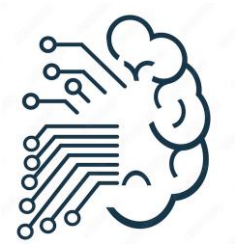
Early Detection of Diabetes: A Machine Learning Approach Using Health Indicators

Ram Bagaria

Shrobanti Banerjee

Introduction

Objective: Predict diabetes using health indicators.



Preprocessing: PCA+ Oversampling technique

Approach: Machine Learning models + Neural networks.

Dataset Overview

Source: BRFSS 2015 Dataset

Features: Health indicators (BMI, Smoking, Physical Activity, etc.)

Target Variable: Diabetes_012

0: No Diabetes

1: Prediabetes

2: Diabetes

Problem Statement

- Build ML models to **predict diabetes** in patients.
- Work with a dataset that has a **class imbalance** (fewer diabetic cases).
- Ensure the model performs well for **3 classes**, especially the minority (pre-diabetic) class.
- Explore and compare:
  - **ML models** (Random Forest, XGBoost)
  - **Neural Networks**

Purpose of the code

- Apply various **preprocessing techniques**:
  - Scaling
  - PCA (dimensionality reduction)
- Use **data balancing methods**:
  - SMOTE
  - ADASYN
- Optimize model performance using **Bayesian Optimization**.
- Goal: Achieve a **balanced, accurate** model that handles class imbalance effectively.

Preprocessing Steps

NULL VALUES HANDLED, PLACEHOLDERS CLEANED.

STANDARDSCALER FOR NORMALIZATION.

PCA FOR DIMENSIONALITY REDUCTION.

SMOTE & ADASYN TO BALANCE CLASSES.

ML Models & Workflow

1. Split → StandardScaler → PCA → SMOTE

2. Models: Random Forest, XGBoost

3. Hyperparameter tuning: Bayesian Optimization

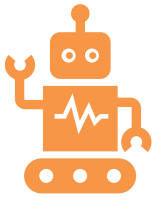
Neural Network Experiments

- Trained directly on imbalanced data — poor prediction for class 1.
- SMOTE applied on training data (No Scaling, No PCA) → Better performance.
- ADASYN applied on training data (No Scaling, No PCA) → Improved minority class prediction.

Challenges

Class Imbalance impacted model performance

Oversampling improved class detection.



Scaling & PCA boosted RF & XGBoost, skipped in NN experiments.

Conclusion- Performance Evaluation Metrics

MODEL	OVER SAMPLING	ACCURACY	PRECISION	RECALL	F1- SCORE
Random Forest	SMOTE	64	85	64	71
XGBoost	SMOTE	64	84	64	71
Neural Networks	SMOTE	70	83	70	73
Neural Networks	ADASYN	73	81	73	75
Neural Networks	Imbalance	85	80	85	81

Links

- Github - Github - Diabetes_Prediction_Files

- Youtube – Diabetes_Prediction

Thank you!