Individual Coursework Submission Form

Specialist Masters Programme

Surname: Rawat	**First Name: Nikhil**
MSc in: Business Analytics	**Student ID number: 230044335**
Module Code: SMM636	
Module Title: Machine Learning	
Lecturer: Dr Rui Zhu	**Submission Date: 22/03/2024**

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:** **%**

Introduction

In the age of data-driven healthcare, Coronary Heart Disease (CHD) remains a principal global health challenge. Accurate early detection is paramount to reducing its mortality rate. Traditional diagnostic methods grapple with the complex interplay of medical data variables. This research uses machine learning and historical patient data to enhance CHD prediction and inform healthcare strategies.

Objective

The research aims to evaluate key machine learning models and features for precise CHD prediction.

Dataset

Dataset encompasses 462 entries and 10 features: 9 clinical and lifestyle independent features and 1 target variable for CHD prediction.

Data Pre-processing

In the initial data pre-processing phase, I checked for missing values and converted 'famhist' from text to numerical format.

Exploratory Data Analysis (EDA)

Descriptive Statistics

The summary table 1 highlights significant variability in features such as tobacco use, low-density lipoprotein (LDL) cholesterol, and alcohol consumption, as denoted by their high standard deviation.

Table 1 Statistical Characteristics

Statistics	Systolic Blood Pressure	Tobacco	LDL Cholesterol	Adiposity	Type A Behaviour	Obesity	Alcohol	Age
count	462	462	462	462	462	462	462	462
mean	138.33	3.64	4.74	25.41	53.10	26.04	17.04	42.82
std	20.50	4.59	2.07	7.78	9.82	4.21	24.48	14.61
min	101	0	0.98	6.74	13	14.7	0	15
25%	124	0.05	3.28	19.78	47.00	22.99	0.51	31
50%	134	2.00	4.34	26.12	53.00	25.81	7.51	45
75%	148	5.50	5.79	31.23	60.00	28.50	23.89	55
max	218	31.20	15.33	42.49	78.00	46.58	147.19	64

Family History and CHD Events

Fig 1 analysis indicates a potential familial or genetic link to CHD risk, with a higher incidence of CHD events among individuals with a family history of the condition.

Fig 1. Family History vs CHD Event

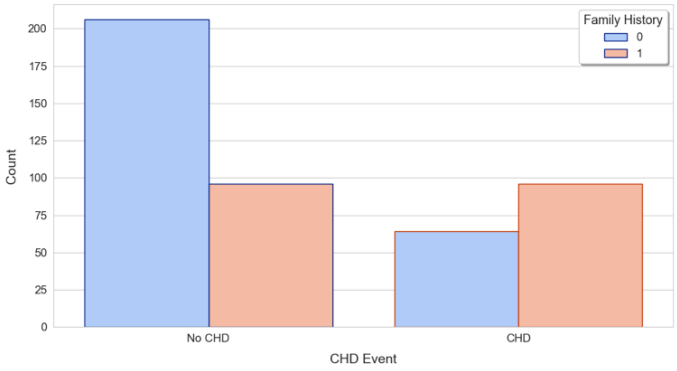
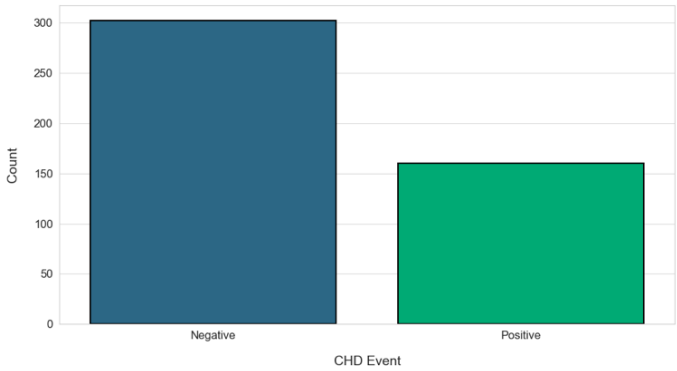


Fig 2. Frequency of CHD Events



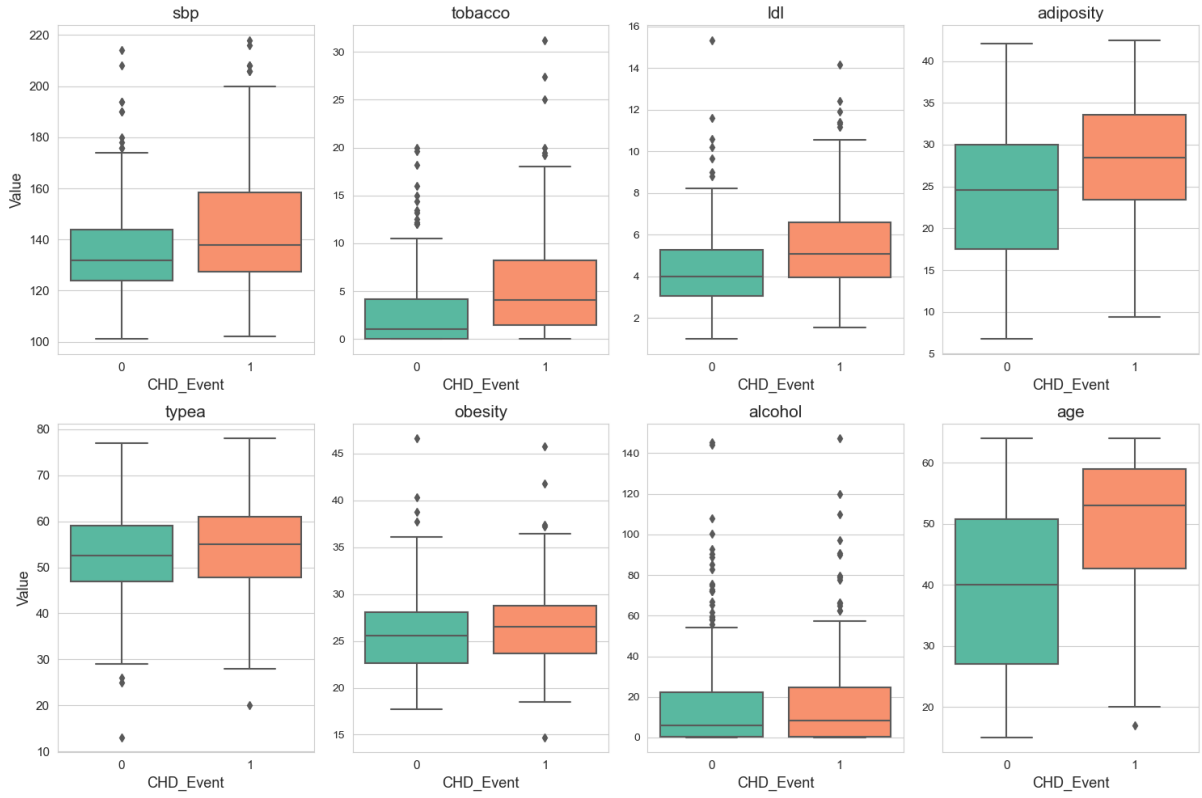
CHD Event Frequency

There is an observed class imbalance in the dataset, with a higher prevalence of individuals without CHD events, suggesting the need for specialized machine learning approaches to handle this imbalance (Fig 2).

Boxplot Analysis

Fig 3 comparative boxplot analysis shows individuals with CHD tend to have higher readings of systolic blood pressure, tobacco usage, LDL cholesterol, and other risk factors, with notable outliers in the CHD group. These outliers may indicate an increased risk of CHD and must be carefully validated to ensure they accurately inform the predictive model without skewing results.

Fig. 3. Boxplot Representing the Distribution of Data Features with respect to CHD Event



Feature Importance and Selection

Insights from Random Forest

The Random Forest model identifies age, tobacco use, LDL cholesterol, adiposity, and systolic blood pressure as crucial predictors of CHD. In contrast, family history, Type-A behaviour, and alcohol consumption show lower predictive power. (Please refer to Fig. 4 in the appendix)

Correlation Analysis

Pearson's Correlation Coefficient uncovers a strong positive correlation (0.72) between adiposity and obesity, suggesting redundant information due to multicollinearity. Additionally, age is moderately correlated with both systolic blood pressure and adiposity, hinting at the relationship between age, blood pressure, and body fat. (Please refer to Fig. 5 in the appendix)

Decision on Feature Exclusion

Given the high correlation between 'obesity' and 'adiposity', 'obesity' is excluded to avoid multicollinearity, which can obscure the distinct impact of individual variables. By keeping 'adiposity', which is more indicative of CHD according to the model, I aim to streamline the predictors and enhance model performance.

Framework

This approach utilizes a pre-processed dataset, omitting 'obesity', and integrates a machine learning pipeline for feature scaling and class imbalance correction with SMOTE variants. I refined models using GridSearchCV, then assessed them with weighted metrics: accuracy, precision, recall, F1 score, and specificity. These metrics ensure accurate CHD diagnosis, minimising false positives and negatives even in the presence of class imbalance, which is essential for effective patient care and reducing unnecessary treatments. This comprehensive evaluation framework enhances CHD prediction by addressing class imbalances and ensuring robust detection across diverse cases.

Logistic Regression with Ridge Penalty

In my research, logistic regression with a ridge (L2) penalty emerged as a significant model for CHD classification. The optimal model configuration was established using GridSearchCV, which determined the 'newton-cg' solver with an 'l2' penalty at a regularization strength of 0.01 as the most effective. This model not only performed

well across standard metrics, but also provided a balanced prediction of both CHD presence and absence, as evidenced by a cross-validation score around 0.708 and a test set accuracy of approximately 0.753. The chosen model exhibited a promising ability to generalize, an essential trait for practical applications in healthcare settings.

Classifier Exploration :

In developing a predictive model for Coronary Heart Disease (CHD), several classifiers were optimised through GridSearchCV, each tailored to match the dataset's unique aspects.

Table 2 "A Guide to Machine Learning Models for CHD Prediction"

Model Name	Description	Hyperparameters
Support Vector Machine (SVM)	Good for capturing non-linear patterns, and it can be finely tuned to balance model complexity with prediction accuracy, which is ideal for datasets with complex relationships.	Kernel: rbf, C: 14, gamma: 0.01
K-Nearest Neighbours (KNN)	Leverages similarity between data points to make predictions, making it suitable for datasets where CHD patterns are closely related to clusters of similar cases.	Metric: minkowski, Neighbors: 21, p: 2
Multilayer Perceptron (MLP) Neural Network	Can model complex, non-linear relationships and interactions between features, beneficial for datasets with many variables affecting CHD.	Layers: 128 & 64, Activation: relu, Learning rate: 0.01
Bernoulli Naive Bayes	Assumes feature independence and can be effective if the dataset features that influence CHD are independent of each other, often used for binary features.	Binarization threshold: 0.15, Alpha: 0.0005
Quadratic Discriminant Analysis (QDA)	Handles non-linear data well due to its ability to model the variance of each class, which is useful if CHD is affected by complex, non-linear feature interactions.	reg_param: 0.25
Linear Discriminant Analysis (LDA)	Good for datasets where classes are linearly separable and benefits from its efficiency in reducing feature space dimensionality, aiding in clear class separability.	Solver: lsqr, Shrinkage: 0.925

Results:

Table 3 "Classifier Efficacy in CHD Diagnosis"

Classifier	Mean of 10 fold CV	Accuracy	F1 Score	Precision	Recall	Specificity
Logistic Regression	0.7075	0.7527	0.7557	0.7630	0.7527	0.7627
Support Vector Machine (SVM)	0.6968	0.7849	0.7835	0.7828	0.7849	**0.8475***
K-Nearest Neighbours (KNN)	0.7128	0.7419	0.7419	0.7419	0.7419	0.7966
Multilayer Perceptron (MLP) Neural Network	0.7047	0.7634	0.7667	0.7763	0.7634	0.7627
Bernoulli Naive Bayes	0.6992	0.7204	0.7185	0.7173	0.7204	0.7966
Quadratic Discriminant Analysis (QDA)	**0.7209***	**0.7957***	**0.7974***	**0.8007***	**0.7957***	0.8136
Linear Discriminant Analysis (LDA)	0.6885	0.7097	0.7150	0.7479	0.7097	0.6610

In the comparative analysis of machine learning classifiers for Coronary Heart Disease(CHD) prediction, the Quadratic Discriminant Analysis (QDA) classifier stands out with stellar performance across several key metrics. According to the 10-fold cross-validation data, QDA attains the top mean score of 0.7209, an accuracy of 79.57%, and demonstrates excellent balance with a precision of 80.07%, recall of 79.57%, and the highest specificity of 81.36%. This indicates QDA's robust capability of accurately identifying both true positives and true negatives for CHD.

Support Vector Machine (SVM) classifier follows closely, boasting the highest specificity at 84.75%, a critical feature for reducing the likelihood of false positives in clinical predictions. Though SVM is less accurate than QDA, its correct identification of non-CHD patients is notable.

The Naive Bayes classifier showed a mean 10-fold cross-validation score of 0.6992 with an accuracy of 72.04%. While its performance metrics, such as precision, recall, and specificity, are not the highest compared to QDA and SVM, its results are still notable.

The Multilayer Perceptron (MLP) Neural Network and the K-Nearest Neighbours (KNN) classifiers show commendable accuracy in the range of 74-76%. Their uniform performance across precision, recall, and F1 score indicates a good foundational model that could be beneficial for datasets with complex patterns, albeit not as high-performing as QDA.

Logistic regression shows moderate 75.27% accuracy and recall, indicating fair performance. The Linear Discriminant Analysis (LDA) classifier, while exhibiting the lowest specificity at 66.10%, may still hold value, especially when its propensity for false positives can be balanced by the high specificity of a model like SVM.

Conclusion:

This analysis confirms the potential of machine learning for enhancing CHD prediction. The integration of classifiers, particularly the standout QDA and SVM models, suggests a path towards a more accurate detection system when combined. Critical to my approach is strategic feature selection, which highlights the importance of clinical indicators such as tobacco use and LDL cholesterol. This research explores the synergistic use of these classifiers and applies these findings to a broader clinical context, paving the way for more personalised and timely healthcare interventions.

Word count: 993 words

Appendix:

Fig 4. Feature Importance using Random Forest

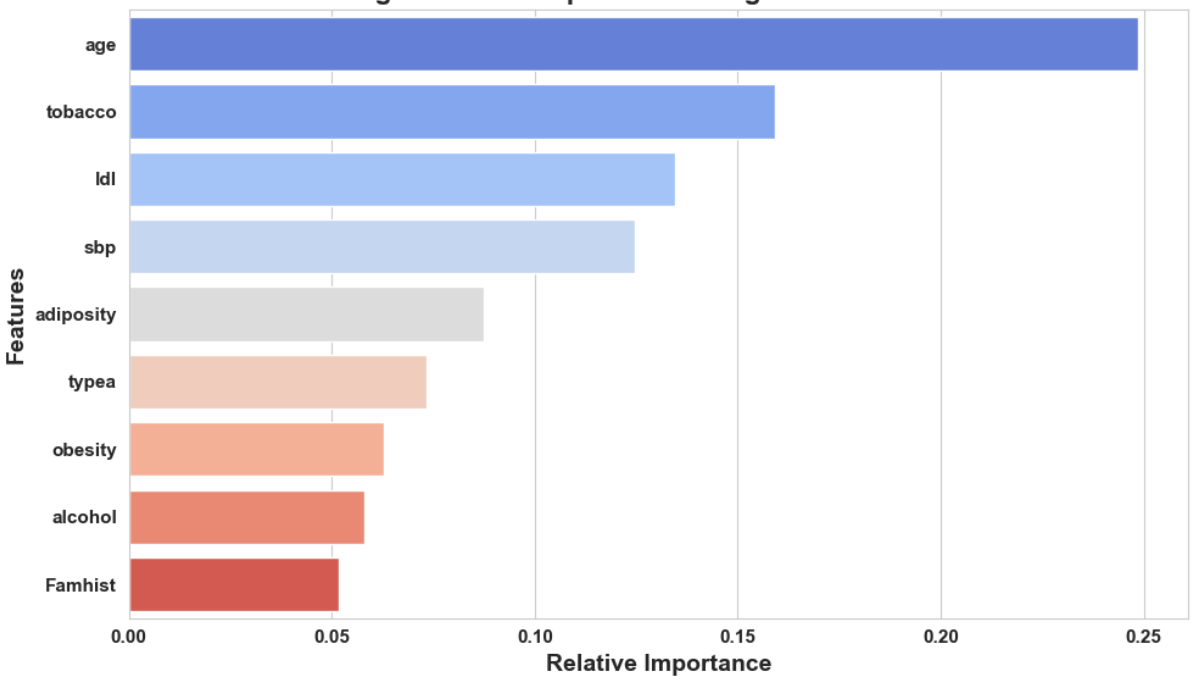


Fig 5. Pearson Correlation Coefficient Matrix