BAYES BUSINESS SCHOOL
CITY ST GEORGE'S UNIVERSITY OF LONDON

Individual Coursework Submission Form

Specialist Masters Programme

Surname: **Ahn**	**First Name:** **Hyun**
MSc in: Mathematical Trading and Finance	**Student ID number:**
Module Code: **SMM748**	
Module Title: **Machine Learning for Quantitative Professionals**	
Lecturer: **Rui Zhu**	**Submission Date:** **16/March/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: [%]

0. Information About Coursework Conduction

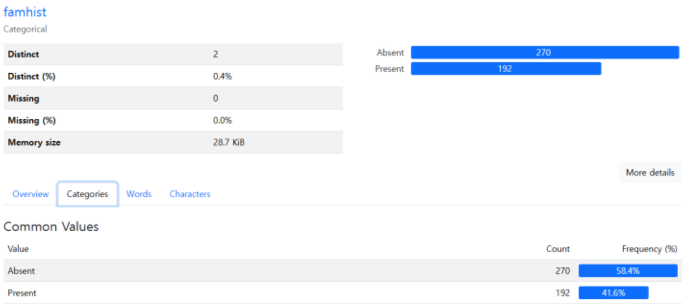
1. This report presents the results of the first individual coursework for the 2025 **Machine Learning for Quantitative Professionals** course.
2. The assignment was completed using Python, and the details of the libraries used are specified at the end of this document.

1. Introduction

This report aims to predict **coronary heart disease (CHD) for males** in a high-risk heart disease region of the **Western Cape, South Africa**, as specified in the coursework requirements. Instead of providing an extensive description of the given dataset, this report will primarily focus on the methodologies used and the rationale behind their selection.

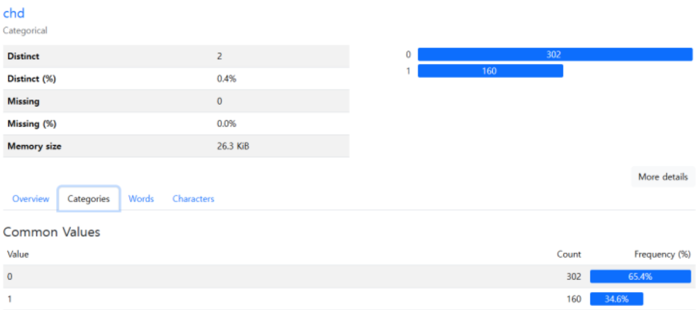
2. Exploratory Data Analysis (EDA)

To perform exploratory data analysis, the **ydata-profiling** library was utilized. The key findings deemed significant are summarized below:



2-1. Binary Representation of the famhist Feature

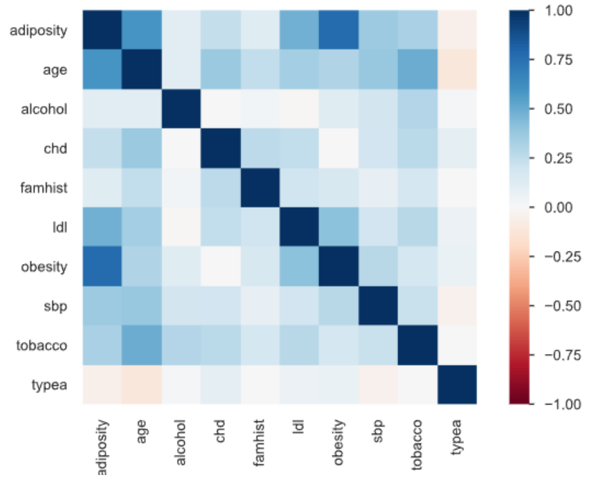
- The famhist feature represents a family history of heart disease and is stored as categorical values **"Absent"** or **"Present",** which are not suitable for direct model analysis.



2-2. Class Imbalance in the Target Variable (chd)

- The **chd** variable, which is the dependent variable to be predicted, exhibits a significant class imbalance:
  - **Positive cases (CHD = 1):** 34.6%
  - **Negative cases (CHD = 0):** 65.4%

- Given this imbalance, evaluating model performance using only **accuracy** is insufficient. Additional techniques such as **undersampling, oversampling**, or alternative evaluation metrics should be considered to address this issue.



2-3. Correlation Analysis and Multicollinearity

- The **Correlation Heatmap** reveals five significant relationships between independent variables:
  - **Adiposity – Age**
  - **Adiposity – Ldl**
  - **Adiposity – Obesity**
  - **Age – Tobacco**
  - **Ldl – Obesity**
- Given that the dataset contains **only 462 observations** but includes **nine independent variables**, there is a potential risk of **multicollinearity**. Applying **Principal Component Analysis (PCA)** may help mitigate this issue.

3. Data Preprocessing

3-1. Binarization of the famhist Feature

- To facilitate model training, the famhist feature was converted into a binary format:
  - "Present" → **1**
  - "Absent" → **0**

3-2. Principal Component Analysis (PCA)

- Since there are **9 independent variables** and only **462 observations**, the risk of **multicollinearity** is high.
- **PCA was applied to five features** (adiposity, age, ldl, obesity, tobacco) to reduce dimensionality.

- The first **two principal components** accounted for **71.70% of the variance** in these five features.

4. Logistic Regression with Ridge Penalty

- As per the coursework requirements, a **logistic regression model with a ridge penalty** was implemented using a fixed **C value of 1**.
- The key results are as follows:

```
                           Logit Regression Results                           
==============================================================================
Dep. Variable:                    chd   No. Observations:                  323
Model:                          Logit   Df Residuals:                      316
Method:                           MLE   Df Model:                            6
Date:                Sun, 16 Mar 2025   Pseudo R-squ.:                  0.1731
Time:                        16:08:40   Log-Likelihood:                -172.38
converged:                       True   LL-Null:                       -208.47
Covariance Type:            nonrobust   LLR p-value:                 1.456e-13
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.6854      0.653     -4.114      0.000      -3.965      -1.406
sbp            0.7849      0.745      1.053      0.292      -0.676       2.246
famhist        0.9010      0.267      3.380      0.001       0.379       1.423
typea          2.2073      0.909      2.428      0.015       0.425       3.989
alcohol       -0.4644      0.787     -0.590      0.555      -2.006       1.078
PC1            0.5327      0.103      5.147      0.000       0.330       0.736
PC2            0.4200      0.132      3.173      0.002       0.161       0.679
==============================================================================
```

- The model did not produce **overly high coefficients**, suggesting that **scaling and PCA** helped stabilize variable importance.
- Most features showed statistically significant **p-values**, except for sbp and alcohol. However, this could be due to the **simplicity of logistic regression**, so further evaluation was postponed.

4-1. Justification for Performance Metrics

The model prediction results are as follows:

accuracy	f1-score	roc_auc	recall
0.7410071942446043	0.6170212765957447	0.8070054945054945	0.6041666666666666

These four metrics were chosen for evaluating classifiers:

- **Accuracy:** Commonly used to measure model performance.
- **F1 Score:** Particularly useful for imbalanced datasets, balancing precision and recall.
- **ROC_AUC:** Evaluates how well the model differentiates between classes across different thresholds.
- **Recall:** Especially relevant in medical prediction tasks, as **identifying positive cases (CHD patients) is crucial for healthcare companies**.
- Given these results, **logistic regression** provides a reasonably reliable performance, with **accuracy of 74.1%** and a **robust ROC_AUC score**. However, the **recall rate of 60.4%** may

be concerning from a business perspective, as it suggests that some **CHD-positive cases are being misclassified**.

5. Other Classifiers

The following classifiers were tested using **GridSearchCV** with **5-fold cross-validation** to determine optimal hyperparameters based on **F1 Score**:

Model	**Hyperparameters**
Decision Tree	ccp_alpha: [1, 0.1, 0.01, 0.001, 0.0001]
Random Forest	max_features: [5, 10, 20, 30, 40, 50, "sqrt"]
AdaBoost	learning_rate: [0.001, 0.01, 0.1, 1]
Gradient Boosting	learning_rate: [0.001, 0.01, 0.1, 1]
k-Nearest Neighbors	n_neighbors: [3, 5, 7, 9]
LDA, QDA, GaussianNB	No hyperparameters
SVC	gamma: [1, 1e-1, 1e-2, 1e-3, 1e-4], C: [1, 10, 100, 1000]

- The **best model in terms of accuracy was LDA (Linear Discriminant Analysis)**, as required by the coursework.

6. Results and Discussion

6-1. LDA Performance Metrics

accuracy	f1-score	roc_auc	recall
0.7338129496402878	0.6021505376344086	0.8129578754578755	0.5833333333333334

6-2. Comparison with Logistic Regression

- LDA achieved a **higher ROC_AUC** than **logistic regression**, suggesting better class separation.
- However, **F1 Score, Accuracy, and Recall were lower**, making it less effective for **CHD detection**.

6-3. Business Perspective Analysis

- Medical companies prioritize **recall (identifying all positive cases)** over accuracy.
- Since LDA had a lower **recall** than logistic regression, it is **not a suitable replacement**.
- The **accuracy-based selection criterion** in the coursework may have led to the **incorrect conclusion that LDA is the best model**.

- Given the small dataset (**<500 observations**), a **simple model like logistic regression performed the best**, as seen in the **lower accuracy of hyperparameter-tuned models** due to **data splitting**.