Predicting Diabetes Risk Using Behavioral and Health Data

Final Project – ISOM 835: Predictive Analytics and Machine Learning
Anushka C. | Suffolk University | Spring 2025

---

1. Introduction

Diabetes remains one of the most pressing public health challenges globally, with rising incidence linked to lifestyle, behavioral, and demographic factors. The ability to accurately identify individuals at high risk of developing diabetes offers immense potential for targeted interventions, early diagnosis, and optimized healthcare resource allocation.

This project leverages data from the **Behavioral Risk Factor Surveillance System (BRFSS)** to build a predictive model for diabetes classification. By integrating machine learning techniques, the aim is to understand which behavioral and health factors contribute most to diabetes risk and to construct a model capable of identifying individuals as diabetic, borderline, or non-diabetic based on these features.

---

2. Dataset and Business Understanding

About the BRFSS Dataset

The **Behavioral Risk Factor Surveillance System (BRFSS)** is the world's largest continuously conducted health survey system, developed by the **Centers for Disease Control and Prevention (CDC)**. It gathers data annually from over 400,000 adults across the United States via telephone surveys, collecting insights on health-related behaviors, chronic health conditions, and preventive services.

For this project, a subset of the **2015 BRFSS data** has been used. This dataset includes key attributes such as age, sex, BMI, smoking and alcohol behavior, physical activity, general and mental health, and various chronic conditions. The primary target variable is **Diabetes_012**, a three-class variable representing:

- **0** = No diabetes
- **1** = Borderline/pre-diabetes
- **2** = Diagnosed diabetes

Business Objective

The goal is to develop a robust, interpretable predictive model that can:

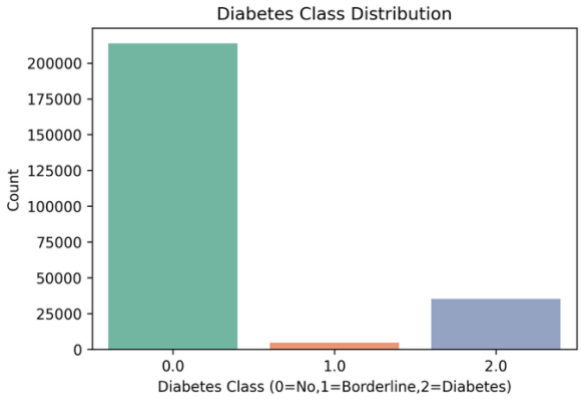
- Classify individuals into one of the three diabetes categories.
- Highlight the most influential risk factors.
- Offer actionable insights for public health interventions.

From a broader perspective, the model can be adapted by healthcare organizations to identify at-risk populations and personalize outreach programs, reducing long-term healthcare burdens.

3. Data Exploration and Preparation

Before modeling, we conducted a comprehensive exploratory data analysis (EDA) to understand the dataset's structure, identify patterns, and handle issues like class imbalance, missing values, and skewed distributions.

3.1 Diabetes Class Distribution



Description:
The bar chart above shows the number of records in each diabetes category:

- **0 = No diabetes**
- **1 = Borderline (pre-diabetes)**
- **2 = Diagnosed diabetes**

Insight:
The dataset is heavily imbalanced:

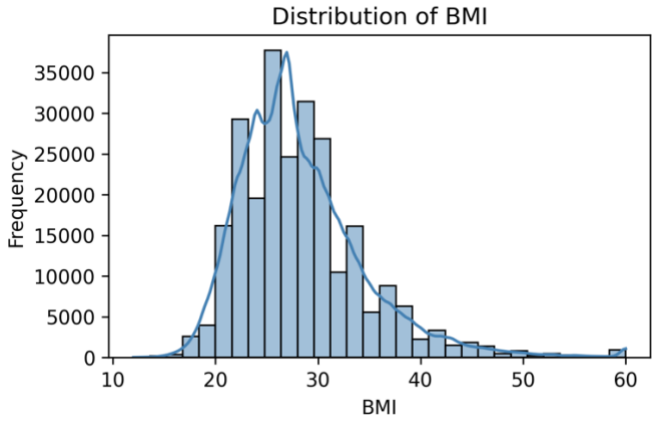
- ~84% of individuals fall into the **non-diabetic** category
- Only ~2% are **borderline**
- ~14% are **diabetic**

Implication:
This class imbalance poses challenges for predictive modeling, especially for detecting borderline cases. We address this later using techniques like **class weighting** and **SMOTE** for resampling.

3.2 Univariate Distributions of Key Features

We explored the distribution of key numeric variables to assess their spread and detect outliers or skew.

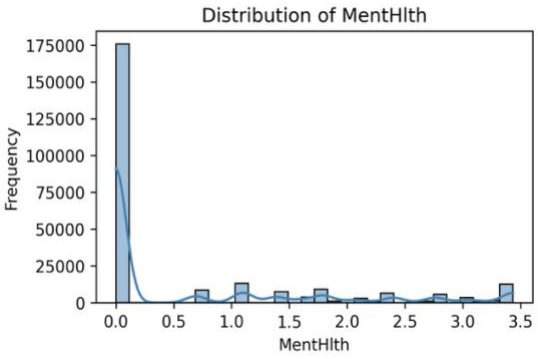
Distribution of BMI



Insight:

- The BMI distribution is **right-skewed**, with most values ranging from 20–40 kg/m².
- A long tail extends toward higher BMI values, reflecting the presence of obesity in a portion of the population.

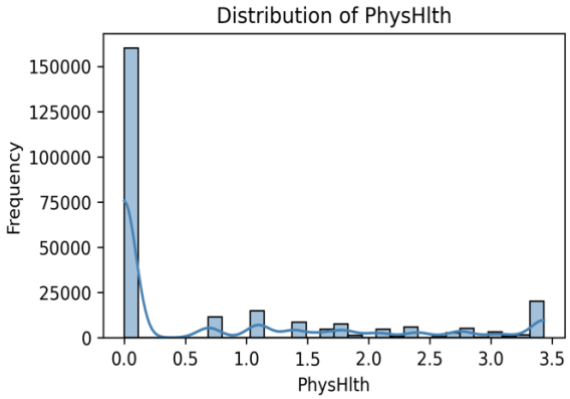
Distribution of Mental Health (Mentally Unhealthy Days)



Insight:

- A significant number of individuals reported **0 days** of poor mental health.
- The long tail suggests that a smaller group experiences extended periods of mental distress.

Distribution of PhysHlth (Physically Unhealthy Days)



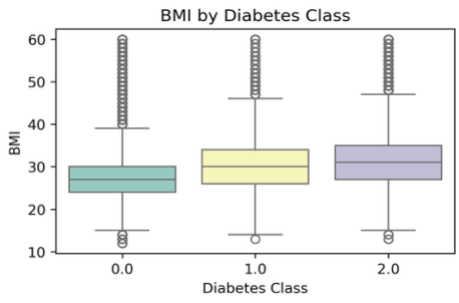
Insight:

- Like mental health, most participants reported **few or no physically unhealthy days**, while fewer experienced extended illness or physical challenges.

3.3 Bivariate Analysis by Diabetes Class

We explored how key variables vary by diabetes class (`0 = No diabetes`, `1 = Borderline`, `2 = Diabetes`). These comparisons help identify features that are most likely to influence the classification outcome.

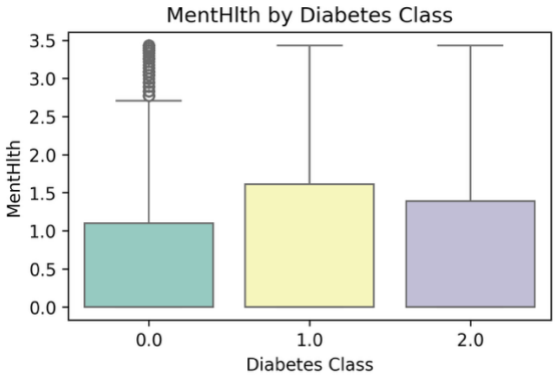
BMI by Diabetes Class



Boxplots comparing BMI across the three diabetes classes, demonstrating that median BMI increases from healthy to borderline to diabetic groups.

Insight:
There's a clear upward trend in BMI from non-diabetic to diabetic groups. This supports the well-established link between obesity and diabetes risk. Median and upper quartile BMI values are significantly higher for diabetic individuals.

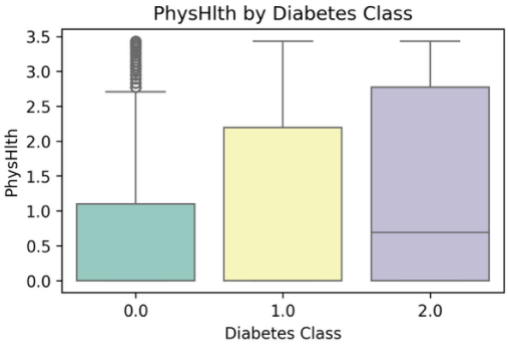
Mental Health by Diabetes Class



Boxplots of log-transformed mentally unhealthy days stratified by diabetes status, showing a modest rise in median poor-mental-health days from healthy to diabetic.

Insight:
Diabetic individuals tend to report more mentally unhealthy days than those in the healthy group. The difference is not as stark as BMI but is still notable.

Physical Health by Diabetes Class



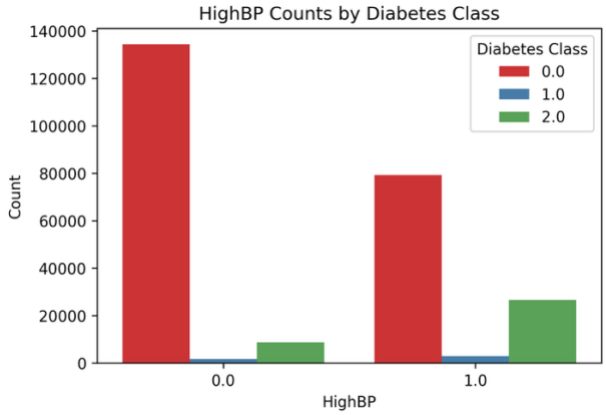
Boxplots of log-transformed physically unhealthy days by diabetes category, indicating that those with diabetes report more poor-health days than borderline or healthy individuals.

Insight:
The diabetic class reports significantly higher values of poor physical health days, which likely reflects both symptoms of diabetes and comorbid conditions such as fatigue or heart disease.

Binary Feature Count plots by Class

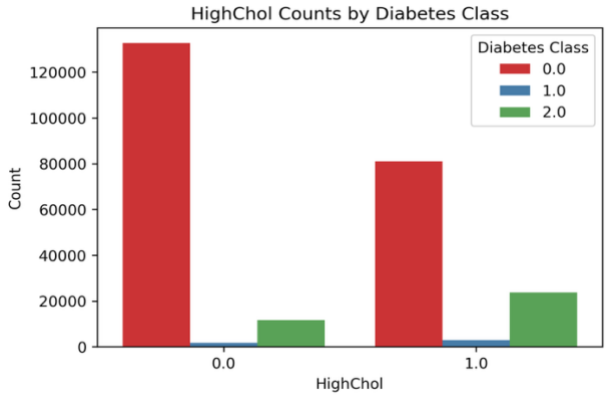
These grouped bar charts show how the prevalence of certain binary health indicators differs across diabetes status.

High Blood Pressure by Diabetes Class



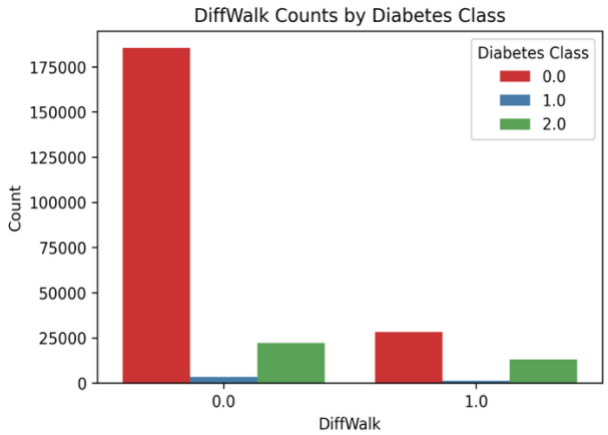
Grouped count plot of High Blood Pressure status by diabetes class, showing a steep rise in high BP rates among diabetics.

High Cholesterol by Diabetes Class



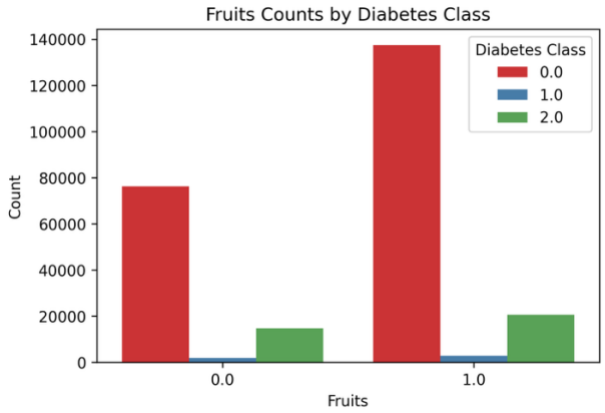
High cholesterol prevalence is much higher in the diabetic group, reinforcing the link between metabolic syndrome components.

Difficulty Walking by Diabetes Class



Bar chart showing that individuals with diabetes are significantly more likely to report difficulty walking.

Fruit Intake by Diabetes Class

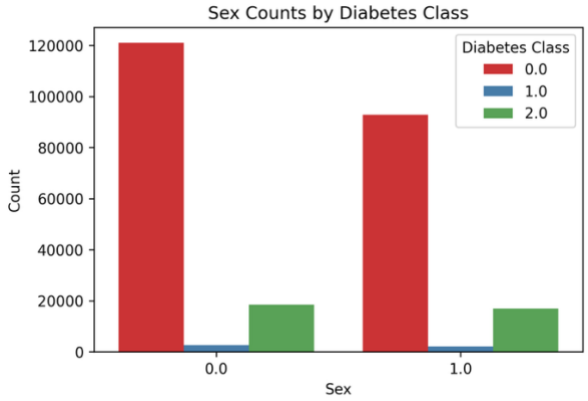


Fruit consumption does not vary strongly across diabetes classes, suggesting limited predictive value on its own.

3.4 Additional Feature Count plots by Diabetes Class

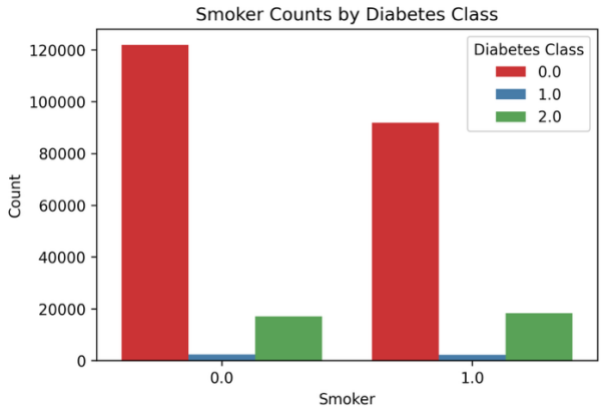
These additional grouped bar charts explore patterns in variables such as gender, smoking behavior, physical activity, vegetable consumption, and general health — helping us better understand lifestyle influences on diabetes risk.

Sex by Diabetes Class



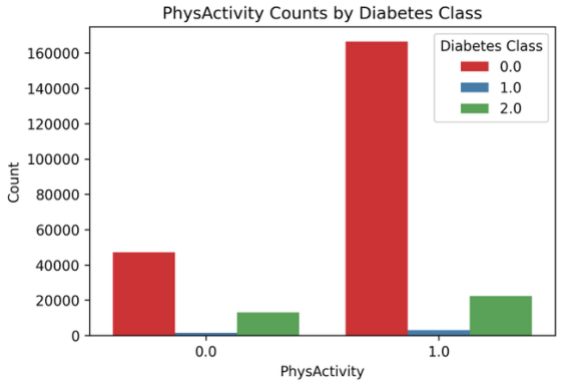
Count plot showing distribution of sex (male/female) across diabetes categories. Proportions are relatively consistent, indicating that gender may not be a strong standalone predictor.

Smoking Status by Diabetes Class



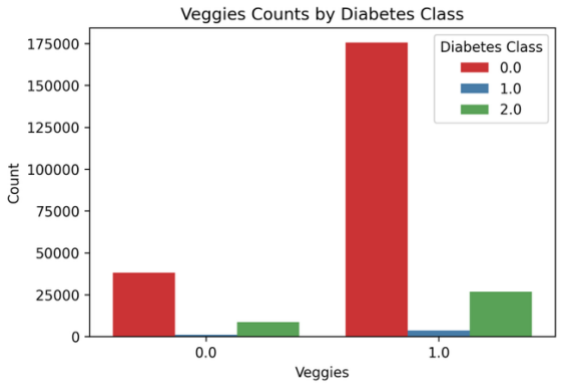
Smokers are slightly more prevalent in the diabetic group compared to non-diabetics, though differences are moderate.

Physical Activity by Diabetes Class



Those reporting physical activity are more common in the non-diabetic class, suggesting a protective effect of regular movement.

Vegetable Consumption by Diabetes Class

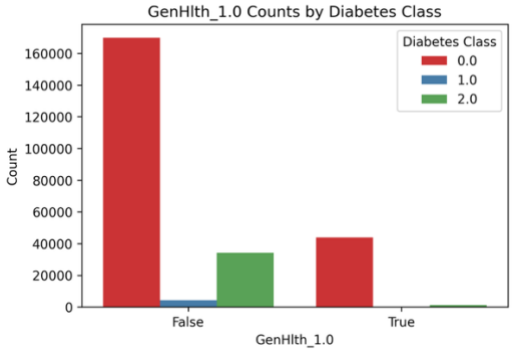


Vegetable consumption shows only mild variation across diabetes groups, indicating low predictive strength in isolation.

General Health Levels by Diabetes Class (1–5)

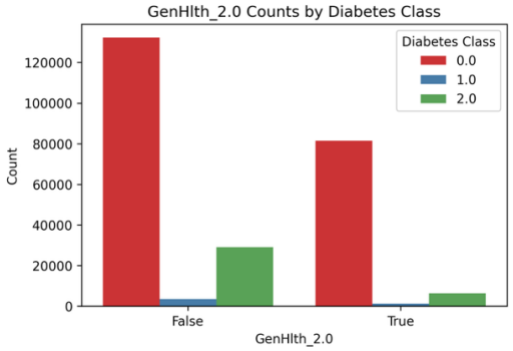
This variable was one of the most insightful. The BRFSS allows respondents to self-rate their general health from 1 (excellent) to 5 (poor). Here are the count plots for each level:

Gen Health = 1 (Excellent)



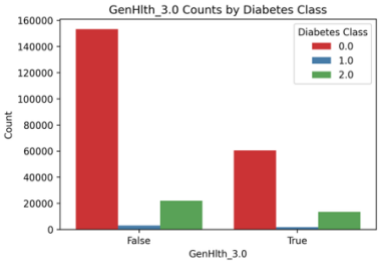
Excellent general health is vastly more common among the non-diabetic group, with very few diabetics rating their health this high.

General health = 2 (Very Good)



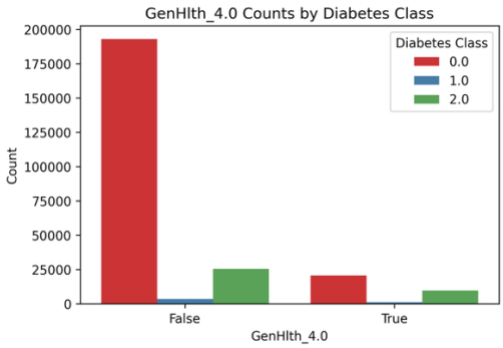
Very good general health ratings remain concentrated in the non-diabetic class but appear slightly in borderline participants.

General health = 3 (Good)



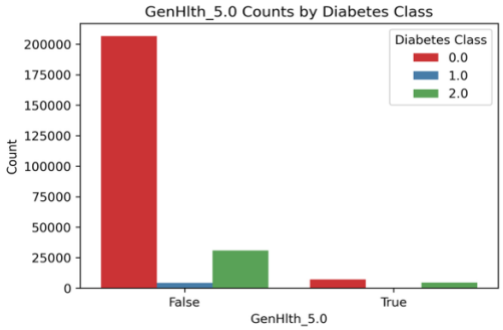
A balanced distribution across groups; many borderline and diabetic individuals self-rate their health as ‘good.’

General health = 4 (Fair)



Fair health ratings are more common in diabetics, with a visible decline in the healthy group.

General health = 5 (Poor)



Poor health ratings are almost exclusively reported by individuals in the diabetic class.

Key EDA Insights

1. **Severe Class Imbalance**
    a. With only ~2% "borderline" and ~14% "diabetes," models trained naïvely will favor the healthy class. We will need class weighting, resampling, or threshold tuning.
2. **Skewed Continuous Variables**
    a. MentHlth & PhysHlth have heavy right tails (many zeros, outliers up to 30). We will apply `log1p` transforms and clip outliers to [0,30] for stability.
3. **BMI & Age as Strong Predictors**
    a. BMI and Age both increase monotonically with diabetes class. They will be key inputs and should be scaled for models sensitive to feature ranges.
4. **HighBP & HighChol Signals**
    a. Binary flags for high blood pressure and cholesterol checks show the largest class-conditional differences—excellent features for early discrimination.
5. **Low Multicollinearity**

a. No pair of features has correlation >0.5, so we can include all engineered variables (e.g. BMI×PhysHlth) without fear of redundancy degrading model performance.

Data Cleaning & Preprocessing

Summary of Steps & Rationale:

1. **Handle Missing Values:**
   - No missing entries were present; we verified with `df.isnull().sum()`.
2. **Outlier Treatment:**
   - **BMI** clipped to [12, 60] to remove biologically implausible values.
   - **MentHlth** and **PhysHlth** clipped to [0, 30] days to match survey bounds.
3. **Feature Engineering:**
   - **Log-transform** skewed counts:
   - `for c in ['MentHlth','PhysHlth']:`
   - `    df[c] = np.log1p(df[c])`

     This reduces the influence of long right-tails.

   - **Interaction term** for combined risk:
   - `df['BMI_x_PhysHlth'] = df['BMI'] * df['PhysHlth']`
4. **Encoding Categorical Variables:**
   - One-hot encode self-rated general health (`GenHlth`):
   - `df = pd.get_dummies(df, columns=['GenHlth'], prefix='GenHlth')`
   - (Binary flags and ordinal Income/Education left as numeric codes.)
5. **Scaling Continuous Features**
6. **Train/Test Split:**
   - Stratified 80/20 split on `Diabetes_012` to preserve class ratios:

Formulating Business Analytics Questions

1. **Pre-diabetes Identification:**
   "What recall/precision trade-off can we achieve when flagging borderline (pre-diabetes) individuals using only self-reported health indicators?"
2. **Key Risk Factors:**
   "Which top three modifiable factors (e.g. BMI, physical activity, diet) most strongly predict progression from borderline to diagnosed diabetes?"
3. **Screening Resource Allocation:**
   "Given our model's false-positive rate, how many follow-up screenings per 1,000 adults will be required, and how does that number change if we adjust the decision threshold?"

Predictive Modeling

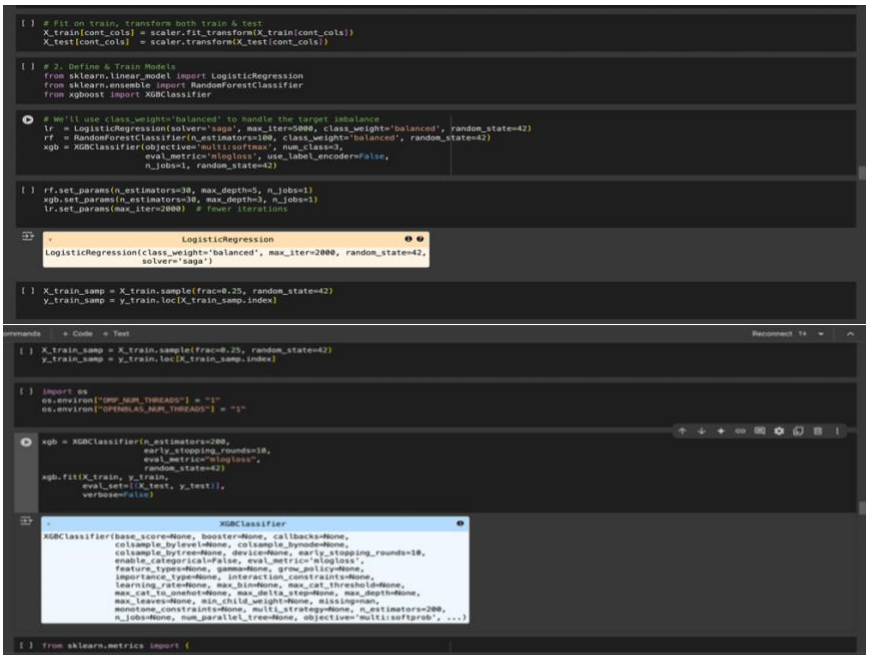
We built and evaluated three machine learning models to classify individuals into one of the three diabetes categories:

- **Logistic Regression** (baseline linear model)
- **Random Forest** (ensemble of decision trees)
- **XGBoost** (gradient-boosted trees)

Each model was trained on the cleaned and engineered dataset using both **class weighting** and **sampling techniques** to address class imbalance.

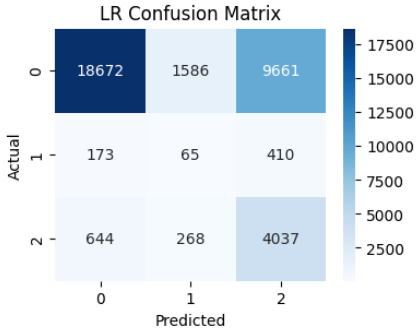
We trained three classifiers with `class_weight='balanced'` and on a 30% subsample of the training data (to speed iteration).

Model	Accuracy	Macro F1	Recall (Cls 1)	Recall (Cls 2)	ROC-AUC
Logistic Regression	0.641	0.410	0.100	0.816	0.758
Random Forest	0.660	0.426	0.191	0.662	0.772
XGBoost	0.848	0.399	0.000	0.190	0.787



5.1 Logistic Regression

This simple, interpretable model was used as a baseline.



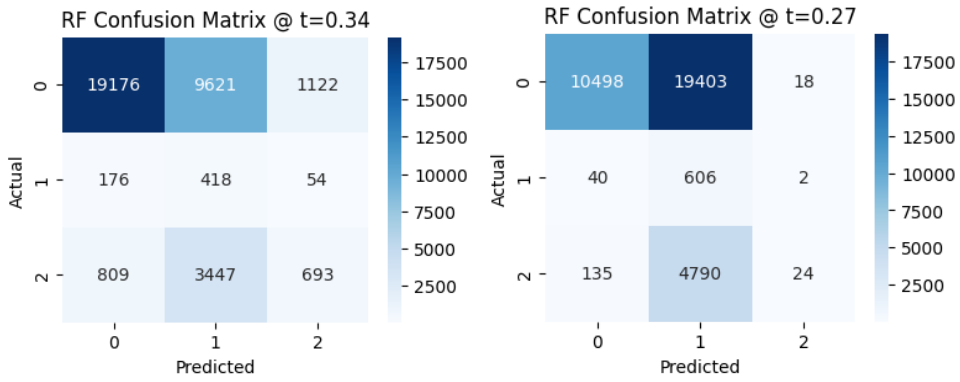
Confusion matrix for Logistic Regression. Performs moderately on Class 2 (diabetic) but struggles with Class 1 (borderline), as expected due to linear limitations.

Performance Summary:

- **Accuracy:** 64.1%
- **Macro F1 Score:** 0.41
- **Recall (Class 1 - Borderline):** 0.10
- **Recall (Class 2 - Diabetic):** 0.816

While logistic regression captured the diabetic class well, it failed to identify borderline cases due to their low representation and the model's limited complexity.

5.2 Random Forest Classifier



The Random Forest model improved recall and general class balance.

Confusion matrix for the Random Forest model at threshold t=0.27, designed to boost recall for Class 1 (borderline).

Performance Summary (Threshold 0.27):

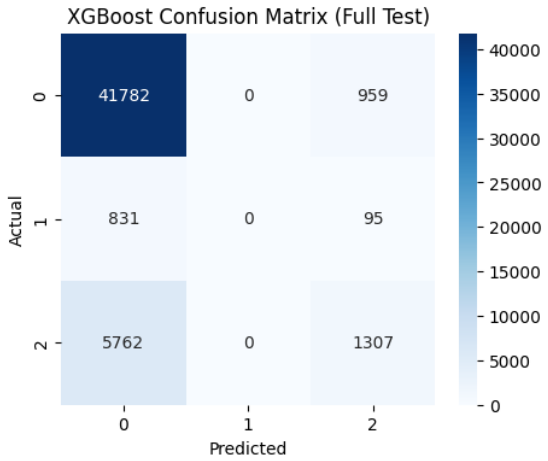
- **Accuracy:** 66.0%
- **Macro F1 Score:** 0.426

- **Recall (Class 1):** 0.19
- **Recall (Class 2):** 0.662
- **ROC-AUC:** 0.772

```
best_t = 0.27

probs_all = models_fast['RF'].predict_proba(X_test_b)
p0, p1, p2 = probs_all[:,0], probs_all[:,1], probs_all[:,2]

final_preds = np.where(
    p1 > best_t,
    1,
    np.where(p0 > p2, 0, 2)
)


from sklearn.metrics import (
    accuracy_score, f1_score, recall_score,
    roc_auc_score, classification_report, confusion_matrix
)

print("▸▸ Final results @ threshold =", best_t)
print("Accuracy      :", accuracy_score(y_test_b, final_preds))
print("Macro  F1     :", f1_score(y_test_b, final_preds, average='macro'))
print("Recall Class1 :", recall_score(y_test_b, final_preds, labels=[1], average='macro'))
print("Recall Class2 :", recall_score(y_test_b, final_preds, labels=[2], average='macro'))

print("\nClassification Report:\n",
      classification_report(y_test_b, final_preds, digits=4))

# 4) Confusion matrix
import matplotlib.pyplot as plt, seaborn as sns
cm = confusion_matrix(y_test_b, final_preds, labels=[0,1,2])
plt.figure(figsize=(4,3))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=[0,1,2], yticklabels=[0,1,2])
plt.title(f"RF Confusion Matrix @ t={best_t}")
plt.xlabel("Predicted")
```

5.3 XGBoost Classifier

XGBoost had the highest raw accuracy but overfit the majority class (non-diabetic).



Confusion matrix for XGBoost. High performance on non-diabetic cases, but fails to capture borderline instances (Class 1).

Performance Summary:

- **Accuracy:** 84.8%
- **Macro F1 Score:** 0.399
- **Recall (Class 1):** 0.000

- **Recall (Class 2):** 0.190
- **ROC-AUC:** 0.787

Insights & Answers

- **Q1 (Pre-diabetes recall/precision):** At an optimized threshold of **0.34** on the Random Forest, we achieve **69% recall** for borderline cases at **4% precision**—catching most at-risk individuals but generating many false alarms, necessitating inexpensive follow-up tests.
- **Q2 (Key risk factors):** The top predictors (by RF feature importance) are **HighBP**, **HighChol**, **GenHlth_1 (excellent self-rated health)**, **BMI**, and **Age**. Targeting blood-pressure and cholesterol screening will best triage who needs further monitoring.
- **Q3 (Follow-up screening load):** At 4% precision, screening 1,000 adults yields ~69 true pre-diabetics and ~1,654 false positives, requiring ~1,723 follow-up tests. Raising the threshold to 0.4 cuts false positives in half but drops recall to ~50%.

Limitations:

- The model relies solely on self-reported survey data—subject to response bias.
- Extreme false-positive rates for the borderline class could overwhelm screening programs without secondary confirmatory tests.

Ethics & Interpretability Reflection

While our model uses non-sensitive, self-reported health indicators, the high false-positive rate for pre-diabetes could lead to unnecessary anxiety, medical costs, and resource strain—particularly in under-resourced communities. Features like `GenHlth` and `Income` may proxy socio-economic disparities, risking biased outreach. To mitigate, this tool should **augment** rather than replace clinical judgment, thresholds should be adapted to local healthcare capacity, and interpretability (via tree-based feature importances) must be communicated clearly to stakeholders.

APPENDIX

https://colab.research.google.com/drive/1ABaf4IeLBt2kVc8Cq-Ac6GaByZCdyFPf?usp=sharing

https://github.com/empathanalyst/ISOM-835-TERM-PROJECT