Diabetes Detection Project

Nick Sleeper*

Department of Computational Mathematics, Science and Engineering

Michigan State University, East Lansing, MI 48824

(Dated: November 4, 2025)

Abstract

Diabetes is a dangerous chronic disease whose complications can lead to severe cardiovascular, neurological, and renal conditions if not detected early. As its global prevalence continues to rise, the ability to identify individuals at high risk has become an essential step toward effective prevention and management. This project aims to develop a machine learning model that predicts the likelihood of diabetes diagnosis using demographic, behavioral, and clinical indicators.

The dataset used—*Diabetes Health Indicators Dataset* (100,000 records, 31 features)—is a synthetic dataset modeled after the CDC Behavioral Risk Factor Surveillance System (BRFSS). A supervised classification approach is employed. After exploratory analysis and preprocessing (encoding, scaling, and leakage removal), a logistic regression model was trained as a baseline. Preliminary results achieved an accuracy of 0.861 and F1-score of 0.885, showing strong predictive potential even with a simple linear model.

The project will extend this baseline by comparing three models of increasing complexity (Logistic Regression, Random Forest, Gradient Boosting) and reporting their performance under a consistent evaluation framework. The expected outcome is a clear, reproducible, and interpretable machine learning pipeline for diabetes-risk prediction.

Repository: https://github.com/sleepernj/cmse492_project

BACKGROUND AND MOTIVATION

Diabetes is a chronic metabolic disease characterized by elevated blood glucose levels that can lead to serious health complications including cardiovascular disease, neuropathy, kidney failure, and vision impairment. It affects hundreds of millions of people worldwide, and its prevalence continues to rise each year. This project holds personal significance to me as a Type 1 diabetic who has experienced firsthand the lifelong challenges of managing blood sugar, medication, and long-term health risks. Understanding how data-driven methods can improve early detection and prevention is both an academic and personal motivation for pursuing this work.

Why this problem is important

Early detection and risk assessment are critical for preventing or delaying the onset of diabetes complications. Traditional diagnostic methods rely on blood tests performed intermittently, which can miss individuals at early risk stages. A predictive model that integrates behavioral and physiological indicators could allow for continuous risk assessment and earlier intervention.

Who cares about this problem

Public health organizations, healthcare providers, and insurers can use predictive analytics to identify high-risk individuals before complications develop. For individuals, such models can provide early warnings and actionable insights. For data scientists, this represents a practical healthcare application of interpretable ML methods.

Consequences of solving the problem

Accurate risk prediction can reduce the societal and financial burden of diabetes, guide resource allocation, and promote preventive care. From a clinical standpoint, models that integrate diverse health indicators can complement traditional screening by identifying subtle patterns not captured by simple thresholds.

Previous work

Existing work (e.g., on the Pima Indians Diabetes dataset and BRFSS data) has applied models like logistic regression, random forests, and support vector machines for risk prediction. These studies validate ML's potential but are often limited by small sample sizes or missing data. The Kaggle synthetic dataset used here offers a complete, privacy-safe, large-scale foundation for building reproducible and interpretable models.

DATA DESCRIPTION

Data Origin and Provenance

The dataset is the *Diabetes Health Indicators Dataset*, published on Kaggle by Mohan Krishna Thalla (2023). It was designed to simulate realistic health patterns based on the CDC's Behavioral Risk Factor Surveillance System (BRFSS), a nationwide survey program that monitors health-related risk behaviors, chronic health conditions, and use of preventive services. BRFSS collects data via telephone surveys; the synthetic dataset mirrors BRFSS-like distributions while preserving privacy. The target variable is `diagnosed_diabetes` (0 = No, 1 = Yes).

Dataset Characteristics

- **Samples:** 100,000
- **Features:** 31 total (24 numerical, 7 categorical)
- **Target:** `diagnosed_diabetes`
- **Missing Values:** None detected in any column

Data Quality and Key Properties

A missing-values heatmap (Fig. 1) confirmed all 31 features are complete. The target shows moderate imbalance (∼60% diabetic) (Fig. 2). Additional EDA finds physiologically consistent relationships: higher glucose and HbA1c associate with diabetes status, BMI is centered near 25–26, and diabetics tend to be older on average.

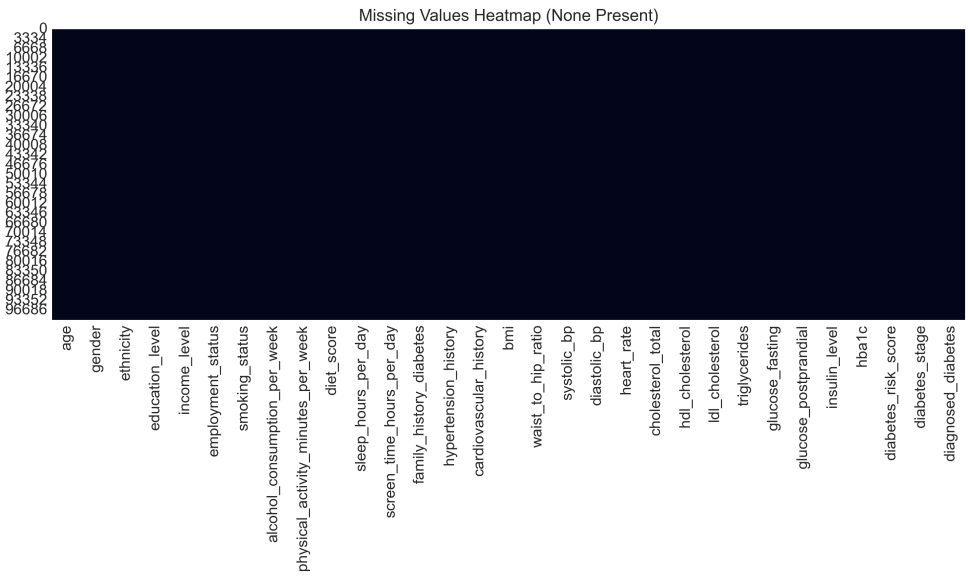

FIG. 1: Heatmap confirming no missing values across all variables.

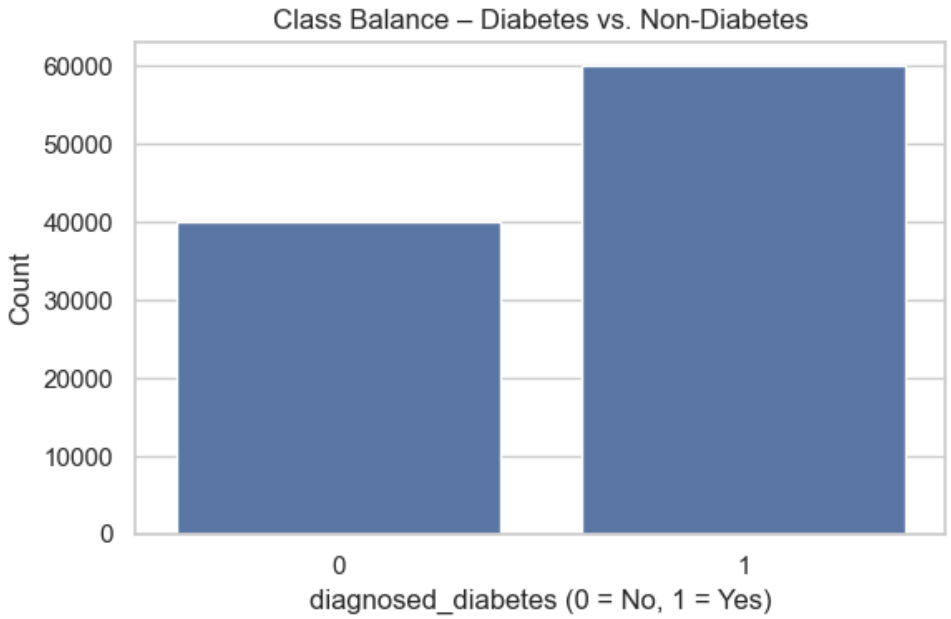

FIG. 2: Class balance for the diabetes diagnosis target variable (0 vs 1).

Additional EDA Figures Referenced in Text

Fig. 3 shows the glucose–HbA1c relationship by diagnosis; Fig. 4 shows BMI distribution; Fig. 5 compares age by diagnosis; Figs. 6–7 summarize correlations.

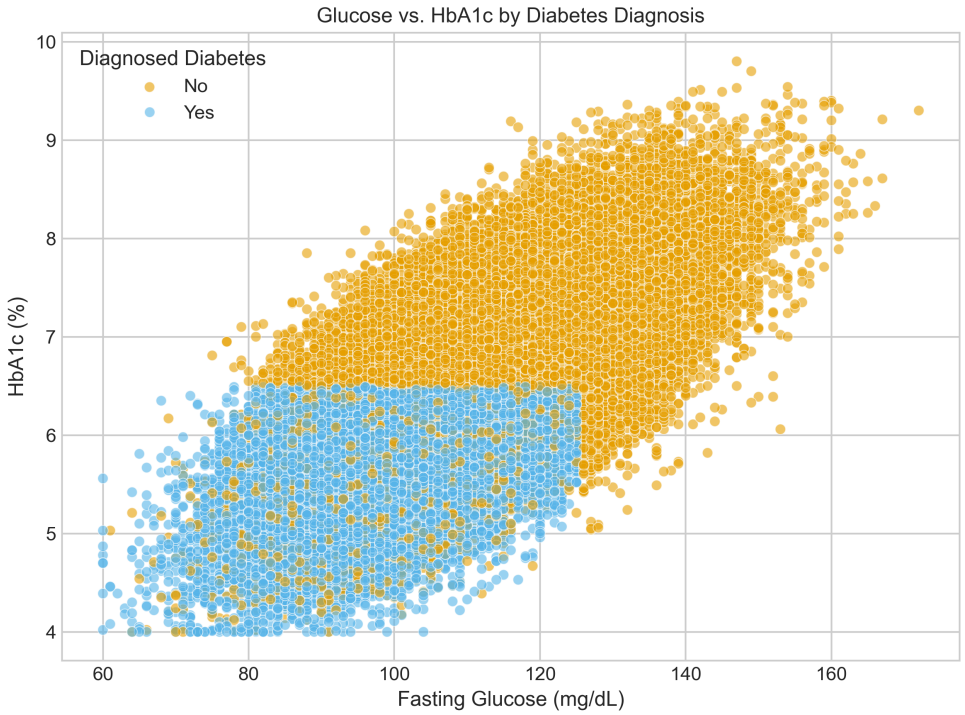

FIG. 3: Fasting glucose vs. HbA1c by diabetes diagnosis.

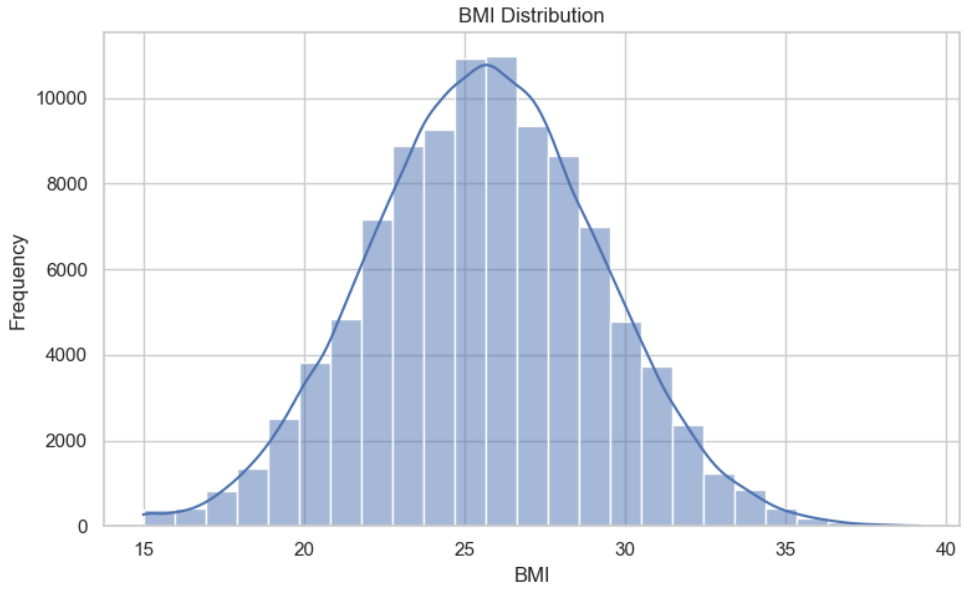

FIG. 4: Distribution of BMI across the population.

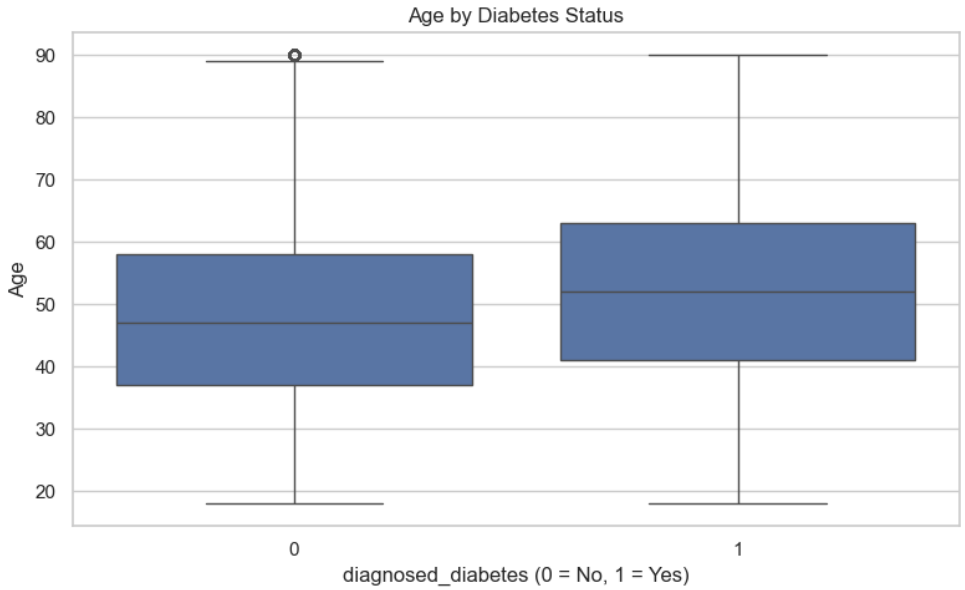

FIG. 5: Age by diabetes status (0 = No, 1 = Yes).

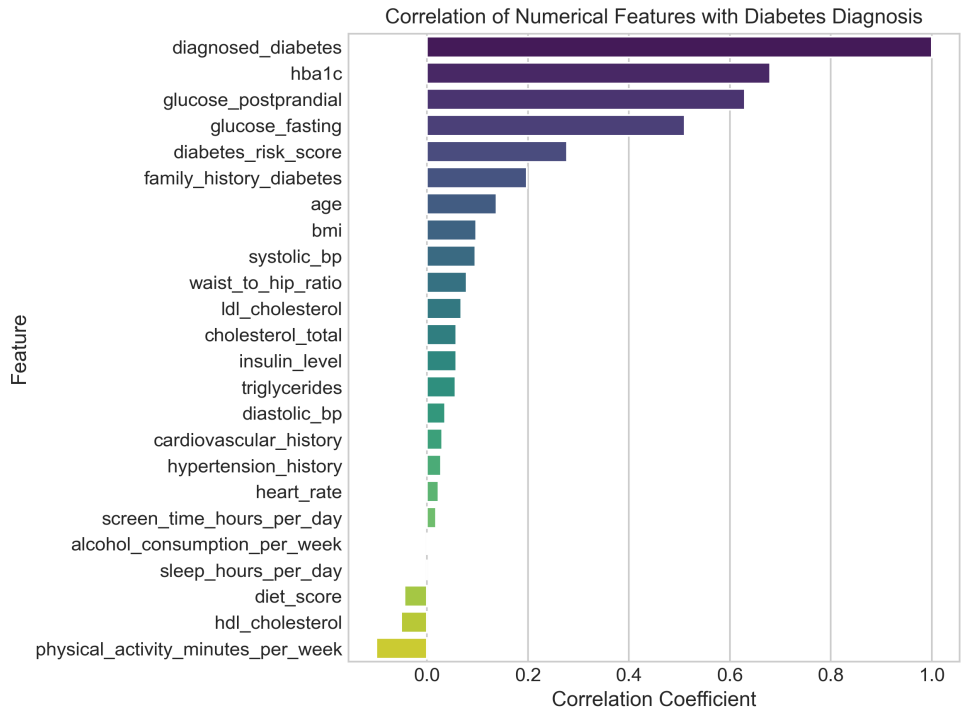

FIG. 6: Correlation of numerical features with diabetes diagnosis.

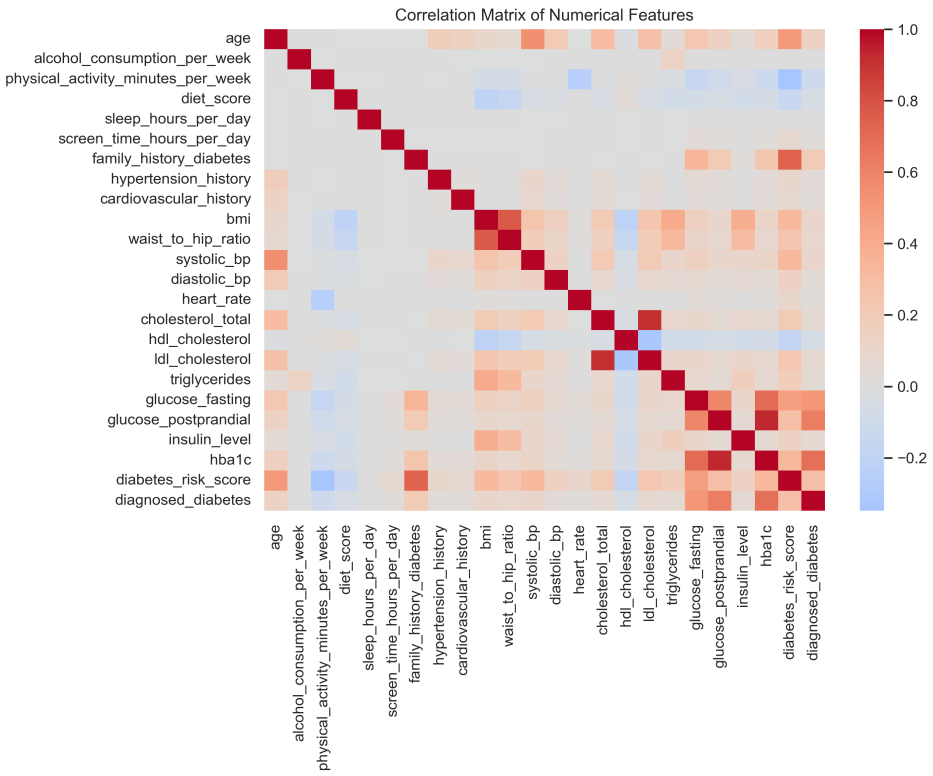

FIG. 7: Correlation matrix of numerical health indicators.

PROPOSED METHODOLOGY

This is a **supervised binary classification** task to predict whether an individual is diagnosed with diabetes (`diagnosed_diabetes`=1) or not (0). The approach will systematically compare three models of increasing complexity using identical preprocessing and evaluation protocols.

Preprocessing (implemented):

1. Remove leakage features: `diabetes_stage`, `diabetes_risk_score`.

2. One-hot encode categorical variables (`gender`, `ethnicity`, `education_level`, `income_level`, `employment_status`, `smoking_status`).

3. Standardize numerical features.

4. Stratified 80/20 train/test split.

Model family (increasing complexity) & justification

1. **Logistic Regression (Linear Baseline).** *Complexity:* Linear decision boundary; $O(pd)$ parameters. *Why:* Interpretable coefficients, fast to train, strong baseline.

2. **Random Forest (Nonlinear Bagging Ensemble).** *Complexity:* $O(T \cdot \text{tree\_size})$ with T trees. *Why:* Captures nonlinear interactions and reduces overfitting.

3. **Gradient Boosting (Nonlinear Boosting Ensemble).** *Complexity:* Sequential trees with learning rate; highest representational power. *Why:* Strong performance for structured data with mixed types.

Loss functions / training objective

Logistic Regression: Regularized log-loss

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)] + \lambda\|\mathbf{w}\|_2^2$$

with $\hat{p}_i = \sigma(\mathbf{w}^\top \mathbf{x}_i)$. **Random Forest:** Minimizes Gini impurity; combines trees via majority vote. **Gradient Boosting:** Fits trees to negative gradients of log-loss sequentially.

Planned hyperparameters and search

- Logistic: penalty = L2, $C \in \{0.1, 1, 10\}$, max_iter = 1000.
- Random Forest: $n\_estimators \in \{200, 400\}$, max_depth $\in \{8, 12, 16\}$.
- Gradient Boosting: learning_rate $\in \{0.05, 0.1\}$, $n\_estimators \in \{200, 400\}$.

5-fold cross-validation will identify the optimal hyperparameters.

Comparison of at least 3 models

All three models will be compared using identical splits, metrics, and preprocessing. Complexity ordering: Logistic (linear), Random Forest (nonlinear bagging), Gradient Boosting (sequential boosting). Metrics include Accuracy, Precision, Recall, F1, and ROC–AUC.

Methodological flow (required description)

The methodological flow proceeds as follows: **data ingestion → preprocessing → model training → cross-validation → evaluation → interpretability analysis**. Each model receives identical preprocessed inputs to ensure fairness, and evaluation results will guide final model selection.

EVALUATION FRAMEWORK

Metrics (with justification): F1 (primary; balances precision/recall under moderate class imbalance), Accuracy (overall), Precision and Recall (clinical trade-offs), and ROC–AUC (threshold-agnostic separability).

Data split: Stratified 80/20 train/test, with 5-fold CV on the training data.

Baselines: Majority-class predictor and Logistic Regression (implemented).

Preliminary baselines:

- Majority-Class: Accuracy = 0.600, F1 = 0.750
- Logistic Regression: Accuracy = 0.861, F1 = 0.885

Success criteria: Gradient Boosting should improve F1 by $\geq$0.01 and maintain Recall $\geq$0.88. If performance gain is marginal, the simpler logistic model will be preferred for interpretability.

Model-selection rationale: The model achieving the highest mean F1-score across cross-validation folds, with stable variance and strong ROC–AUC, will be chosen as the final classifier for deployment.

TABLE I: Planned comparison of models (final metrics to be reported in the completed project).

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC–AUC**
Majority Class	0.600	–	–	0.750	–
Logistic Regression	0.861	0.88	0.90	0.885	(TBD)
Random Forest	(TBD)	(TBD)	(TBD)	(TBD)	(TBD)
Gradient Boosting	(TBD)	(TBD)	(TBD)	(TBD)	(TBD)

TIMELINE AND MILESTONES

The project spans November–December 2025 and follows the CMSE 492 structure:

- **Weeks 10–11:** Data acquisition, cleaning, and EDA (*complete*).
- **Weeks 11–12:** Model development and pipeline construction.
- **Week 13:** Model comparison and tuning.
- **Week 14:** Interpretation (feature importance, SHAP) and presentation prep.
- **Week 15:** Final report and GitHub submission (Dec 8).

Critical path: finalize preprocessing → train → tune → evaluate → interpretability → report. EDA and writing proceed in parallel; Week 13 buffer time covers re-training or data issues.

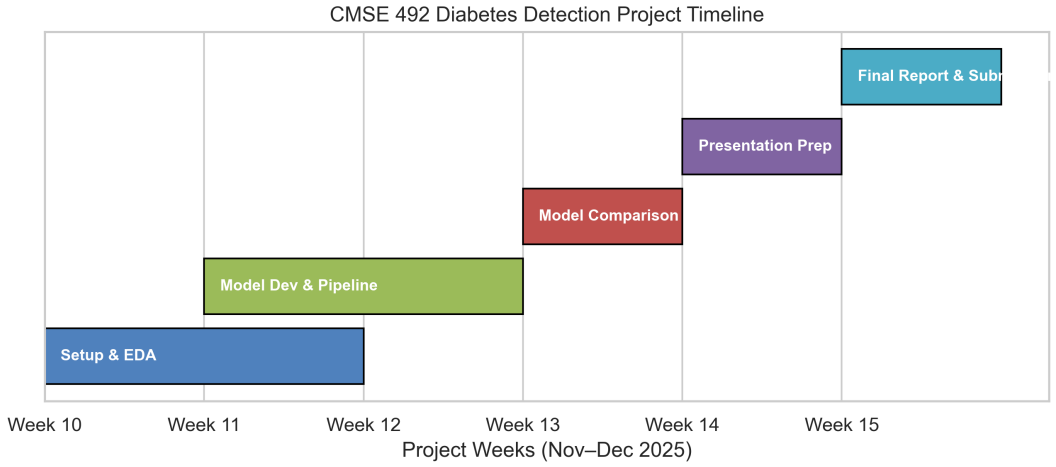

FIG. 8: Planned project timeline and milestone sequence (Nov–Dec 2025).

CONCLUSION

The exploratory analysis confirmed strong data quality and meaningful feature relationships. Baseline Logistic Regression achieved 0.861 accuracy and 0.885 F1, validating the modeling approach. The next phase will train and compare Random Forest and Gradient Boosting using identical pipelines, evaluate using F1 and ROC–AUC, and produce feature-importance interpretations via SHAP. This project aims to deliver a reproducible,

interpretable ML system supporting early diabetes detection and preventive healthcare analytics.

REFERENCES

* sleepern@msu.edu

[1] M. K. Thalla, "Diabetes Health Indicators Dataset," *Kaggle*, 2023. https://www.kaggle.com/datasets/mohankrishnathalla/diabetes-health-indicators-dataset

[2] Centers for Disease Control and Prevention (CDC), "Behavioral Risk Factor Surveillance System (BRFSS)," 2023. https://www.cdc.gov/brfss/