REPORT

Comparative Performance Evaluation of Logistic Regression, Random Forest, and XGBoost Models for Diabetes Prediction on the CDC BRFSS Dataset:

1. Background

Diabetes mellitus is one of the most prevalent chronic diseases worldwide, affecting more than 537 million adults and projected to reach 783 million by 2045. In the United States alone, approximately 38 million adults live with diabetes, many of whom are undiagnosed until complications arise. Early detection is therefore critical for prevention and management.

Traditional diagnostic methods, such as fasting plasma glucose and HbA1c are accurate but reactive, identifying the disease only after symptoms appear. This creates a gap for proactive, data-driven screening methods capable of identifying at-risk individuals earlier and more efficiently.

Advancements in **machine learning (ML)** and **artificial intelligence (AI)** have enabled predictive modelling using population health data. This study applies and compares three ML algorithms **Logistic Regression**, **Random Forest**, and **XGBoost** to the **CDC Diabetes Health Indicators dataset** from the **Behavioural Risk Factor Surveillance System (BRFSS)**. The objective is to assess their effectiveness in predicting diabetes using non-clinical, self-reported data and to determine the most influential risk factors.

2. Objectives

The study aims to:

1. Develop and evaluate three machine learning models for diabetes prediction.
2. Compare their performance in terms of **accuracy**, **precision**, **recall**, and **F1-score**.
3. Identify the most influential features associated with diabetes risk.
4. Assess the potential of ML for early detection and public health applications.

Key research questions included:

- Which features best predict diabetes using BRFSS data?
- Which model performs best for multi-class classification (no diabetes, pre-diabetes, diabetes)?
- Can ensemble methods mitigate the challenge of class imbalance in health survey data?

3. Methodology

3.1 Dataset

The study utilized the **2015 CDC Diabetes Health Indicators dataset**, sourced from the **UCI Machine Learning Repository**. It includes **253,680 records** and **21 variables** encompassing demographic, behavioural, and health-related factors.

The target variable, **`Diabetes_012`**, contains three categories:

- 0 = No diabetes (87% of cases)
- 1 = Pre-diabetes (3.4%)

- 2 = Diabetes (9.6%)

The severe **class imbalance** presented a major challenge for model performance and fairness.

3.2 Data Processing

Data cleaning was minimal due to the pre-processed nature of the dataset. Continuous variables (e.g., BMI, age) were normalized for Logistic Regression, while Random Forest and XGBoost used unscaled features. Exploratory analysis showed strong correlations between diabetes and factors such as **high blood pressure**, **general health**, **BMI**, and **cholesterol**.

3.3 Model Development

Three supervised classification algorithms were trained using an **80/20 stratified train-test split**:

- **Logistic Regression:** Linear baseline model emphasizing interpretability.
- **Random Forest:** Ensemble of decision trees reducing overfitting and capturing non-linear patterns.
- **XGBoost (Extreme Gradient Boosting):** Sequentially optimized model minimizing prior errors with built-in regularization for robustness.

3.4 Evaluation Metrics

Performance was measured using:

- **Accuracy** (overall correctness)
- **Precision**, **Recall**, and **F1-score** (per-class and macro averages)
- **Confusion Matrix** (error distribution)
- **Feature Importance** (for tree-based models)

4. Results

4.1 Model Performance

Model	Accuracy	F1 (Macro)	Strengths	Weaknesses
Logistic Regression	**64.9%**	0.43	Simple, interpretable	Poor minority class detection.
Random Forest	**84.3%**	0.40	Robust, handles non-linearity	Weak on pre-diabetes.
XGBoost	**84.97%**	0.40	Best accuracy, balanced predictions	Limited pre-diabetes sensitivity.

The **XGBoost model** achieved the highest accuracy approximately (85%) and the most balanced classification across classes, outperforming Logistic Regression by over 20 percentage points.

Although all models struggled to identify **pre-diabetes** (due to extreme imbalance), XGBoost slightly improved detection for confirmed diabetes cases.

4.2 Feature Importance

XGBoost identified the following as the top predictors:

1. **High Blood Pressure (HighBP)**
2. **General Health (GenHlth)**
3. **High Cholesterol (HighChol)**

4. **Cholesterol Check Frequency (CholCheck)**
5. **BMI and Age**

Behavioural factors such as physical inactivity, walking difficulty, and alcohol consumption also contributed meaningfully. These results are consistent with established clinical knowledge, reinforcing the reliability of the models.

5. Discussion

5.1 Model Insights

The study confirms that **ensemble learning** (Random Forest, XGBoost) significantly outperforms linear models for predicting diabetes from complex, non-linear data. XGBoost's gradient boosting mechanism and regularization produced superior accuracy and generalization while mitigating overfitting.

However, the limited performance in identifying pre-diabetic individuals suggests that **class imbalance correction** (e.g., SMOTE or reweighting) is necessary for fairer and more sensitive predictions.

5.2 Feature and Risk Analysis

Feature importance analysis aligns with existing medical research: individuals with **hypertension**, **poor general health**, **high cholesterol**, or **obesity (high BMI)** face elevated diabetes risk. Socioeconomic and behavioural attributes (education, income, exercise) further shape outcomes, highlighting the value of integrating **social determinants of health** into predictive models.

5.3 Implications for Healthcare

Machine learning models trained on survey-based, non-clinical data can serve as early warning tools in public health systems. They can be embedded in:

- **Mobile health (mHealth)** and **telemedicine** applications for population screening.
- **Electronic Health Records (EHR)** systems for automated risk flagging.
- **Policy-driven resource allocation**, targeting underserved or high-risk communities.

These applications can help reduce healthcare costs and improve early intervention through preventive care strategies.

5.4 Limitations

- **Class imbalance** significantly reduced detection accuracy for pre-diabetes.
- **Self-reported data** introduce potential bias and recall errors.
- **Temporal constraint:** the 2015 dataset may not represent current health behaviours.
- **Model transparency:** while XGBoost performs best, it operates as a "black box" unless supplemented with explainable AI methods like **SHAP** or **LIME**.

6. Conclusion and Recommendations

This study demonstrates the viability of using **machine learning** for diabetes risk prediction using **non-invasive, self-reported data**. Among the three tested models, **XGBoost achieved the highest performance approximately (85% accuracy)** and most effectively balanced predictive precision with generalization.

Key findings include:

- High blood pressure, general health, and cholesterol are the strongest predictors of diabetes.
- Ensemble models outperform traditional statistical methods for complex health datasets.
- ML models can complement traditional diagnostics by enabling **early, scalable, and cost-effective screening**.

Recommendations for Future Work

1. **Data Enhancement:** Incorporate multi-year BRFSS data and clinical biomarkers for improved temporal and biological relevance.
2. **Imbalance Handling:** Apply **oversampling (SMOTE)** and **class-weight adjustments** to improve minority-class sensitivity.
3. **Advanced Models:** Explore **deep learning** architectures and **hybrid ensembles** for better feature interaction modelling.
4. **Explainability:** Integrate **XAI tools** to enhance interpretability and trust in AI-driven health decisions.
5. **Deployment:** Develop web or mobile-based predictive dashboards for public health monitoring.