MASINDE MULIRO UNIVERSITY OF SCIENCE AND TECHNOLOGY

SCHOOL OF COMPUTING AND INFORMATICS

DEPT. OF COMP. SCIENCE

BCS 413: MACHINE LEARNING GROUPWORK

GROUP MEMBERS:

1. RUFUS CHICHILI — COM/B/01-00116/2019
2. BENJAMIN WATHINDA - COM/B/01-04562/2020
3. MATHIU GERALD — COM/B/01-02171/2016

Question 1

(a) The above stated problem is a binary classification problem. The goal is to predict whether a patient is at a high risk of developing a heart disease or not. In binary classification, there are two possible classes and the model's task is to categorise each patient into one of these classes. In this case, the classes that have been used are "high risk of heart disease" and "not at a high risk of heart disease".

(b) Challenges in classification:

(i) Data quality - Ensuring the quality and accuracy of the data, including medical history, lifestyle habits and so on.

(ii) Feature selection - Identifying which features (variables) are most relevant for predicting heart disease is essential.

(iii) Imbalanced data - It's common for medical datasets to be imbalanced meaning one class (e.g. high risk) may have fewer samples than the other. Handling class imbalance is important to prevent the model from being biased.

(iv) Model selection - Choosing the appropriate machine learning algorithms for classification, such as logistic regression, decision tree, neural networks and fine tuning them for optimal performance.

(v) Data pre-processing - This includes data scaling, normalization, missing values and encoding categorical variables.

vi) Ethical and privacy considerations - Dealing with sensitive patient data requires adherence to privacy regulations like HIPAA and ethical considerations.

vii) Interpretability - It's important to create models that are interpretable, especially in healthcare, to explain the predictions to healthcare professionals and patients.

(i) Data Split - Split the dataset into training, validation and test set.

(ii) Model training - Train the classification model on the training dataset using the chosen algorithm and hyperparameters.

(iii) Hyperparameter tuning - Optimize the model's hyperparameters to achieve the best performance.

(iv) Evaluation metrics - Use appropriate evaluation metrics to assess the model's performance on the test set.

v) Cross-validation: Perform k-fold cross validation to ensure the model's robustness and assess generalization performance.

vi) Reporting and documentation - Document the results, model performance and and limitations.

vii) Model deployment - Deployed if it meets the desired performance criteria.