1. **Project Title:** Predicting Diabetes Risk Using Health Indicators

Group 11: Nishant Chacko, Rajat Gusain, Senay Hagos, Miguel Morales Gonzalez, Harold Xue, Rosy Zhou

2. Dataset

Name: CDC diabetes Dataset (UCI Machine Learning Repository) (Link: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators)

Or dataset also available on Kaggle (Link: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset)

Description: The Diabetes Health Indicators Dataset contains healthcare statistics and lifestyle survey information about people in general along with their diagnosis of diabetes. In total there are over 250,000 instances. The dataset contains 21 features consisting of demographics, lab test results, and answers to health-related survey questions for each patient. The target variable for classification is whether a patient is diabetic/pre-diabetic, or healthy.

Variables: The dataset includes 21 input features, a mixture of continuous and binary features – diabetes, high blood pressure, high cholesterol, cholesterol check, BMI, smoker, stroke, heart diseases/attack, physical activity, fruits, vegies, heavy alcohol consumption, any health care, lack of physician care due to cost, general health, mental health, physical health, walk, sex, age, education, and income.

3. Hypothesis

Lifestyle and clinical factors (such as exercise frequency, BMI, blood pressure, smoking habits, and diet) significantly influence diabetes risk. Machine learning models can be used to accurately predict whether an individual is diabetic/pre-diabetic, or healthy.

4. Proposed Approach

- **Data Understanding & Preprocessing**
    - Explore dataset: demographics, lab results, survey responses.
    - Handle missing values, normalize continuous features (e.g., cholesterol, BMI).
    - Encode categorical variables (e.g., gender, smoking status).
    - Split into **training (70%)** and **test (30%)** sets.

- **Model Selection**

    - Since the target has two classes (diabetic/pre-diabetic or healthy), this is a Binary classification problem.
    - Logistic Regression → interpretable baseline.
    - Random Forest / Gradient Boosting → strong performance, feature importance analysis.
    - Support Vector Machines (SVM) → good for smaller datasets, but slower on large ones.
- **Evaluation Metrics**
    - Accuracy: overall correctness.
    - Precision, Recall, F1-score: especially important for imbalanced classes.
    - Confusion Matrix: visualize misclassifications.
    - ROC-AUC (macro/micro): measure separability across classes.
- **Interpretability**
    - Use feature importance (Random Forest).
    - Highlight which risk factors (e.g., BMI, cholesterol, smoking) most influence diabetes classification.
- **Validation**
    - Perform cross-validation to ensure robustness.
    - Compare models and select the one that balances accuracy and interpretability.