Predicting Diabetes Risk Using Machine Learning & Deep Learning Methods

Course: ALY6140 – Analytics System Technology

Team: Immortals (Sai Ashwin Anumula, Vamshi Krishna Korutla, Vyas Kadiyala, Xiaoxi Li)

Date: May 15, 2025

Introduction

Diabetes remains one of the most serious global health concerns, impacting millions and placing substantial pressure on healthcare systems. The aim of this project is to identify and understand key demographic and lifestyle factors that contribute to diabetes and prediabetes risk.

Our key questions are:
- What lifestyle and demographic factors are most strongly associated with diabetes and prediabetes?
- How do income levels and healthcare access affect the likelihood of developing diabetes?
- Can we develop a reliable model that predicts diabetes status from health indicators?

To answer these questions, we applied advanced data analysis and predictive modeling techniques. Our methodology included data extraction, cleaning, exploratory data analysis (EDA), data visualization, and model building using logistic regression, random forest, XGBoost, SVM, and neural networks. These models were further optimized using SMOTE, undersampling, and class weighting.

Data Extraction

We used the Diabetes Health Indicators Dataset, publicly available on Kaggle. This dataset includes 253,680 records with 22 features related to health status, lifestyle choices, demographics, and healthcare access. The primary target variable is Diabetes_012, where:
- 0 = No diabetes
- 1 = Prediabetes
- 2 = Diabetes

The question we aimed to answer was: What variables most accurately predict the presence of diabetes, and can we build a predictive model to classify individuals accordingly?

Data Cleanup

Initial data processing was done in Jupyter Notebook using Python libraries such as pandas, numpy, and sklearn. The cleanup included:
- Null value checks: The dataset had no missing values.
- Encoding check: All categorical data was already numerically encoded.
- Outlier detection and removal using the Interquartile Range (IQR) method.
- Balanced sampling techniques like SMOTE, undersampling, and class weighting to address class imbalance.

Exploratory Data Analysis (EDA)

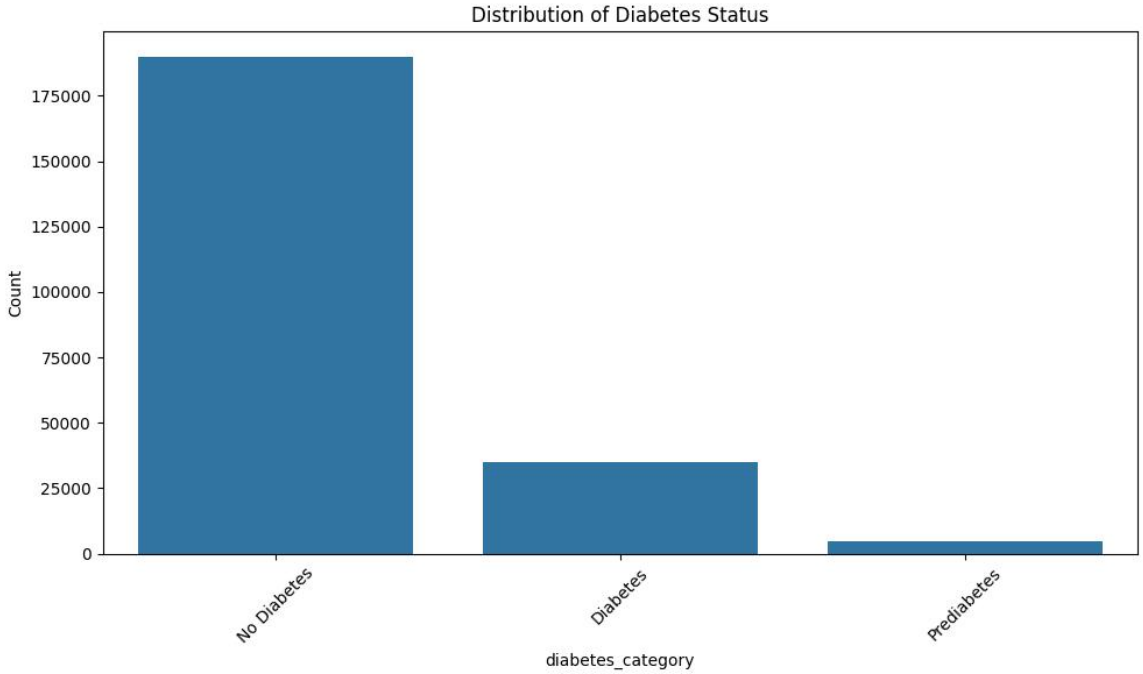
We conducted exploratory data analysis (EDA) to understand feature distributions and correlations.

Key insights:
- Most participants had no diabetes (~175,000), followed by ~35,000 with diabetes, and ~5,000 prediabetics.
- Obesity Class III had the highest percentage of diabetes cases.
- Individuals with 3–5 risk factors were significantly more likely to have diabetes.
- People with healthier lifestyle scores (3–5) had lower diabetes prevalence.
- Poor general health was strongly correlated with diabetes.

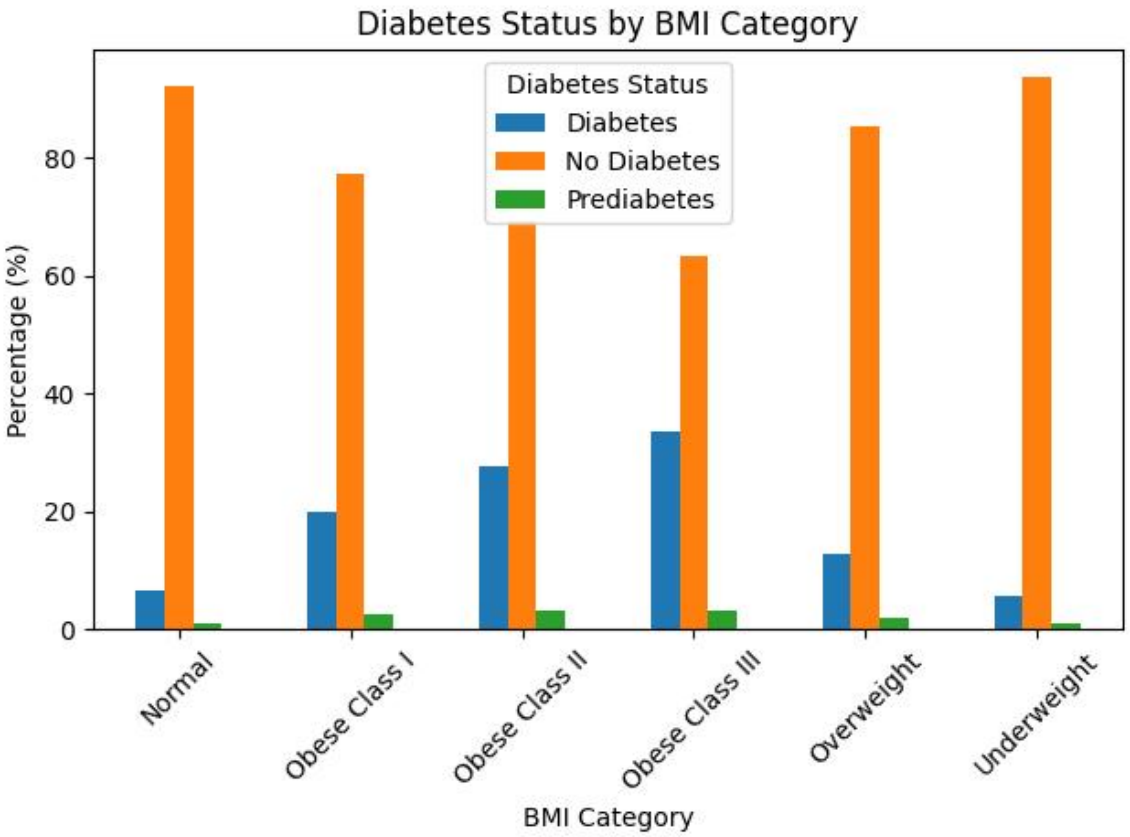
We also used correlation heatmaps to visualize feature relationships with diabetes, identifying high BP and BMI as notable predictors.

Data Visualization

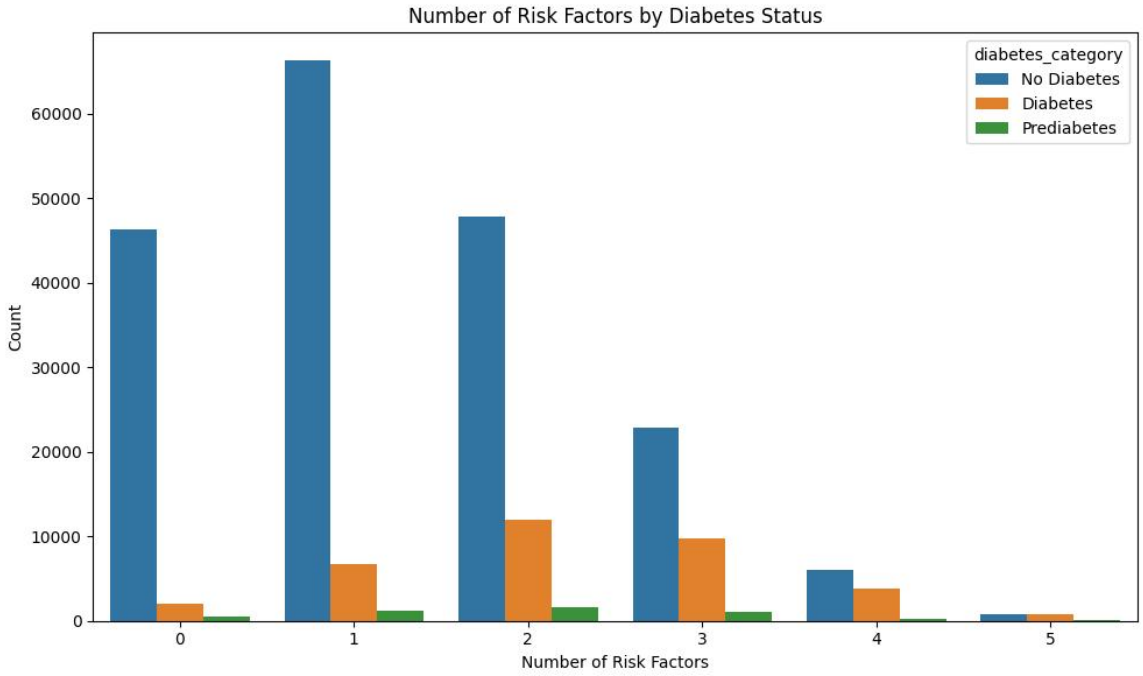
We used matplotlib and seaborn to explore:
- Diabetes status distribution



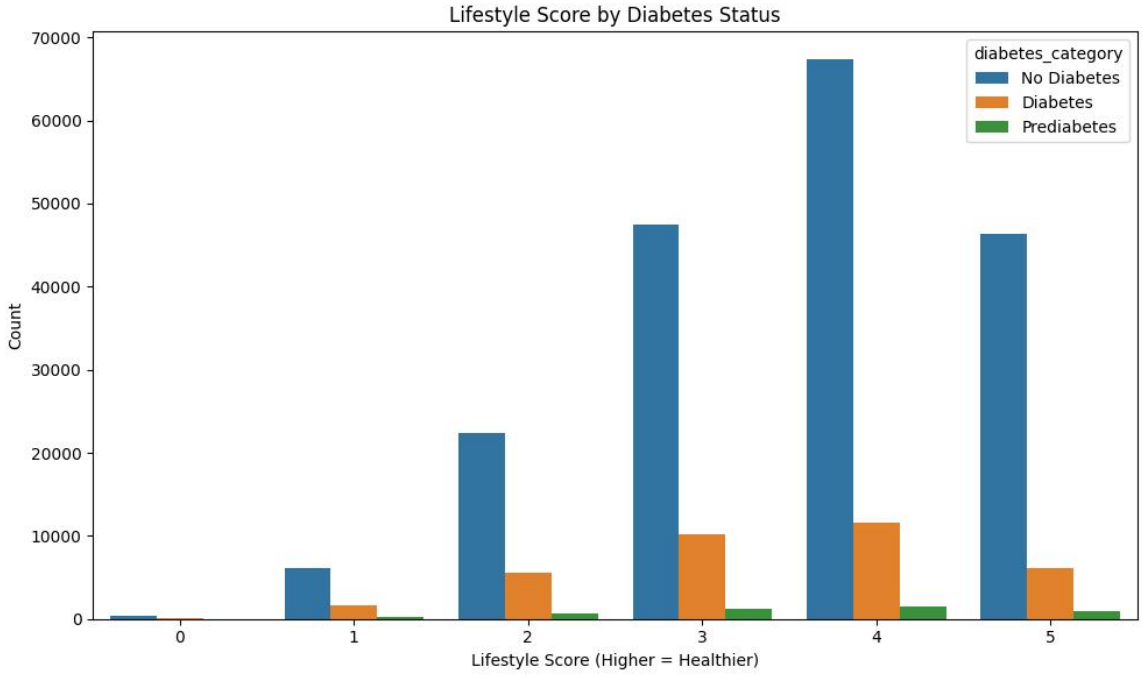
- BMI vs. Diabetes



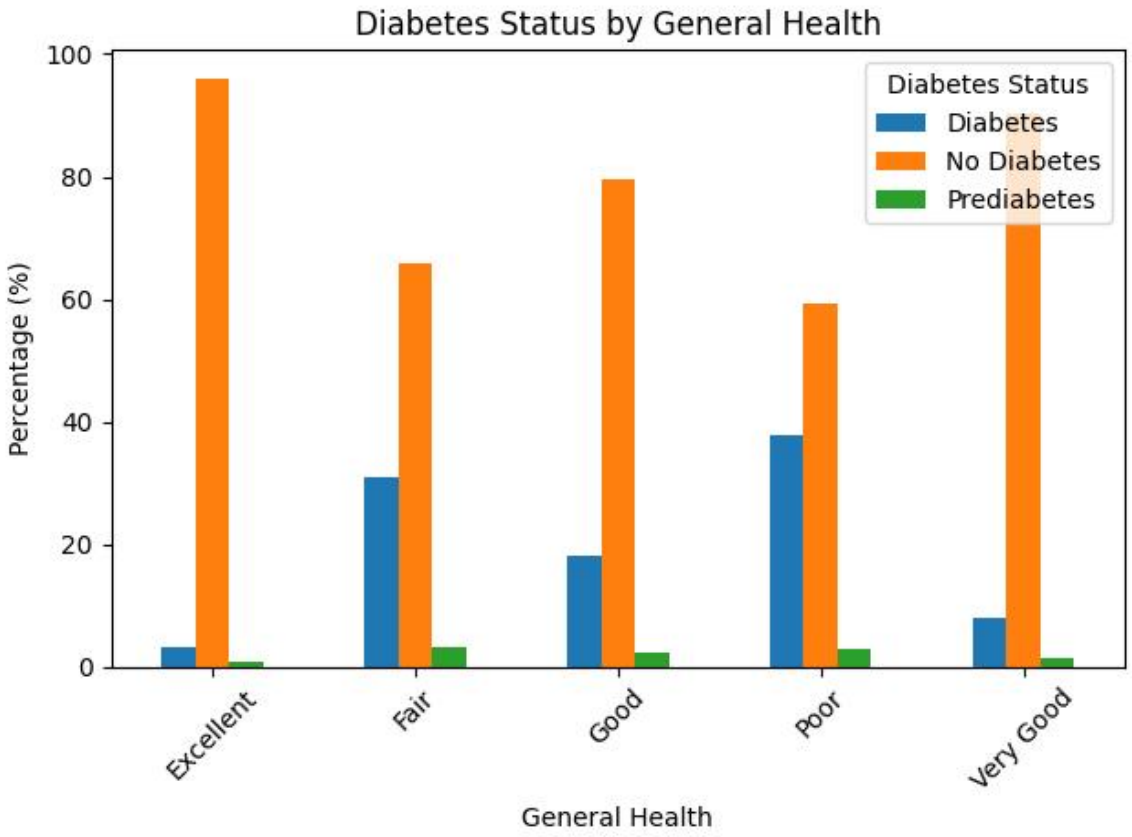
- Risk factors



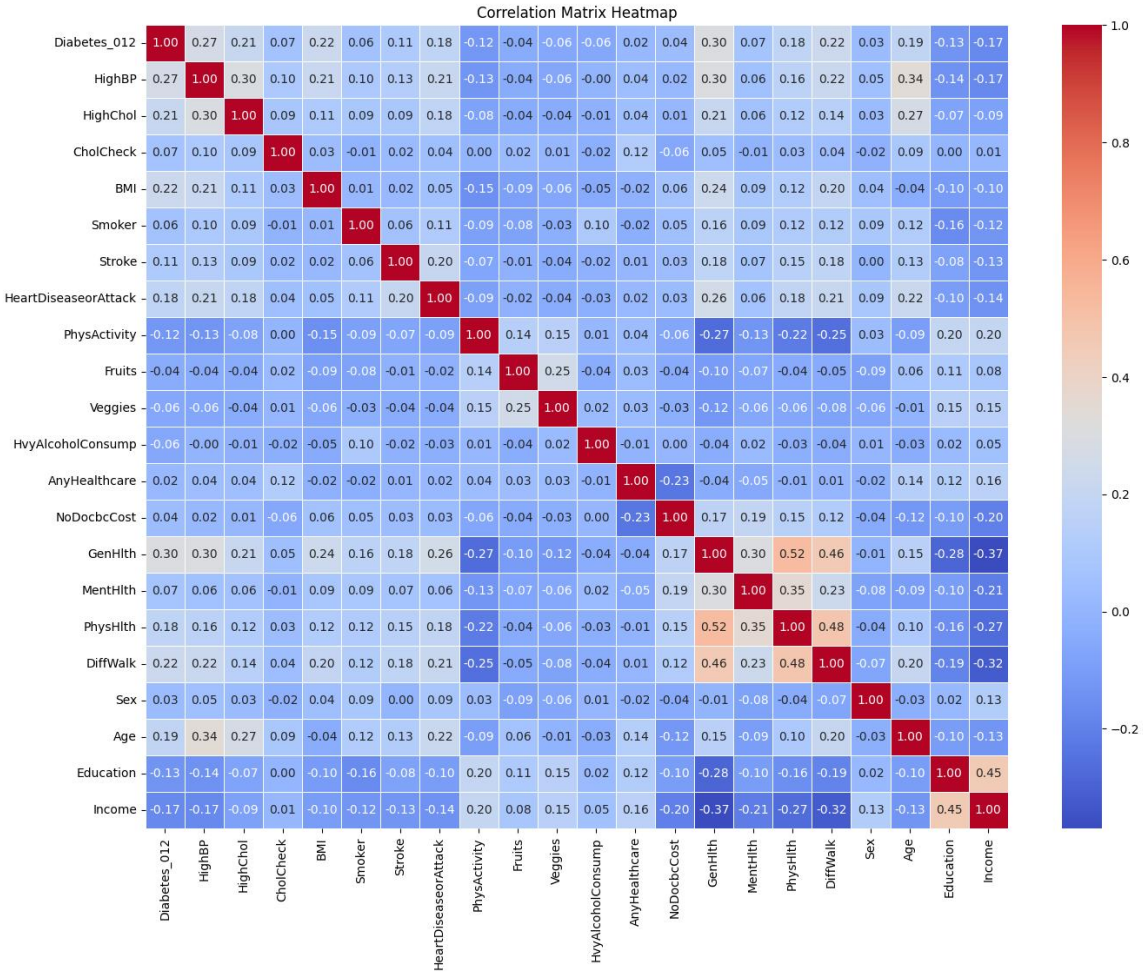
- Lifestyle scores



- General health



- Correlation heatmaps



Predictive Models (Statistical/Predictive Analysis)

We built and compared multiple classification models:

1. Logistic Regression:

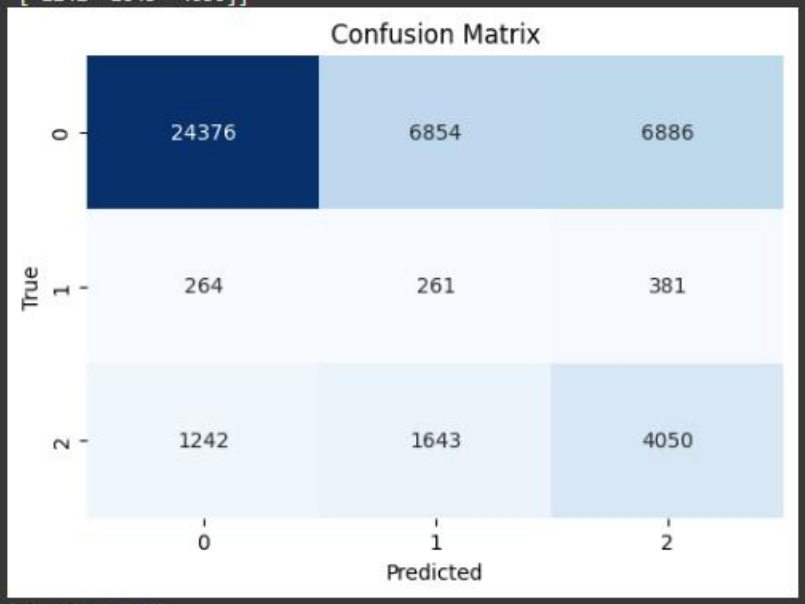
- Interpretable model, decent performance using SMOTE.

Using Smote Sampling for Logistic Regression:

```
n_iter_i = _check_optimize_result(
```

	precision	recall	f1-score	support
0.0	0.94	0.64	0.76	38116
1.0	0.03	0.29	0.05	906
2.0	0.36	0.58	0.44	6935
accuracy			0.62	45957
macro avg	0.44	0.50	0.42	45957
weighted avg	0.84	0.62	0.70	45957

```
[[24376  6854  6886]
 [  264   261   381]
 [ 1242  1643  4050]]
```




```
Accuracy: 0.62
Precision: 0.84
Recall: 0.62
F1 Score: 0.70
MSE: 0.91
RMSE: 0.95
```

The confusion matrix shows the model identifies the "No Diabetes" group well, with high precision and decent recall. However, it performs poorly on the "Diabetes" group and only moderately on "Prediabetes." With 62% accuracy and a 0.70 F1 score, the model favors the majority class, likely due to class imbalance.

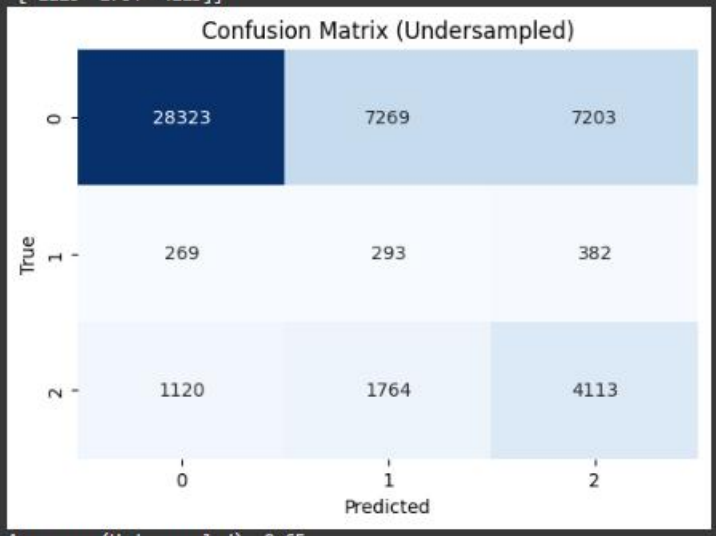
Using Undersampling for Logistic regression:

```
              precision    recall  f1-score   support

         0.0       0.95      0.66      0.78     42795
         1.0       0.03      0.31      0.06       944
         2.0       0.35      0.59      0.44      6997

    accuracy                           0.65     50736
   macro avg       0.45      0.52      0.43     50736
weighted avg       0.85      0.65      0.72     50736

[[28323  7269  7203]
 [  269   293   382]
 [ 1120  1764  4113]]
```




```
Accuracy (Undersampled): 0.65
Precision (Undersampled): 0.85
Recall (Undersampled): 0.65
F1 Score (Undersampled): 0.72
MSE (Undersampled): 0.85
RMSE (Undersampled): 0.92
```

After undersampling, the model shows slightly better balance, with accuracy rising to 65% and an F1 score of 0.72. It still predicts "No Diabetes" well, while recall for "Diabetes" and "Prediabetes" improves slightly. However, detecting minority classes remains difficult.

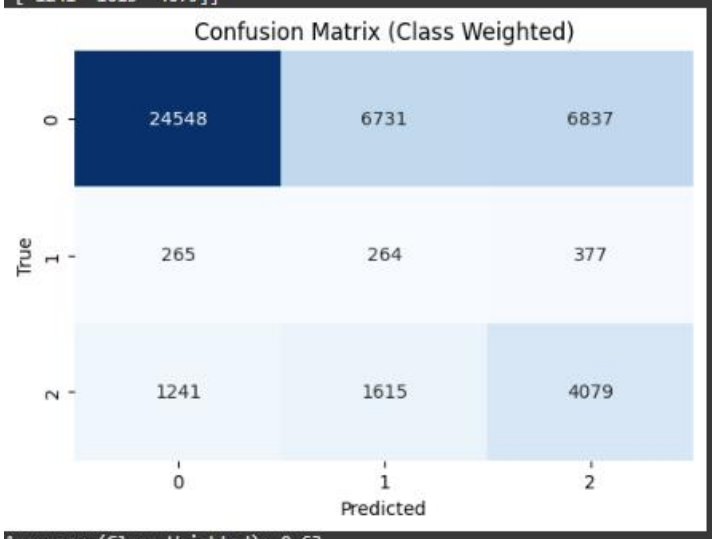
Using Class Weighing Sampling for Logistic regression:

```
              precision    recall  f1-score   support

         0.0       0.94      0.64      0.77     38116
         1.0       0.03      0.29      0.06       906
         2.0       0.36      0.59      0.45      6935

    accuracy                           0.63     45957
   macro avg       0.44      0.51      0.42     45957
weighted avg       0.84      0.63      0.70     45957

[[24548  6731  6837]
 [  265   264   377]
 [ 1241  1615  4079]]
```




```
Accuracy (Class Weighted): 0.63
Precision (Class Weighted): 0.84
Recall (Class Weighted): 0.63
F1 Score (Class Weighted): 0.70
MSE (Class Weighted): 0.90
RMSE (Class Weighted): 0.95
```

The class-weighted model reaches 63% accuracy and a 0.70 F1 score, similar to earlier models. It predicts "No Diabetes" well but continues to struggle with "Diabetes" due to low precision and recall. "Prediabetes" shows moderate improvement. Class weighting helps overall balance but doesn't fully fix the issue of poor minority class detection.

2. Random Forest:

- Outperformed other models, especially with class weighting. Highlighted key features like HighBP, BMI, and Age.

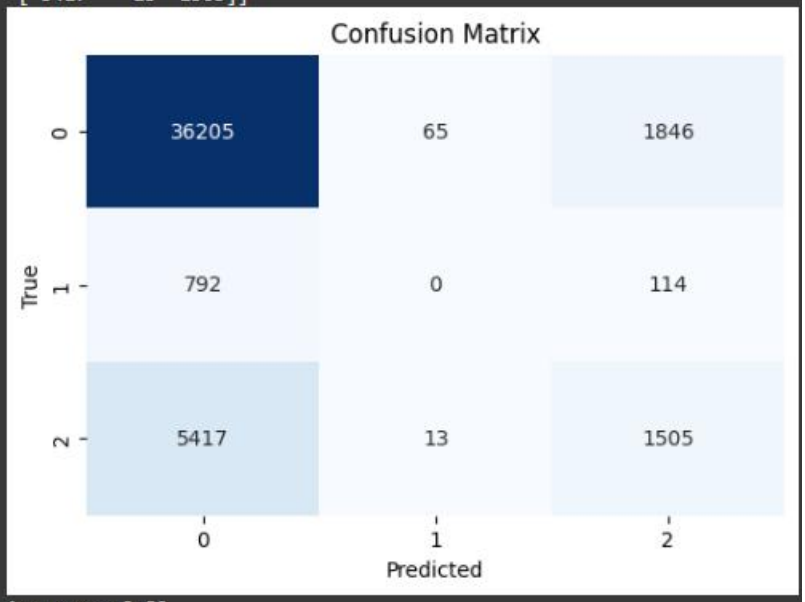
Using Smote sampling for RandomForest:

```
              precision    recall  f1-score   support

         0.0       0.85      0.95      0.90     38116
         1.0       0.00      0.00      0.00       906
         2.0       0.43      0.22      0.29      6935

    accuracy                           0.82     45957
   macro avg       0.43      0.39      0.40     45957
weighted avg       0.77      0.82      0.79     45957

[[36205    65  1846]
 [  792     0   114]
 [ 5417    13  1505]]
```




```
Accuracy: 0.82
Precision: 0.77
Recall: 0.82
F1 Score: 0.79
MSE: 0.65
RMSE: 0.81
```

This model achieves 82% accuracy and a 0.79 F1 score, mainly due to strong performance on the "No Diabetes" group with high recall. However, it completely fails to detect "Diabetes" cases and performs poorly on "Prediabetes" as well. Despite strong overall metrics, the model heavily favors the majority class, limiting its usefulness in real-world healthcare.

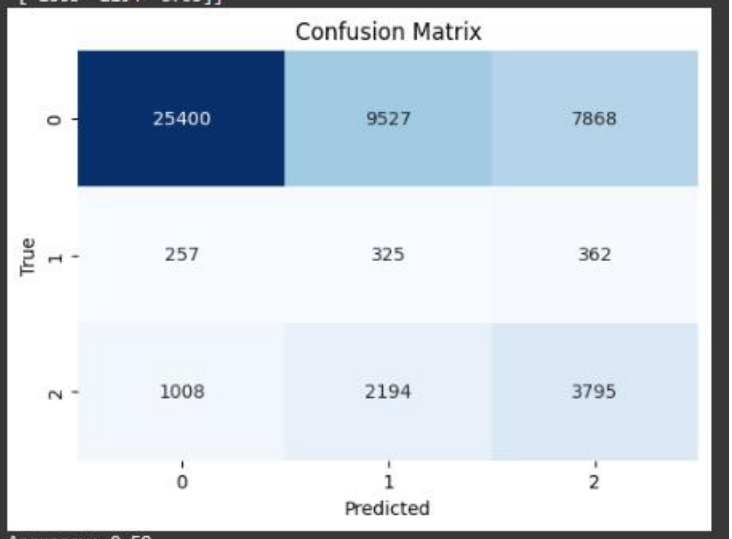
Using Understanding for Random Forest

```
              precision    recall  f1-score   support

         0.0       0.95      0.59      0.73     42795
         1.0       0.03      0.34      0.05       944
         2.0       0.32      0.54      0.40      6997

    accuracy                           0.58     50736
   macro avg       0.43      0.49      0.39     50736
weighted avg       0.85      0.58      0.67     50736

[[25400  9527  7868]
 [  257   325   362]
 [ 1008  2194  3795]]
```




```
Accuracy: 0.58
Precision: 0.85
Recall: 0.58
F1 Score: 0.67
MSE: 0.94
RMSE: 0.97
```

This model has a lower accuracy of 58% but still predicts the "No Diabetes" class well, with high precision but moderate recall. It struggles with the "Diabetes" class, showing very low precision and recall, while "Prediabetes" has slightly better recall but poor precision. Overall, the model favors the majority class and remains weak at detecting minority groups, limiting its real-world reliability.

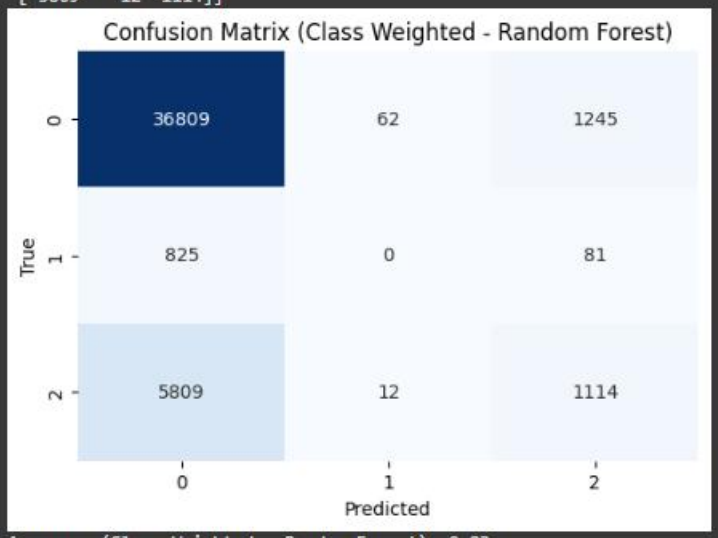
Class Weighing for Randomforest:

```
              precision    recall  f1-score   support

         0.0       0.85      0.97      0.90     38116
         1.0       0.00      0.00      0.00       906
         2.0       0.46      0.16      0.24      6935

    accuracy                           0.83     45957
   macro avg       0.43      0.38      0.38     45957
weighted avg       0.77      0.83      0.78     45957

[[36809    62  1245]
 [  825     0    81]
 [ 5809    12  1114]]
```



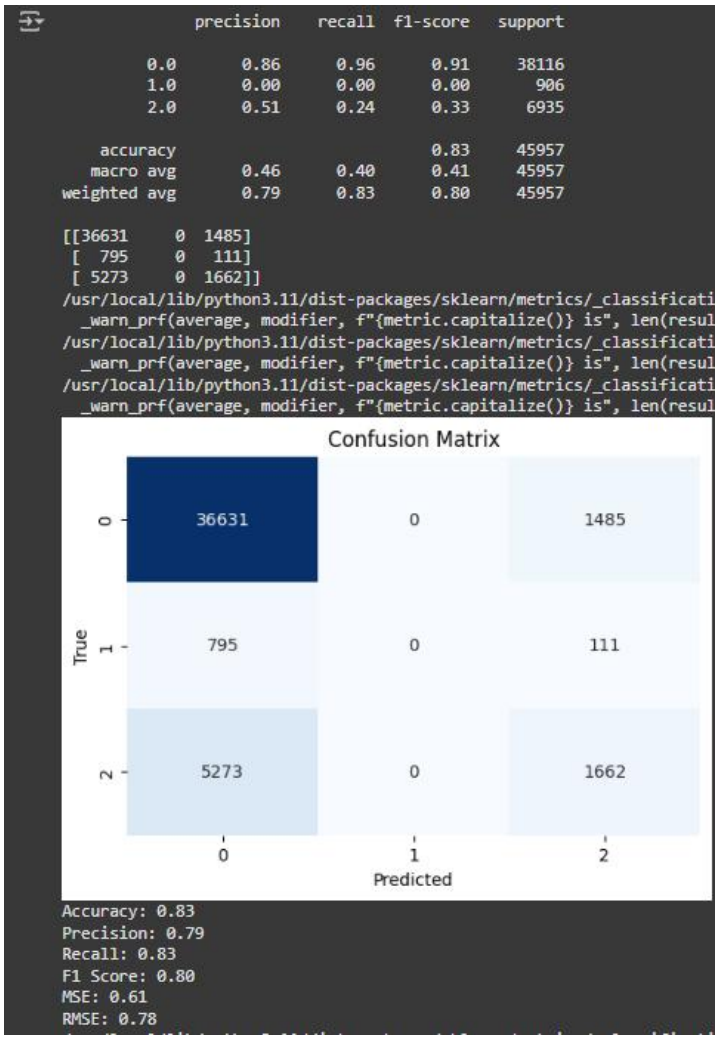

```
Accuracy (Class Weighted - Random Forest): 0.83
Precision (Class Weighted - Random Forest): 0.77
Recall (Class Weighted - Random Forest): 0.83
F1 Score (Class Weighted - Random Forest): 0.78
MSE (Class Weighted - Random Forest): 0.64
RMSE (Class Weighted - Random Forest): 0.80
```

This class-weighted Random Forest model reaches 83% accuracy and a 0.78 F1 score, mainly due to strong recall for "No Diabetes." However, it fails to detect any "Diabetes" cases and performs poorly on "Prediabetes" as well. Despite good overall metrics, the model is heavily biased toward the majority class, making it less effective for balanced healthcare predictions.

3. XGBoost:

- High-performing model with good precision and recall using class weighting.

Using Smote Sampling for XG boost:



This model shows 83% accuracy and a strong F1 score of 0.80, driven by high recall for the "No Diabetes" class. However, it fails to detect any "Diabetes" cases and performs poorly on "Prediabetes," with low recall. The confusion matrix highlights a strong bias toward the majority class, making the model unreliable for real-world healthcare use where detecting all conditions accurately is essential.

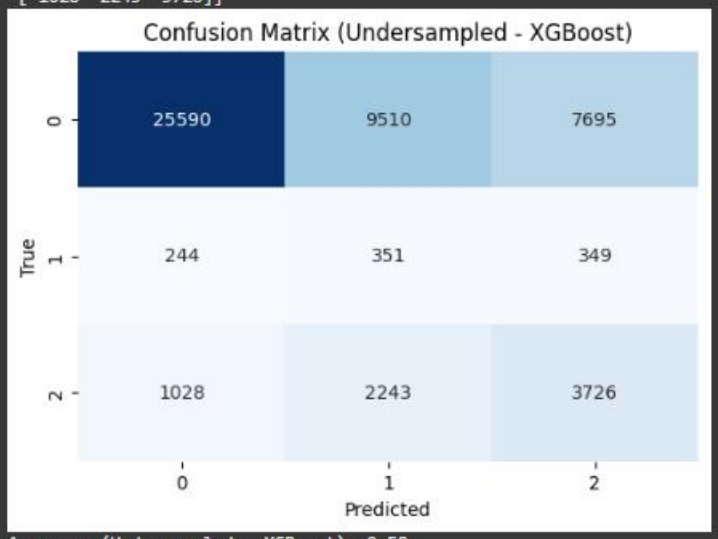
Undersampling for XG boosting:

```
              precision    recall  f1-score   support

         0.0       0.95      0.60      0.73     42795
         1.0       0.03      0.37      0.05       944
         2.0       0.32      0.53      0.40      6997

    accuracy                           0.58     50736
   macro avg       0.43      0.50      0.40     50736
weighted avg       0.85      0.58      0.68     50736

[[25590  9510  7695]
 [  244   351   349]
 [ 1028  2243  3726]]
```




```
Accuracy (Undersampled - XGBoost): 0.58
Precision (Undersampled - XGBoost): 0.85
Recall (Undersampled - XGBoost): 0.58
F1 Score (Undersampled - XGBoost): 0.68
MSE (Undersampled - XGBoost): 0.93
RMSE (Undersampled - XGBoost): 0.96
```

The undersampled XGBoost model reaches 58% accuracy and an F1 score of 0.68. It predicts the "No Diabetes" group well but only slightly improves recall for "Diabetes" and "Prediabetes." Precision for these classes remains low. Undersampling improves balance somewhat, but the model still favors the majority class and needs further refinement for reliable detection across all groups.

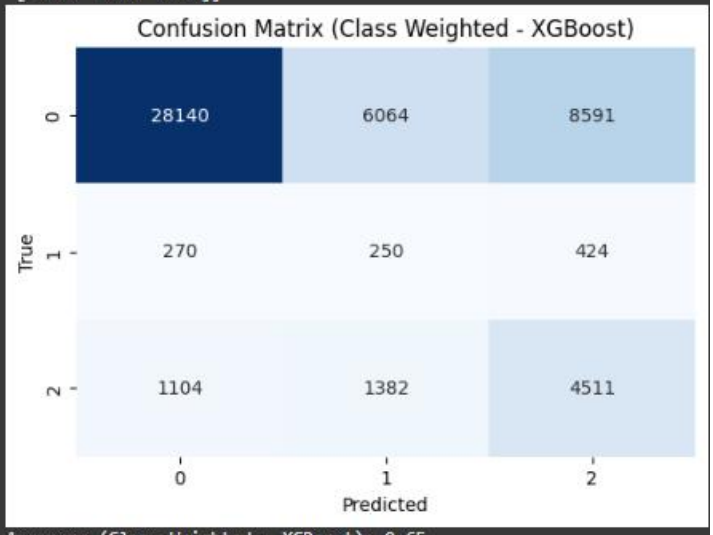
Class Weighing for XG boosting:

```
  warnings.warn(smsg, UserWarning)
              precision    recall  f1-score   support

         0.0       0.95      0.66      0.78     42795
         1.0       0.03      0.26      0.06       944
         2.0       0.33      0.64      0.44      6997

    accuracy                           0.65     50736
   macro avg       0.44      0.52      0.43     50736
weighted avg       0.85      0.65      0.72     50736

[[28140  6064  8591]
 [  270   250   424]
 [ 1104  1382  4511]]
```



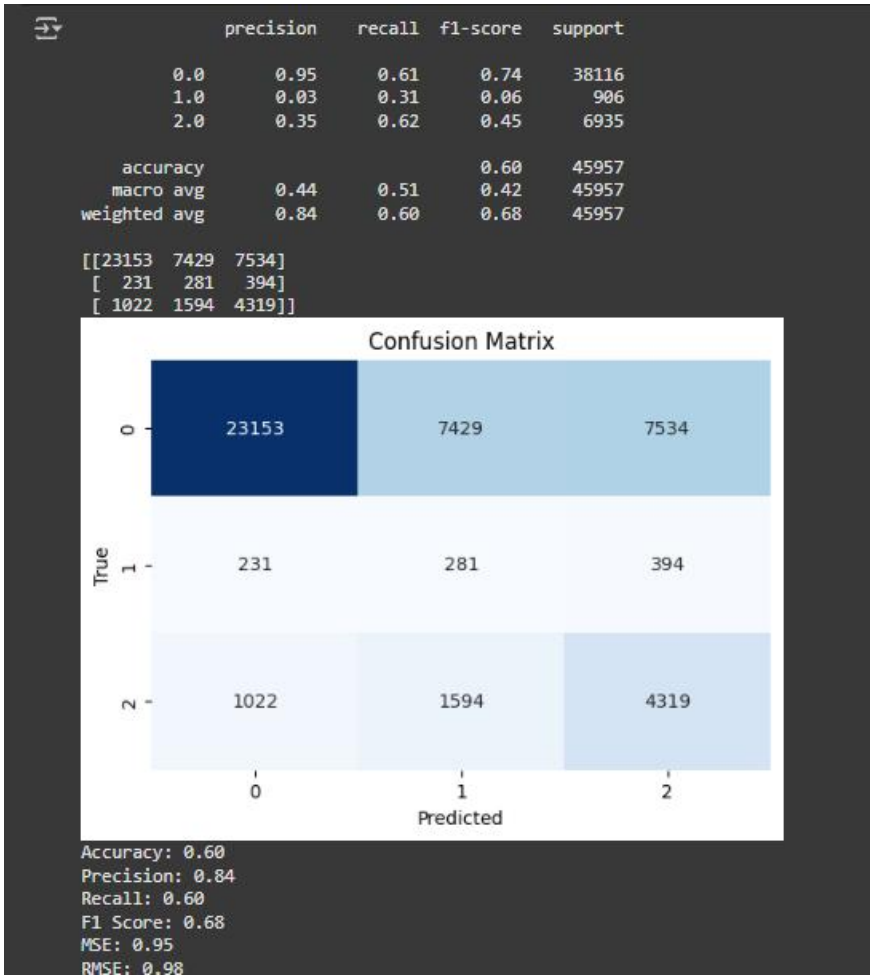

```
Accuracy (Class Weighted - XGBoost): 0.65
Precision (Class Weighted - XGBoost): 0.85
Recall (Class Weighted - XGBoost): 0.65
F1 Score (Class Weighted - XGBoost): 0.72
MSE (Class Weighted - XGBoost): 0.92
RMSE (Class Weighted - XGBoost): 0.96
```

The class-weighted XGBoost model shows 65% accuracy and an F1 score of 0.72, with more balanced results across classes. It maintains strong precision for "No Diabetes" and improves recall for "Prediabetes" to 0.64. However, performance on "Diabetes" remains weak, with low recall and precision. While class weighting helps, the model still favors the majority class, and further adjustments are needed for better overall balance.

4. SVM:

- Effective on balanced data, but resource-intensive.

Using smote sampling for SVM:



This model reaches 60% accuracy and an F1 score of 0.68. It performs well on "No Diabetes" with high precision but moderate recall. "Prediabetes" shows fair recall but low precision, while "Diabetes" remains poorly detected with very low scores. Overall, the model handles the majority class well but continues to underperform on minority class detection.

5. Neural Networks:

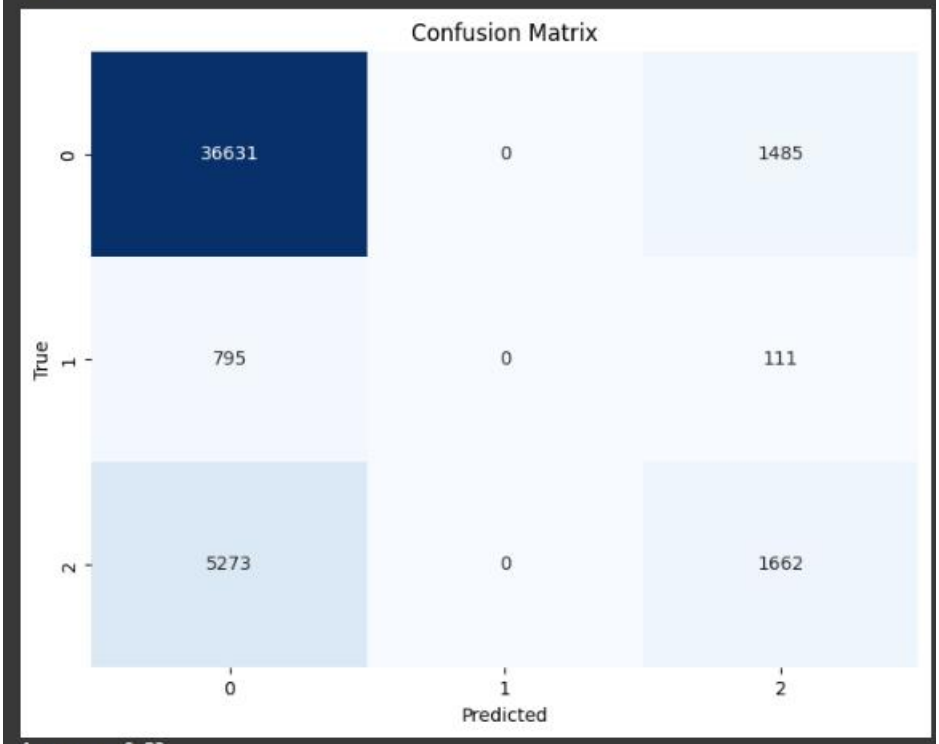
- Used Keras, performed reasonably well but required more resources and was less interpretable.

Smote Sampling for neural network:

```
Classification Report for 15 epochs:
              precision    recall  f1-score   support

         0.0       0.89      0.81      0.85     38116
         1.0       0.02      0.06      0.03       906
         2.0       0.38      0.50      0.43      6935

    accuracy                           0.75     45957
   macro avg       0.43      0.46      0.44     45957
weighted avg       0.80      0.75      0.77     45957
```




```
Accuracy: 0.83
Precision: 0.79
Recall: 0.83
F1 Score: 0.80
MSE: 0.61
RMSE: 0.78
```

This model delivers 83% accuracy and a strong F1 score of 0.80, largely due to its high performance on the "No Diabetes" group. However, it fails to detect "Diabetes" cases, with nearly zero precision and recall. The "Prediabetes" class performs slightly better but remains weak. Despite strong overall metrics, the model is still biased toward the majority class and needs improvement in identifying minority groups.

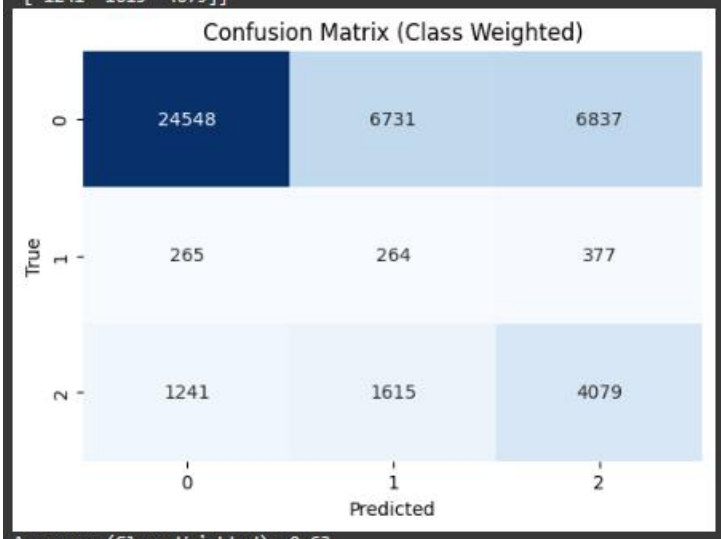
Class Weighing for Neural Network:

```
  n_iter_i = _check_optimize_result(
              precision    recall  f1-score   support

         0.0       0.94      0.64      0.77     38116
         1.0       0.03      0.29      0.06       906
         2.0       0.36      0.59      0.45      6935

    accuracy                           0.63     45957
   macro avg       0.44      0.51      0.42     45957
weighted avg       0.84      0.63      0.70     45957

[[24548  6731  6837]
 [  265   264   377]
 [ 1241  1615  4079]]
```




```
Accuracy (Class Weighted): 0.63
Precision (Class Weighted): 0.84
Recall (Class Weighted): 0.63
F1 Score (Class Weighted): 0.70
MSE (Class Weighted): 0.90
RMSE (Class Weighted): 0.95
```

This class-weighted model reaches 63% accuracy and a 0.70 F1 score. It performs well on the "No Diabetes" group and shows moderate improvement for "Prediabetes" with better recall. However, it continues to struggle with detecting "Diabetes" due to low precision and recall. While class weighting helps a bit, the model still leans toward the majority class and underperforms on minority groups.

Evaluation metrics included accuracy, precision, recall, F1 score, and ROC-AUC where applicable. The best model was Random Forest with class weighting. It achieved strong overall accuracy and high performance in predicting the majority class. However, like many others, it still struggled to correctly identify minority classes, particularly individuals with diabetes. While class weighting helped improve balance slightly, most models, including XGBoost and neural networks, showed a clear bias toward the "No Diabetes" group. This highlights the need for further tuning or advanced techniques to boost minority class detection without sacrificing overall performance.

Interpretation and Conclusions

Our analysis confirmed that demographic, lifestyle, and healthcare access variables significantly influence diabetes risk. High BMI, poor general health, and high blood pressure are the strongest indicators, while higher income and education offer protection.

Random Forest with class weighting proved to be the best-performing model. This can be used to support early intervention programs.
Based on the findings, several recommendations can help improve outcomes. First, launching preventive programs that promote healthy lifestyles may reduce health risks over time. Second, making healthcare more accessible can ensure earlier detection and better management of conditions. Lastly, including more detailed information, like genetic data and long-term health history, could improve the accuracy of future predictions and analyses.

References

1. Teboul, A. (2022). Diabetes Health Indicators Dataset. Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
2. American Diabetes Association. (2023). Standards of Medical Care in Diabetes—2023. Diabetes Care, 46(Supplement 1), S1–S272.
3. World Health Organization. (2021). Global report on diabetes. WHO.