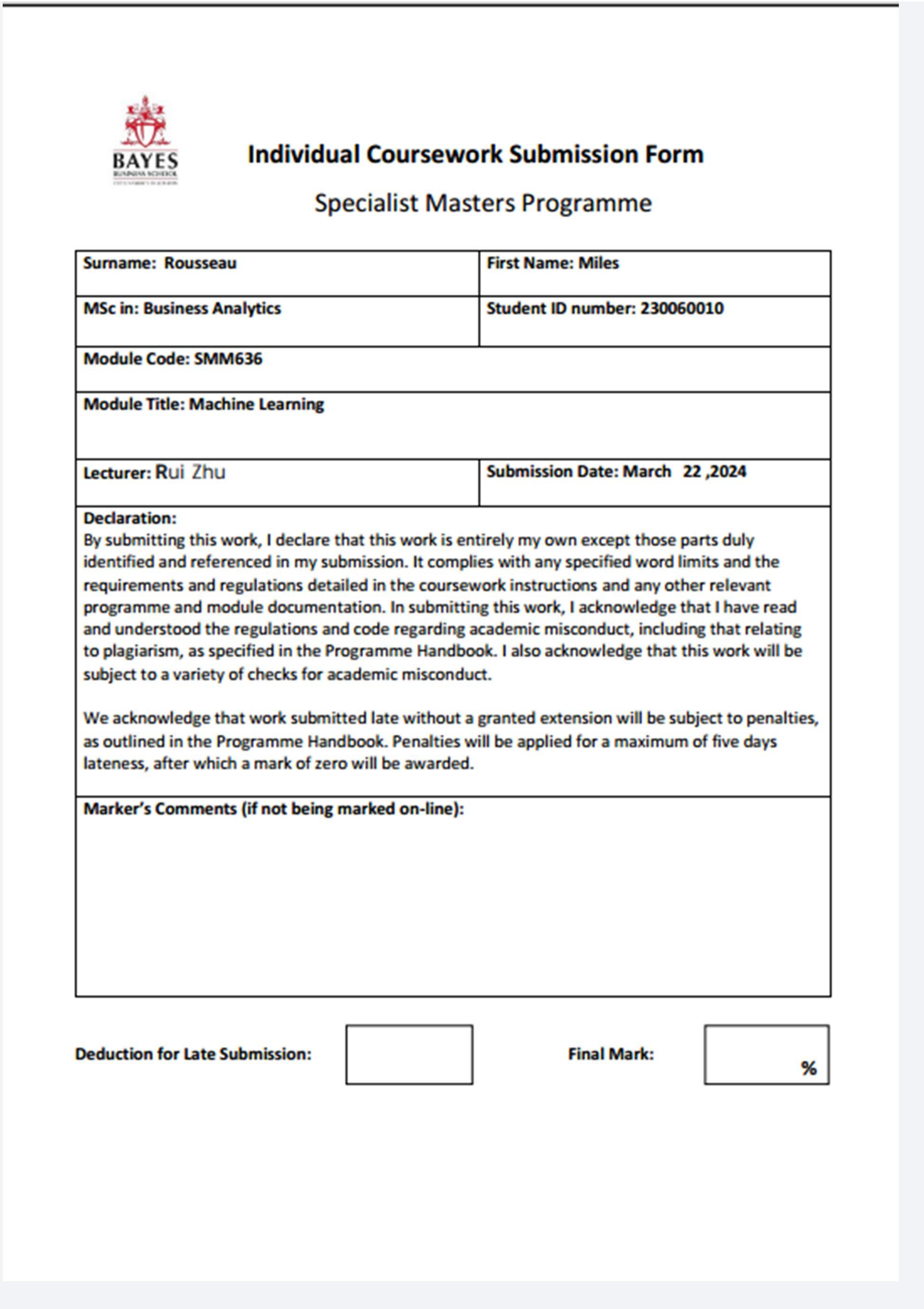


BAYES
BUSINESS SCHOOL

# Individual Coursework Submission Form

## Specialist Masters Programme

| | |
|---|---|
| **Surname: Rousseau** | **First Name: Miles** |
| **MSc in: Business Analytics** | **Student ID number: 230060010** |
| **Module Code: SMM636** | |
| **Module Title: Machine Learning** | |
| **Lecturer:** Rui Zhu | **Submission Date: March 22 ,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:** %

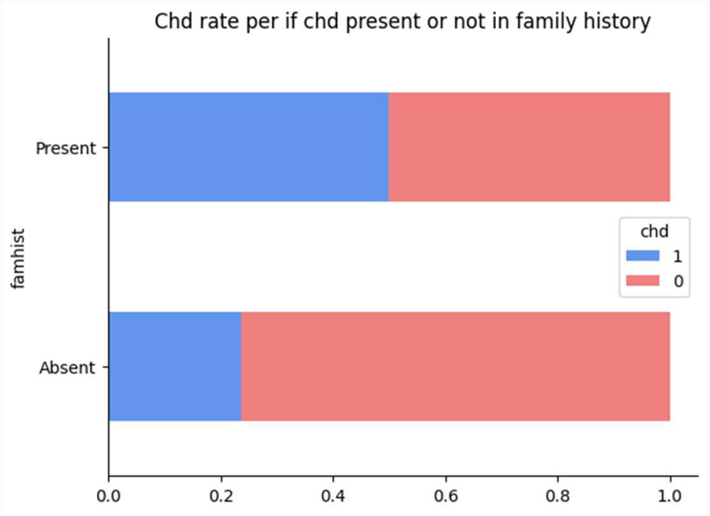
# SMM636 Machine Learning – Individual Assignment

**Introduction:**

In this assignment I will be using machine learning classifiers learnt in this module to predict coronary heart disease (chd: 1/0) for males in a heart-disease high-risk region of the Western Cape, South Africa. The dataset comprises nine features, including systolic blood pressure (sbp), tobacco usage (tobacco), low-density lipoprotein cholesterol (ldl), adiposity, family history of heart disease (famhist), type-A behavior (typea), obesity, alcohol consumption (alcohol), and age.

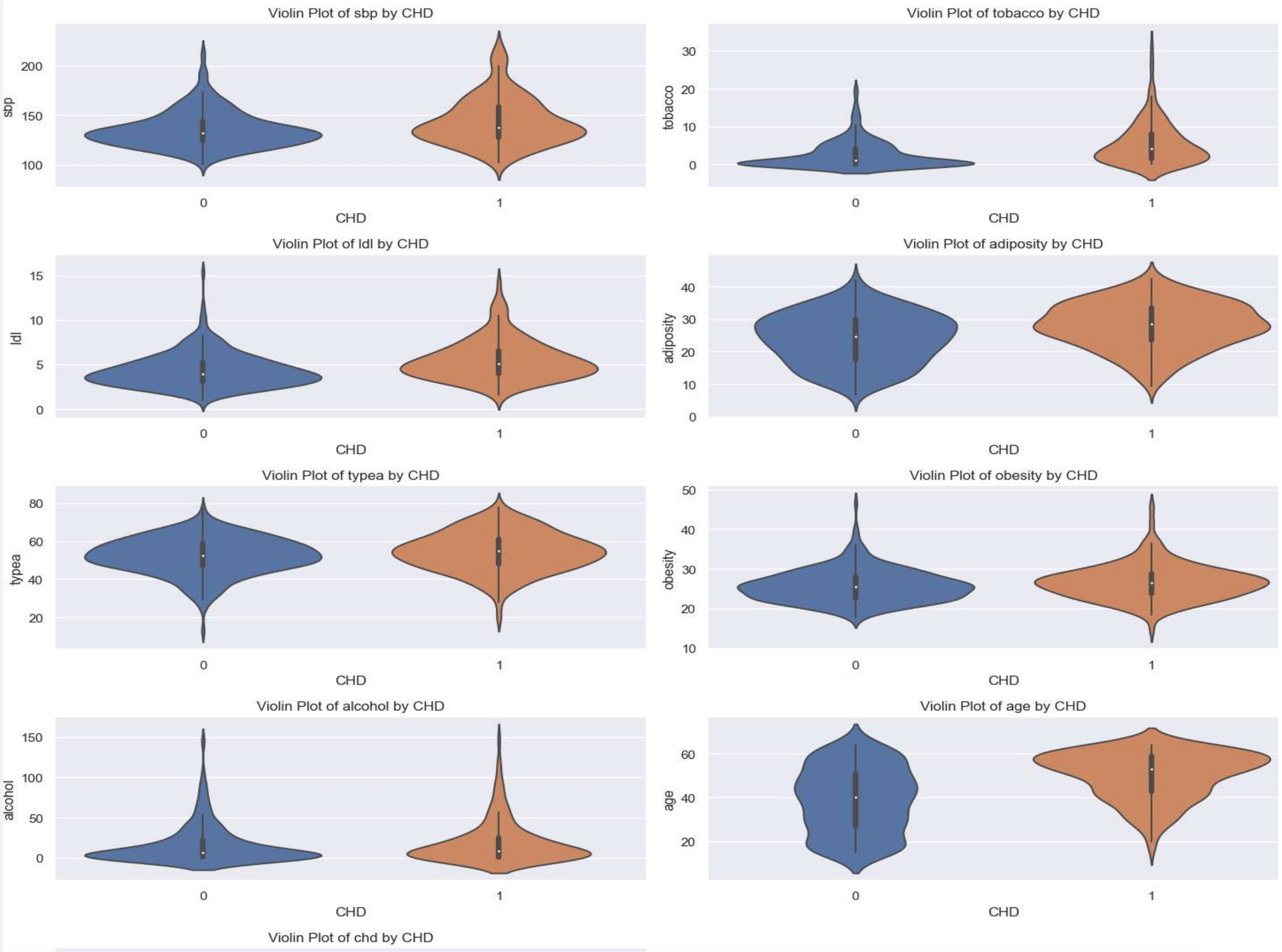
**Exploratory Data Analysis:**

The dataset provided contains 462 rows of data with 9 features and one target column(CHD). The dataset was clean with no nulls or duplicates. There was a slight imbalance in the CHD target column and most of the features had a significant number of outliers, both discoveries influenced the models selected and the scalers used on the features.



One of the most significant discoveries of the EDA process was that there was a higher rate of CHD in patients who had a family history of CHD than those that didn't. Several genetic health conditions that a baby inherits from 1 or both parents can cause congenital heart disease (NHS, 2017). Research from the NHS explains the correlation between CHD and a family history of CHD that was seen in the data.

nhs.uk. (2017). Congenital heart disease - Causes. [online] Available at: https://www.nhs.uk/conditions/congenital-heart-disease/causes/#:~:text=Several%20genetic%20health%20conditions%20that.



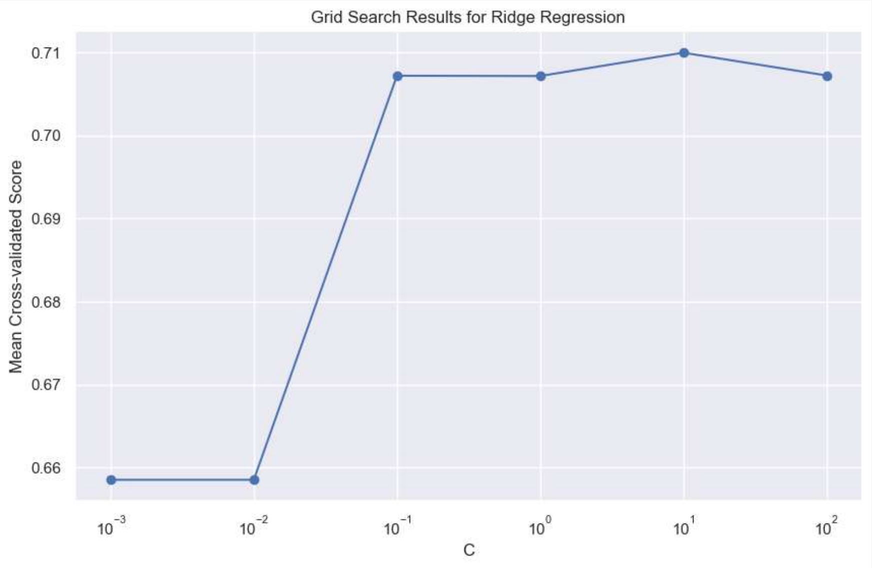
The image above uses violin plots to explore how the various features are distributed based on whether CHD is present (brown graphs) or not present (blue graphs). Age and tobacco usage graphs indicate a relationship between higher age and tobacco use with higher CHD prevalence. Sbp, ldl, adiposity, typea, obesity and alchohol only show very slight positive correlation with CHD prevalence.

**<u>Pre-Processing and scalers</u>**:

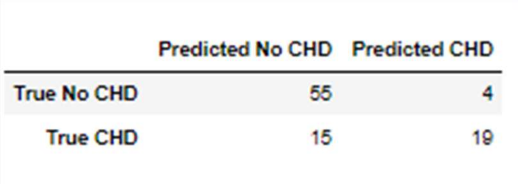
I split the data into 80% train and 20% test. Additionally, I scaled the X data using a Quantile Transformer which greatly improved test accuracy when comparing it to the Robust and Standard scaler. The quantile transformer method transforms the features to follow a uniform or normal distribution which spreads out the most frequent values. I used this method because as previously mentioned, there are many outliers in the features and the quantiles method reduces the impact of these outliers on the quality of the model produced.

## Logistic regression with ridge penalty:

To start, I fit a logistic regression with ridge penalty to classify the patients. To perform this, I used 5-fold cross validation to tune my hyperparameter C which controls the strength of regularization. The best resulting model resulted in a test accuracy of 79.5% using C equal to 10.



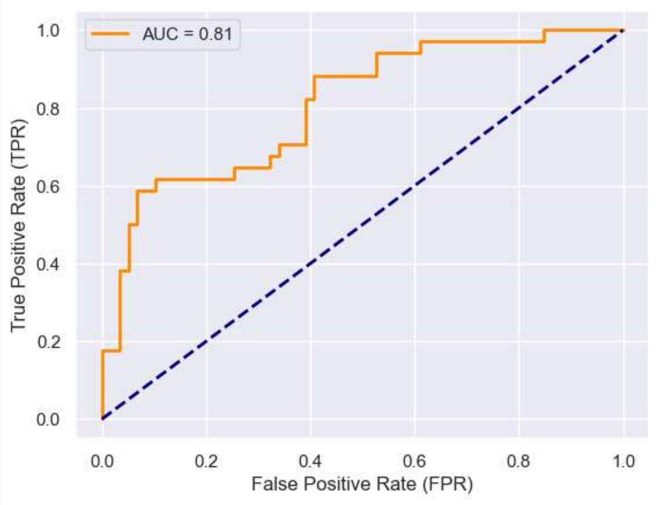
Above is an image of how the model performed as the regularization strength was varied.



| | Predicted No CHD | Predicted CHD |
|---|---|---|
| True No CHD | 55 | 4 |
| True CHD | 15 | 19 |

The table above displays the confusion matrix which the log-reg model produced from the test results. The model had a high precision score of 81.6%, meaning that when we predict CHD we can be fairly confident there is CHD present. On the other hand, the model had a relatively low recall score of 55.9%, meaning that we are missing a lot of the CHD present cases, and they are being predicted to be no CHD. In the medical industry, this will be very damaging as patients will leave having a false sense of security.

The area under the ROC curve was 0.808. AUC values in this range are generally considered good. It indicates that the classifier performs significantly better than random guessing. AUC summarizes the ROC curve into a single value, providing a measure of the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. In our case, an AUC of 0.808 would not be good enough to be considered for decision making in the medical field.



Lastly, I looked at the coefficients of the log-reg model. Interestingly, obesity and ldl had the highest impact on the model with magnitudes of 0.149 and 0.103 respectively. Alcohol and sbp had the least impacts on the model's predictions. Doctors can use this information to focus more on the relationship between obesity and ldl with CHD.

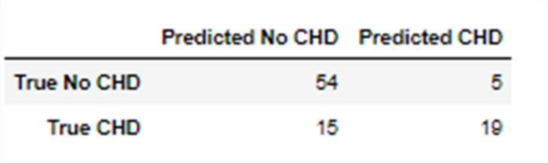
**Classifier Exploration:**

| Classifier | Hyperparameters | Test Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|---|
| KNN | N_neighbours: 17 | 0.774193548 | 0.809524 | 0.5 | 0.77 | 0.788385 |
| Decsion Tree | Max Depth: 5 | 0.688172043 | 0.575758 | 0.558824 | 0.69 | 0.73006 |
| SVM | Linear, C: 1 | 0.784946237 | 0.791667 | 0.558824 | 0.78 | 0.800598 |
| Fisher Discriminant | Solver: svd | 0.774193548 | 0.76 | 0.558824 | 0.77 | 0.802592 |

**SVM:**

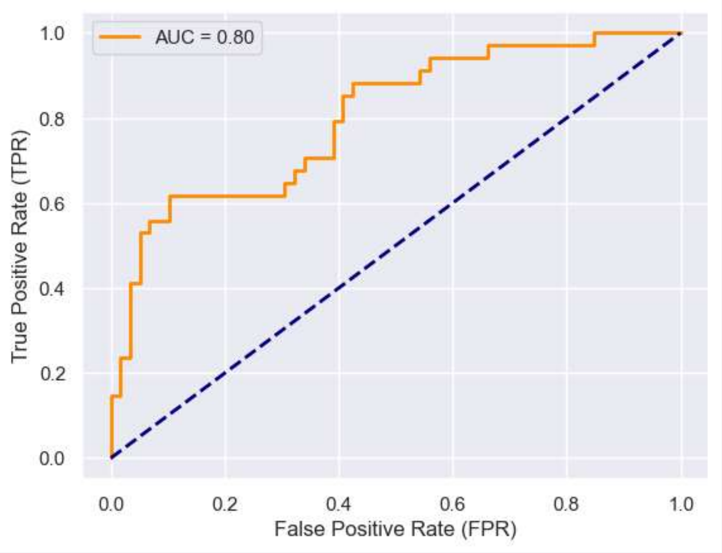
I fit the SVM to classify the patients. Aside from having the highest test accuracy, I chose a SVM model because it is an alternative approach to classification that potentially offers competitive performance. Additionally, SVMs perform well even in the presence of class imbalance, which is the case for our dataset. By penalizing misclassifications based on the margin, SVMs can effectively handle imbalanced datasets. Lastly, SVMs are particularly well-suited for small to medium-sized datasets, and with only 462 rows there is not much data to work with in our assignment.

To perform the modelling, I used 5-fold cross validation to tune my hyperparameter C which controls the strength of regularization like in the logistic regression model. The best resulting model from the cross validation resulted in a test accuracy of 78.5% using C equal to 1 and the kernel used was linear.



| | Predicted No CHD | Predicted CHD |
|---|---|---|
| True No CHD | 54 | 5 |
| True CHD | 15 | 19 |

The SVM classifier with a linear kernel and C=1 demonstrates reasonably good performance across various evaluation metrics. It achieves a relatively high accuracy of 78.5%, indicating that the model correctly classifies a significant portion of instances. The precision score of 79.2% suggests that the classifier maintains a good balance between true positive predictions and false positive predictions. However, the recall score of 55.9% indicates that the classifier may have difficulty in identifying all true positive instances, potentially leading to patients not learning about their condition and leaving with a false sense of security. The AUC score of approximately 0.801 indicates the classifier's ability to distinguish between positive and negative instances. The F1-score, which is the harmonic mean of precision and recall, is calculated to be 0.78 which is a good score and an appropriate metric to consider for this dataset since it is slightly imbalanced.


AUC = 0.80
True Positive Rate (TPR)
False Positive Rate (FPR)