Project Proposal (Milestone 1)

E. Mutimba & E. Toobian

1. Project Title

Predicting Diabetes Diagnosis Using Supervised Learning Models

2. Research Questions

Primary Research Question:

How do different supervised learning modeling approaches compare in terms of predictive performance and interpretability in the prediction of diabetes diagnosis using patient-level health and risk factors?

Secondary Research Question:

Which patient-level health and risk factors are most influential in predicting diabetes diagnosis across different modeling approaches?

3. Dataset Source & Description

This section describes the source, structure, and context of the dataset used for supervised learning analyses in this project.

URL / Upload location:

This dataset was obtained from Kaggle and authored by Rakesh Kolipaka and Ranjith Kumar Digutla:

Diabetes Health Indicator Dataset[3]

Variables:

The dataset contains a total of *31 variables*, including one binary outcome variable and a mix of numerical and categorical predictors capturing demographics, lifestyle, medical history, and clinical health factors. Variables are grouped below by their clinical and contextual role.

- Outcome Variable:

  Binary indicator of whether an individual has been diagnosed with diabetes (0 = No, 1 = Yes):

  - `diagnosed_diabetes`

- Demographic and Socioeconomic Variables:

  Basic population characteristics describing individual background and status:

  - *Numerical Demographic*: `age`
  - *Categorical Demographics*: `gender`, `ethnicity`
  - *Categorical Socioeconomic Factors*: `education_level`, `income_level`, `employment_status`

- Lifestyle and Behavioral Variables:

  Indicators of health-related behaviors associated with diabetes risk:

  - *Categorical Behavior Indicator*: `smoking_status`

  - *Numerical Lifestyle Measures*: `alcohol_consumption_per_week`, `physical_activity_minutes_per_week`, `sleep_hours_per_day`, `screen_time_hours_per_day`
  - *Numerical Quality Measure (higher = healthier)*: `diet_score`

- Medical History Indicators:

  Binary variables capturing known Risk Factors and comorbidities:

  - *Family*: `family_history_diabetes`
  - *Personal*: `hypertension_history`, `cardiovascular_history`

- Clinical and Physiological Measurements:

  Quantitative health measurements commonly used in diabetes screening and diagnosis:

  - *Anthropometric Measures*: `bmi`, `waist_to_hip_ratio`
  - *Blood Pressure and Cardiovascular Measures*: `systolic_bp`, `diastolic_bp`, `heart_rate`
  - *Lipid Profile*: `cholesterol_total`, `hdl_cholesterol`, `ldl_cholesterol`, `triglycerides`
  - *Glycemic Markers*: `glucose_fasting`, `glucose_postprandial`, `insulin_level`, `hba1c`

- Derived Risk and Severity Measures:
  - *Composite Numerical Risk Score*: `diabetes_risk_score`
  - *Categorical Indicator of Disease Severity*: `diabetes_stage`

Size:

The dataset contains **100,000 observations** and **31 variables**. Initial inspection shows *no missing values* and *no duplicate rows*. The outcome variable exhibits an approximate **60/40 class split**, with about 60% of observations labeled as diagnosed with diabetes.

Domain context:

This dataset represents a synthetically generated population-level health record designed to model known demographic, lifestyle, family medical, and clinical risk factors associated with diabetes. Variables such as BMI, blood glucose measurements, HbA1c, lipid levels, blood pressure, and family history reflect clinically meaningful predictors commonly used in diabetes risk assessment. Variables were generated using statistical distributions informed by real-world medical research and public health sources, including the International Diabetes Federation (IDF), Centers for Disease Control (CDC), and World Health Organization (WHO)[3] . Although all values are synthetic, they fall within medically realistic ranges and reflect established relationships. The dataset is intended for educational and research use where patient privacy must be preserved.

Documentation Notes:

During initial exploration, we observed inconsistencies between the dataset documentation and the distributed CSV file, such as references to variables not present in the data or inconsistent descriptions of variable distributions. For this project, the CSV file itself is treated as the authoritative source for variables utilizing included data and observed values.

4. Motivation

Diabetes continues to pose a significant and escalating challenge to both global and national health systems. As one of the most rapidly expanding non-communicable diseases identified by the World Health Organization [4], its rise is driven by a complex interplay of factors including aging populations, increasingly sedentary

lifestyles, dietary changes, and persistent disparities in access to preventive care. In the United States alone, recent estimates from the CDC suggest that tens of millions of people are living with diabetes or prediabetes, many without a formal diagnosis[2]. This underscores the critical need for earlier identification and more targeted, proactive public health interventions.

This project aims to investigate the underlying risk factors for diabetes using individual-level data from the Diabetes Health Indicators Dataset, a resource that integrates demographic, behavioral, and clinical variables while maintaining strict privacy safeguards. Through the application of statistical analysis and machine learning techniques, the objective is to detect meaningful patterns in known risk factors that could enhance early detection capabilities and inform evidence-based public health policies. These efforts are closely aligned with the strategic priorities outlined by both the WHO and the CDC, particularly in the advancement of data-informed approaches to the prevention and management of chronic diseases [1],[5].

5. Planned ESL Methods

We will apply a set of complementary classification models drawn from the ESL framework, chosen in order to balance interpretability, flexibility, and predictive performance.

- **Method 1: Logistic Regression (Baseline and Regularized)**
  Logistic regression will serve as an interpretable baseline model, with Ridge and LASSO regularization used to address multicollinearity and assess variable importance. Model performance will be evaluated using standard classification metrics such as accuracy and ROC-AUC.
  Relevant ESL Chapters: Chapters 4 and 6

- **Method 2: Random Forest (Tree-Based Ensemble)**
  Random forests will be used to capture nonlinear relationships and interactions among predictors. This method builds an ensemble of trees using bootstrap aggregation and feature randomness, offering improved predictive performance and robustness over single trees.
  Relevant ESL Chapter: Chapters 9 and 15

- **Method 3: Gradient Boosting (XGBoost)**
  Gradient boosting methods will be explored as a high-capacity ensemble approach that sequentially improves model performance by focusing on previously misclassified observations. This model will allow comparison of bias-variance tradeoffs relative to random forests and logistic regression.
  Relevant ESL Chapter: Chapter 10

- **Method 4 (Optional): Multilayer Perceptron (MLP) Neural Network**
  If time permits, we will explore a shallow multilayer perceptron (MLP) as a nonlinear extension beyond traditional ESL models. This model will serve as an exploratory comparison to whether increased model flexibility provides meaningful performance gains. Emphasis will be placed on a controlled architecture size and regularization to avoid overfitting.
  Relevant ESL Chapter: Chapter 11

6. Expected Challenges

Several methodological and practical challenges are anticipated in this project given the size, structure, and nature of the dataset, as well as the range of modeling approaches under consideration.

- **Synthetic Data and Documentation Limitations**
  Although the dataset is designed to reflect realistic health patterns, all observations are synthetically generated. As a result, conclusions must be carefully interpreted, particularly with respect to real-world implications and generalizations. Additionally, minor inconsistencies between the dataset documentation and the distributed CSV file require treating the observed data as the authoritative source during analysis.

- **Class Distributions and Evaluation Considerations**
  The binary outcome variable exhibits an approximate 60/40 class split. While not severely imbalanced, this motivates the use of evaluation metrics beyond accuracy, such as AUC and confusion matrices, to ensure robust model comparison.

- **Correlated Predictors and Interpretability**
  Many clinical and behavioral variables, including BMI, physical activity, age, and comorbid conditions, are inherently correlated. To address potential multicollinearity and support stable estimation and interpretation in regularized logistic regression methods such as Ridge and LASSO will be employed.

- **Model Complexity and Overfitting Risk**
  Flexible models such as gradient boosting and neural networks offer strong predictive power but may overfit when interactions are complex or predictors are highly correlated. Cross-validation, regularization, and controlled model complexity will be used to help ensure stable generalization performance within the scope of our dataset and modeling framework.

- **Feature Scaling and Heterogeneous Variable Types**
  Because the dataset contains heterogeneous variable types, including binary, ordinal, and continuous features, preprocessing steps such as feature scaling will be used for models that are sensitive to predictor scale, including logistic regression and neural networks.

References

[1] Centers for Disease Control and Prevention. Diabetes prevention and control, 2022. Accessed 2026-01-27.

[2] Centers for Disease Control and Prevention. National diabetes statistics report, 2023. Accessed 2026-01-27.

[3] Rakesh Kolipaka and Ranjith Kumar Digutla. Diabetes health indicators dataset, 2025.

[4] World Health Organization. Noncommunicable diseases, 2022. Accessed 2026-01-27.

[5] World Health Organization. Diabetes, 2023. Accessed 2026-01-27.