BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Individual Coursework Submission Form

Specialist Masters Programme

Surname: Kiefer	**First Name: Marius**
MSc in: Business Analytics	**Student ID number: 240052822**
Module Code: SMM636	
Module Title: Machine Learning	
Lecturer: Dr Rui Zhu	**Submission Date: 24/03/2024**

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:** %

Individual Coursework

SMM636 Machine Learning

Marius Kiefer

March 2025

Words: 1038

Introduction

In this report, I analyze a heart disease dataset to predict coronary heart disease (CHD: 1/0) among males in a high-risk region of the Western Cape, South Africa. The dataset comprises nine features reflecting various aspects of an individual's health and lifestyle, and my goal is to develop machine learning models that accurately classify disease status based on the available data.

Exploratory Data Analysis

Before selecting machine learning classifiers, it is essential to examine the nature of the dataset. Since different models rely on varying data assumptions, I conduct an Exploratory Data Analysis (EDA) to evaluate key attributes such as missing values, feature distributions, class balance, normality of features, the correlation of features with coronary heart disease (chd), inter-feature correlations, and outlier detection.

The dataset is complete with no missing values, and an examination of the mean, median, maximum, and minimum values for each feature reveals no anomalies. **Figure 1** illustrates an imbalance in the target feature, with class 0 (Absent) comprising 65% of the data and class 1 (Present) accounting for 35%. Inspection of the feature distributions (see **Figure 2**) reveals that most features are skewed. In particular, the features sbp, tobacco, idl, and alcohol exhibit a pronounced left-skew, with a heavy concentration of values on the left side of the x-axis. An additional noteworthy finding is the disparity in the age distributions between classes 0 and 1. For class 0, the age values are fairly evenly distributed, whereas for class 1 there is a clear inclination towards older ages. The remaining features show similar distribution patterns across both classes. Furthermore, **Figure 3** employs Q-Q plots to illustrate that the majority of the variables deviate from a normal distribution. The data suggests that certain features are interrelated (see **Figure 4**). I compute the Variance Inflation Factor (VIF) for each predictor and the results support the presence of multicollinearity. The feature correlating the most with the target feature is age (see **Figure 5**). Finally, I detect outliers in the dataset using the interquartile range (IQR) method (see **Figure 6**).

In summary, the dataset, which consists of 462 observations, is characterized by imbalanced classes, significant multicollinearity, non-normally distributed features, and the presence of outliers.

Methodological Framework

Before fitting any models, I partition the data into training and test sets using a 90/10 split to maximize the data available for training given the limited observations. I then apply stratified 10-fold cross-validation on the training set to preserve class distribution and ensure robust evaluation while mitigating overfitting. I choose to keep outliers because of the limited amount of observations and focus on models that handle this characteristic well.

I develop my classification models using pipelines that incorporate only the necessary data adjustments, promoting simplicity and reproducibility, especially when scaling to larger datasets. For logistic regression (LR), I apply L2 regularization and explore a range of C values to balance model complexity, using solvers such as liblinear and newton-cholesky for their efficiency with small- to medium-sized datasets. StandardScaler ensures feature stability, and PCA is used to address multicollinearity by reducing redundant information.

Besides LR, I select RandomForest (RF) and SVM to remain as close as possible to the original data, as they require minimal transformation and effectively capture its inherent characteristics. Additionally, I implement LDA as a generative method to leverage fine-tuning of its covariance structure, with the goal of further improving performance.

For RandomForest, I develop a pipeline that preserves the natural data structure with minimal transformation. This non-parametric, ensemble-based method requires no scaling or distributional assumptions, making it inherently robust to outliers and multicollinearity. By aggregating multiple decision trees, RandomForest effectively captures complex interactions within the data while mitigating noise, which is particularly valuable given the dataset's irregularities.

In contrast, SVM is more sensitive to feature scaling and class imbalance. To address these issues, the SVM pipeline incorporates StandardScaler to standardize the features and SMOTE to balance the classes. Moreover, SVM's ability to leverage various kernel functions (such as linear, polynomial, RBF, and sigmoid) enables it to adapt to non-linear relationships present in the data, offering a flexible approach to modeling complex decision boundaries.

Lastly, I implement a pipeline for LDA. I choose this generative method because it models the covariance structure among predictors and provides extensive fine-tuning capabilities through shrinkage (using the lsqr solver) and PCA optimization to control multicollinearity. While Naive Bayes relies on the simplistic assumption of feature independence and QDA's flexibility can lead to overfitting by estimating separate covariance matrices for each class, LDA offers a balanced approach that allows for more precise control of model complexity and ultimately achieves robust performance for CHD classification.

Evaluation Metrics

In evaluating my models, I focus on accuracy, sensitivity, and ROC AUC, as these metrics provide critical insights for this task. Given that a majority class classifier (one that always predicts the negative class) would achieve an accuracy of 0.65, with zero precision, sensitivity, and F1 score, and an ROC AUC of 0.5, these baselines highlight the challenge of detecting true positive cases. Moreover, both a random classifier and a classifier that predicts outcomes based solely on class distribution yield an accuracy of 0.545, with precision, recall, and F1 scores around 0.35 and an ROC AUC of 0.5. These figures underscore that the fitted models must considerably outperform these baselines, especially in terms of sensitivity, to be clinically useful.

Figure 7 displays the ROC curves of the selected models, and **Table 1** summarizes their performance metrics. My LR model, for instance, achieved an accuracy of 0.77, sensitivity of 0.69, and an ROC AUC of 0.84, clearly surpassing the baseline performance. In contrast, RF recorded an accuracy of 0.64, sensitivity of 0.31, and an ROC AUC of 0.71, indicating its limited ability to detect positive cases reliably. The SVM model demonstrates improved performance with 0.70 accuracy, 0.81 sensitivity, and an ROC AUC of 0.87, while the LDA model matches LR in accuracy (0.77) and achieves an even higher sensitivity of 0.81 along with an ROC AUC of 0.86. These results justify my focus on models such as SVM and LDA, as their superior sensitivity ensures that true CHD cases are detected. However, while these results may enhance efficiency, they are not reliable enough for a diagnosis without a doctor's supervision.

Figures and Tables

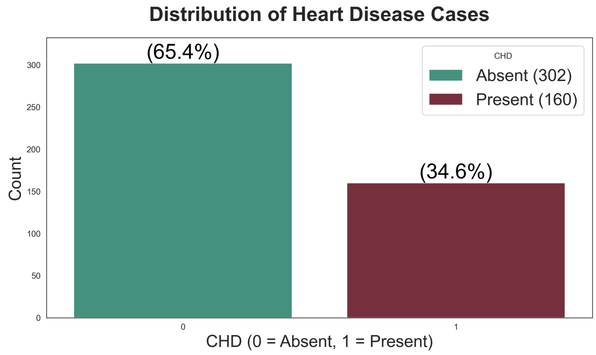

Figure 1: Class Distribution

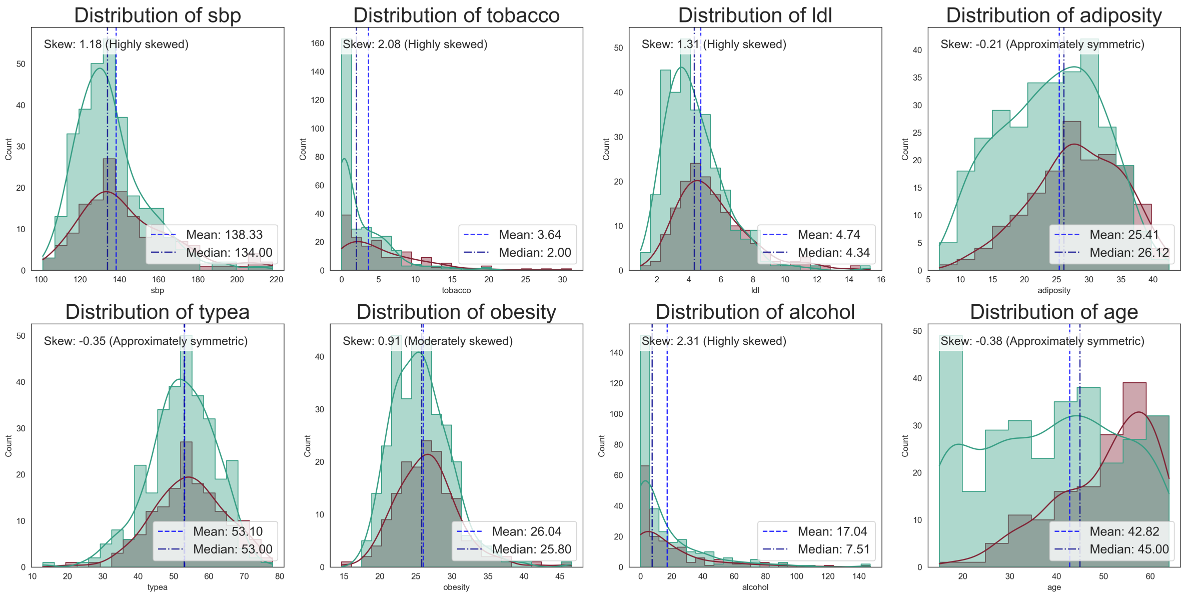

Figure 2: Feature Distribution

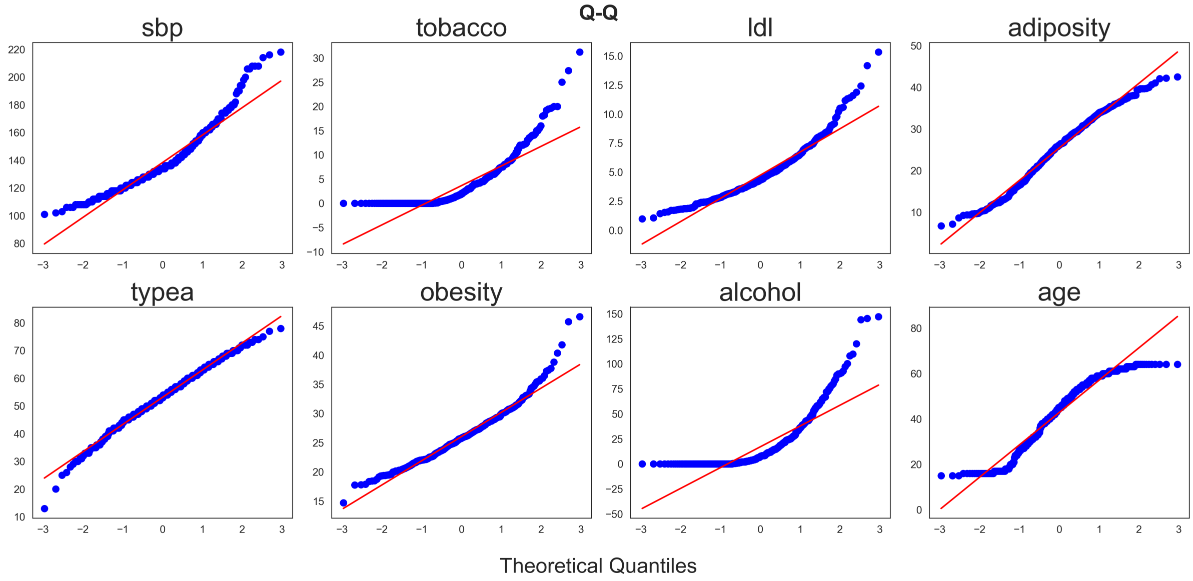

Figure 3: Proof of Normality

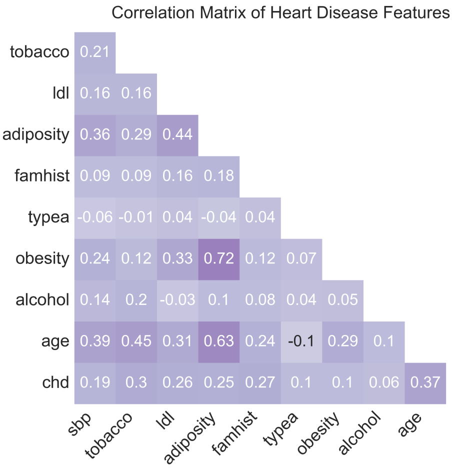

Figure 4: Correlation Matrix of Features

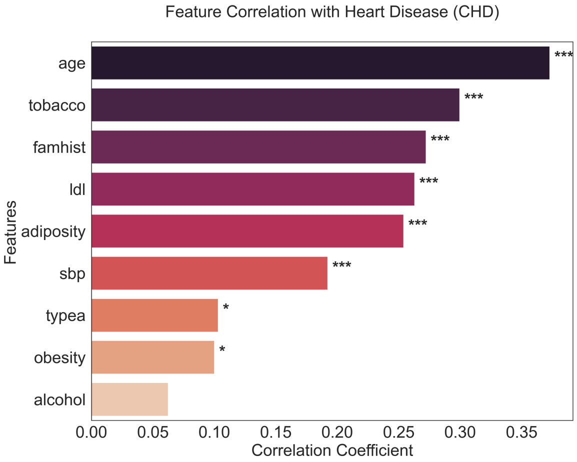

Figure 5: Feature Correlation with Target Feature

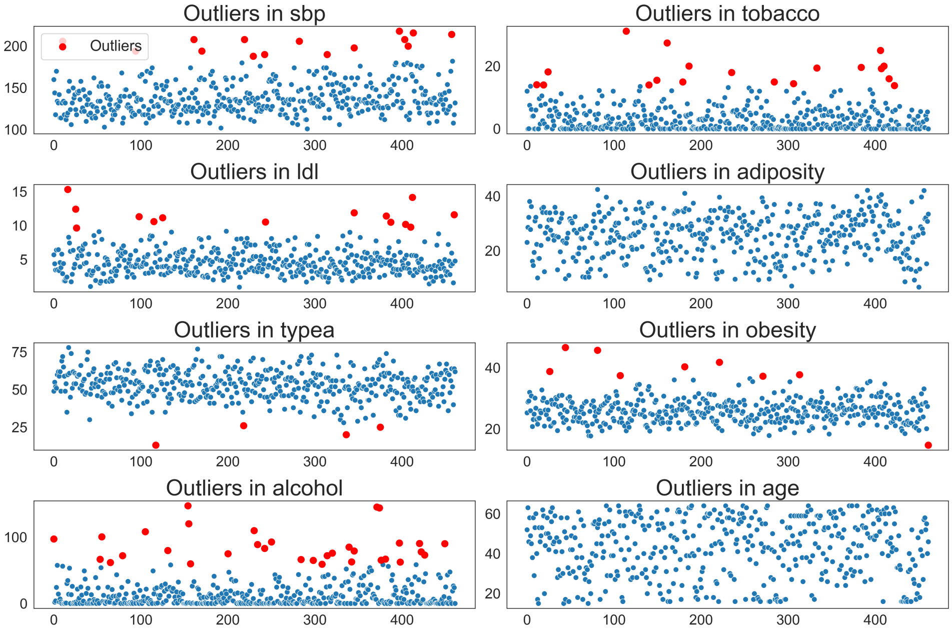

Figure 6: Outlier Detection

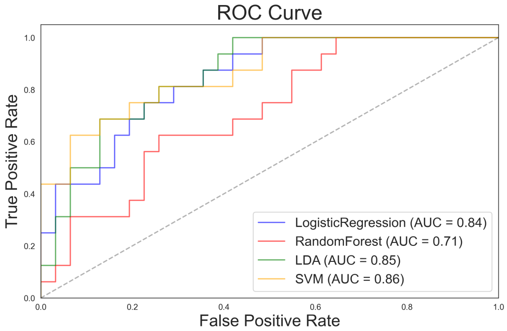

Figure 7: ROC Curve

Model	Acc.	Prec.	Rec.	F1	AUC	Sens.	Spec.
LR	0.77	0.65	0.69	0.67	0.84	0.69	0.81
RF	0.64	0.46	0.31	0.37	0.71	0.31	0.81
SVM	0.70	0.54	0.81	0.65	0.87	0.81	0.65
LDA	0.77	0.62	0.81	0.70	0.86	0.81	0.74

Table 1: Test Set Performance Comparison for CHD Classification