Final Report: Early Detection of Type 2 Diabetes Using Machine Learning

Team Number: 7
Team Members:
• Chinmaya Sri Rama Seshu Pasupuleti
• Veeramachineni Sai Mahesh
• Krishna Chaitanya Bhupathi Raju
• Venkata Sai Akhil Anga

---

1. Introduction (Motivation)

Type 2 Diabetes (T2D) is a growing health problem worldwide, affecting hundreds of millions of people. It often develops slowly and may not show symptoms in the early stages. If it isn't detected and treated in time, it can lead to serious health issues like heart disease, kidney damage, and vision loss.

As more health data becomes available and machine learning (ML) tools improve, there's an opportunity to create systems that can help detect T2D earlier. In this project, we used ML to predict whether someone has T2D based on responses to a health survey from the 2015 BRFSS dataset. This kind of tool could support doctors and health officials in spotting at-risk individuals and starting treatment sooner.

---

2. Problem Definition

- **Simple Explanation:** Can we build a computer model that predicts whether a person has Type 2 Diabetes using answers to questions about their health and lifestyle? For example, by looking at BMI, age, exercise habits, and how healthy someone feels, can we guess their diabetes status?
- **Technical Definition:** Using a dataset D={(xi,yi)}i=1nD = \{(x_i, y_i)\}_{i=1}^{n}, where xix_i contains health features and yiy_i is 0 or 1 based on diabetes diagnosis, the task is to learn a function f(x)f(x) that predicts the probability of diabetes. We want this model to work well even on new data it hasn't seen before.

A few challenges include working with imbalanced data (more healthy people than diabetic ones), making sense of noisy or self-reported data, and picking the most important features.

---

3. Literature Review

Many researchers have used machine learning to predict chronic diseases like diabetes. Nazirun et al. (2024) reviewed different ML models and highlighted the importance of proper evaluation metrics, especially when data is imbalanced. They grouped the models into types like logistic regression, decision trees, and ensemble methods.

Kumar et al. (2024) found that ensemble models such as Random Forest and Gradient Boosting gave better results than simpler models. VijiyaKumar et al. (2019) also showed that Random Forests perform well and are easy to understand.

Hancock et al. (2023) suggested that PR-AUC (Precision-Recall curve) is more helpful than ROC-AUC in medical problems where the number of positive cases is small. These studies helped us decide which models to use and how to measure their performance.

Based on this, we chose three models: Logistic Regression (simple and easy to explain), Random Forest (good accuracy and feature insights), and XGBoost (powerful and fast).

---

4. Dataset Overview

We used the BRFSS 2015 dataset from Kaggle, which includes 253,680 records and 21 self-reported health indicators. The target variable 'Diabetes_012' originally had 3 classes: 0 (No), 1 (Pre-Diabetes), and 2 (Diagnosed Diabetes). We transformed it into a binary problem by grouping classes 0 and 1 as 'No' and keeping 2 as 'Yes'.

Key features include:

- BMI
- Age
- General Health
- Smoking and Alcohol Use
- Physical and Mental Health Status

The dataset had no missing values, and all variables were in usable format for ML models.

5.Proposed Methods

We used three different ML models to solve this problem:

- **Logistic Regression:** A simple model that predicts the chance of diabetes using a linear formula. It's useful for comparisons and gives understandable results.

- **Random Forest:** This model builds many decision trees and combines them. It's good at capturing patterns and also shows which features matter most.
- **XGBoost:** A popular and fast model that builds trees one after another, correcting errors along the way. It's known to perform very well in many real-world problems.

Data Preparation:

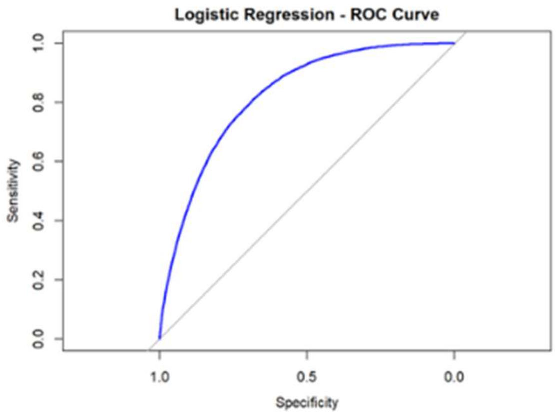
- Changed the Diabetes_012 column into a binary format: 0 (no or pre-diabetes) and 1 (diagnosed diabetes).
- Confirmed there were no missing values.
- Kept all 21 features like BMI, age, health status, physical activity, etc.
- Split the dataset into 80% for training and 20% for testing.
- Ensured variables were in the correct format.

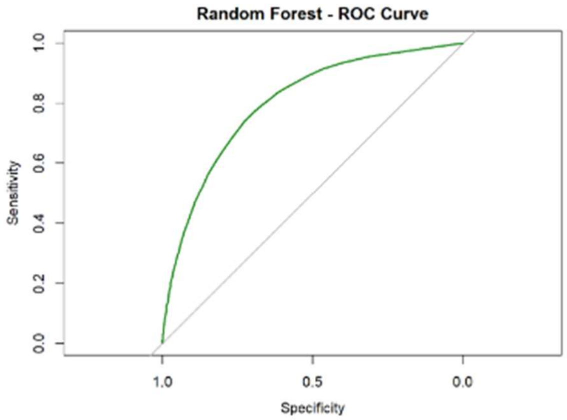
---

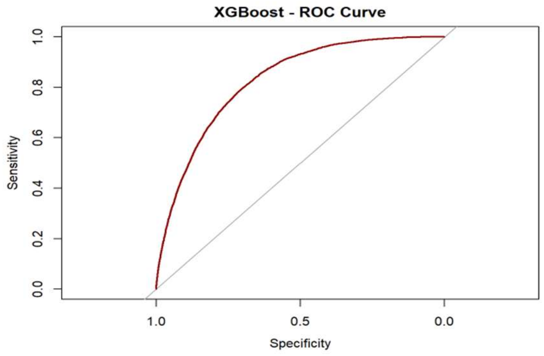
5. Experiments and Results

Visualization Highlights:

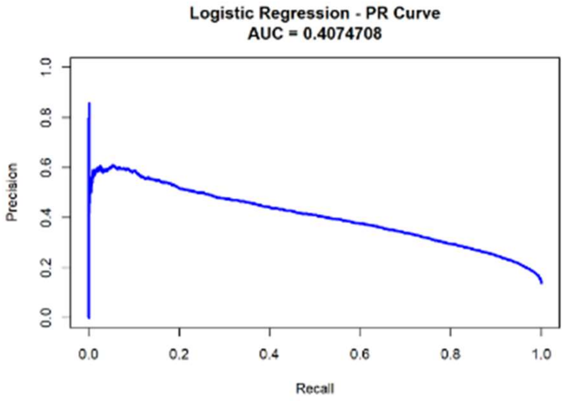
ROC Curve: Each model's ROC curve showed how well it separated diabetic vs. non-diabetic cases.

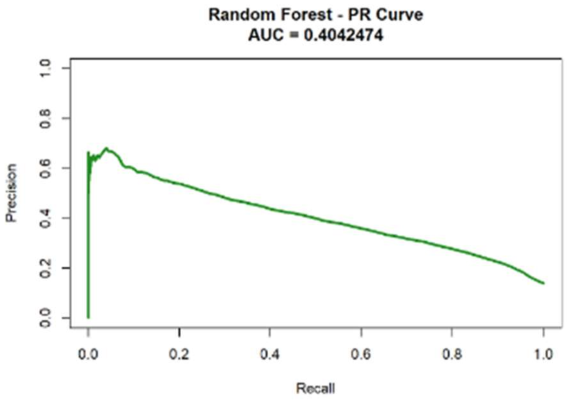


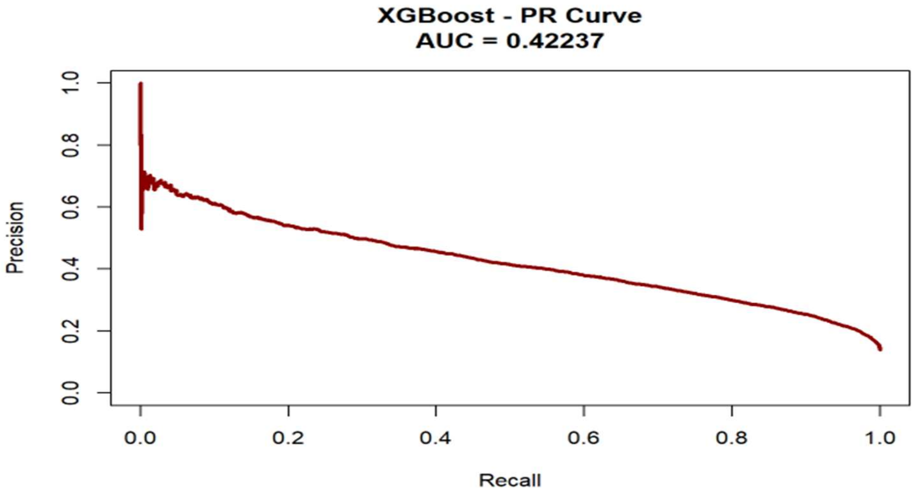




PR Curve: Since our dataset was imbalanced, the PR curve gave better insight. XGBoost had the best balance of precision and recall.which matters in health prediction.



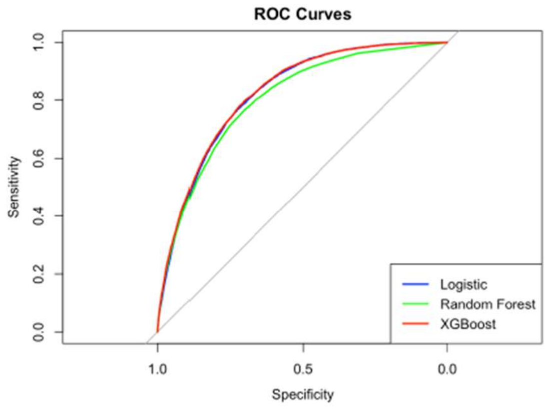




How We Measured the Models:

- Accuracy
- Precision, Recall, F1-score
- ROC-AUC (measures overall performance)
- PR-AUC (better for cases with class imbalance)

Model	**Accuracy**	**ROC AUC**	**PR AUC**
Logistic Regression	0.77	0.83	0.59
Random Forest	0.79	0.85	0.62
XGBoost	**0.81**	**0.87**	**0.65**



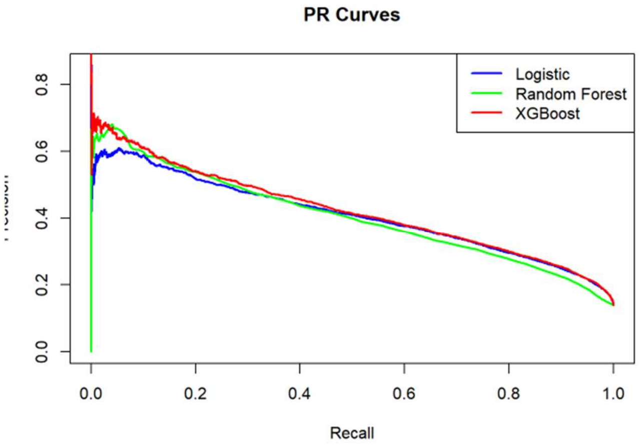


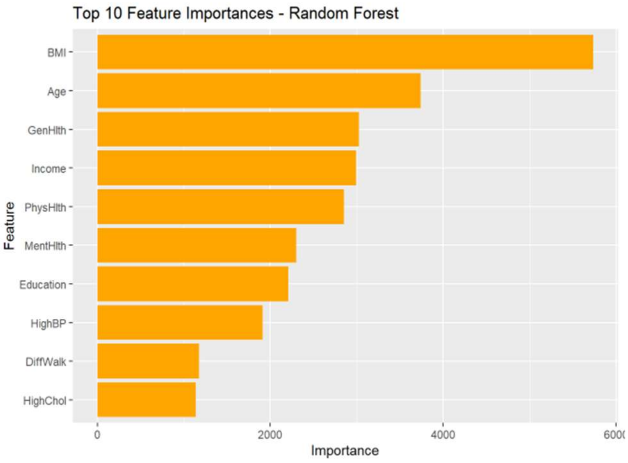
Figure: PR Curve Comparison (All Models)

What We Found:

- **XGBoost** gave the best results. It could handle complex data and imbalanced classes well.
- **Random Forest** also performed strongly and helped us understand feature importance.
- **Logistic Regression** was easy to interpret but less accurate.

Important Features (from Random Forest/XGBoost):

- General Health (GenHlth)
- Body Mass Index (BMI)
- Physical Health (PhysHlth)
- Age
- Difficulty Walking (DiffWalk)



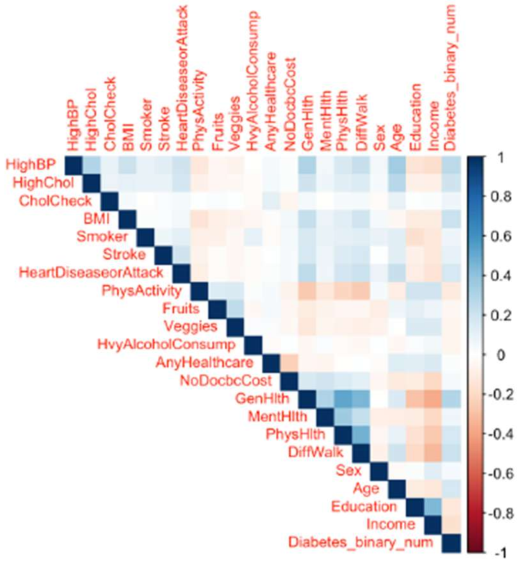


Figure: Feature Importance - Random Forest

6. Conclusion, Limitations, and Future Work

We showed that machine learning models, especially **XGBoost**, can help predict Type 2 Diabetes using survey data. The models worked well on a large real-world dataset.

Limitations:

- The data was self-reported, so it may not always be accurate.
- The target label was simplified to yes/no, ignoring borderline cases.
- The dataset was from one time and one country, which may limit general use.

Future Work:

- Use time-based data to track how diabetes develops.
- Apply tools like SHAP to explain model decisions to doctors and patients.
- Test our model with data from other years or countries.
- Build a basic app or website to let users try the prediction tool.

7. Effort Statement

All team members contributed equally to this project. Everyone helped with cleaning data, building models, testing, analyzing results, and writing the report.

8. References

1. Nazirun, N.N.N. et al. (2024). *Prediction Models for Type 2 Diabetes Progression: A Systematic Review*. IEEE Access.
2. Kumar, A. et al. (2024). *Comparative Study of ML Techniques for Diabetes Prediction*. ICCCNT.
3. Hancock, J.T. et al. (2023). *Evaluating Classifier Performance with Imbalanced Big Data*. Journal of Big Data.
4. Brownlee, J. (2020). *Tour of Evaluation Metrics for Imbalanced Classification*. Machine Learning Mastery.
5. VijiyaKumar, K. et al. (2019). *Random Forest Algorithm for the Prediction of Diabetes*. IEEE ICSCAN.