**1. Research Title**

**Early Detection of Diabetes Using Machine Learning on NHANES 2021–2022 Survey Data**

**2. Research Aim**

To develop and evaluate machine learning models that can predict the presence of diabetes in individuals using non-invasive demographic, dietary, physiological, and biochemical variables from the NHANES 2021–2022 survey.

**3. Research Question**

**Primary RQ**:

How accurately can machine learning models predict diabetes using lifestyle, biometric, and biochemical data from the NHANES 2021–2022 dataset?

**4. Research Objectives**

|  |  |
| --- | --- |
| **Objective No.** | **Description** |
| **O1** | To extract and clean relevant demographic, clinical, and lifestyle data from NHANES 2021–2022 .XPT files |
| **O2** | To define a target variable based on self-reported diabetes status and HbA1c thresholds |
| **O3** | To engineer meaningful features (e.g., BMI, waist circumference, calorie intake ratios) |
| **O4** | To train and compare the performance of various machine learning models (e.g., Logistic Regression, Random Forest, XGBoost, ~~CatBoost~~) |
| **O5** | To evaluate models using classification metrics such as ROC-AUC, precision, recall, and feature importance |
| **O6** | To interpret key predictors of diabetes and explore their clinical relevance using SHAP values |

**5. Research Methodology**

**Data Source**

* NHANES 2021–2022 publicly available .XPT files from the CDC.

**Data Collection**

* Downloaded via pandas.read\_sas() using direct CDC links.
* Merged datasets using the SEQN identifier.

**Preprocessing Steps**

* Handle missing data with imputation (median strategy)
* Encode categorical variables (e.g., gender, ethnicity)
* Normalize numerical features

**Feature Engineering**

* Derived variables such as BMI categories, macronutrient ratios
* Included both lab-based (e.g., HbA1c) and survey-based predictors

**Model Training**

* Split data (80/20)
* Trained Logistic Regression, Random Forest, XGBoost, and CatBoost models

**Evaluation Metrics**

* Accuracy
* Precision, Recall, F1-score
* ROC-AUC
* SHAP analysis for interpretability

**6. Module & Variable Mapping**

|  |  |  |
| --- | --- | --- |
| **Module** | **Prefix** | **Key Variables** |
| **Demographics** | DEMO\_L | RIDAGEYR (Age), RIAGENDR (Gender), RIDRETH1 (Race) |
| **Body Measures** | BMX\_L | BMXBMI (BMI), BMXWAIST (Waist Circ.) |
| **Lab – Glucose** | GLU\_L | LBXGLU (Fasting Glucose) |
| **Lab – HbA1c** | GHB\_L | LBXGH (HbA1c %) |
| **Lab – Insulin** | INS\_L | LBXIN (Insulin) |
| **Questionnaire** | DIQ\_L | DIQ010 (Self-reported diabetes) |
| **Physical Activity** | PAQ\_L | PAQ605, PAQ620 (activity levels, sitting) |
| **Dietary Intake** | DR1TOT\_L | DR1TKCAL (Calories), DR1TPROT, DR1TSUGR, DR1TFAT |