UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING

SEMESTER 1, SESSION 2025/2026

PROJECT PROPOSAL

OBESITY LEVEL CLASSIFICATION

SECB3203 : PROGRAMMING FOR BIOINFORMATICS

SECTION 02

GROUP MEMBER:

1. MUHAMMAD FARIHIN BIN SALEH A25CS0102
2. MUHAMMAD MIRZA HASIF BIN MOHD FAHMI A25CS0108
3. MUHAMMAD NAWFAL BIN MOHD SHAIFUDDIN A25CS0109

LECTURER NAME **:** DR. SEAH CHOON SEN

GROUP **:** GROUP 07

TABLE OF CONTENTS

1.0 INTRODUCTION

Obesity is a complex and multifactorial health condition associated with increased risks of chronic diseases such as diabetes, cardiovascular disorders, and certain cancers. With rising global obesity rates, there is a growing need for accurate and automated classification systems that can assist in early detection and risk assessment. This project, **Obesity Level Classification** aims to develop a multi-class classification system to predict obesity levels based on individual lifestyle, dietary, and physiological attributes. Using machine learning models including **Random Forest**, **XGBoost**, **Logistic Regression**, and **Support Vector Machines**, we intend to create a reliable tool for obesity risk stratification, contributing to personalized health interventions and public health planning.

1.1 PROBLEM BACKGROUND

Obesity classification traditionally relies on body mass index (BMI) and other anthropometric measurements, which may not fully capture the influence of behavioural and metabolic factors. Manual assessment and self-reported data can be subjective and prone to error. Recent advances in machine learning offer opportunities to integrate diverse data sources such as physical activity, eating habits, and demographic information which can be used to improve classification accuracy. The Obesity Levels dataset from Kaggle provides a structured set of features that can be leveraged to train and evaluate predictive models, offering a data-driven approach to obesity assessment.

1.2 PROBLEM STATEMENT

Current methods for obesity classification often lack integration of multidimensional lifestyle data and may not adapt well to individual variability. There is a need for an automated, accurate, and scalable system that can classify obesity levels using a comprehensive set of features. This project addresses this gap by developing and comparing multiple machine

learning models to identify the most effective approach for obesity level prediction, thereby supporting healthcare professionals in early intervention and personalized care.

1.3 OBJECTIVES

1. **Preprocess and explore the Obesity Levels dataset**
   Clean data, encode categorical variables, normalize features, and conduct exploratory data analysis (EDA).

2. **Implement and train four classification models**
   Build and train Random Forest, XGBoost, Logistic Regression, and SVM models for obesity level prediction.

3. **Evaluate and compare model performance**
   Assess models using accuracy, precision, recall, F1-score, and confusion matrices.

4. **Identify the best-performing model**
   Compare results to select the most accurate and reliable model for obesity classification.

5. **Document workflow and maintain GitHub repository**
   Keep code and documentation organized on GitHub for reproducibility and collaboration.

1.4 SCOPES & LIMITATIONS

Scopes:

1. **Dataset Utilization**
   The project will use the "Obesity Levels" dataset from Kaggle, which includes lifestyle, dietary, and physiological attributes for classification.

2. **Model Implementation**

   Four machine learning models which are Random Forest, XGBoost, Logistic Regression, and SVM thatwill be implemented and trained for multi-class obesity level classification.

3. **Performance Comparison**

   Models will be evaluated and compared using standard classification metrics (accuracy, precision, recall, F1-score, confusion matrices).

4. **Tool and Environment**

   The project will be developed using Python in a local environment, utilizing libraries such as Pandas, Scikit-learn, and Matplotlib/Seaborn.

5. **Documentation**

   All code, results, and documentation will be maintained and regularly updated on a GitHub repository for transparency and collaboration.

Limitation:

1. **Local Execution Only**

   The project will not utilize cloud platforms (Microsoft Azure or AWS), limiting scalability and real-time processing capabilities.

2. **Dataset Specificity**

   The model's performance may be constrained by the size, quality, and demographic scope of the Kaggle dataset, affecting generalizability.

3. **Model Interpretability**

   Some models, such as Random Forest and XGBoost, may act as "black boxes," making it difficult to interpret how specific predictions are made.

4. **Lack of Clinical Validation**

   The study is computational and does not include real-world clinical testing or integration with healthcare systems.

1.5 CONCLUSION

This project will leverage machine learning to classify obesity levels using a well-curated Kaggle dataset. By comparing multiple models, we aim to identify a robust classification approach that can contribute to data driven health assessments. The findings will be documented in a comprehensive report and shared via GitHub, adhering to the course requirements for collaboration and transparency. This work aligns with current bioinformatics trends in disease classification and offers practical experience in Python-based data science.