HEART DISEASE PREDICTION USING MACHINE LEARNING

CAPSTONE PROJECT - PREDICTIVE ANALYTICS

PRESENTED BY JANE FOO

DATE 29 AUG 2023

AGENDA

BUSINESS PROBLEM

DATA PREPROCESSING

DATA MODELING

MODEL PERFORMANCE EVALUATION

CONCLUSION

BUSINESS PROBLEM

Background

Heart disease is a leading cause of mortality globally. Based on World Heart Federation, this disease has caused over 18.6 million deaths per year worldwide

51k death per day

Objective

The goal is to develop a machine learning model that can effectively predict the presence or absence of heart disease

This would allow the medical team to perform early intervention to prevent cardiac arrest and death

DATA PREPROCESSING

DATA UNDERSTANDING

CORRELATION ANALYSIS

DATA PREPARATION

DATA UNDERSTANDING

Dataset provided was a subset of Parkway Pantai's patient biodata (2018 - 2023)

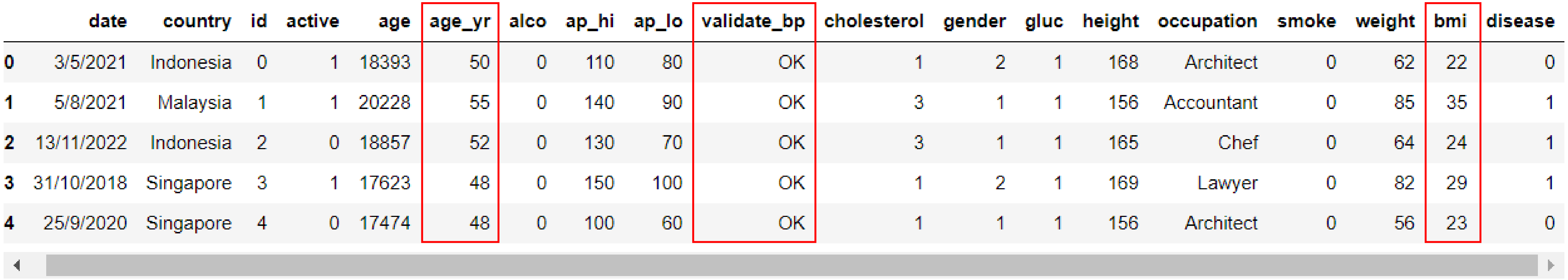
- Consists of patients from India, Indonesia, Malaysia and Singapore

Raw data consists of 70k rows and 16 columns

- Include presence or absence of heart disease (target)
- Contains dirty data
  - Data cleaning using Microsoft Excel
    - Blood pressure (bp)
      - Systolic bp within 60 to 200
      - Diastolic bp within 40 to 100
      - Systolic bp > Diastolic bp
    - Weight within 40 kg to 200 kg
- No duplicates, no null and no missing data

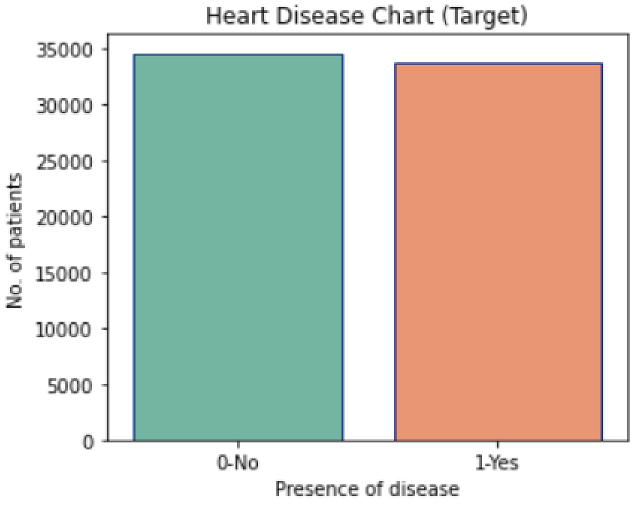
DATA UNDERSTANDING

- Custom columns were created to fine-tune the dataset (boxed in red below)
  - age_yr – to convert patients' age from days to years
  - validate_bp – to ensure that systolic bp (ap_hi) > diastolic bp (ap_lo)
  - bmi – to calculate the body mass index to determine overweight patients
- Performed 2.4% of data cleaning with a balance of 68k rows



	date	country	id	active	age	age_yr	alco	ap_hi	ap_lo	validate_bp	cholesterol	gender	gluc	height	occupation	smoke	weight	bmi	disease
0	3/5/2021	Indonesia	0	1	18393	50	0	110	80	OK	1	2	1	168	Architect	0	62	22	0
1	5/8/2021	Malaysia	1	1	20228	55	0	140	90	OK	3	1	1	156	Accountant	0	85	35	1
2	13/11/2022	Indonesia	2	0	18857	52	0	130	70	OK	3	1	1	165	Chef	0	64	24	1
3	31/10/2018	Singapore	3	1	17623	48	0	150	100	OK	1	2	1	169	Lawyer	0	82	29	1
4	25/9/2020	Singapore	4	0	17474	48	0	100	60	OK	1	1	1	156	Architect	0	56	23	0

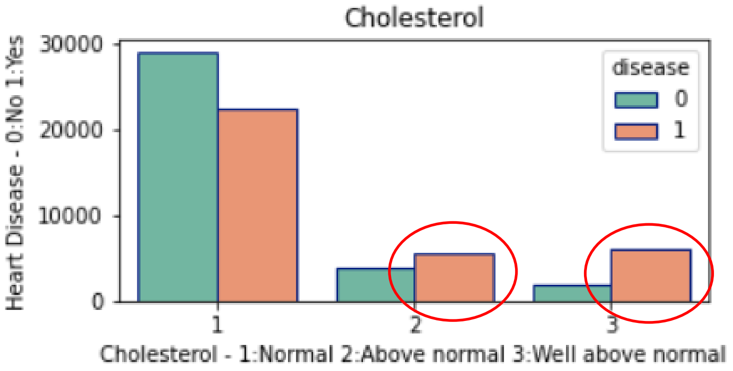
CATEGORICAL VARIABLES

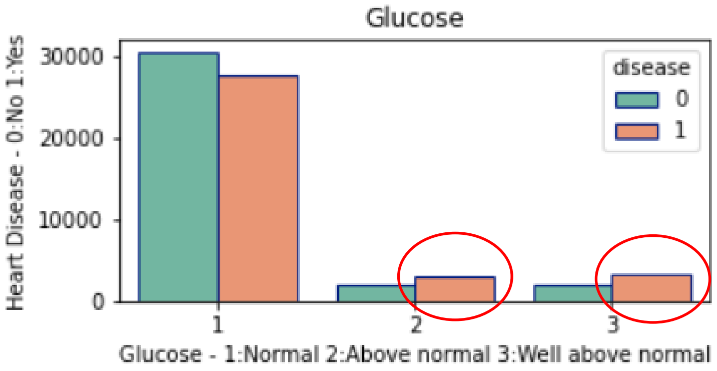


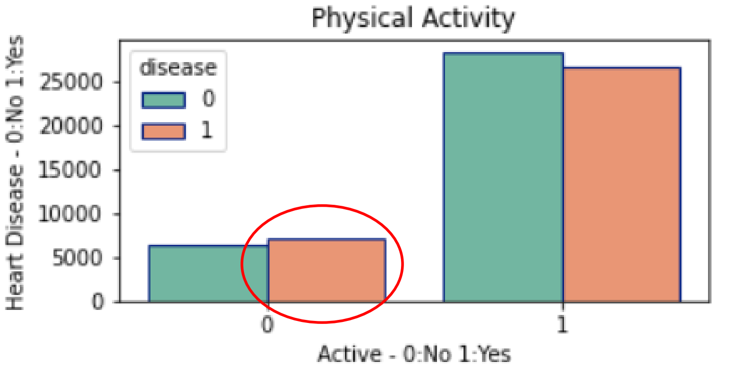
(Target) Balanced Distribution of patients with heart disease and no heart disease

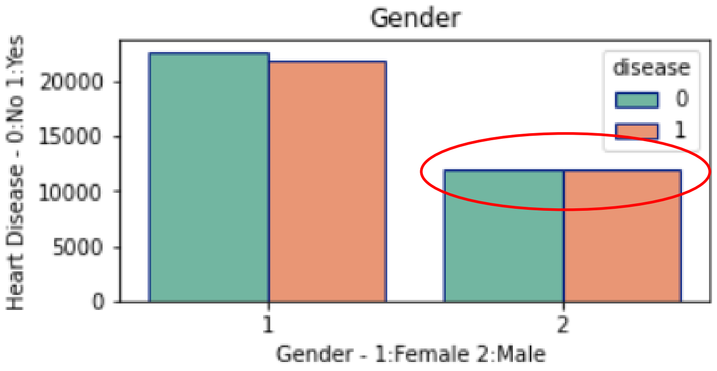
(Predictors)

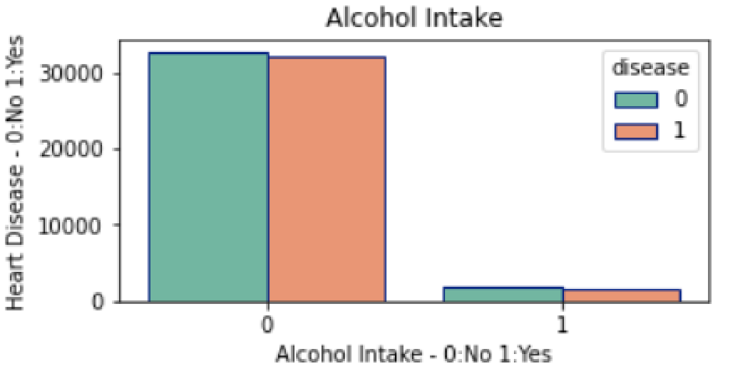
- Imbalanced distributions of inputs
- Risk factors
  - High cholesterol or glucose level
  - Physically inactive
  - Likely to affect male than female
- Not related to heavy drinker or smoker

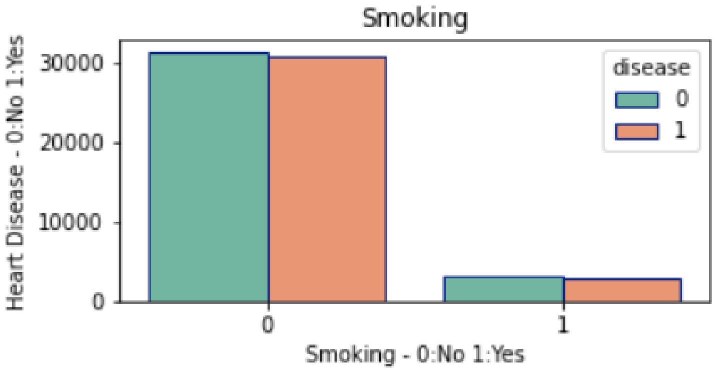




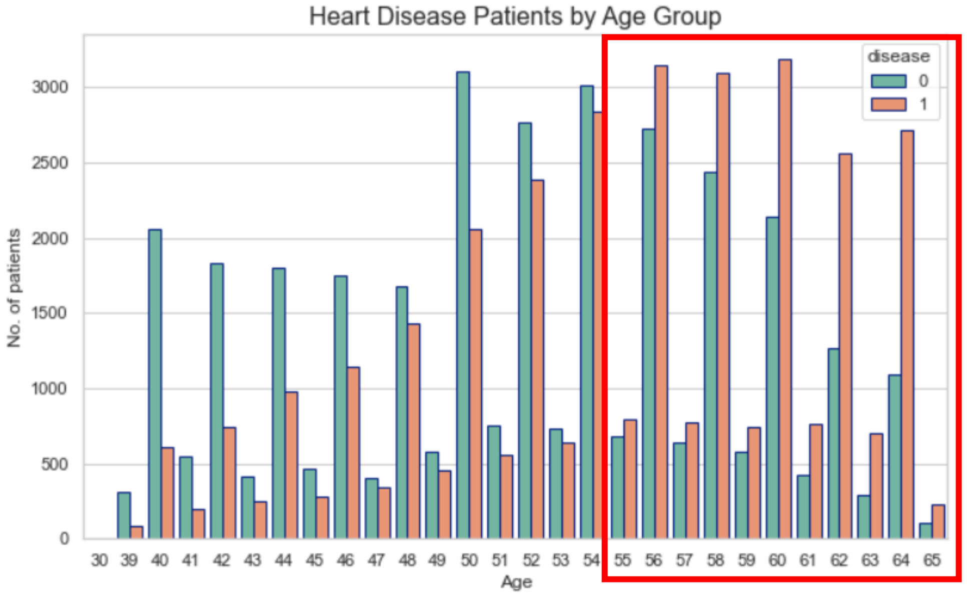


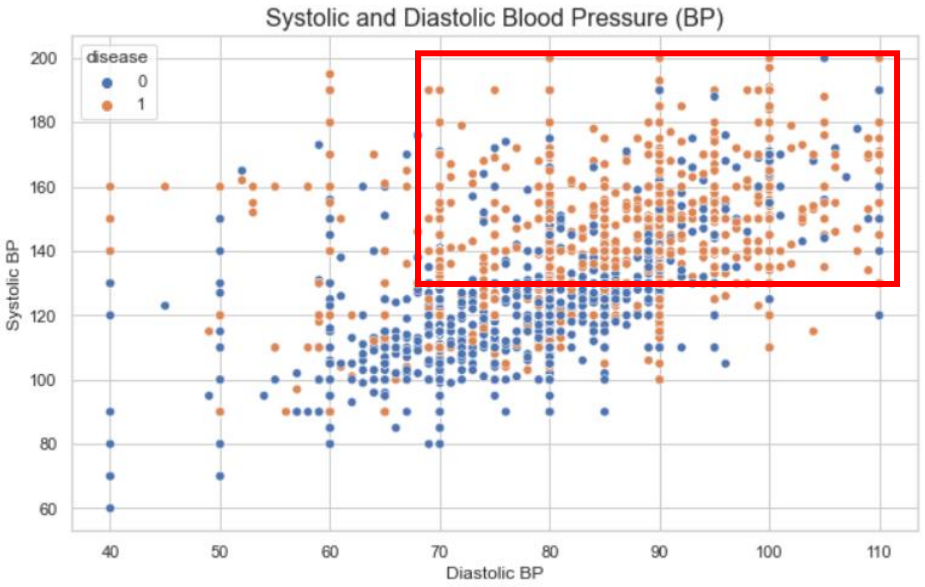


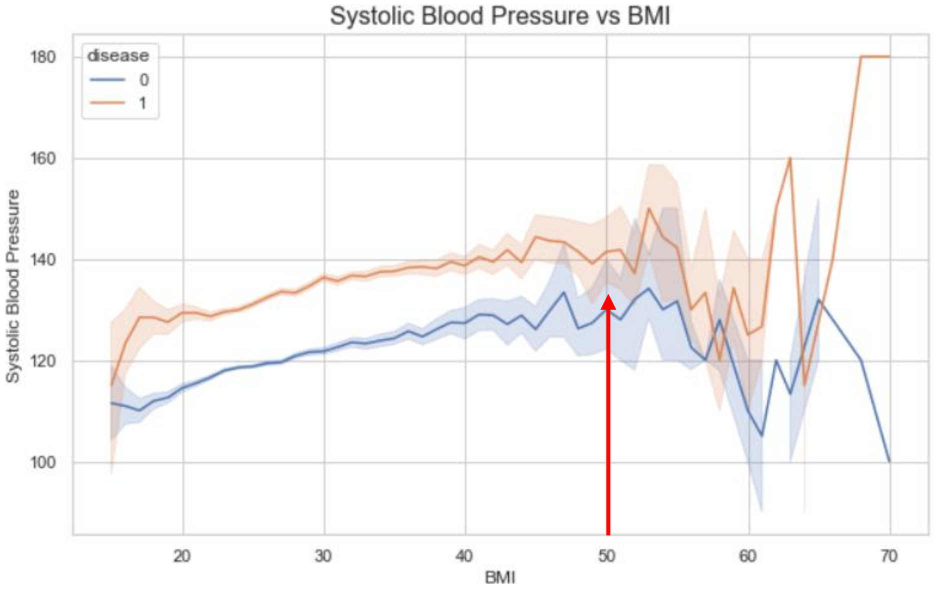




NUMERICAL VARIABLES


Heart Disease Patients by Age Group


Systolic and Diastolic Blood Pressure (BP)


Systolic Blood Pressure vs BMI

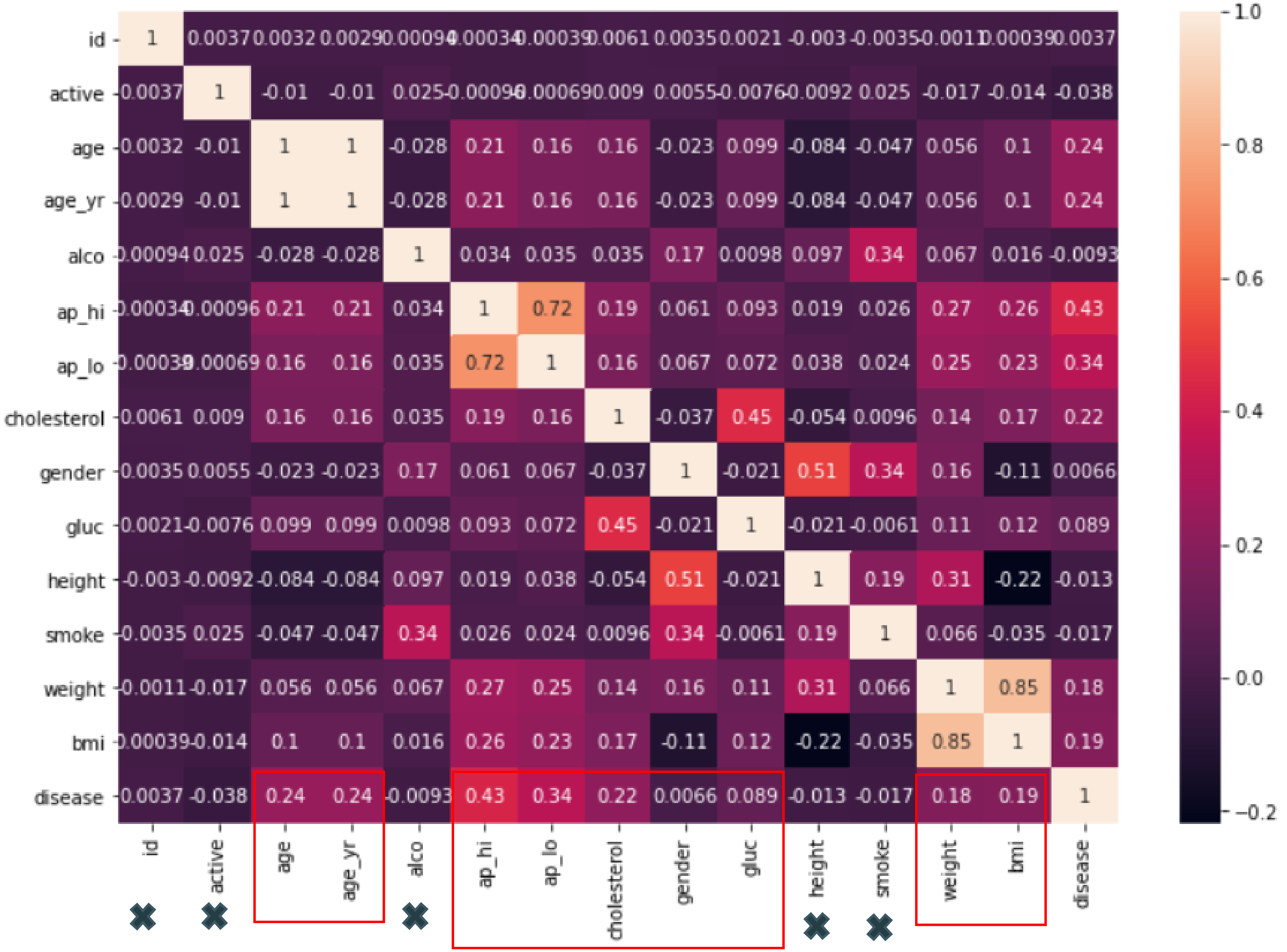
- Older group of patients were more likely to get a heart disease, i.e. 55 years old and above (boxed in red)

- Heart disease risks
  - Systolic bp 130 and above (boxed in red)
  - Diastolic bp 70 and above (boxed in red)
  - Obese patients especially those with bmi 50 and above (red arrow)

CORRELATION ANALYSIS

- Predictors with positive correlation with heart disease (boxed in red)

- Predictors with negative correlation with heart disease
  - Patient identification
  - Physically active
  - Alcohol drinker
  - Patient height
  - Smoker



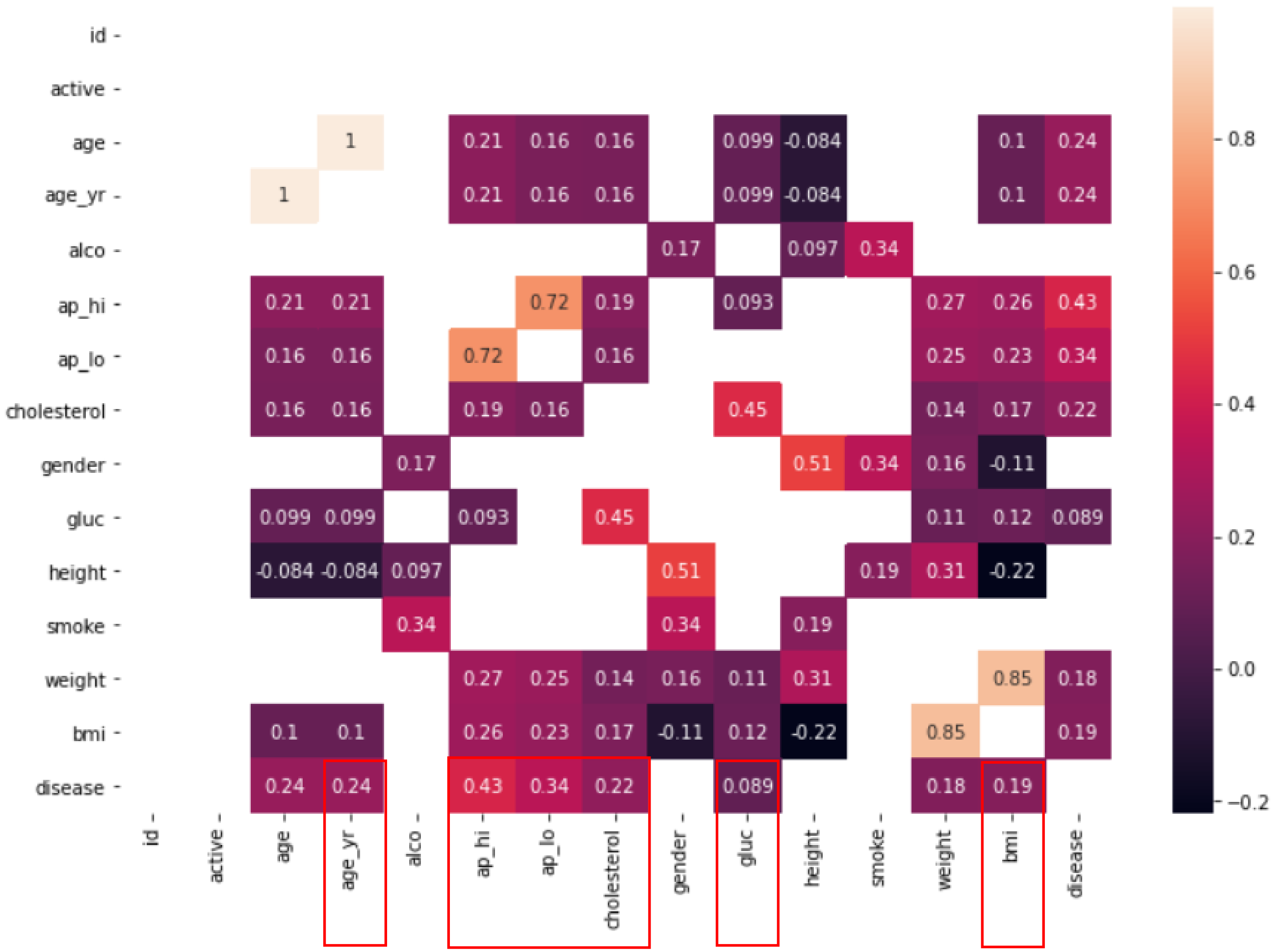
	id	active	age	age_yr	alco	ap_hi	ap_lo	cholesterol	gender	gluc	height	smoke	weight	bmi	disease
id	1	0.0037	0.0032	0.0029	0.00094	0.00034	0.00039	0.0061	0.0035	0.0021	-0.003	-0.0035	-0.0011	0.00039	0.0037
active	0.0037	1	-0.01	-0.01	0.025	-0.00096	-0.00069	0.009	0.0055	-0.0076	-0.0092	0.025	-0.017	-0.014	-0.038
age	0.0032	-0.01	1	1	-0.028	0.21	0.16	0.16	-0.023	0.099	-0.084	-0.047	0.056	0.1	0.24
age_yr	0.0029	-0.01	1	1	-0.028	0.21	0.16	0.16	-0.023	0.099	-0.084	-0.047	0.056	0.1	0.24
alco	0.00094	0.025	-0.028	-0.028	1	0.034	0.035	0.035	0.17	0.0098	0.097	0.34	0.067	0.016	-0.0093
ap_hi	0.00034	-0.00096	0.21	0.21	0.034	1	0.72	0.19	0.061	0.093	0.019	0.026	0.27	0.26	0.43
ap_lo	0.00039	-0.00069	0.16	0.16	0.035	0.72	1	0.16	0.067	0.072	0.038	0.024	0.25	0.23	0.34
cholesterol	0.0061	0.009	0.16	0.16	0.035	0.19	0.16	1	-0.037	0.45	-0.054	0.0096	0.14	0.17	0.22
gender	0.0035	0.0055	-0.023	-0.023	0.17	0.061	0.067	-0.037	1	-0.021	0.51	0.34	0.16	-0.11	0.0066
gluc	0.0021	-0.0076	0.099	0.099	0.0098	0.093	0.072	0.45	-0.021	1	-0.021	-0.0061	0.11	0.12	0.089
height	-0.003	-0.0092	-0.084	-0.084	0.097	0.019	0.038	-0.054	0.51	-0.021	1	0.19	0.31	-0.22	-0.013
smoke	-0.0035	0.025	-0.047	-0.047	0.34	0.026	0.024	0.0096	0.34	-0.0061	0.19	1	0.066	-0.035	-0.017
weight	-0.0011	-0.017	0.056	0.056	0.067	0.27	0.25	0.14	0.16	0.11	0.31	0.066	1	0.85	0.18
bmi	0.00039	-0.014	0.1	0.1	0.016	0.26	0.23	0.17	-0.11	0.12	-0.22	-0.035	0.85	1	0.19
disease	0.0037	-0.038	0.24	0.24	-0.0093	0.43	0.34	0.22	0.0066	0.089	-0.013	-0.017	0.18	0.19	1

DATA PREPARATION

- Feature Selection
  - Predictors that had a **0.08** and above correlation with heart disease

- Feature Scaling
  - Normalised all features to a standard scale to prevent feature with larger magnitude from dominating the learning process

- Feature Engineering
  - Age in year format
  - BMI



DATA MODELING AND EVALUATION

MODEL SELECTION

PREDICTION PERFORMANCE

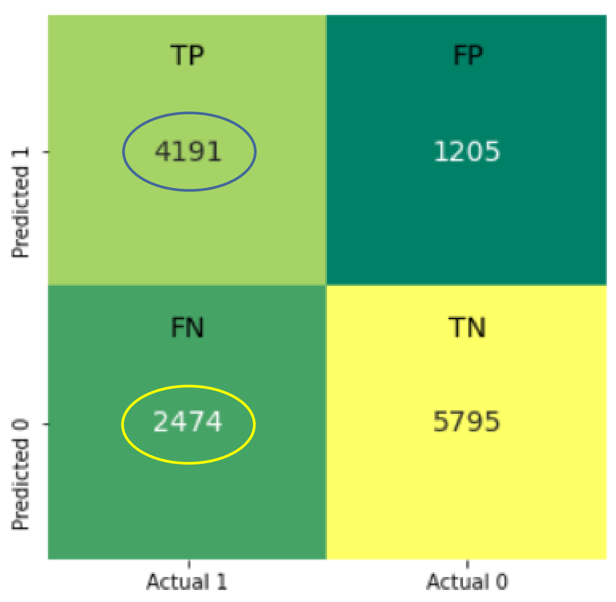
CONFUSION MATRIX AND CLASSIFICATION REPORT

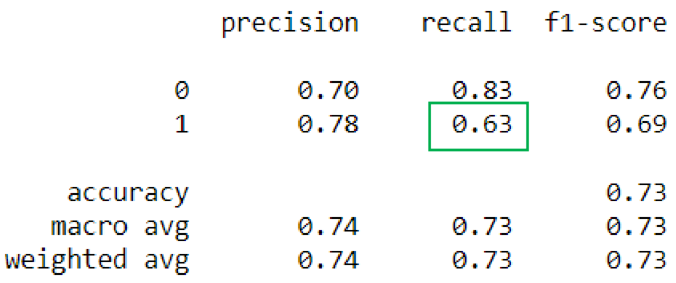
MODEL SELECTION

SUPERVISED MACHINE LEARNING - CLASSIFIER MODEL	PREDICTION PERFORMANCE	SETTING DATA SPLIT: 20% TEST, 80% TRAIN
Support Vector Machine	★ 73.08%	Linear kernel, Regularisation strength 1.0
K-Nearest Neighbour	★ 73.21%	Used 27 nearest neighbours for prediction, Manhattan distance metric
Logistic Regression	★ 73.03%	Used grid searching key for other hyperparameter setting same accuracy
Decision Tree	68.07%	Default

CONFUSION MATRIX AND CLASSIFICATION REPORT

Confusion Matrix - Support Vector Machine



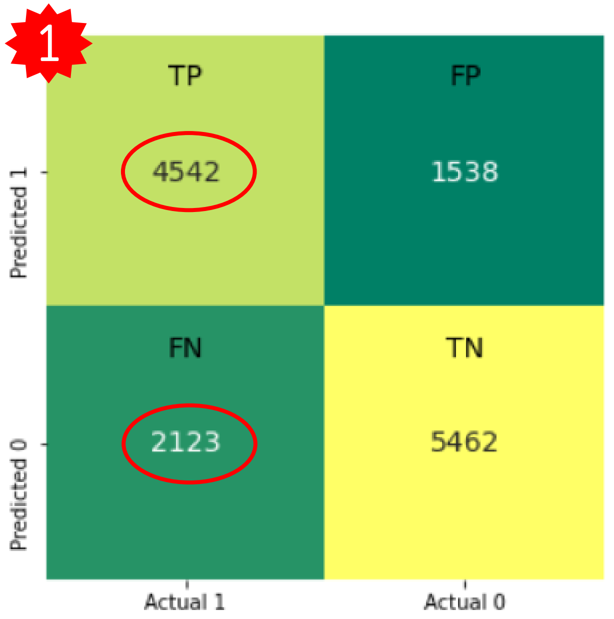


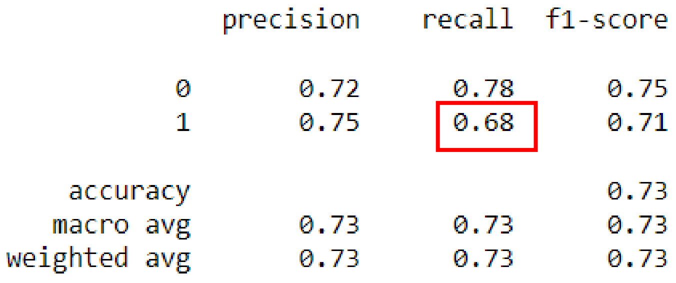
```
              precision    recall  f1-score

           0       0.70      0.83      0.76
           1       0.78      0.63      0.69

    accuracy                           0.73
   macro avg       0.74      0.73      0.73
weighted avg       0.74      0.73      0.73
```

Confusion Matrix - K-Nearest Neighbour



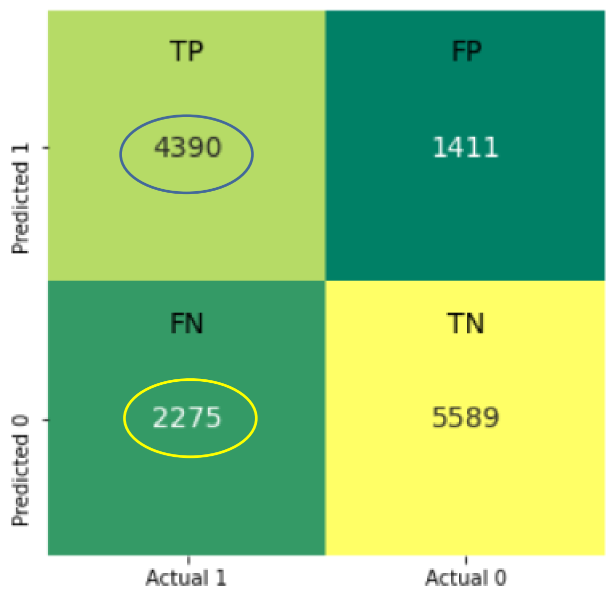


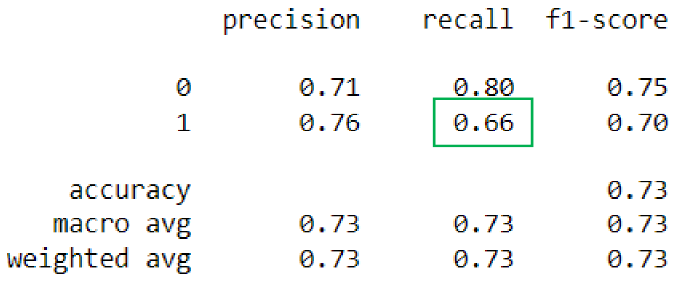
```
              precision    recall  f1-score

           0       0.72      0.78      0.75
           1       0.75      0.68      0.71

    accuracy                           0.73
   macro avg       0.73      0.73      0.73
weighted avg       0.73      0.73      0.73
```

Confusion Matrix - Logistic Regression





```
              precision    recall  f1-score

           0       0.71      0.80      0.75
           1       0.76      0.66      0.70

    accuracy                           0.73
   macro avg       0.73      0.73      0.73
weighted avg       0.73      0.73      0.73
```

CONCLUSION

K-Nearest Neighbour (KNN) is the right model

Accurately predicted the highest number of heart disease patients

Lowest misclassification score as compared to the other two models

THANK YOU

JANE FOO