Early Detection of Type 2 Diabetes Using Machine Learning

TEAM NUMBER 007

1. Chinmaya Sri Rama Seshu Pasupuleti 2. Veeramachineni Sai Mahesh
3. Krishna Chaitanya Bhupathi Raju 4. Venkata Sai Akhil Anga

Introduction

The Type 2 Diabetes (T2D) is a growing global health issue. Many people remain undiagnosed due to the absence of early symptoms. Delayed diagnosis can lead to severe complications. Our project aimed to use machine learning (ML) techniques to predict T2D early using the BRFSS2015 survey dataset.

Dataset Overview

- **Source: BRFSS 2015 (Kaggle/UCI)**
- **Records: 253,680**
- **Features: 21 health indicators**
- **Target: Binary (0 = No Diabetes/Pre-Diabetes, 1 = Diagnosed Diabetes)**

• Key features include:
1. BMI
2. Age
3. General Health
4. Smoking and Alcohol Use
5. Physical and Mental Health Status

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.

Proposed Methods

- **We used three different ML models to solve**
- **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.**

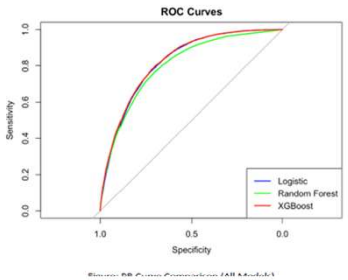
Data Preparation:

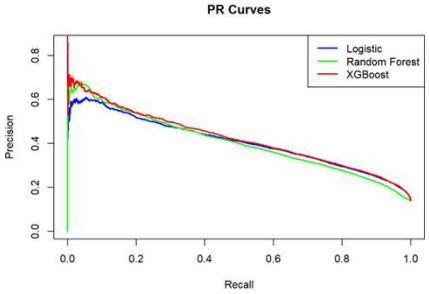
- **Changed the Diabetes_012 column into a binary format: 0 (no or pre-diabetes) and 1 (diagnosed diabetes).**
- **Confirmed there were no missing values.**
- **Kept all 21 features like BMI, age, health status, physical activity, etc.**
- **Split the dataset into 80% for training and 20% for testing.**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

Experiments and Results

Visualization Highlights:

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





Unamcorper efficitur sed in nulla.

Results

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

How We Measured the Models:

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

Conclusion

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.**

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.