STAT 5104 Data Mining
Final Report

Study on Predictive Factors of Myocardial Infarction Risk

Group Members:

WONG Tin Yan Tim 1155110207

[redacted]

Table of Contents

1. Introduction

Myocardial infarction (MI), commonly known as heart attack, is a major manifestation of ischemic heart disease (IHD) and poses a significant burden on healthcare systems worldwide. In Hong Kong alone, there were 7,700 inpatient discharges and deaths related to MI in 2020 (Ho, 2022). The economic impact is substantial, with the direct healthcare cost of IHD reaching 10,064 International Dollars per patient annually in 2019, representing a notable share of Hong Kong's healthcare expenditure (Rittiphairoj et al., 2025).

In recent years, the concept of personalized medicine has gained traction across medical disciplines, emphasizing the tailoring of medical care to individual patient profiles. In cardiology, accurately identifying individuals at high 10-year risk of atherosclerotic cardiovascular disease (ASCVD) is crucial (Arnett et al., 2019). Early identification enables targeted lifestyle interventions and screening, which can help prevent the onset or progression of advanced ASCVD and improve quality of life. Conversely, precise risk stratification helps avoid unnecessary procedures for low-risk individuals, reducing both financial and resource burdens on the healthcare system.

To address this need, we employed a machine learning approach to analyze a wide range of potential risk factors associated with MI. By leveraging a real-world dataset representative of the Asian population, we aimed to generate clinically relevant insights for the Hong Kong setting, while accounting for possible racial and genetic differences.

2. Objectives

The objective of this project was to facilitate personalized medicine by uncovering potential ASCVD predictive factors and to assist physicians in individualizing treatment for individuals with different ASCVD risk levels.

3. Methodology

3.1 Data Collection

We used the dataset "Heart Attack Prediction in Indonesia" from Kaggle.com (HA.csv). The original dataset contains 158,355 entries with its 27 predictor variables being categorized into five aspects, including 1) demographics, 2) clinical risk factors and comorbidities, 3) lifestyle and behavioral factors, 4) environmental and social factors and 5) clinical measurement, medical history and health system factors. Continuous, nominal and ordinal variables are all present in the dataset.

3.2 Data Cleaning

To facilitate the coming data exploration and modelling for the purpose of this study, we cleaned the data with the following steps:

(i) Handling Missing Values

Only the *alcohol_consumption* column contained missing values, which were removed from the dataset. Given the dataset's comprehensive medical measurements and minimal missing data, we assumed the results were not generated from surveys and the loss of values were random. Hence, removal of observations with missing value will not introduce skewness. After cleaning, 63,507 observations with 27 predictors and 1 response variable remained.

(ii) Standardizing Text Data in Lower Case

Apart from missing data, among the existing data, there are inconsistencies and upper case letters. For better handling and to avoid issues arising from case sensitivity, all string columns were converted to lowercase.

(iii) Creating a New Feature

Non-HDL cholesterol, a key cardiovascular risk factor know to have predictive power in ASCVD risk, was not provided in the dataset. Since total cholesterol and high-density lipoprotein cholesterol (HDL-C) were available, we created a new column, *cholesterol_nonhdl*, by subtracting the HDL-C from total cholesterol.

(iv) Dropping Irrelevant Columns

The variable "hypertension" was a binary variable solely dependent on the level of systolic and diastolic blood pressure (SBP and DBP). The columns *blood_pressure_systolic* and *blood_pressure_diastolic* would contain all information regarding a subject's health in terms of blood pressure, hence it is removed to reduce multicollinearity.

To reduce multicollinearity between *cholesterol_hdl*, *cholesterol_ldl* and *total_cholesterol*, the column *cholesterol_hdl* was removed as the association of HDL-C was lower than that of low-density lipoprotein cholesterol (LDL-C) in myocardial infarction (Arnett et al., 2019).

(v) Encoding Categorical Variables

To facilitate analysis and modelling, categorical variables in strings including *income_level*, *region, gender, smoking_status, alcohcol_consumption, physical_activity, dietary_habits, air_pollution_exposure, stress_level* and *EKG_results* were encoded into numerical values. The mappings are all stored in a list of dictionaries and deployed in the data set.

(vi) Rearranging Columns

To improve readability and logical grouping, the new column *cholesterol_nonhdl* was repositioned after *cholesterol_level* and before *cholesterol_ldl*.

3.3 Data Description and Structure

The cleaned data set consists of 63,507 entries and 26 columns, described as followed:

Column Name	Data Type	Description
age	Numeric	Age of the patient in years
gender	Nominal, binary	Patient's gender (0 = Female, 1 = Male)
region	Nominal, binary	Region of residence (0 = Rural, 1 = Urban)
income_level	Ordinal, categorical	Economic status (0 = Low, 1 = Medium, 2 = High)
diabetes	Nominal, binary	Presence of diabetes (0 = No, 1 = Yes)
obesity	Nominal, binary	Presence of obesity (0 = No, 1 = Yes)
waist_circumference	Continuous	Waist circumference measurement in centimeters
family_history	Nominal, binary	Family history of heart disease (0 = No, 1 = Yes)
smoking_status	Nominal, categorical	Smoking habits (0 = Never, 1 = Former, 2 = Current)
alcohol_consumption	Ordinal, binary	Regular alcohol consumption (0 = Moderate, 1 = High)
physical_activity	Ordinal, categorical	Level of physical activity (0 = Low, 1 = Moderate, 2 = High)
dietary_habits	Nominal, binary	Healthy dietary habits (0 = Poor, 1 = Good)
air_pollution_exposure	Ordinal, categorical	Level of exposure to air pollution (0 = Low, 1 = Medium, 2 = High)
stress_level	Ordinal, categorical	Level of reported stress (0 = Low, 1 = Medium, 2 = High)
sleep_hours	Continuous	Average daily sleep hours
blood_pressure_systolic	Continuous	Systolic blood pressure measurement (mmHg)
blood_pressure_diastolic	Continuous	Diastolic blood pressure measurement (mmHg)
fasting_blood_sugar	Continuous	Fasting blood glucose level (mg/dL)
cholesterol_level	Continuous	Total cholesterol level (mg/dL)
cholesterol_nonhdl	Continuous	Non-HDL cholesterol level (mg/dL)
cholesterol_ldl	Continuous	LDL cholesterol level (mg/dL)
triglycerides	Continuous	Triglycerides level (mg/dL)
EKG_results	Nominal, binary	Electrocardiogram results (0 = Normal, 1 = Abnormal)
previous_heart_disease	Nominal, binary	History of heart disease (0 = No, 1 = Yes)
medication_usage	Nominal, binary	Current use of cardiovascular medications (0 = No, 1 = Yes)
participated_in_free_screening	Nominal, binary	Participation in preventive screening (0 = No, 1 = Yes)
heart_attack	Nominal, binary	Occurrence of myocardial infarction (0 = No, 1 = Yes) - **Target variable**

4. Data Visualization and Analysis

4.1 Continuous Variables

Exploratory data analysis was performed following the cleaning of the dataset. For continuous variables such as *age*, *waist_circumference*, *cholesterol_level,* etc., boxplots were created to explore their distributions with respect to the target variable, *heart_attack*:

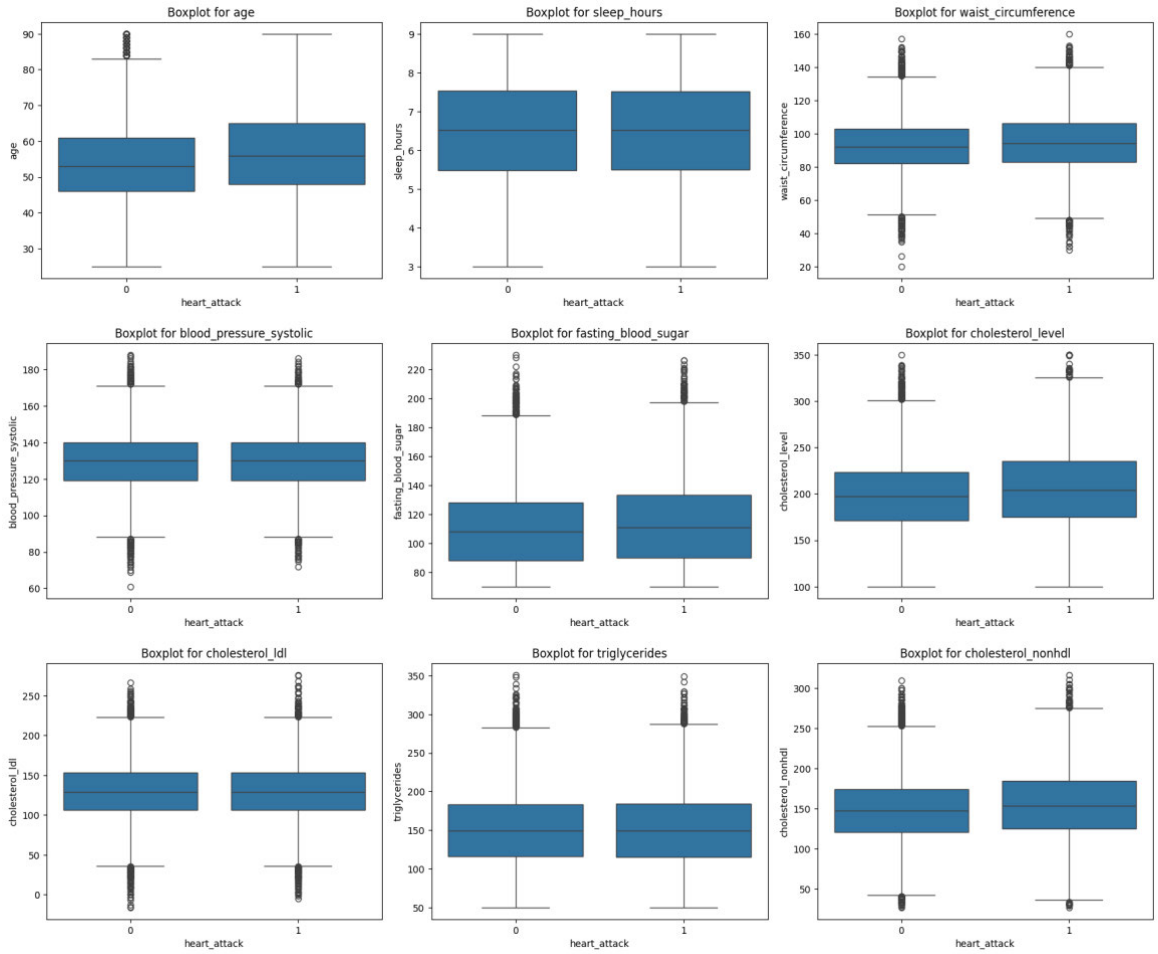


Fig 4.1 Exploratory plots of continuous variables in subjects with and without MI

Among the subjects experienced with MI, it is observed that they have a high median of total cholesterol level and systolic blood pressure. The same observations were observed in ***waist_circumference*** and ***age***, suggesting their potential in predicting myocardial infarction. The remaining variables do not show a pattern within the two groups of subjects.

4.2 Categorical Variables

Bar charts were created to compare the distribution of variables with more than two categories between those suffered or not suffered myocardial infarction.

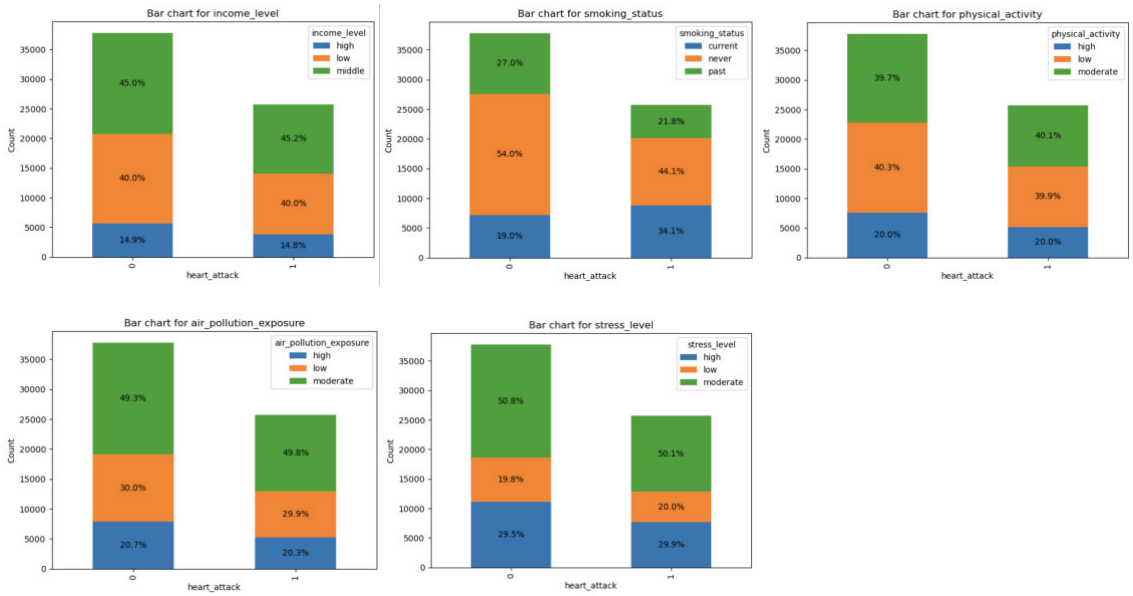
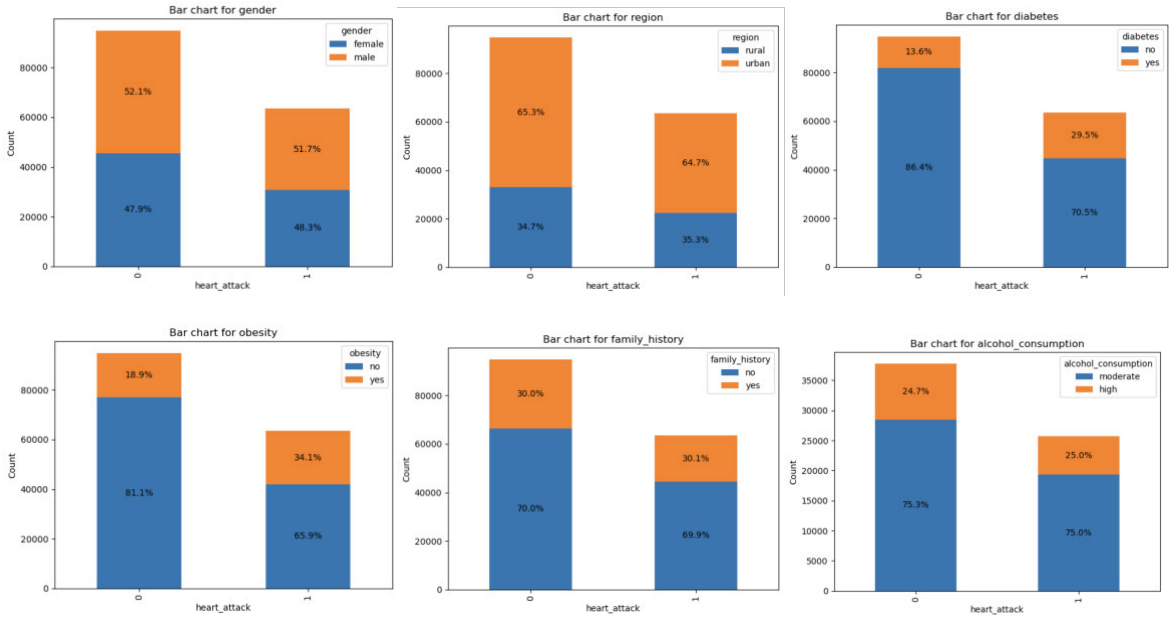


Fig 4.2 Exploratory plots of categorical variables in subjects with and without MI

Among the individuals suffered myocardial infarction, a larger proportion (34.1%) of them were current smoker compared to those who hadn't suffered (19.0%). In the without myocardial infarction, over half of them were non-smokers. It signals the association between **smoking behavior** and MI, potentially attributable to the detrimental effect to the cardiovascular system (Wilhelmsson et al., 1975). The distribution of categories showed a comparable pattern in the remaining variables such as income level, physical activity, air pollution exposure and stress level among the two groups of subjects

4.3 Binary Variables

We further analyzed the binary variables and their distribution within subjects with and without myocardial infarction.



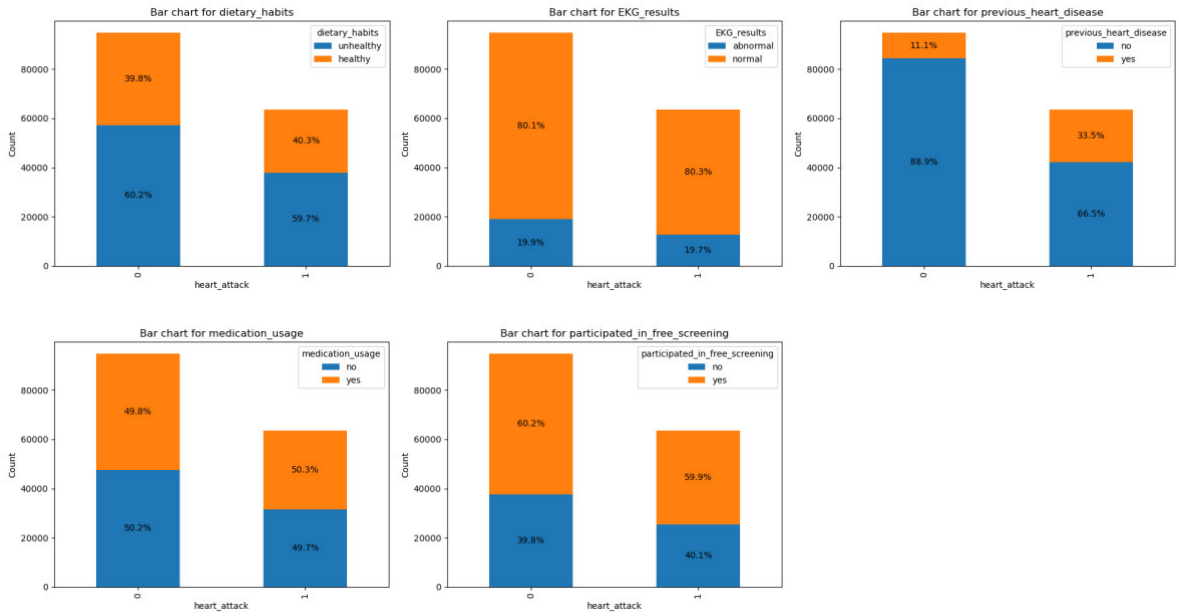


Fig 4.3 Exploratory plots of binary variables in subjects with and without MI

It was observed that the proportion of patients with the comorbidity **diabetes** (29.5%) in subjects with MI more than doubled those without (13.6%). Similar observation was found in the proportion of obesity. In addition, the proportion of subjects with the **history of previous heart disease** more than tripled than those without, evidencing the association between heart disease history and MI. The remaining variables did not display a particular difference among the two groups

5. Prediction Models and Optimization

We selected Gaussian Naïve Bayes, Logistic Regression and Decision Tree Classifier as our three machine learning models. Decision Tree Classifier could be employed first to estimate useful variables followed by the two other models. To optimize model performance, we performed feature selection and data preprocessing with the aid of Mahalanobis Distance (for outlier detection), MinMax Scaling and Hyperparameter Tuning.

The dataset was split into training (80%) and testing (20%) data and the three models were subsequently trained. F1 score, the harmonic mean of precision and recall were picked in our classification problem. False negative rate was also chosen to evaluate the specificity, because the consequence of missing a high risk MI individual (false negative) was larger than wrongly classifying a low-risk individaul as high-risk (false positive).

Feature importance was determined using model-specific methods:

- Gaussian Naive Bayes - Permutation importance, which quantifies the decrease in model performance when a feature's values are randomly shuffled.
- Logistic Regression - Magnitude of model coefficients.

- Decision Tree Classifier - The model's intrinsic feature_importances_ attribute.

In the beginning, we computed the F1 score and false negative rate of each model in the training dataset. Then, we optimized the models with the above techniques and computed the new F1 scores and false negative rates after training.

Finally, we repeated the process with an integrated workflow that combined all three optimization steps: Mahalanobis distance for outlier removal, MinMax scaling for feature normalization, and hyperparameter tuning. This comprehensive preprocessing and optimization approach was expected to achieve the highest predictive performance.

5.1 Mahalanobis Distance

While the MinMax scaling approach normalizes each feature independently to a fixed range, the Mahalanobis distance approach addresses multivariate outliers by considering the correlation structure between variables, providing a more holistic data cleaning method.

The predictive performance of the three models, Gaussian Naive Bayes, Logistic Regression, and Decision Tree Classifier, were evaluated using the fok gull feature set.:

```
Training Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.608979             0.420412
1      LogisticRegression  0.539422             0.514413
2  DecisionTreeClassifier  1.000000             0.000000

Testing Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.610648             0.420332
1      LogisticRegression  0.527279             0.532356
2  DecisionTreeClassifier  0.533139             0.464516
```

Then the outliers were using Mahalanobis distance and the models were retrained.

```
mean = df.mean()
cov = df.cov()

# calculate Mahalanobis distance
mahalanobis_distances = []
for i, row in df.iterrows():
    mahalanobis_distance = distance.mahalanobis(row, mean, np.linalg.inv(cov))
    mahalanobis_distances.append(mahalanobis_distance)

# identify outliers (e.g., above 3 standard deviations)
outlier_threshold = 7
outliers = np.array(mahalanobis_distances) > outlier_threshold

# remove outliers
clean_df = df[~outliers]
```

```
Training Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.606070             0.426348
1      LogisticRegression  0.557451             0.508027
2  DecisionTreeClassifier  1.000000             0.000000

Testing Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.613611             0.418537
1      LogisticRegression  0.562894             0.503610
2  DecisionTreeClassifier  0.539689             0.452098
```

Improvement in F-1 score and false negative rate were observed in Gaussian Naive Bayes and Logistic Regression. Since the decision tree method was robust to outlier, there was no improvement observed after the correction.

5.2 Minmax Scaling

Given that the dataset contains features with vastly different scales and units (e.g., age in years, blood pressure in mmHg, cholesterol in mg/dL), the MinMax scaler was applied to all non-binary features. This normalization step mapped feature values to a 0–1 range, preventing those with larger numeric scales to gain a disproportionately heavy weight and influence the models.

We evaluated the predictive performance of the three models—Gaussian Naive Bayes, Logistic Regression, and Decision Tree Classifier—using the unscaled, full feature set.:

```
Training Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.608979             0.420412
1      LogisticRegression  0.539422             0.514413
2  DecisionTreeClassifier  1.000000             0.000000

Testing Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.610648             0.420332
1      LogisticRegression  0.527279             0.532356
2  DecisionTreeClassifier  0.530322             0.467449
```

We looked for the non-binary features, applied MinMax scaling, and retrained all three models:

```
bin_col = ['gender','region','diabetes','obesity','family_history','alcohol_consumption','dietary_habits',
           'EKG_results','previous_heart_disease','medication_usage','participated_in_free_screening']

scaler = MinMaxScaler()

df_scaled = df.copy()

for col in df.columns:
    if col not in bin_col:
        df_scaled[col] = scaler.fit_transform(df_scaled[[col]])
```

```
Training Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.608979             0.420412
1      LogisticRegression  0.580899             0.484616
2  DecisionTreeClassifier  1.000000             0.000000

Testing Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.610648             0.420332
1      LogisticRegression  0.577621             0.489345
2  DecisionTreeClassifier  0.532473             0.467840
```

Improvement in F-1 score and false negative rate were observed in Logistic Regression, suggesting its susceptibility in non-scaled data points. Decision Tree did not show improvement as it was not affected the distance. It identified cutoffs in various variables to create rectangular classification zones.

5.3 Parameter Tuning

Hyperparameter tuning was employed to optimize the models, specifically the hyperparameters that were not learned before training. Bayesian optimization (via BayesSearchCV) was performed to efficiently search through complex hyperparameter spaces. The model used here focused on Logistic Regression and Decision Tree and the optimization process uses 5-fold cross-validation with F1 score as the optimization metric. 10 iterations were performed to identify the optimal hyperparameters.

We evaluated the predictive performance of the three models, Gaussian Naive Bayes, Logistic Regression, and Decision Tree Classifier, using the full feature set.:

```
Training Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.608979             0.420412
1      LogisticRegression  0.539422             0.514413
2  DecisionTreeClassifier  1.000000             0.000000

Testing Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.610648             0.420332
1      LogisticRegression  0.527279             0.532356
2  DecisionTreeClassifier  0.527794             0.470968
```

Two models with their respective hyperparameter spaces were defined:

- Logistic Regression: regularization strength, penalty type, iteration limits, convergence tolerance, and class weighting
- Decision Tree: maximum depth and feature selection method

Optimized models result were as follows:

```
Training Results:
                    model  f1_score  false_negative_rate  \
0      LogisticRegression  0.629068             0.347180
1  DecisionTreeClassifier  0.623618             0.452587

                                        best_params
0  {'C': 3.640786167835674, 'class_weight': 'bala...
1          {'max_depth': 12, 'max_features': 'log2'}

Testing Results:
                    model  f1_score  false_negative_rate  \
0      LogisticRegression  0.622285             0.358553
1  DecisionTreeClassifier  0.561800             0.509482

                                        best_params
0  {'C': 3.640786167835674, 'class_weight': 'bala...
1          {'max_depth': 12, 'max_features': 'log2'}
```

5.4 Integrated use of All Optimization Tools

In this trial, all three optimizing tools from section 5.1 to 5.3 were applied concomitantly.

Prediction results with full unscaled data set and untuned models:

```
Training Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.608979             0.420412
1      LogisticRegression  0.539422             0.514413
2  DecisionTreeClassifier  1.000000             0.000000

Testing Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.610648             0.420332
1      LogisticRegression  0.527279             0.532356
2  DecisionTreeClassifier  0.527794             0.470968
```

Prediction results with optimized, scaled data set and tuned models:

```python
# Define models and hyperparameters
models = [
    {'model': LogisticRegression(),
    'params': {'C': (0.1, 10),
               'penalty': ['l2'],  # Change to l2 or none
               'max_iter': (100, 1000),
               'tol': (0.0001, 0.1),
               'class_weight': ['balanced', None]}},
    {'model': DecisionTreeClassifier(),
    'params': {'max_depth': (3, 30),
               'max_features': ['sqrt', 'log2'],
               }}
]
```

```
Training Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.608979             0.420412
1      LogisticRegression  0.539422             0.514413
2  DecisionTreeClassifier  1.000000             0.000000

Testing Results:
                    model  f1_score  false_negative_rate
0              GaussianNB  0.610648             0.420332
1      LogisticRegression  0.527279             0.532356
2  DecisionTreeClassifier  0.535721             0.459042
```

The decision tree draw and top 10 most important features selected from each model:

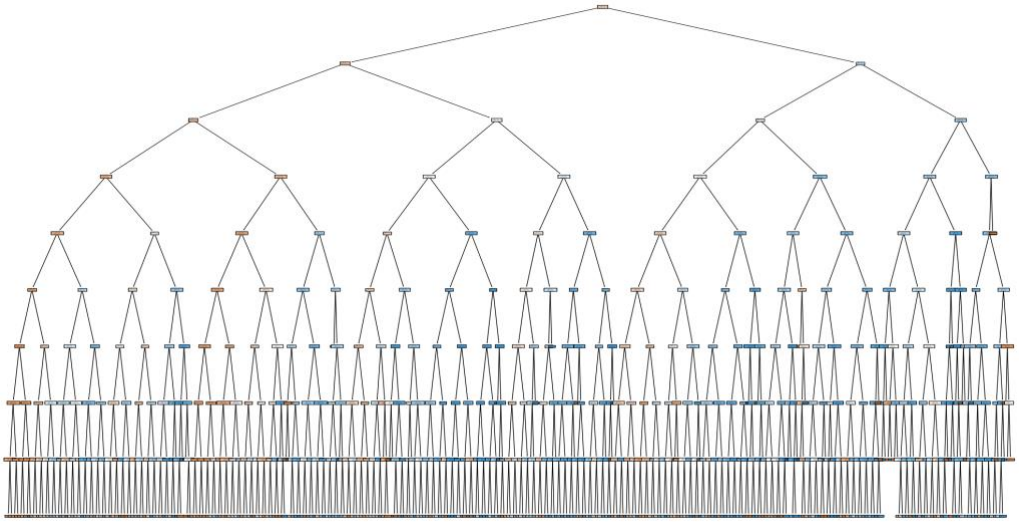


Fig 5.1 Decision Tree obtained

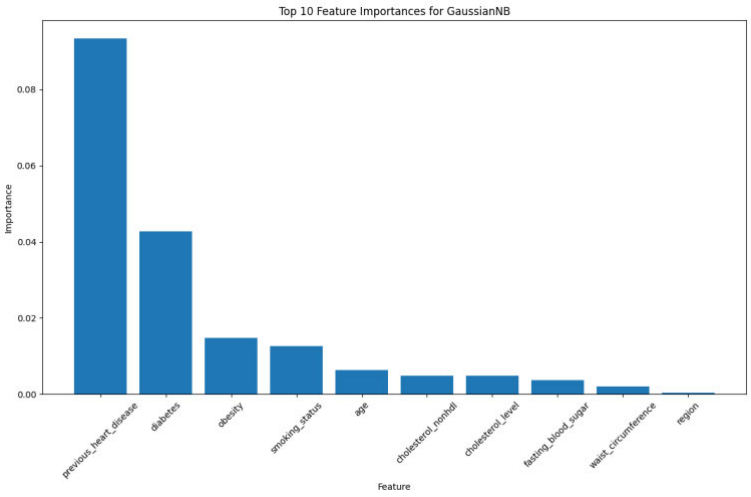


Fig 5.2 Predictive factors ranked by Naïve Bayes

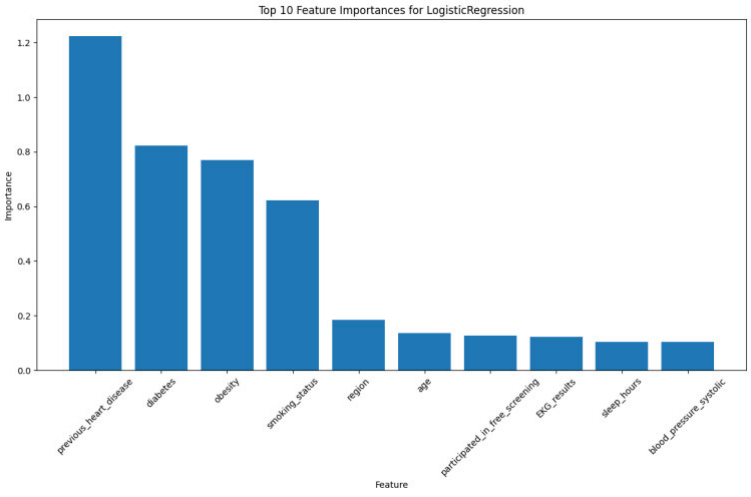


Fig 5.3 Predictive factors ranked by Logistic Regression

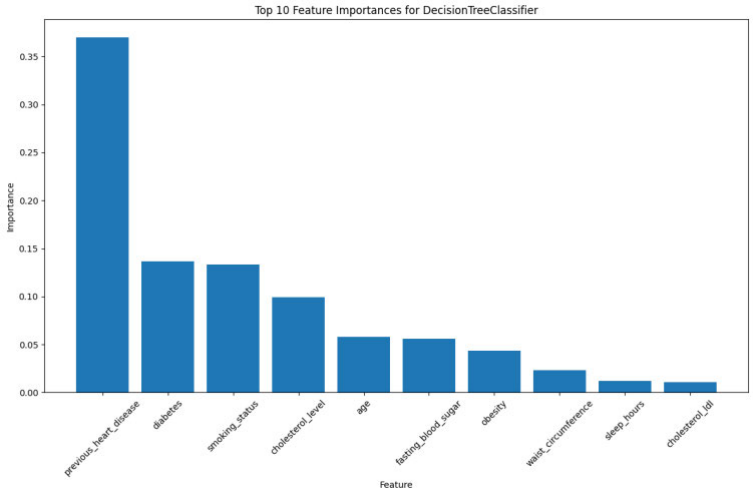


Fig 5.3 Predictive factors ranked by Decision Tree

All three models confirmed that the history of previous heart disease followed by diabetes was strongly associated with the occurrence of myocardial infarction. Interestingly, the decision tree model disagreed with the naïve bayes and logistic regression models in the third place. The tree model put obesity in the seventh place while naïve bayes and logistic regression placed it in in the third. All three models picked smoking status as the next feature to follow, further confirming our observation in the exploratory data analysis. The history of previous heart disease was a non-modifiable risk factor, yet diabetes, obesity and smoking were all risk factors that could be eliminated with preventive lifestyle modifications. These factors also echoed with the health advice provided by the Center for Health Protection.

Apart from these factors, the variables that followed such as age, region and cholesterol levels played a significantly lesser role in terms of the risk of MI. It would be worthy to further look into the relationship between other physiological measurements and MI.

6. Conclusion

In this project, we systematically applied machine learning methodologies to a real-world myocardial infarction dataset with an aim to identify key risk factors to myocardial infarction and performed ASCVD risk stratification. The three machine learning models, Gaussian Naïve Bayes, Logistic Regression and Decision Tree were used. Besides normal training procedure, tools of data preprocessing, feature selection, and model optimization were employed and successfully improved the predictive performance. Logistic Regression showed notable improvement after the application of data preprocessing and parameter optimization techniques. The analysis highlighted previous heart disease, diabetes, obesity and smoking as the top risk factors and confirmed the importance of lifestyle modifications considering these were the top risk factors.

7. Limitations

7.1 Timeframe of the data

Firstly, the dataset is a cross-sectional data certain timepoint. In real-life, the outcome of the risk factors were measured in a unit of 10-years, as the effects of risk factors might be detectable after a long period. The quality of evidence would be higher if the data were from a longitudinal study instead of a cross-sectional study.

7.2 Racial Difference

The dataset were from Indonesia and the racial or genetic difference between Chinese and Southeast Asians might pose an effect on difference susceptibility on risk factor in myocardial infarction. The difference in food consumption or healthcare system might also play a role in the prediction. Owing to the requirement of this project (>= 20 variables with >= 1000 observations), we were unable to identify a dataset that satisfies this requirement and had a similar descent as Chinese. In the future, we will relaxing the criteria and pick a data from a population with closer descent.

7.3 Incomplete Data in Alcohol Consumption

In the column *alcohol_consumption,* only 63507 out of 158355 (40%) observations contained non-null value. It was abnormal as the data in other columns were complete. Alcohol_consumption was an ordinal variable with "high" and "moderate" results. We were unable to know whether the missing values were all "low" in consumption or were null values. The potential next step would be performing an imputation, for example a logistic regression of alcohol consumption against other variables.

8. Appendix

See the attached data set and Jupyter Notebooks for the codes and exploratory results.

9. Reference

Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., Himmelfarb, C. D., Khera, A., Lloyd-Jones, D., & McEvoy, J. W. (2019). 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Journal of the American College of cardiology*, *74*(10), e177-e232.

Ho, D. R. (2022). About Heart Attack. In C. f. H. Protection (Ed.).

Rittiphairoj, T., Bulstra, C., Ruampatana, C., Stavridou, M., Grewal, S., Reddy, C. L., & Atun, R. (2025). The economic burden of ischaemic heart diseases on health systems: a systematic review. *BMJ Global Health*, *10*(2).

Wilhelmsson, C., Elmfeldt, D., Vedin, J., Tibblin, G., & Wilhelmsen, L. (1975). Smoking and myocardial infarction. *The Lancet*, *305*(7904), 415-420.