Individual Coursework Submission Form

Specialist Masters Programme

Surname: Gehlot	**First Name: Purvi**
MSc in: Business Analytics	**Student ID number: 240025662**
Module Code: SMM636	
Module Title: Machine Learning	
Lecturer: Dr Rui Zhu	**Submission Date: 24/03/2025**

Declaration:
By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission: | **Final Mark:** **%**

Introduction

Coronary Heart Disease (CHD) is a major health issue, especially in high-risk regions. This study aims to predict CHD occurrence in males using machine learning models based on clinical and lifestyle features. The dataset includes 462 observations with 9 predictors. CHD is treated as a binary classification problem (0 = No CHD, 1 = CHD). The study follows a structured approach:

1. Exploratory Data Analysis

1.1 Data Summary

Initial inspection showed the dataset had no missing values. The categorical variable famhist was marked for binary encoding, and Tobacco and Alcohol displayed skewed distributions requiring transformation. It also revealed class imbalance in target variable.

1.2 Skewness Analysis

To evaluate the distribution of each continuous variable, histograms were plotted.

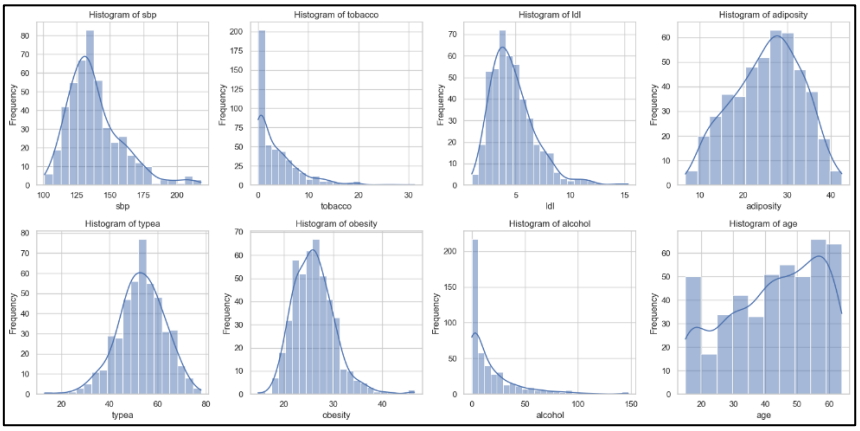

Figure (1): Histograms for continuous variables

Both Tobacco and Alcohol were found to be highly right-skewed, indicating the presence of extreme values in the dataset. Therefore, logarithmic transformation was planned for these features during preprocessing to normalize their distribution. Adiposity, Type-A, SBP and Obesity follow a normal distribution and require no transformation.

1.3 Outlier Detection

To identify potential outliers, boxplots were created for all numerical variables.

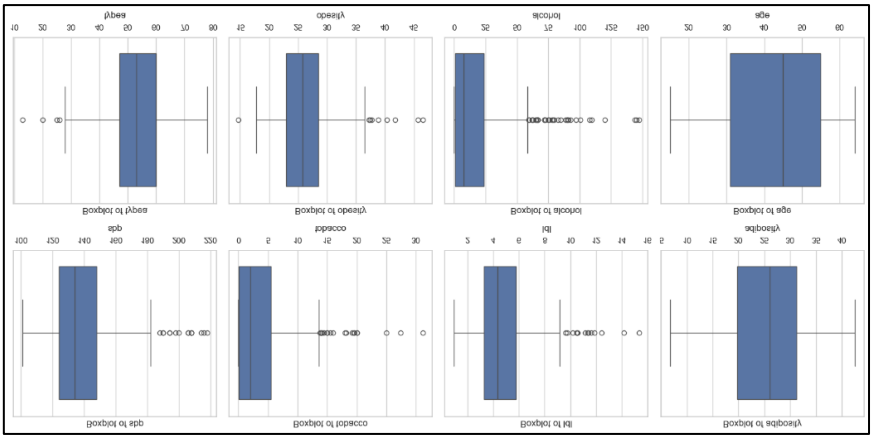

Figure (2): Boxplots for continuous variables

Outliers were clearly visible in Tobacco, LDL, Alcohol, SBP and Obesity. These extreme values were considered medically significant and possibly indicative of high-risk individuals within the dataset.

Unlike conventional methods that remove outliers, this analysis retained all extreme values due to the nature of healthcare data—where outliers often reflect real medical conditions rather than noise. Removing them could exclude high-risk individuals crucial to CHD prediction. To minimize their impact while preserving them, standardization using StandardScaler was applied to bring all features to a comparable scale.

1.4 Correlation Analysis

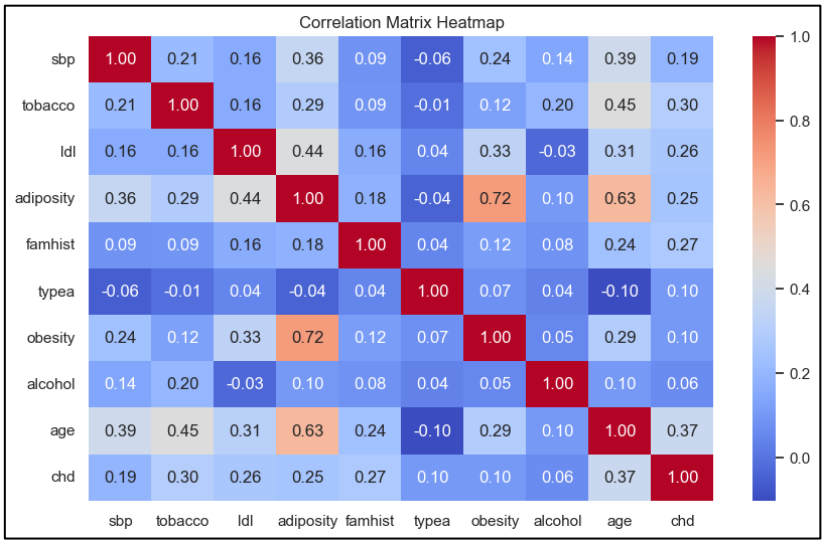

Figure (3): Correlation Matrix

Age, Tobacco, LDL, and Famhist showed strong correlation with CHD, making them key predictors. While Adiposity and Obesity were highly correlated, both were retained due to their distinct clinical roles—Obesity reflects overall body mass, while Adiposity captures visceral fat, more directly linked to CHD. Their inclusion allows the model to better differentiate risk based on fat distribution, supporting clinical relevance and prediction accuracy.

2. Data Preprocessing and Model Building

Following EDA, preprocessing was performed to ensure the data met the assumptions of Logistic Regression with Ridge Penalty and to maximize model accuracy and generalizability.

2.1 Encoding Categorical Variables: The dataset contained a categorical feature—famhist—with two categories: "Absent" and "Present". As ML models cannot process categorical string values directly, this feature was encoded into numerical format using binary mapping: “Absent”- 0 and “Present”- 1.

2.2 Log Transformation: Tobacco and Alcohol were highly right-skewed, so a log transformation was applied to normalize their distributions, reduce the effect of extreme values, and improve logistic regression’s stability and interpretability.

2.3 Feature Scaling: Logistic regression is sensitive to feature scale, especially with Ridge regularization, so the data was standardized—centering features at mean 0 and scaling to unit variance—to prevent variables with larger ranges (like sbp) from dominating model learning.

Note: *A baseline Logistic Regression model was first evaluated as a benchmark. Class imbalance was addressed using balanced class weights, after which the tuned model slightly reduced accuracy (69% vs 71%) but greatly improved CHD recall (0.81 vs 0.58), making it more reliable for detecting high-risk patients where minimizing false negatives is crucial. Thus, the tuned model is preferred.*

2.4 Model Training-The dataset was split with balanced CHD representation. A Grid Search with 5-fold Cross-Validation tuned the regularization parameter C, balancing bias and

variance. The optimal C = 1.0 provided the best performance, helping prevent overfitting while preserving key patterns.

2.5 Model Evaluation: Once trained, the model was evaluated on the test set using several performance metrics to assess classification quality, discrimination ability, and error types.

a. **Accuracy and ROC-AUC- Test Accuracy:** 69.0% – The model correctly classified roughly 69 out of every 100 cases in the test data. **ROC-AUC Score:** 0.77 – This indicates good discriminative ability, with the model effectively distinguishing between CHD and non-CHD cases across a range of thresholds.

b. **Confusion Matrix**

Actual/Predicted	No CHD (0)	CHD (1)
No CHD (0)	57 (TN)	34 (FP)
CHD (1)	9 (FN)	39 (TP)

Table (1): Confusion Matrix

TN = 57: Correctly identified as not having CHD
FP = 34: Incorrectly flagged as CHD
FN = 9: Missed CHD cases
TP = 39: Correctly identified CHD patients

c. **Classification Report**

Class 0 (No CHD): Precision (86%) – High confidence in predicting non-CHD cases. Recall (63%) – Some actual non-CHD cases misclassified. F1-Score (73%) – Balanced performance on non-CHD class

Class 1 (CHD): Precision (53%) – Moderate; some over-prediction of CHD**.** Recall (81%) – Strong ability to detect true CHD patients**.** F1-Score (64%) – Reflects a solid trade-off between sensitivity and precision

Interpretation: The model is highly effective in detecting CHD cases, prioritizing recall, which is essential to avoid missing true patients. The moderate precision is an acceptable compromise in medical settings.

d. ROC and Precision-Recall Curve Analysis

ROC Curve: It shows that the model reliably differentiates between CHD and non-CHD cases across varying thresholds. AUC of 0.77 confirms strong overall classification quality, especially valuable for risk prediction tasks in clinical settings.

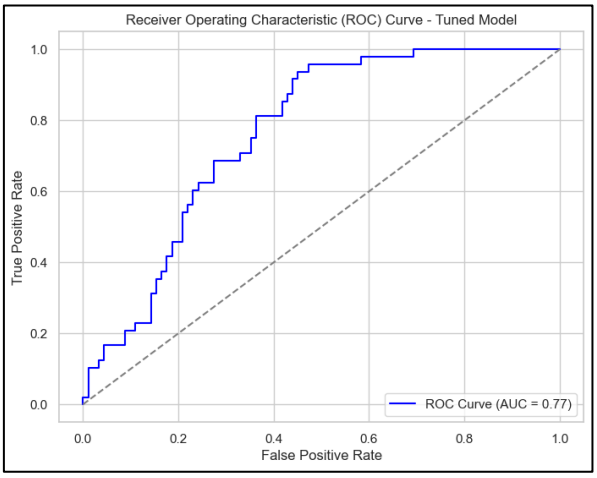

Figure (4) ROC Curve

Precision-Recall Curve: Focused on the CHD class, the curve shows high precision at low recall, decreasing as recall increases. This reflects the clinical trade-off—prioritizing recall to reduce false negatives, which is crucial in CHD diagnosis where missing true cases poses greater risk than over-predicting.

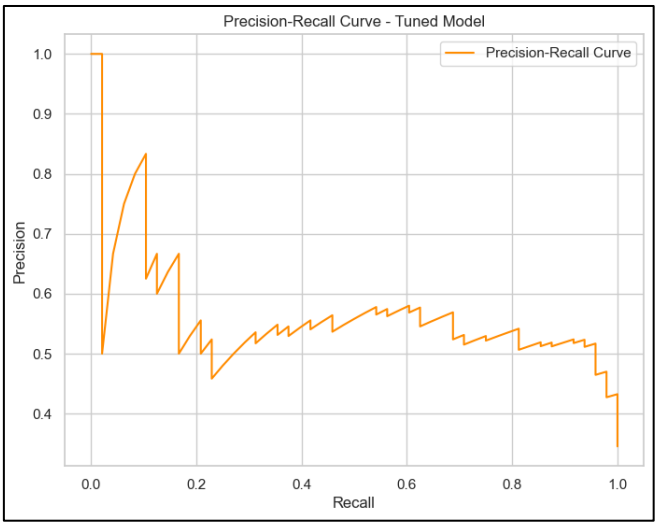

Figure (5): Precision Recall Curve

e. Coefficient Analysis

Logistic regression revealed key CHD predictors: Age (+0.737), Famhist (+0.624), LDL (+0.533), and Tobacco (+0.386) showed strong positive associations with CHD risk. Adiposity (−0.286) and Alcohol (−0.037) had weaker negative effects. These coefficients align with medical literature, reinforcing the model's interpretability and clinical value.

3. Selecting Best Classifier

Naïve Bayes was the most clinically suitable model for CHD prediction, achieving the highest recall (0.78), correctly identifying 78% of CHD cases—critical in healthcare where missing true cases can lead to serious outcomes.

Metric	Value
Test Accuracy	72.04 %
CHD Precision (1)	0.57
ROC-AUC Score	0.7597

Table (2): Model Performance Matrix

Confusion Matrix: Only 7 out of 32 CHD patients were misclassified as healthy, indicating strong sensitivity.

This model strikes an effective balance between sensitivity (recall) and overall discrimination power (ROC-AUC), making it well-suited for early CHD risk detection. Although the precision for CHD is moderate (0.57), this trade-off is clinically acceptable when the priority is not to miss any high-risk patients.

Naïve Bayes further proves to be: Efficient on small datasets (like the current 462 observations), resilient to overfitting, computationally lightweight, and effective despite its independence assumption, which worked well in this medical context.

Given CHD's severity and the need for early intervention, maximizing recall is more important than achieving perfect precision. Therefore, Naïve Bayes is the most reliable model for flagging at-risk individuals, ensuring critical cases are not overlooked.

Appendix

Baseline Logistic Regression Results

Metric	Value	Interpretation
Accuracy	71%	The model correctly classified 71% of all test samples.
ROC-AUC	0.76	Indicates good ability to distinguish between CHD and non-CHD cases.
Recall (CHD- Class 1)	0.58	The model identified 58% of actual CHD cases.
Precision (CHD-Class 1)	0.57	57% of predicted CHD cases were correct.
F1-Score (CHD-Class 1)	0.58	Balanced trade-off between precision and recall.

Table (A1): Baseline logistic regression results

Classifier Comparison

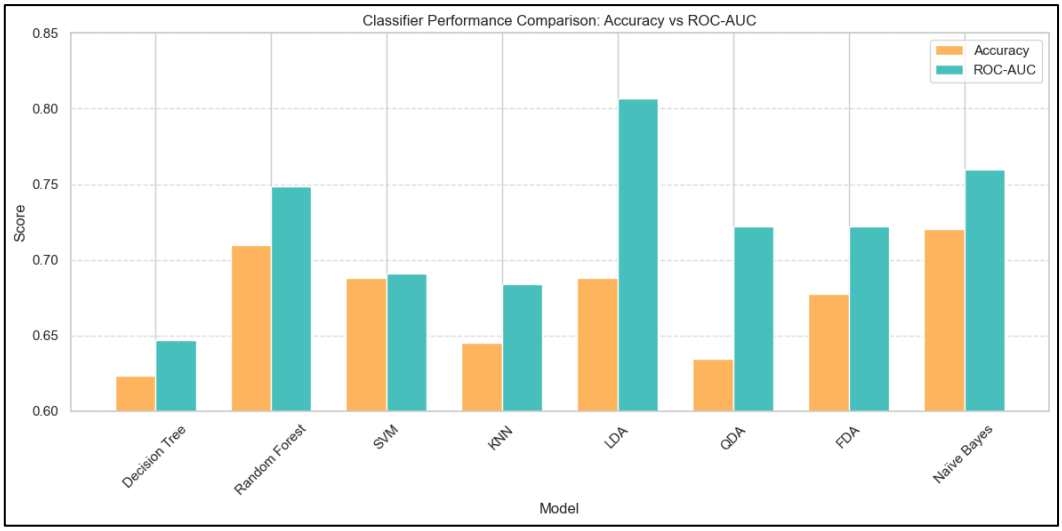

Figure (A1): Classifier Performance Comparison

Naïve Bayes offers the best overall balance of accuracy and ROC-AUC.

LDA shows excellent class separation but slightly lower accuracy.

Random Forest performs reliably across both metrics.

Decision Tree, KNN, and **SVM** show weaker performance.

FDA and **QDA** are moderate but don't outperform Naïve Bayes.