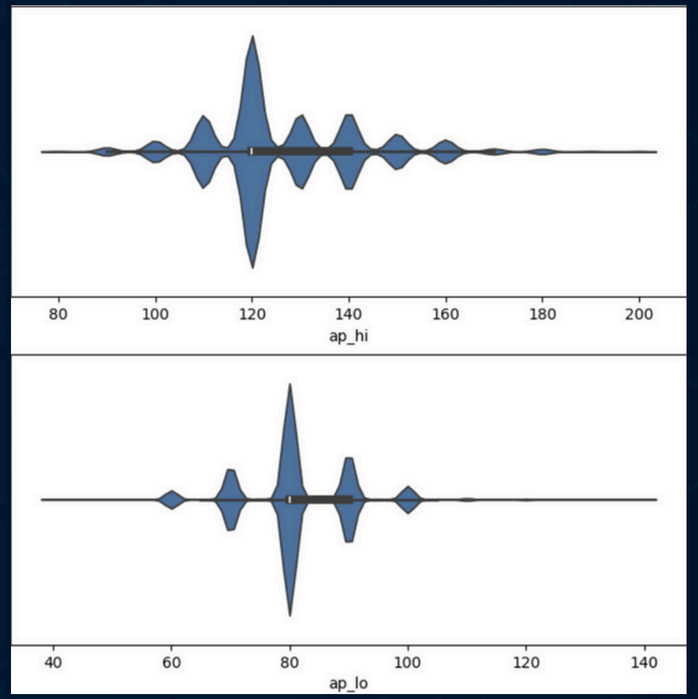
UNIVARIATE ANALYSIS

NUMERICAL FEATURES	CATEGORICAL FEATURES
• VISUALIZING NUMERICAL VARIABLES WITH VIOLIN PLOTS • STATISTICAL DISPERSION AND VARIATION	• PLOTTING COUNTPLOT TO VISUALIZE DATA DISTRIBUTION

EXPLORATORY ANALYSIS

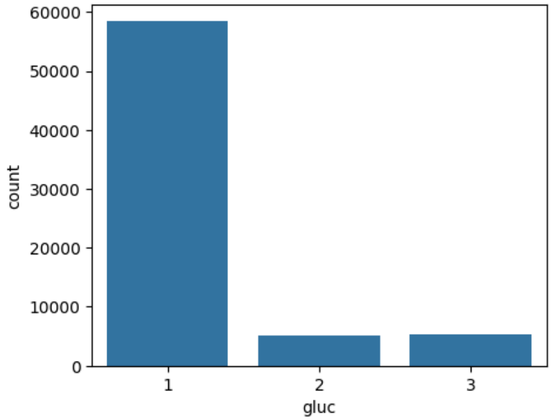
NUMERICAL FEATURES

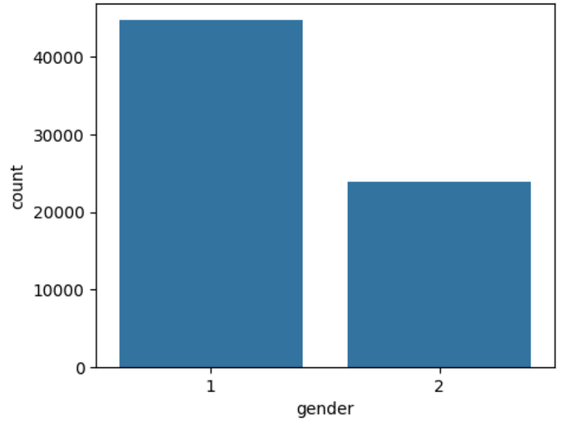


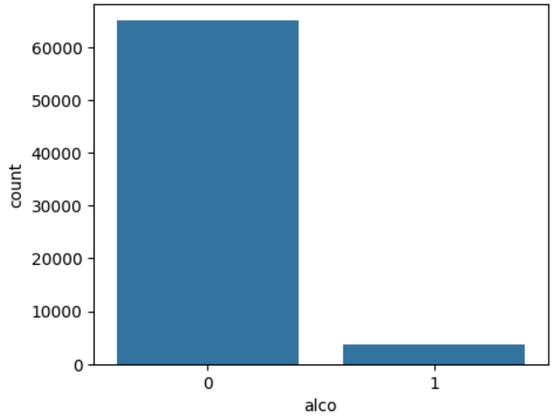


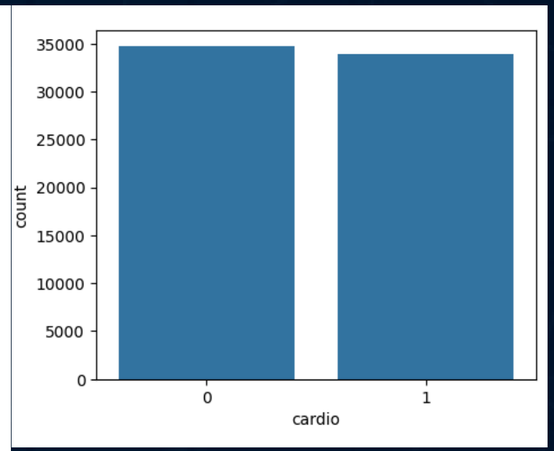
EXPLORATORY ANALYSIS

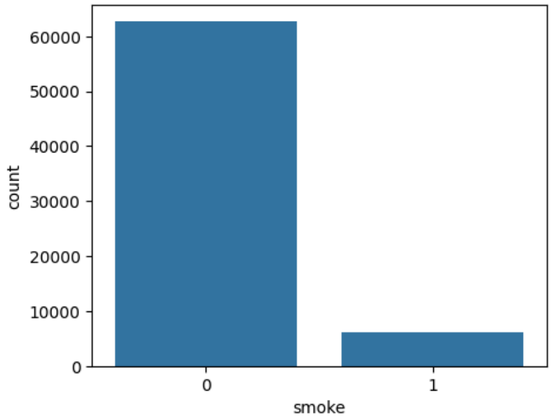
CATEGORICAL FEATURES

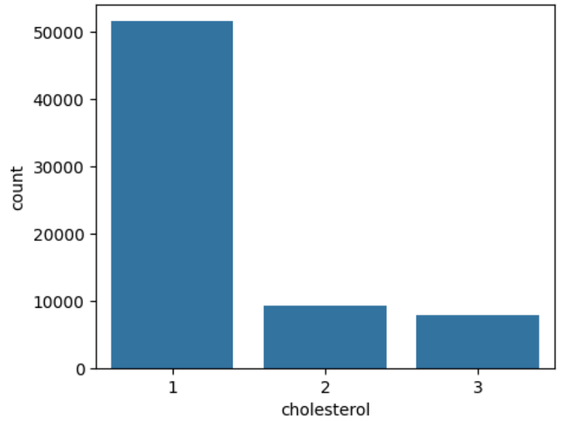


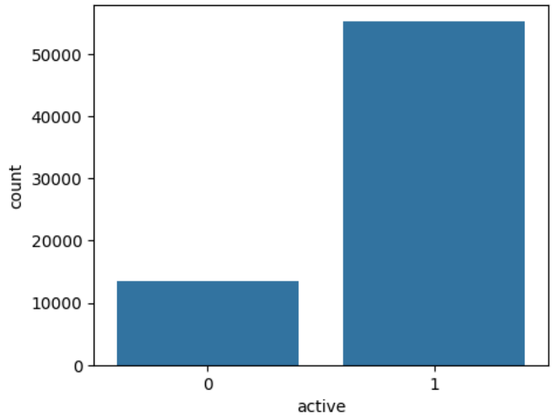






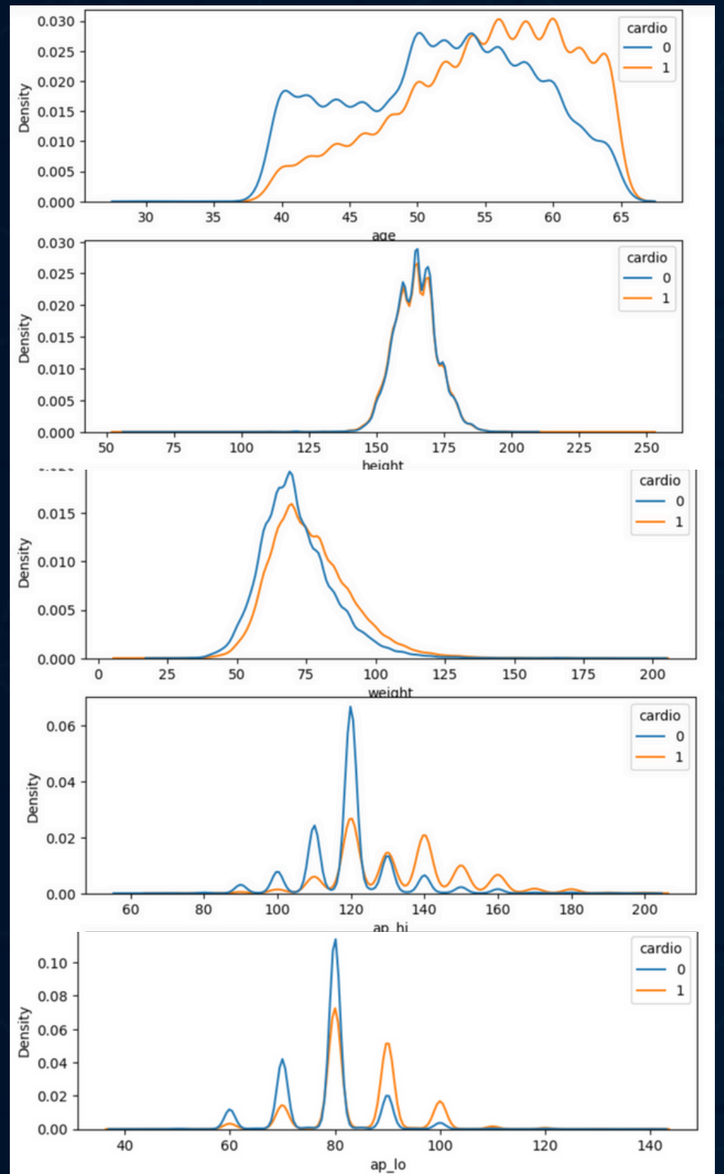


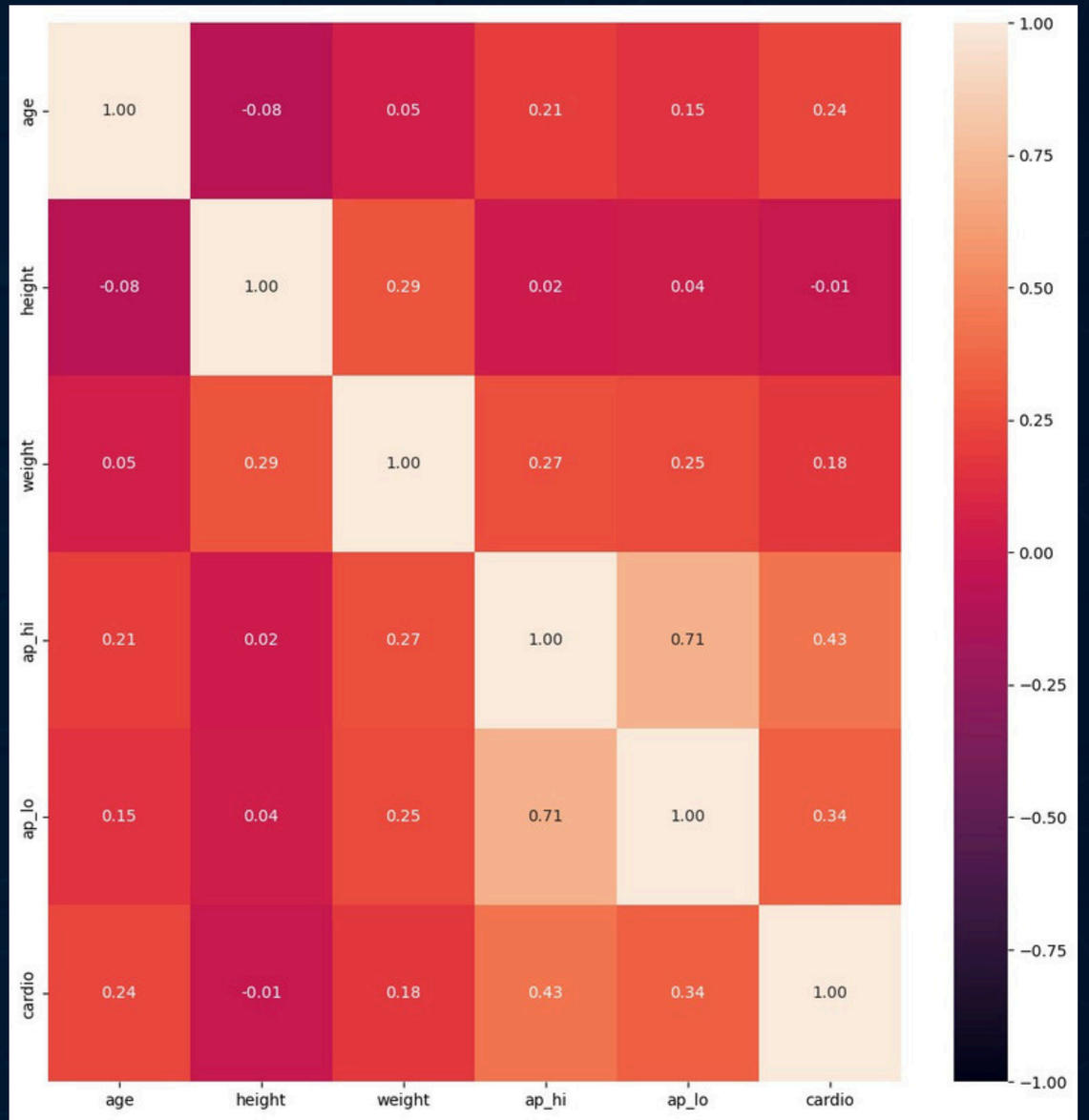




EXPLORATORY ANALYSIS

BI-VARIATE ANALYSIS
NUMERICAL FEATURES

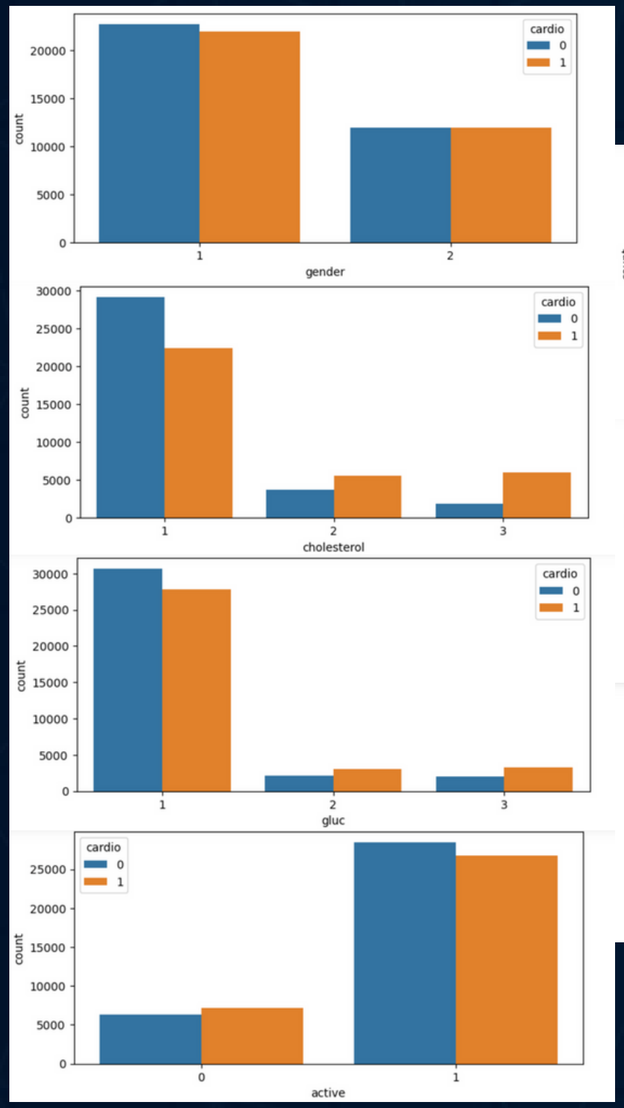


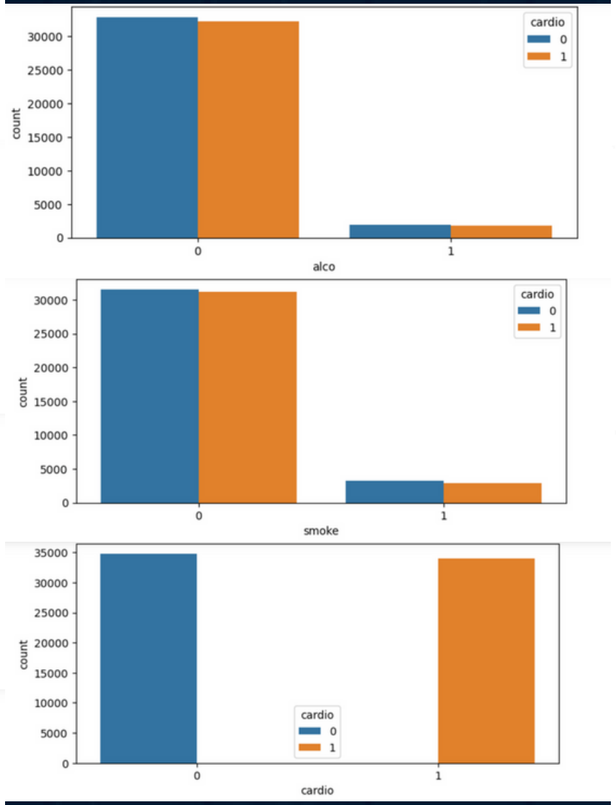


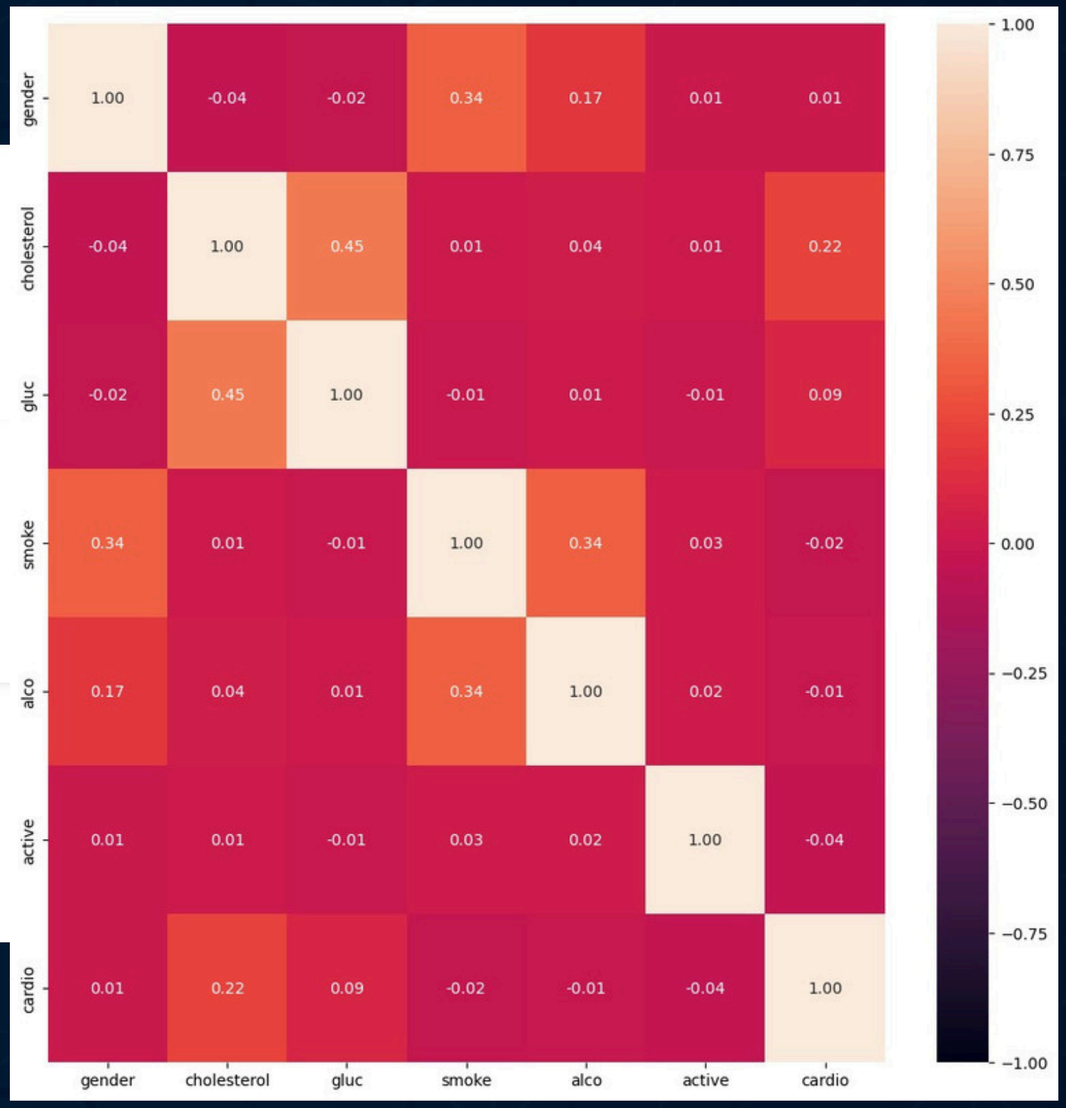
EXPLORATORY ANALYSIS

BI-VARIATE ANALYSIS

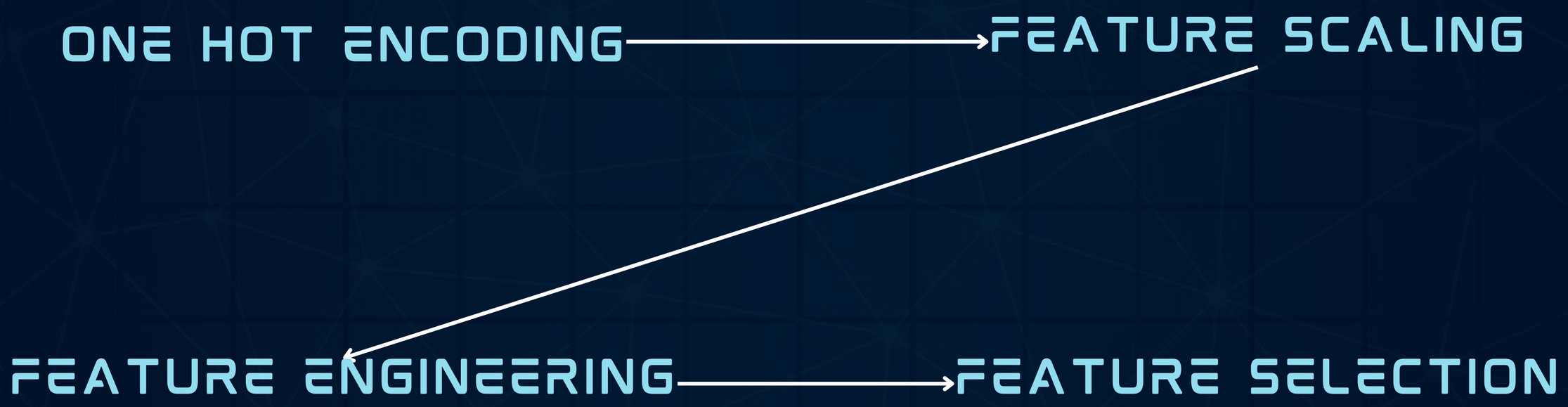
CATEGORICAL ANALYSIS



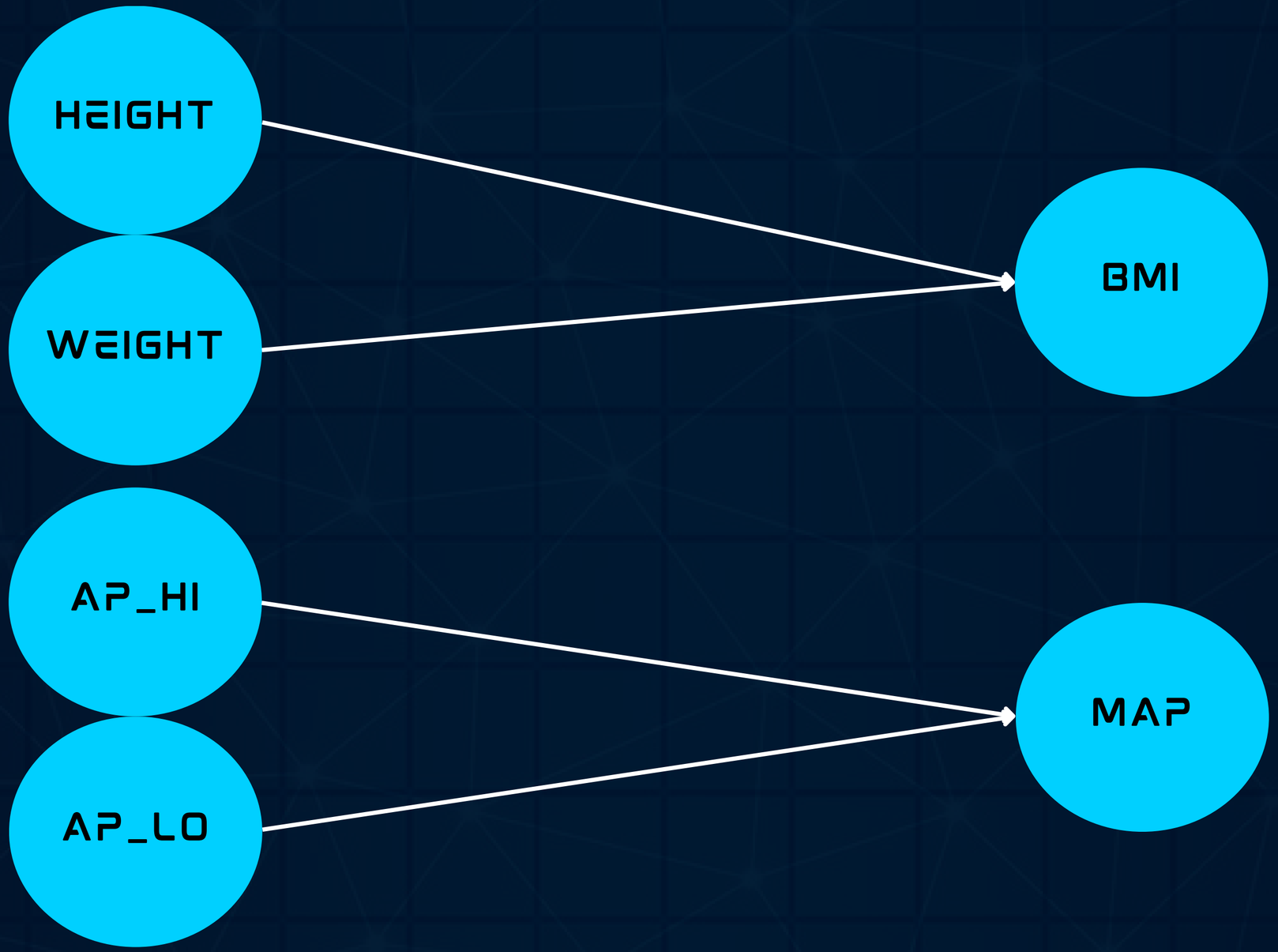




PREPARATORY WORK BEFORE MACHINE LEARNING



FEATURE ENGINEERING



METRIC USED

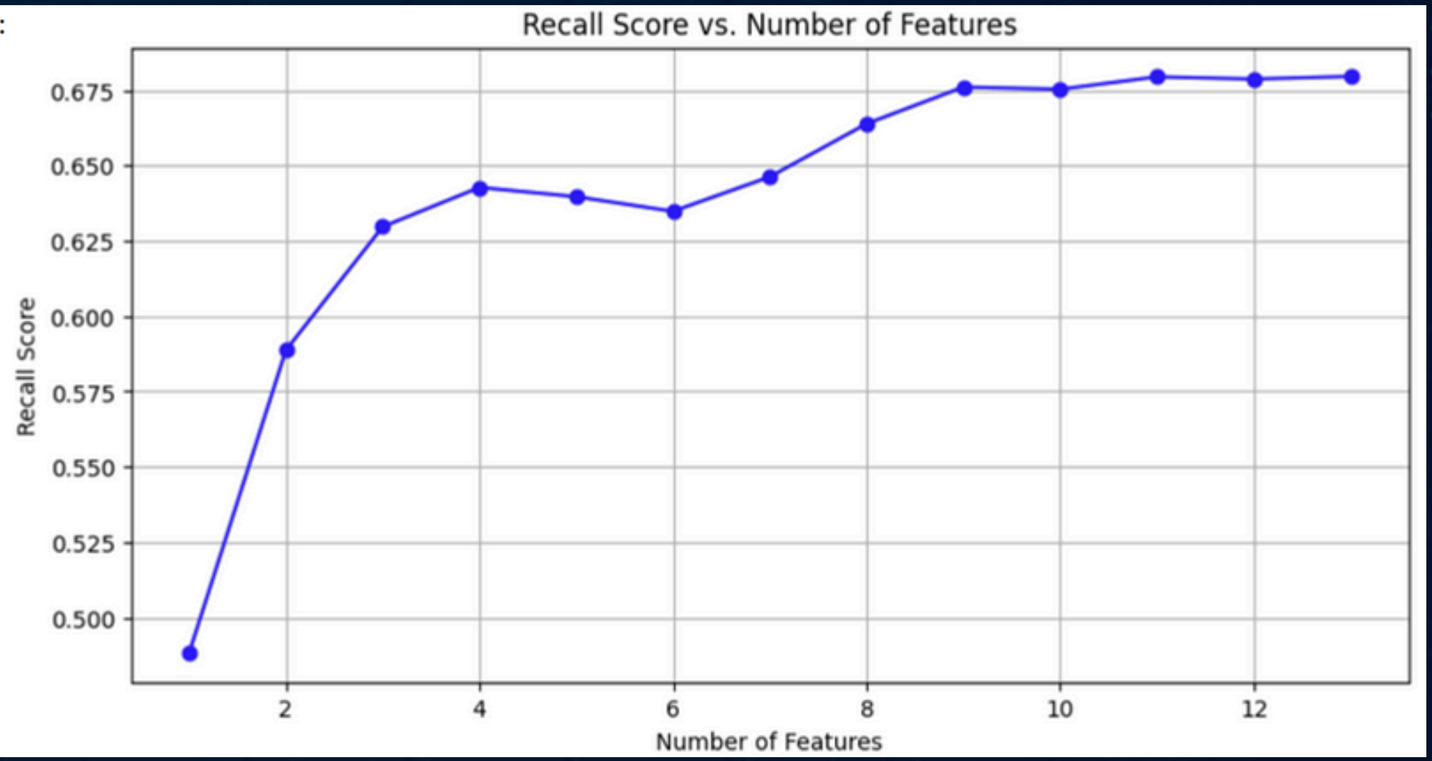
RECALL SCORE

- IT IS THE RATIO: TP/(TP+FN)
- FOCUSES SPECIFICALLY ON THE MODEL'S ABILITY TO IDENTIFY ALL POSITIVE SAMPLES
- MORE RELEVANT TO OUR PROBLEM DEFINITIONS

FEATURE SELECTION

Feature Ranking (1 indicates selected features):

	Feature	Ranking
15	bmi	1
4	ap_hi	2
0	age	3
3	weight	4
2	height	5
16	map	6
1	gender	7
5	ap_lo	8
14	cholesterol_3	9
9	gluc_1	10
8	active	11
12	cholesterol_1	12
6	smoke	13
13	cholesterol_2	14
7	alco	15
10	gluc_2	16
11	gluc_3	17



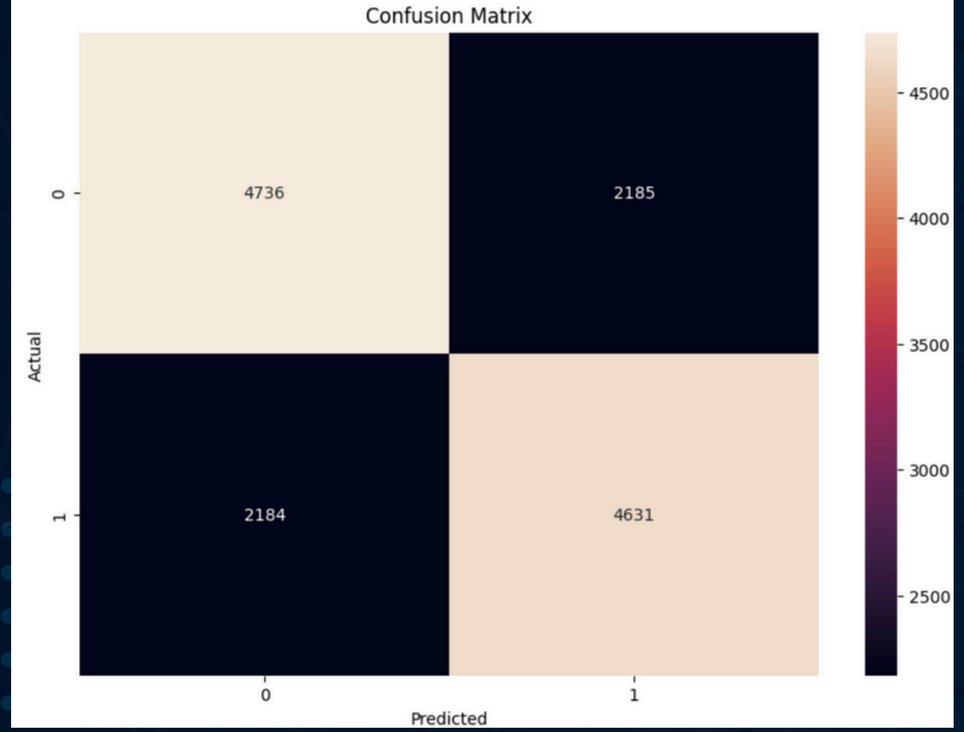
MACHINE LEARNING

MODEL USED
RANDOM FOREST
K-NEAREST NEIGHBOURS
NEURAL NETWORK

RANDOM FOREST

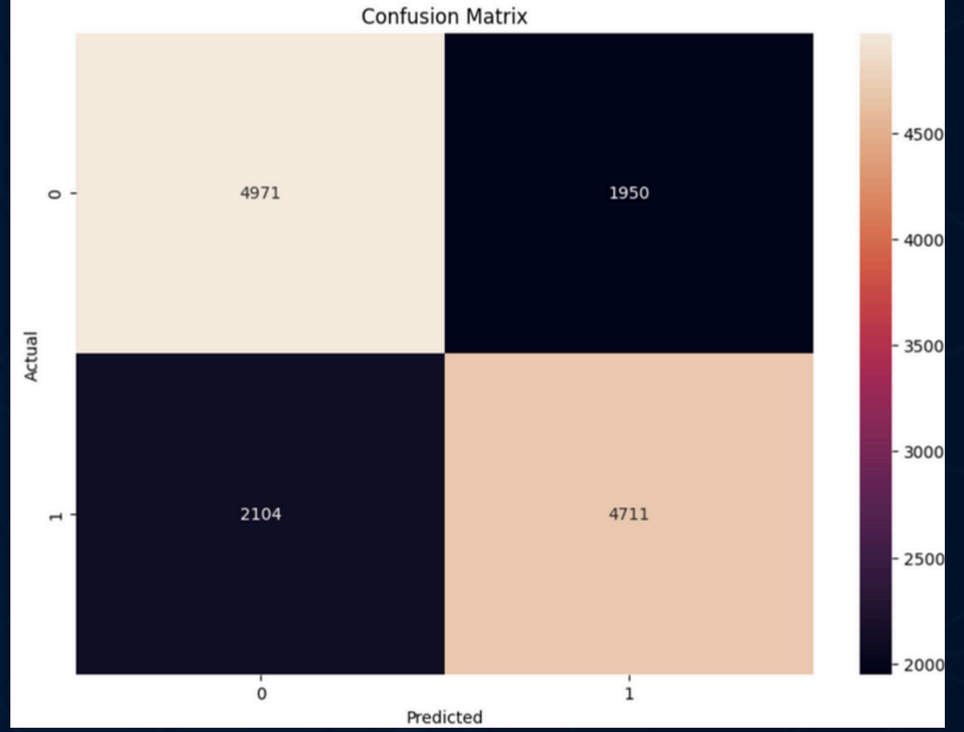
BEFORE GRID SEARCH

	precision	recall	f1-score	support
0	0.68	0.68	0.68	6921
1	0.68	0.68	0.68	6815
accuracy			0.68	13736
macro avg	0.68	0.68	0.68	13736
weighted avg	0.68	0.68	0.68	13736



AFTER GRID SEARCH

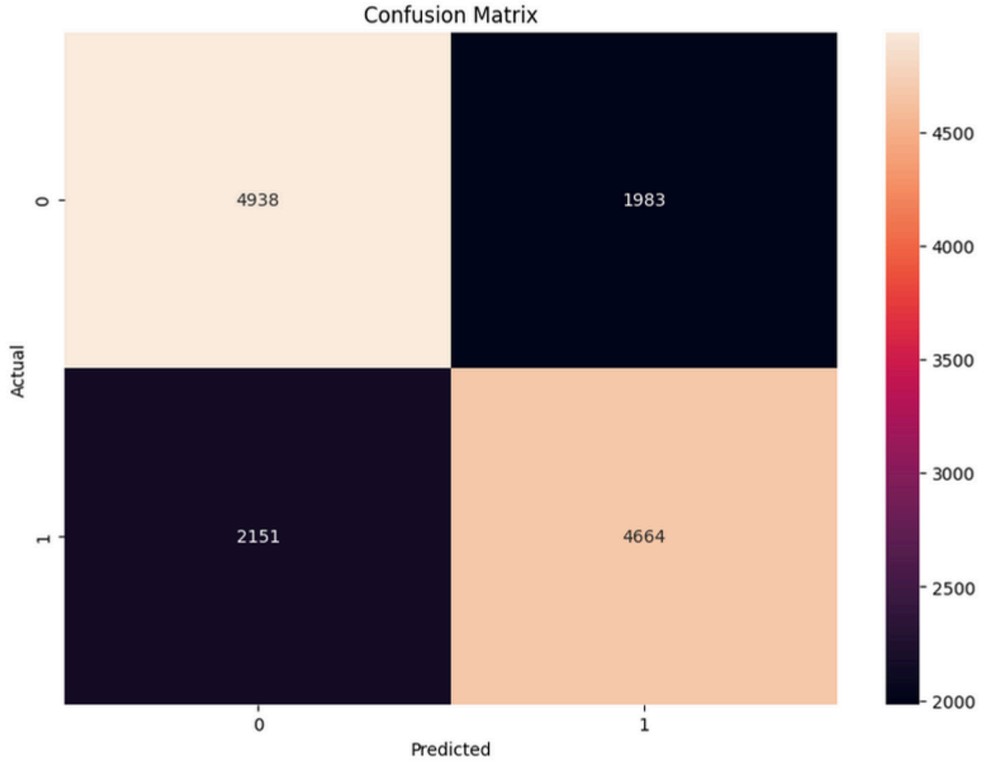
	precision	recall	f1-score	support
0	0.70	0.72	0.71	6921
1	0.71	0.69	0.70	6815
accuracy			0.70	13736
macro avg	0.70	0.70	0.70	13736
weighted avg	0.70	0.70	0.70	13736



K-NEAREST-NEIGHBORS

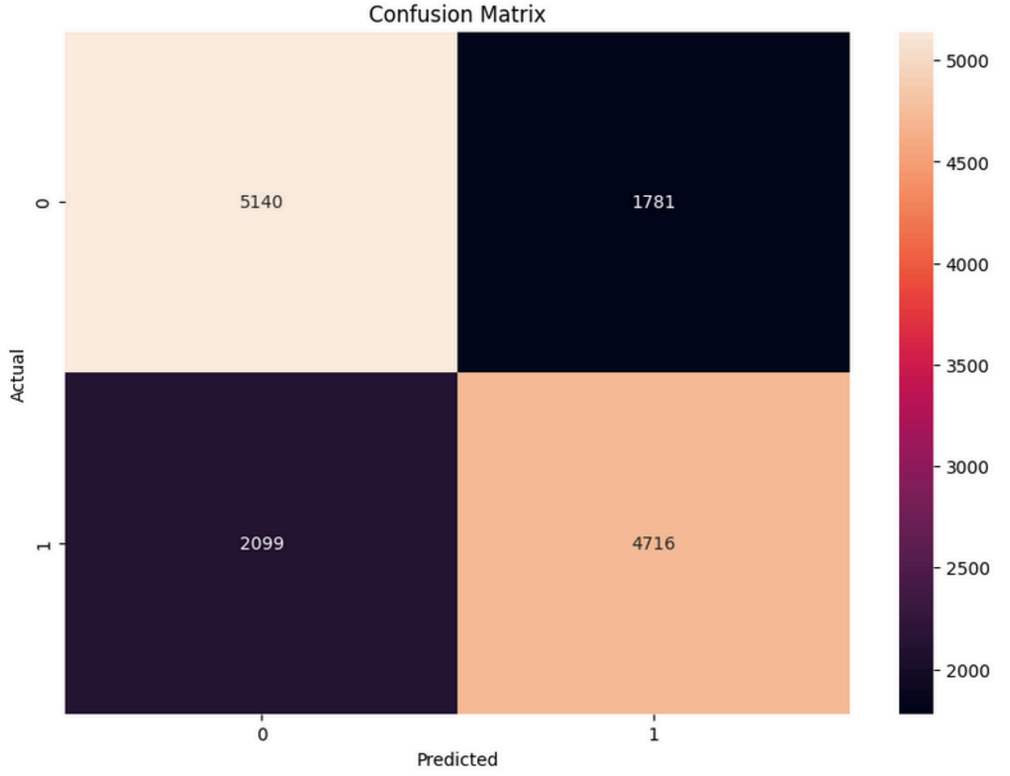
BEFORE GRID SEARCH

	precision	recall	f1-score	support
0	0.70	0.71	0.70	6921
1	0.70	0.68	0.69	6815
accuracy			0.70	13736
macro avg	0.70	0.70	0.70	13736
weighted avg	0.70	0.70	0.70	13736



AFTER GRID SEARCH

	precision	recall	f1-score	support
0	0.71	0.74	0.73	6921
1	0.73	0.69	0.71	6815
accuracy			0.72	13736
macro avg	0.72	0.72	0.72	13736
weighted avg	0.72	0.72	0.72	13736

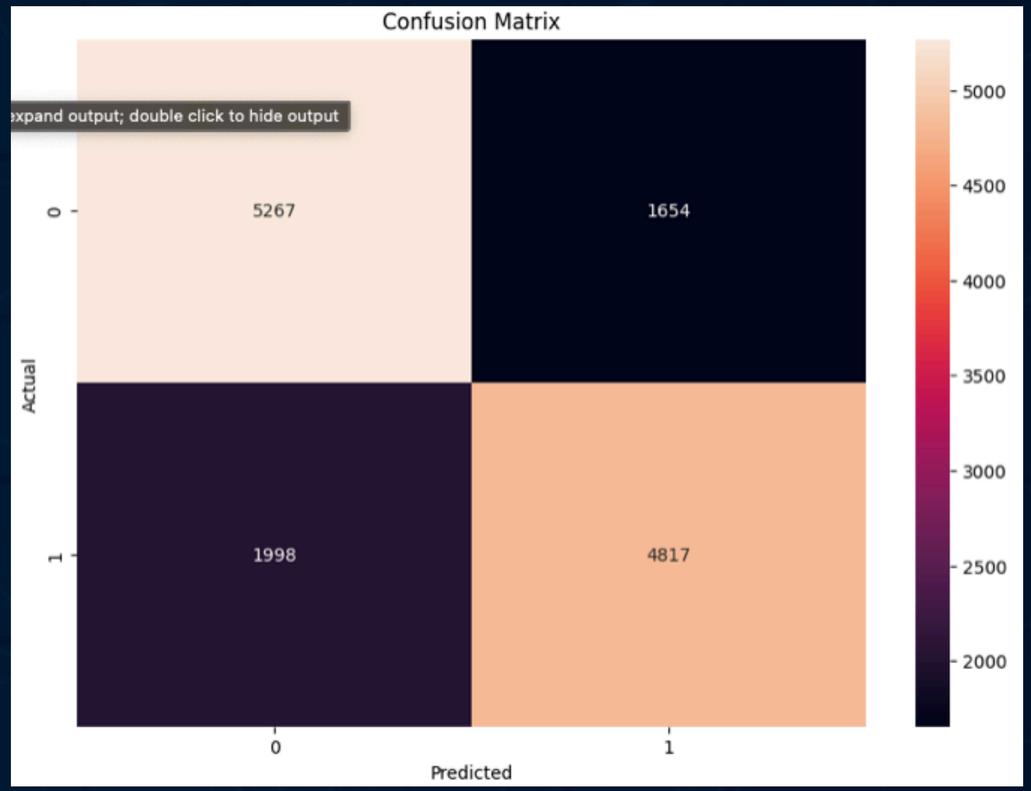


NEURAL NETWORK
TENSORFLOW KERAS

```
Epoch 1/10
1717/1717 ━━━━━━━━━━━━━━━━━━━━ 2s 777us/step - accuracy: 0.6527 - loss: 0.6257
Epoch 2/10
1717/1717 ━━━━━━━━━━━━━━━━━━━━ 1s 656us/step - accuracy: 0.7246 - loss: 0.5547
Epoch 3/10
1717/1717 ━━━━━━━━━━━━━━━━━━━━ 1s 670us/step - accuracy: 0.7267 - loss: 0.5518
Epoch 4/10
1717/1717 ━━━━━━━━━━━━━━━━━━━━ 1s 657us/step - accuracy: 0.7315 - loss: 0.5464
Epoch 5/10
1717/1717 ━━━━━━━━━━━━━━━━━━━━ 1s 652us/step - accuracy: 0.7322 - loss: 0.5443
Epoch 6/10
1717/1717 ━━━━━━━━━━━━━━━━━━━━ 1s 648us/step - accuracy: 0.7275 - loss: 0.5477
Epoch 7/10
1717/1717 ━━━━━━━━━━━━━━━━━━━━ 1s 655us/step - accuracy: 0.7318 - loss: 0.5470
Epoch 8/10
1717/1717 ━━━━━━━━━━━━━━━━━━━━ 1s 659us/step - accuracy: 0.7305 - loss: 0.5454
Epoch 9/10
1717/1717 ━━━━━━━━━━━━━━━━━━━━ 1s 732us/step - accuracy: 0.7333 - loss: 0.5424
Epoch 10/10
1717/1717 ━━━━━━━━━━━━━━━━━━━━ 1s 642us/step - accuracy: 0.7341 - loss: 0.5422
```

	precision	recall	f1-score	support
0	0.72	0.76	0.74	6921
1	0.74	0.71	0.73	6815
accuracy			0.73	13736
macro avg	0.73	0.73	0.73	13736
weighted avg	0.73	0.73	0.73	13736

```
Fit time: 12.810 seconds
Predict time: 0.431 seconds
```



ANALYSIS

```
Random Forest:

Fit time: 14.171 seconds
Predict time: 0.680 seconds
              precision    recall  f1-score   support

           0       0.70      0.72      0.71      6921
           1       0.71      0.69      0.70      6815

    accuracy                           0.70     13736
   macro avg       0.70      0.70      0.70     13736
weighted avg       0.70      0.70      0.70     13736

K-Nearest-Neighbors:

Fit time: 0.126 seconds
Predict time: 1.695 seconds
              precision    recall  f1-score   support

           0       0.71      0.74      0.73      6921
           1       0.73      0.69      0.71      6815

    accuracy                           0.72     13736
   macro avg       0.72      0.72      0.72     13736
weighted avg       0.72      0.72      0.72     13736

Neural Network:

              precision    recall  f1-score   support

           0       0.72      0.76      0.74      6921
           1       0.74      0.71      0.73      6815

    accuracy                           0.73     13736
   macro avg       0.73      0.73      0.73     13736
weighted avg       0.73      0.73      0.73     13736

Fit time: 12.810 seconds
Predict time: 0.431 seconds
```

- K-Nearest-Neighbors has the fastest fit time
- Neural Network has the highest recall score but the slowest fit time
- Time is not an important factor

FUTURE IMPROVEMENTS

- EXPANSION OF DATASET
- VAGUENESS OF OBJECTIVE VARIABLES
- LOCALISING DATASET