**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.

# **Quality Assurance Plan for Diabetes Analysis and Prediction**

### **1. Introduction**

This Quality Assurance (QA) Plan outlines the approach to ensure the integrity, accuracy, and reliability of the synthetic health data generated using feature combinations. The goal of the QA process is to ensure that the data is of high quality and suitable for machine learning model development.

### **2. Objectives**

The key objectives of this QA Plan are:

* To verify that the generated combinations accurately reflect all possible feature permutations.
* To ensure the generated data conforms to the defined constraints and ranges.
* To identify and mitigate any errors or inconsistencies in the generated data.
* To validate that the final dataset is suitable for use in machine learning algorithms.

### **3. Quality Assurance Criteria**

The quality of the generated data will be evaluated based on the following criteria:

* **Completeness**: All valid combinations of the defined feature values are present in the generated dataset.
* **Accuracy**: Each feature value within the combination follows its defined range and characteristics.
* **Consistency**: Data adheres to logical relationships between features, e.g., age categories and BMI values align properly.
* **Efficiency**: The data generation process is free from memory errors and processes data in manageable chunks.

### **4. Quality Assurance Activities**

#### **4.1. Data Validation and Verification**

The first step is to validate and verify the feature definitions and the data generation process:

* **Range Checks**: Ensure that feature values are generated within the defined limits (e.g., BMI values, age categories, etc.).
* **Logical Consistency Checks**: Verify that logical constraints between features are maintained (e.g., an older age group having a higher likelihood of stroke).
* **Data Type Verification**: Ensure that each feature is correctly typed (e.g., integers for categories like age and education, floats for BMI).

#### **4.2. Data Sampling for Manual Review**

Since generating all combinations may result in a very large dataset, a **random sample** of combinations will be selected for manual review:

* Randomly select 1-2% of the data for manual inspection.
* Verify that the sampled data matches expected feature values and combinations.
* Validate whether feature combinations make sense in terms of the target use case.

#### **4.3. Automated Testing**

In addition to manual review, the following automated tests will be applied:

* **Range Testing**: Scripts to confirm that all feature values fall within the specified range.
* **Combination Completeness**: Automated tests to check that all valid combinations of the feature set have been generated.
* **Memory and Efficiency Checks**: Ensure that the batch processing does not run into memory issues and efficiently handles large datasets.

### **5. Tools and Resources**

* **Pandas**: To handle the data and verify its structure.
* **Itertools**: To generate feature combinations and ensure all permutations are covered.
* **Unit Testing**: Python unit test scripts to automate QA checks on data validity.
* **Sampling Tools**: Python’s random sampling libraries to draw samples for manual inspection.

### **6. Roles and Responsibilities**

* **Data Engineer**: Responsible for implementing the data generation process and writing the QA scripts.
* **Quality Assurance Analyst**: Responsible for conducting manual data reviews and ensuring that automated tests are properly implemented and executed.
* **Project Manager**: Responsible for overseeing the QA process, ensuring that the QA criteria are met, and coordinating between teams.

### **7. Continuous Monitoring and Reporting**

* **Progress Reports**: A weekly report on the quality assurance process will be provided, detailing any issues encountered and corrective actions taken.
* **Issue Tracking**: Any data inconsistencies, errors, or issues will be logged in an issue tracker, with priorities assigned for resolution.
* **Final Approval**: Once all tests have passed and the data is deemed accurate and complete, the dataset will be approved for use in the machine learning project.

### **8. Risk Management**

The following risks have been identified, along with mitigation strategies:

* **Memory Overload**: If the data is too large to fit in memory, processing will be done in manageable chunks, and only a subset of combinations will be reviewed manually.
* **Inconsistent Data**: Manual review combined with automated checks will help catch inconsistent or illogical feature combinations.
* **Slow Performance**: The data generation script will be optimized to ensure efficiency, particularly with batch processing.

### **9. Conclusion**

The QA Plan ensures that the synthetic health data generated meets high standards of accuracy, completeness, and consistency. With regular testing, monitoring, and manual reviews, the dataset will be verified as reliable and ready for machine learning use.