**Name: Andrew Soon Ng Yong Liang**

**1. Project Title:**

**Diabetes Analysis and Prediction: A Machine Learning Approach**

**2. Background and Motivation:**

Diabetes is a chronic disease affecting millions of people worldwide, and early detection is crucial for effective management. The **Diabetes Health Indicators Dataset**, available on Kaggle, provides comprehensive data on various health indicators such as physical activity, BMI, blood pressure, and more, all of which are crucial in assessing diabetes risk. Using this dataset, machine learning techniques can be employed to predict the likelihood of diabetes, enabling more effective prevention strategies and healthcare interventions.

**3. Project Objectives:**

The main objectives of this project are:

* To develop a predictive model using machine learning algorithms to identify individuals at high risk of diabetes.
* To analyze the key health indicators contributing to diabetes risk.
* To evaluate the performance of different models (e.g., logistic regression, decision trees, random forests, etc.) for accuracy and interpretability.

Personal objectives of this project including:

* Learning the end-to-end process/steps to perform analysis on data provided
* Learning the tools/approaches/methods for working on data analysis.

**4. Data Overview:**

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

The objective of the BRFSS is to collect uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases in the adult population. Factors assessed by the BRFSS include tobacco use, health care coverage, HIV/AIDS knowledge or prevention, physical activity, and fruit and vegetable consumption. Data are collected from a random sample of adults (one per household) through a telephone survey. 2015 dataset was selected for this project, which contains over 441,000 records with 330 columns. After the data cleaning and preprocessing phase, only 250,000 records with 22 columns representing various health-related attributes such as:

* **Age**, **Gender**, and **BMI**.
* **Smoking** and **Alcohol Consumption**.
* **Physical Activity**, **Diet**, and **Blood Pressure**.
* **Previous Medical History** (e.g., stroke, heart attack).
* **Diabetes Status** (binary target variable).

**5. Methodology:**

* **Data Cleaning and Preprocessing:** Handle missing values, outliers, and normalize the data for effective model training. References: Behavioral Risk Factor Surveillance System 2015 Codebook Report
* **Exploratory Data Analysis (EDA):** Perform EDA to understand the distribution of variables and their relationships with diabetes risk.
* **Feature Engineering:** Generate new features from existing data (e.g., interaction terms between BMI and physical activity).
* **Model Development:** Build and train multiple machine learning models, including logistic regression, decision trees, and ensemble methods such as random forests.
* **Model Evaluation:** Use accuracy, precision, recall, F1-score, and ROC-AUC to evaluate model performance.

**6. Expected Outcomes:**

* A high-performing machine learning model capable of accurately predicting the risk of diabetes.
* Identification of the most influential health indicators in determining diabetes risk.
* Insights that can guide healthcare professionals in prioritizing prevention and intervention efforts.

**7. Timeline:**

The project is expected to be completed in 4 weeks, broken down as follows:

* Week 1: Data cleaning, preprocessing, and exploratory analysis.
* Week 2: Model development and tuning.
* Week 3: Model evaluation and comparison.
* Week 4: Report generation and final model deployment.

**8. Tools and Technologies:**

* Python, with libraries such as Pandas, Scikit-learn, and Matplotlib.
* Kaggle platform for dataset hosting and project collaboration.
* Microsoft Studio for code execution and documentation.

**9. Conclusion:**

This project aims to leverage the Diabetes Health Indicators Dataset to develop a reliable predictive model that will help identify individuals at high risk of diabetes. The insights gained can provide valuable contributions to public health initiatives focused on diabetes prevention and early intervention.