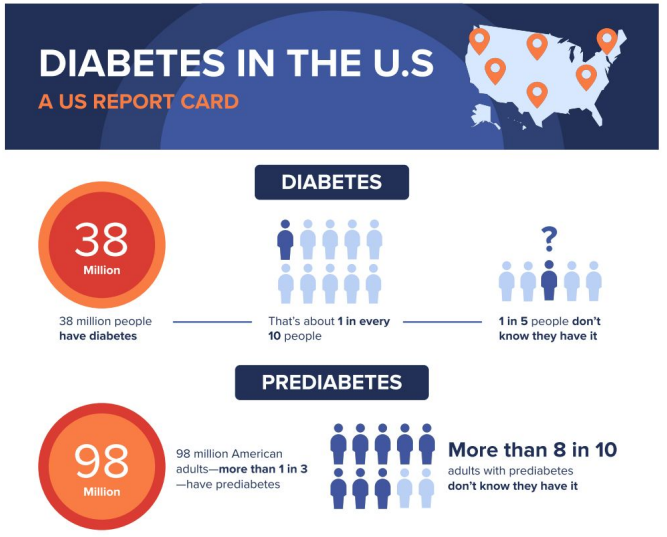
Diabetes Classification

CDC Health Indicators Dataset

Business Problem

- Type 2 Diabetes Mellitus (DM2) creates a major financial burden in the U.S.
- Total annual economic impact: $404B (2017).
- Includes $327B from diagnosed diabetes and $77B from undiagnosed diabetes, prediabetes, and Gestational Diabetes Mellitus (GDM).
- A significant portion of the cost comes from individuals who remain undiagnosed.



Why This Matters to Employers

Rising diabetes prevalence = Rising employer healthcare spending.

- **2x Higher Medical Costs:** Compared to non-diabetic employees.
- **Higher Healthcare Utilization:** More outpatient visits and inpatient stays.
- **Higher Prescription Needs:** Long-term medication costs.
- **Productivity Loss:** More lost workdays due to complications.

Opportunity for Prevention

- Early detection and prevention programs provide measurable ROI.
- NIH research: Participants in the National Diabetes Prevention Program (NDPP) experienced an average **$4,552 reduction in direct medical costs** within 2 years vs non-participants.
- Prevention = clinically effective and financially efficient.

Our Objective

Our predictive model is built to:

Identify

Early identification of individuals at risk of undiagnosed diabetes

Pinpoint

Pinpoint key predictor variables affecting diabetes likelihood to refine screening.

Support

Support employers in targeting prevention resources to reduce long-term medical costs.

Strengthen early identification → reduce healthcare spending → improve workforce health outcomes.

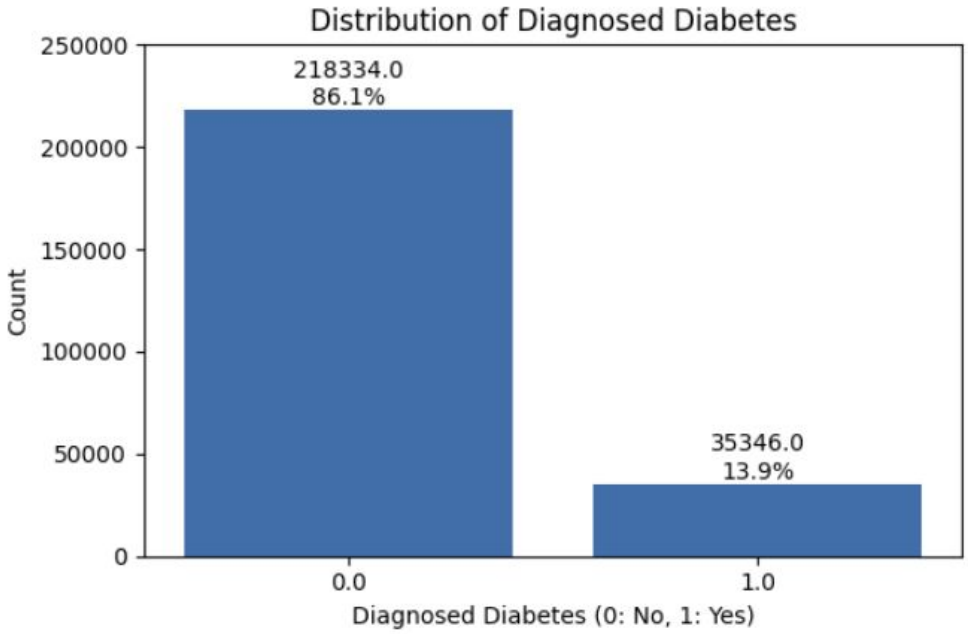
Dataset Introduction

CDC Diabetes Health Indicators

Source: Kaggle
Structure: 253,680 rows x 22 columns
Target: Diabetes_binary (0 = No, 1 = Yes)
Row Meaning: Each row represents a survey response including details related to health-related risk behaviors.
Predictors Example:

- Health Behaviors (BMI, Smoking, Physical Activity)
- Demographic information (Age, Sex, Education)

Target Variable Distribution



The highly imbalanced class of interest (~ 13.9%) presents a significant challenge for the classification models.

Exploratory Data Analysis

Key Risk Factors

Strong associations between DM2 and expected clinical predictors: sedentary lifestyle, prior stroke, hypertension, and hypercholesterolemia.

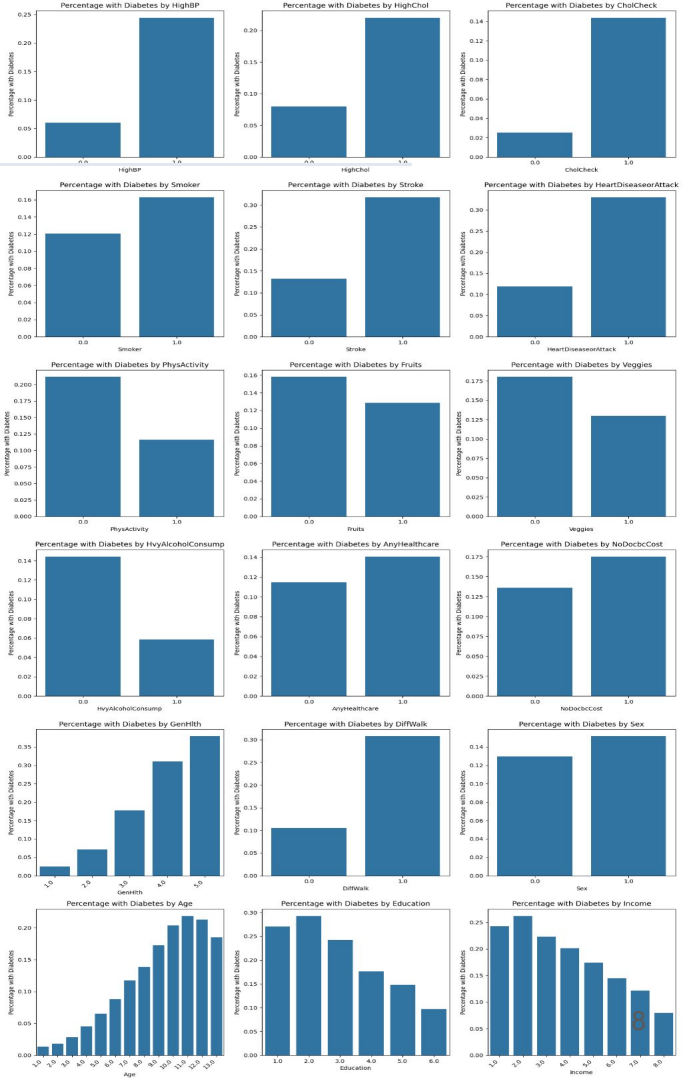
Unexpected Findings

- Smoking status showed no clear relationship.
- Heavy alcohol consumption appeared inversely related to diabetes risk.

Redundancy

Overlap identified between 'Heart Disease' & 'Stroke', and 'Hypercholesterolemia' & 'Chol Check'.

Outliers were removed due to minimal representation and limited impact on overall model integrity.

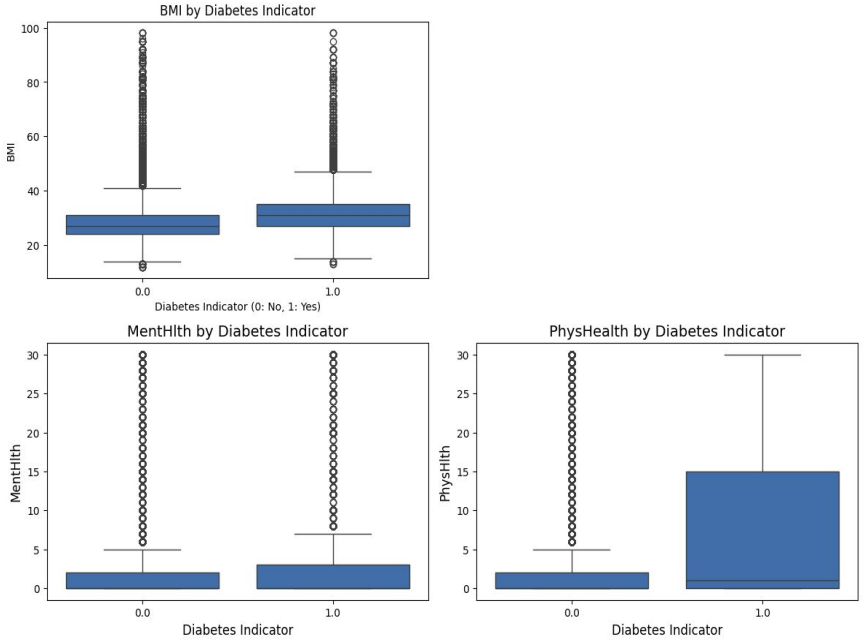


Exploratory Data Analysis

Physical vs. Mental Health

- Our analysis revealed that people with diabetes (Class 1) have a significantly higher median BMI compared to non-diabetics.
- While physical and mental health are moderately correlated (~0.35), only physical health proved to be a strong discriminator for diabetes.

➢ Diabetes' impact may be more related to physiological condition than psychological condition in this dataset.



Model strategy & evaluation approach

Data Partitioning

60% Train / 20% Validation / 20% Test

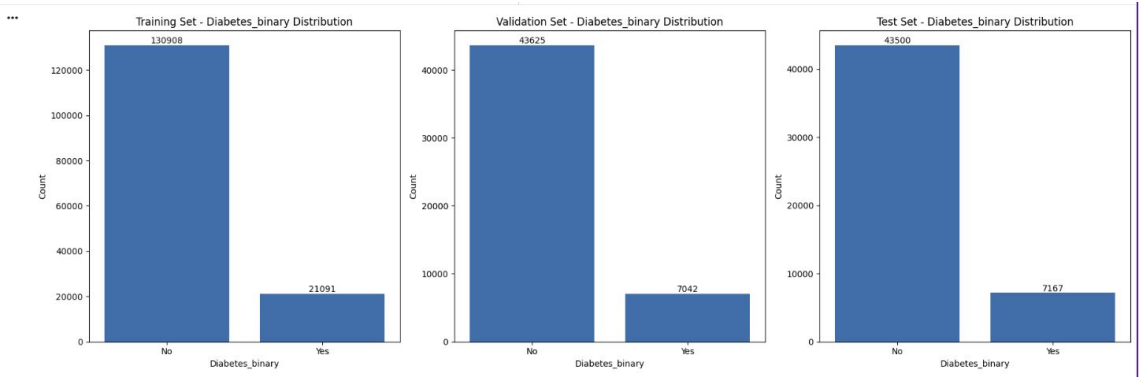
This strict split ensures our evaluation is robust and prevents overfitting to the training data.

Primary Metric: Recall

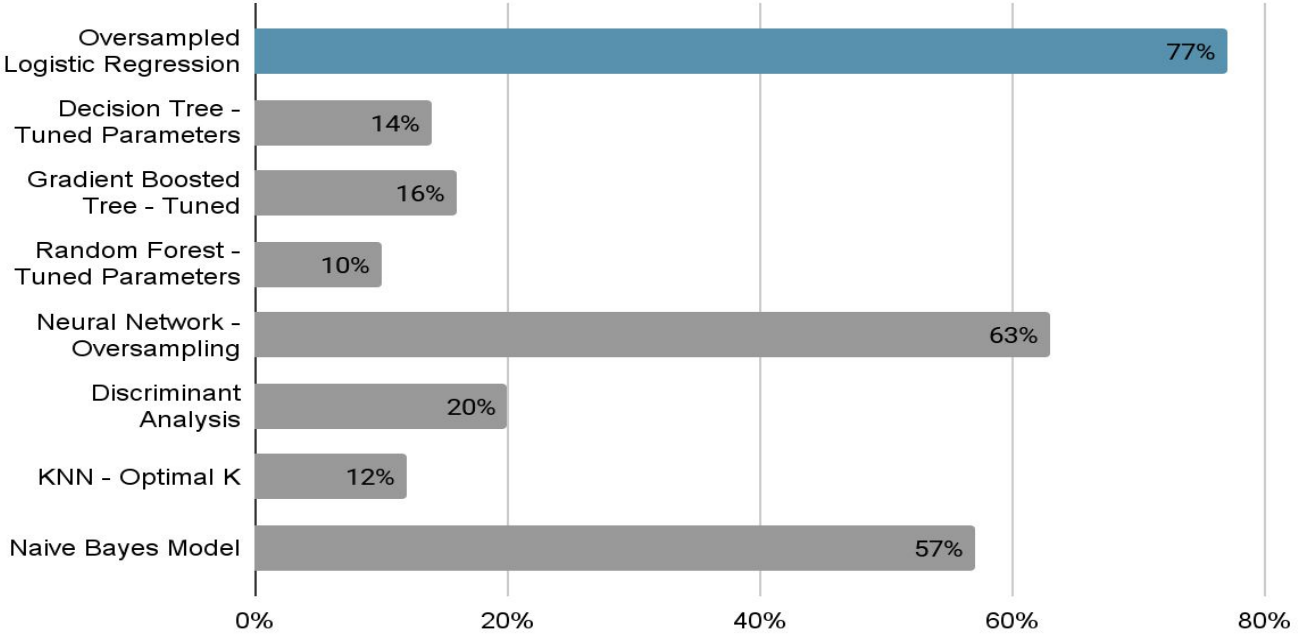
Focus on the Minority Class (1)

Recall = TP/(TP+FN)

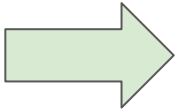
In healthcare, False Negatives are dangerous. We prioritized Recall to maximize the detection of actual diabetic cases, accepting a trade-off in Precision.



Model Comparison: Recall on Class 1



Most of the models failed to detect the minority class. **Oversampling** was the key to detect the actual diabetic cases.

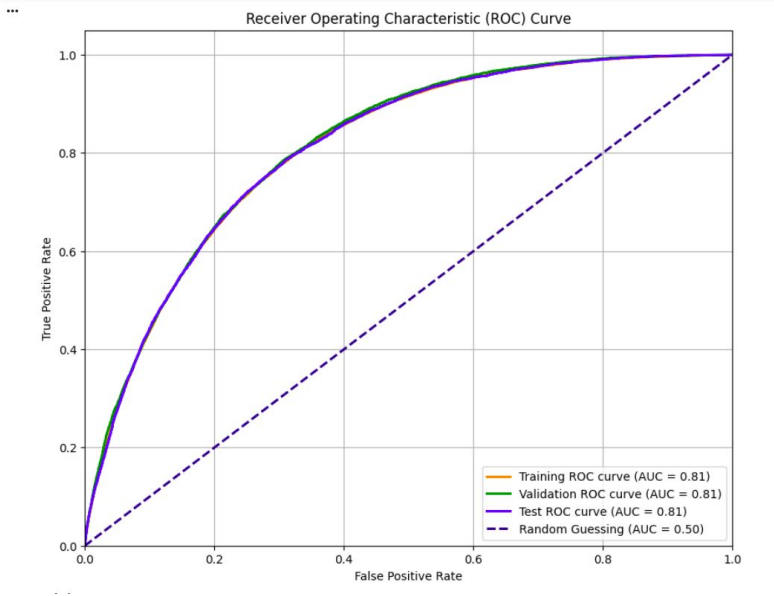


Model	Partition	Non-Target Class (0)			Target Class (1)			Accuracy	AUC
		Precision	Recall	F1	Precision	Recall	F1		
Logistic Regression Model with Selected Features	Train	0.88	0.98	0.92	0.52	0.14	0.22	0.86	0.81
	Validation	0.88	0.98	0.93	0.55	0.15	0.23	0.86	0.81
	Test	0.87	0.98	0.92	0.52	0.13	0.21	0.86	0.81
Logistic Regression Model with Interaction Term	Train	0.88	0.98	0.93	0.54	0.16	0.24	0.86	0.82
	Validation	0.88	0.98	0.93	0.58	0.16	0.25	0.87	0.83
	Test	0.87	0.98	0.92	0.55	0.15	0.23	0.86	0.82
Logistic Regression Model with Oversampling	Train	0.77	0.73	0.75	0.74	0.78	0.76	0.75	0.83
	Validation	0.95	0.73	0.82	0.31	0.76	0.44	0.73	0.83
	Test	0.95	0.73	0.82	0.32	0.77	0.45	0.73	0.82
Decision Tree - Tuned Parameters	Train	0.88	0.98	0.93	0.61	0.17	0.26	0.87	0.84
	Validation	0.88	0.98	0.93	0.54	0.14	0.23	0.86	0.81
	Test	0.87	0.98	0.92	0.53	0.14	0.22	0.86	0.81
Gradient Boosted Tree - Tuned Parameters	Train	0.88	0.98	0.93	0.60	0.17	0.27	0.87	0.83
	Validation	0.88	0.98	0.93	0.59	0.17	0.26	0.87	0.83
	Test	0.88	0.98	0.93	0.58	0.16	0.25	0.86	0.83

Model	Partition	Non-Target Class (0)			Target Class (1)			Accuracy	AUC
		Precision	Recall	F1	Precision	Recall	F1		
Random Forest - Tuned Parameters	Train	0.88	0.99	0.93	0.76	0.15	0.25	0.88	1.00
	Validation	0.87	0.99	0.93	0.64	0.11	0.19	0.87	0.80
	Test	0.87	0.99	0.93	0.60	0.10	0.17	0.86	0.80
Neural Network - Oversampling	Train	0.80	0.81	0.81	0.81	0.80	0.81	0.81	0.90
	Validation	0.93	0.80	0.86	0.34	0.62	0.44	0.78	0.80
	Test	0.93	0.81	0.86	0.35	0.63	0.45	0.78	0.80
Discriminant Analysis	Train	0.88	0.97	0.92	0.51	0.20	0.29	0.86	0.82
	Validation	0.88	0.97	0.93	0.53	0.21	0.30	0.86	0.82
	Test	0.88	0.97	0.92	0.52	0.20	0.29	0.86	0.82
KNN - Optimal K	Train	0.88	0.98	0.93	0.63	0.17	0.26	0.87	0.86
	Validation	0.87	0.98	0.92	0.48	0.12	0.20	0.86	0.78
	Test	0.87	0.98	0.92	0.48	0.12	0.19	0.86	0.77
Naive Bayes Model	Train	0.92	0.81	0.86	0.32	0.57	0.41	0.77	0.78
	Validation	0.92	0.81	0.86	0.33	0.57	0.42	0.78	0.79
	Test	0.92	0.81	0.86	0.33	0.57	0.42	0.77	0.78

Best Model Summary

1. Logistic Regression Model with All Features



Suffered from class imbalance → poor performance on the minority class.

```
Classification Report - Training Set:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.92    130908
         1.0       0.52      0.16      0.24     21091

    accuracy                           0.86    151999
   macro avg       0.70      0.57      0.58    151999
weighted avg       0.83      0.86      0.83    151999


Classification Report - Validation Set:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93     43625
         1.0       0.54      0.17      0.26      7042

    accuracy                           0.86     50667
   macro avg       0.71      0.57      0.59     50667
weighted avg       0.83      0.86      0.83     50667


Classification Report - Test Set:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.92     43500
         1.0       0.52      0.15      0.24      7167

    accuracy                           0.86     50667
   macro avg       0.70      0.57      0.58     50667
weighted avg       0.82      0.86      0.83     50667
```

Best Model Summary

2. Logistic Regression Model with Selected Features

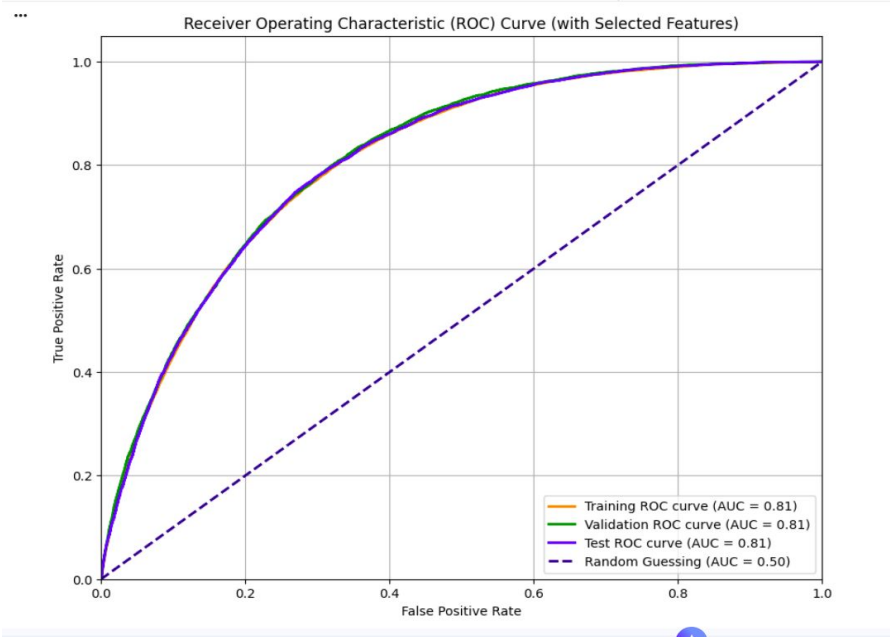
```
Optimization terminated successfully.
         Current function value: 0.318480
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:        Diabetes_binary   No. Observations:               151999
Model:                          Logit   Df Residuals:                   151977
Method:                           MLE   Df Model:                           21
Date:                Tue, 02 Dec 2025   Pseudo R-squ.:                  0.2091
Time:                        14:44:51   Log-Likelihood:                -48409.
converged:                       True   LL-Null:                       -61210.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -8.0143      0.121    -66.119      0.000      -8.252      -7.777
HighBP         0.7386      0.019     38.656      0.000       0.701       0.776
HighChol       0.5749      0.018     32.674      0.000       0.540       0.609
CholCheck      1.2202      0.088     13.869      0.000       1.048       1.393
BMI            0.0688      0.001     54.707      0.000       0.066       0.071
Smoker        -0.0087      0.017     -0.508      0.611      -0.042       0.025
Stroke         0.1686      0.032      5.219      0.000       0.105       0.232
HeartDiseaseorAttack     0.2352      0.023     10.247      0.000       0.190       0.280
PhysActivity  -0.0515      0.019     -2.755      0.006      -0.088      -0.015
Fruits        -0.0249      0.018     -1.402      0.161      -0.060       0.010
Veggies       -0.0546      0.021     -2.656      0.008      -0.095      -0.014
HvyAlcoholConsump       -0.8005      0.050    -15.903      0.000      -0.899      -0.702
AnyHealthcare  0.0434      0.043      1.010      0.313      -0.041       0.128
NoDocbcCost    0.0613      0.030      2.074      0.038       0.003       0.119
GenHlth        0.5267      0.011     50.078      0.000       0.506       0.547
MentHlth      -0.0037      0.001     -3.308      0.001      -0.006      -0.001
PhysHlth      -0.0066      0.001     -6.524      0.000      -0.009      -0.005
DiffWalk       0.0720      0.022      3.269      0.001       0.029       0.115
Sex            0.2477      0.017     14.221      0.000       0.214       0.282
Age            0.1276      0.004     35.199      0.000       0.120       0.135
Education     -0.0325      0.009     -3.602      0.000      -0.050      -0.015
Income        -0.0477      0.005    -10.330      0.000      -0.057      -0.039
==============================================================================
```

Removed 'Smoker', 'Fruits', and 'AnyHealthcare' variables due to non-significance ($p > 0.05$).

Best Model Summary

2. Logistic Regression Model with Selected Features

```
Classification Report - Training Set:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.92    130908
         1.0       0.52      0.14      0.22     21091

    accuracy                           0.86    151999
   macro avg       0.70      0.56      0.57    151999
weighted avg       0.83      0.86      0.83    151999


Classification Report - Validation Set:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93     43625
         1.0       0.55      0.15      0.23      7042

    accuracy                           0.86     50667
   macro avg       0.71      0.56      0.58     50667
weighted avg       0.83      0.86      0.83     50667


Classification Report - Test Set:
              precision    recall  f1-score   support

         0.0       0.87      0.98      0.92     43500
         1.0       0.52      0.13      0.21      7167

    accuracy                           0.86     50667
   macro avg       0.70      0.56      0.57     50667
weighted avg       0.82      0.86      0.82     50667
```



Feature selection based on p-values did not significantly improve the model's ability to handle class imbalance or detect the minority class.

Best Model Summary

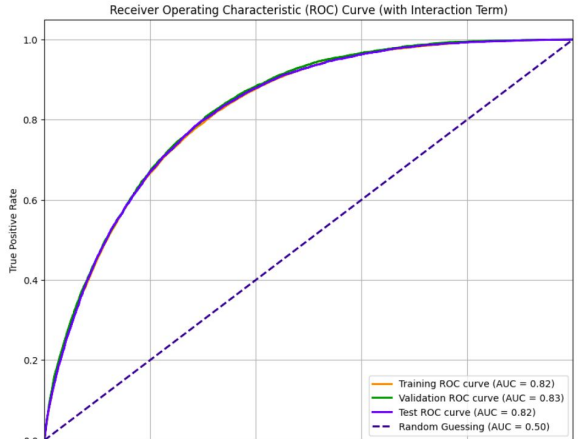
3. Logistic Regression Model with Interaction Term

Effect of walking difficulty on diabetes risk is likely age dependent.

Create the Interaction Term between DiffWalk and Age

The interaction term provided a minor increase in model performance.

```
Classification Report - Training Set (with Interaction Term):
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93    130908
         1.0       0.54      0.16      0.24     21091

    accuracy                           0.86    151999
   macro avg       0.71      0.57      0.59    151999
weighted avg       0.83      0.86      0.83    151999


Classification Report - Validation Set (with Interaction Term):
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93     43625
         1.0       0.58      0.16      0.25      7042

    accuracy                           0.87     50667
   macro avg       0.73      0.57      0.59     50667
weighted avg       0.84      0.87      0.83     50667


Classification Report - Test Set (with Interaction Term):
              precision    recall  f1-score   support

         0.0       0.87      0.98      0.92     43500
         1.0       0.55      0.15      0.23      7167

    accuracy                           0.86     50667
   macro avg       0.71      0.56      0.58     50667
weighted avg       0.83      0.86      0.83     50667
```



Best Model Summary

3. Logistic Regression Model with Interaction Term

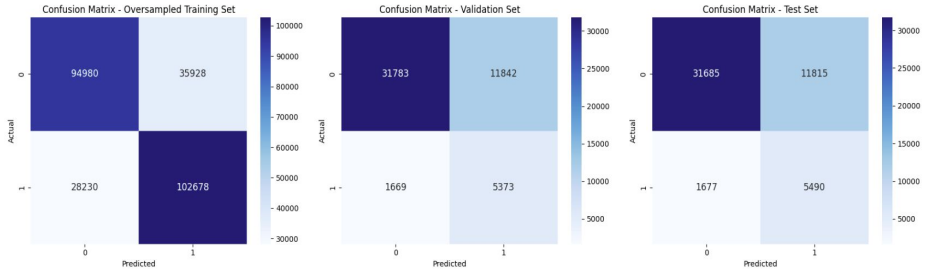
After adding the DiffWalk × Age Interaction Term:

- The interaction term is highly statistically significant (p = 0) → This confirms that the effect of mobility issues on diabetes risk is dependent on the patient's age.
- Most important predictors (HighBP, HighChol, BMI, GenHlth, Age, Sex, etc.) remain strong and significant (p < 0.005).
- NoDocbcCost becomes non-significant (p > 0.005) → its effect is absorbed by stronger predictors. → dropped from the model.
- The model captures a more realistic relationship between mobility limitation and aging.

```
Optimization terminated successfully.
         Current function value: 0.318256
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:        Diabetes_binary   No. Observations:               151999
Model:                          Logit   Df Residuals:                   151979
Method:                           MLE   Df Model:                           19
Date:                Thu, 27 Nov 2025   Pseudo R-squ.:                  0.2097
Time:                        22:03:19   Log-Likelihood:                -48375.
converged:                       True   LL-Null:                       -61210.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                         -8.1079      0.118    -68.462      0.000      -8.340      -7.876
HighBP                         0.7336      0.019     38.411      0.000       0.696       0.771
HighChol                       0.5694      0.018     32.388      0.000       0.535       0.604
CholCheck                      1.2199      0.088     13.874      0.000       1.048       1.392
BMI                            0.0684      0.001     54.421      0.000       0.066       0.071
Stroke                         0.1727      0.032      5.357      0.000       0.110       0.236
HeartDiseaseorAttack           0.2375      0.023     10.384      0.000       0.193       0.282
PhysActivity                  -0.0592      0.019     -3.178      0.001      -0.096      -0.023
Education                     -0.0331      0.009     -3.688      0.000      -0.051      -0.016
Veggies                       -0.0619      0.020     -3.094      0.002      -0.101      -0.023
HvyAlcoholConsump             -0.8018      0.050    -15.969      0.000      -0.900      -0.703
NoDocbcCost                    0.0442      0.029      1.515      0.130      -0.013       0.101
GenHlth                        0.5269      0.011     50.132      0.000       0.506       0.548
MentHlth                      -0.0045      0.001     -4.036      0.000      -0.007      -0.002
PhysHlth                      -0.0071      0.001     -6.963      0.000      -0.009      -0.005
DiffWalk                       0.6823      0.075      9.106      0.000       0.535       0.829
Sex                            0.2457      0.017     14.295      0.000       0.212       0.279
Age                            0.1424      0.004     35.373      0.000       0.135       0.150
Income                        -0.0446      0.005     -9.669      0.000      -0.054      -0.036
DiffWalk_Age_Interaction      -0.0629      0.007     -8.482      0.000      -0.077      -0.048
==============================================================================
```

Best Model Summary

4. Logistic Regression Model with Oversampling



- The dataset remained highly imbalanced, which hurt the model's ability to detect diabetes cases.
- Applied SMOTE oversampling on the training set only.

	Before SMOTE	After SMOTE
Majority class (0)	130,908	130,908
Minority class (1)	21,091	130,908
Training size	151,999	261,816

SMOTE balanced the training dataset to ~130,000 samples per class, allowing the model to learn minority-class patterns more effectively.

Best Model Summary

4. Logistic Regression Model with Oversampling

Oversampling: increased minority class samples to ~130,000.

Recall Improvement: Jumped from 0.15 to 0.77 on the minority class → good at identifying actual diabetes cases

The Trade-off: Precision dropped to 0.31. This is acceptable for a screening tool we'd rather re-test a healthy person than miss a sick one.

Generalization: Consistent performance across Validation and Test sets.

Classification Report - Oversampled Training Set:

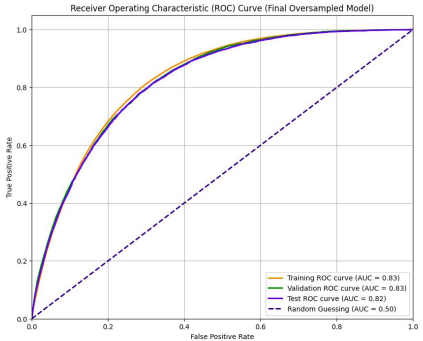
	precision	recall	f1-score	support
0.0	0.77	0.73	0.75	130908
1.0	0.74	0.78	0.76	130908
accuracy			0.75	261816
macro avg	0.76	0.75	0.75	261816
weighted avg	0.76	0.75	0.75	261816

Classification Report - Validation Set:

	precision	recall	f1-score	support
0.0	0.95	0.73	0.82	43625
1.0	0.31	0.76	0.44	7042
accuracy			0.73	50667
macro avg	0.63	0.75	0.63	50667
weighted avg	0.86	0.73	0.77	50667

Classification Report - Test Set:

	precision	recall	f1-score	support
0.0	0.95	0.73	0.82	43500
1.0	0.32	0.77	0.45	7167
accuracy			0.73	50667
macro avg	0.63	0.75	0.64	50667
weighted avg	0.86	0.73	0.77	50667



Best Model Summary

4. Logistic Regression Model with Oversampling

Positive Correlation

- ✓ **High Blood Pressure**
- ✓ **High Cholesterol**
- ✓ **General Health**
- ✓ **BMI**
- ✓ **Stroke**

Negative Correlation

- ✓ **Heavy Alcohol Consumption**
- ✓ **Income**
- ✓ **DiffWalk_Age_Interaction**

- **P-Value:** All key variables shown are statistically significant.

```
Optimization terminated successfully.
         Current function value: 0.504759
         Iterations 7

Statsmodels Summary (Oversampled Model):
                           Logit Regression Results
==============================================================================
Dep. Variable:        Diabetes_binary   No. Observations:               261816
Model:                          Logit   Df Residuals:                   261797
Method:                           MLE   Df Model:                           18
Date:                Tue, 02 Dec 2025   Pseudo R-squ.:                  0.2718
Time:                        14:46:18   Log-Likelihood:            -1.3215e+05
converged:                       True   LL-Null:                   -1.8148e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         -7.5803      0.068   -111.103      0.000      -7.714      -7.447
HighBP                         0.8601      0.011     78.547      0.000       0.839       0.882
HighChol                       0.6667      0.010     64.026      0.000       0.646       0.687
CholCheck                      1.5596      0.048     32.338      0.000       1.465       1.654
BMI                            0.0751      0.001     89.454      0.000       0.073       0.077
Stroke                         0.0719      0.024      3.050      0.002       0.026       0.118
HeartDiseaseorAttack           0.1669      0.016     10.378      0.000       0.135       0.198
PhysActivity                   0.0316      0.012      2.639      0.008       0.008       0.055
Education                     -0.0243      0.006     -4.314      0.000      -0.035      -0.013
Veggies                        0.0368      0.013      2.859      0.004       0.012       0.062
HvyAlcoholConsump             -1.0651      0.029    -36.959      0.000      -1.122      -1.009
GenHlth                        0.6537      0.006    102.063      0.000       0.641       0.666
MentHlth                      -0.0056      0.001     -8.295      0.000      -0.007      -0.004
PhysHlth                      -0.0099      0.001    -15.836      0.000      -0.011      -0.009
DiffWalk                       0.6092      0.048     12.750      0.000       0.516       0.703
Sex                            0.3015      0.010     28.893      0.000       0.281       0.322
Age                            0.1587      0.002     71.016      0.000       0.154       0.163
Income                        -0.0528      0.003    -19.040      0.000      -0.058      -0.047
DiffWalk_Age_Interaction      -0.0607      0.005    -12.657      0.000      -0.070      -0.051
==============================================================================================
```

Relevant Business Findings

- High Blood Pressure, High Cholesterol, and BMI are among the strongest predictors of diabetes

→ These groups represent the highest medical-risk employees → the company should prioritize them for screening and prevention.

- General Health (self-reported) is highly predictive of risk

→ Employees who rate their health as fair/poor tend to generate higher medical costs in the future.

- The DiffWalk × Age interaction is statistically significant

→ Mobility difficulties become much more dangerous as employees get older → older employees with movement limitations should be prioritized for health interventions.

- Income is negatively correlated with diabetes risk

→ Lower-income employees face higher risk → they likely have more barriers to care and should receive additional support.

Recommendations

- **Targeted Screening:** Focus on high-risk employees (High Blood Pressure, High cholesterol, High BMI, Age+Mobility issues) to reduce unnecessary diabetic testing.
- **Wellness Investment:** Subsidize gym memberships nutrition coaching, and meal plan support to combat strong lifestyle predictors.
- **Mobility Support:** Provide ergonomic workstations, physical therapy coverage and standing desks, especially for the older workforce.
- **Reduce Barriers:** Lower copays for screening and offer free annual diabete checks for lower-income groups.

Appendix

Results Comparison Table

Model	Partition	Non-Target Class (0)			Target Class (1)			Accuracy	AUC
		Precision	Recall	F1	Precision	Recall	F1		
Logistic Regression Model - Initial	Train	0.88	0.98	0.92	0.52	0.16	0.24	0.86	0.81
	Validation	0.88	0.98	0.93	0.54	0.17	0.26	0.86	0.81
	Test	0.88	0.98	0.92	0.52	0.15	0.24	0.86	0.81
Logistic Regression Model with Selected Features	Train	0.88	0.98	0.92	0.52	0.14	0.22	0.86	0.81
	Validation	0.88	0.98	0.93	0.55	0.15	0.23	0.86	0.81
	Test	0.87	0.98	0.92	0.52	0.13	0.21	0.86	0.81
Logistic Regression Model with Interaction Term	Train	0.88	0.98	0.93	0.54	0.16	0.24	0.86	0.82
	Validation	0.88	0.98	0.93	0.58	0.16	0.25	0.87	0.83
	Test	0.87	0.98	0.92	0.55	0.15	0.23	0.86	0.82
Logistic Regression Model with Oversampling	Train	0.77	0.73	0.75	0.74	0.78	0.76	0.75	0.83
	Validation	0.95	0.73	0.82	0.31	0.76	0.44	0.73	0.83
	Test	0.95	0.73	0.82	0.32	0.77	0.45	0.73	0.82

Results Comparison Table

Model	Partition	Non-Target Class (0)			Target Class (1)			Accuracy	AUC
		Precision	Recall	F1	Precision	Recall	F1		
Decision Tree - Initial	Train	0.99	1.00	1.00	1.00	0.97	0.98	1.00	1.00
	Validation	0.89	0.87	0.88	0.29	0.33	0.31	0.80	0.60
	Test	0.89	0.87	0.88	0.30	0.33	0.31	0.80	0.60
Decision Tree - Tuned Parameters	Train	0.88	0.98	0.93	0.61	0.17	0.26	0.87	0.84
	Validation	0.88	0.98	0.93	0.54	0.14	0.23	0.86	0.81
	Test	0.87	0.98	0.92	0.53	0.14	0.22	0.86	0.81
Gradient Boosted Tree - Initial	Train	0.88	0.98	0.93	0.58	0.17	0.26	0.87	0.83
	Validation	0.88	0.98	0.93	0.59	0.17	0.26	0.87	0.83
	Test	0.88	0.98	0.93	0.57	0.16	0.25	0.86	0.83
Gradient Boosted Tree - Tuned Parameters	Train	0.88	0.98	0.93	0.60	0.17	0.27	0.87	0.83
	Validation	0.88	0.98	0.93	0.59	0.17	0.26	0.87	0.83
	Test	0.88	0.98	0.93	0.58	0.16	0.25	0.86	0.83

Results Comparison Table

Model	Partition	Non-Target Class (0)			Target Class (1)			Accuracy	AUC
		Precision	Recall	F1	Precision	Recall	F1		
Random Forest - Initial	Train	1.00	1.00	1.00	1.00	0.97	0.98	1.00	1.00
	Validation	0.88	0.97	0.92	0.49	0.16	0.24	0.86	0.80
	Test	0.88	0.97	0.92	0.50	0.16	0.25	0.86	0.80
Random Forest - Tuned Parameters	Train	0.88	0.99	0.93	0.76	0.15	0.25	0.88	1.00
	Validation	0.87	0.99	0.93	0.64	0.11	0.19	0.87	0.80
	Test	0.87	0.99	0.93	0.60	0.10	0.17	0.86	0.80
Neural Network - Initial	Train	0.89	0.98	0.93	0.60	0.22	0.32	0.87	0.56
	Validation	0.88	0.97	0.93	0.54	0.19	0.28	0.86	0.56
	Test	0.88	0.97	0.92	0.54	0.19	0.28	0.86	0.55
Neural Network - Oversampling	Train	0.80	0.81	0.81	0.81	0.80	0.81	0.81	0.90
	Validation	0.93	0.80	0.86	0.34	0.62	0.44	0.78	0.80
	Test	0.93	0.81	0.86	0.35	0.63	0.45	0.78	0.80

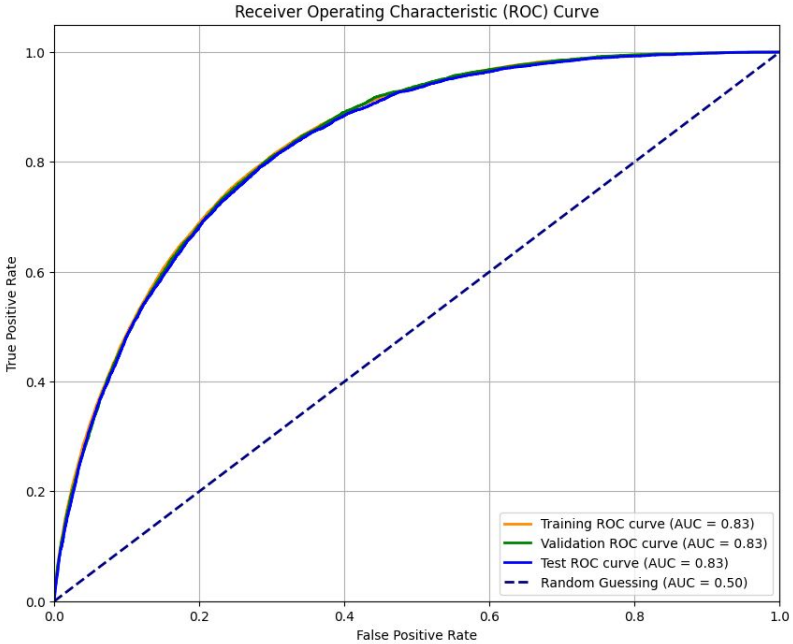
Results Comparison Table

Model	Partition	Non-Target Class (0)			Target Class (1)			Accuracy	AUC
		Precision	Recall	F1	Precision	Recall	F1		
Discriminant Analysis	Train	0.88	0.97	0.92	0.51	0.20	0.29	0.86	0.82
	Validation	0.88	0.97	0.93	0.53	0.21	0.30	0.86	0.82
	Test	0.88	0.97	0.92	0.52	0.20	0.29	0.86	0.82
KNN - Default K	Train	0.90	0.98	0.94	0.68	0.33	0.44	0.89	0.90
	Validation	0.88	0.95	0.92	0.41	0.19	0.26	0.85	0.71
	Test	0.88	0.95	0.91	0.40	0.19	0.26	0.85	0.71
KNN - Optimal K	Train	0.88	0.98	0.93	0.63	0.17	0.26	0.87	0.86
	Validation	0.87	0.98	0.92	0.48	0.12	0.20	0.86	0.78
	Test	0.87	0.98	0.92	0.48	0.12	0.19	0.86	0.77
Naive Bayes Model	Train	0.92	0.81	0.86	0.32	0.57	0.41	0.77	0.78
	Validation	0.92	0.81	0.86	0.33	0.57	0.42	0.78	0.79
	Test	0.92	0.81	0.86	0.33	0.57	0.42	0.77	0.78

Boosted Tree Model

```
Best hyperparameters: {'learning_rate': 0.1, 'max_depth': 3, 'min_samples_split': 10, 'n_estimators': 300}
GradientBoostingClassifier
GradientBoostingClassifier(min_samples_split=10, n_estimators=300,
                           random_state=42)
```

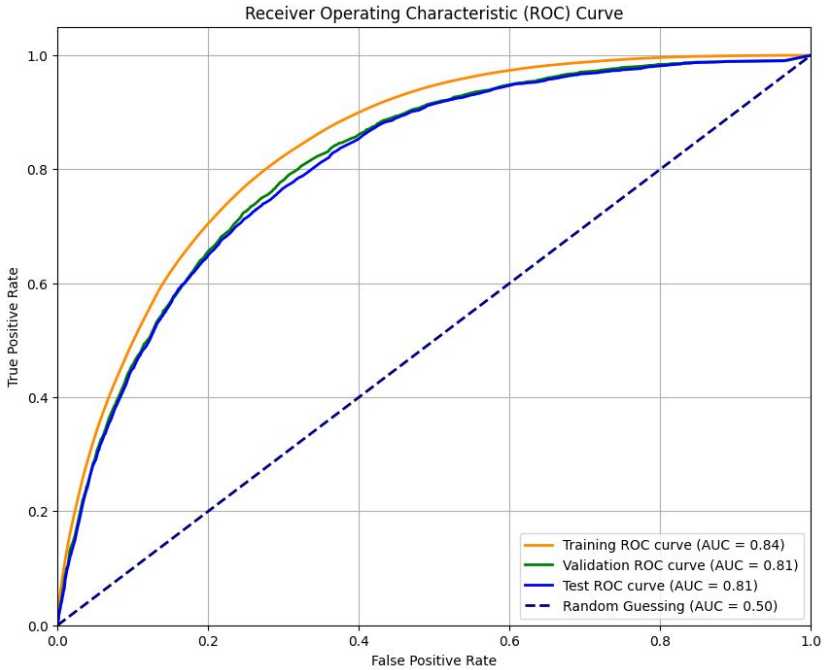
```
Confusion Matrix - Training Set:
[[128434   2474]
 [ 17401   3690]]

Confusion Matrix - Validation Set:
[[42790   835]
 [ 5862  1180]]

Confusion Matrix - Test Set:
[[42652   848]
 [ 6014  1153]]
```




```
Classification Report - Training Set:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93    130908
         1.0       0.60      0.17      0.27     21091

    accuracy                           0.87    151999
   macro avg       0.74      0.58      0.60    151999
weighted avg       0.84      0.87      0.84    151999


Classification Report - Validation Set:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93     43625
         1.0       0.59      0.17      0.26      7042

    accuracy                           0.87     50667
   macro avg       0.73      0.57      0.59     50667
weighted avg       0.84      0.87      0.83     50667


Classification Report - Test Set:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93     43500
         1.0       0.58      0.16      0.25      7167

    accuracy                           0.86     50667
   macro avg       0.73      0.57      0.59     50667
weighted avg       0.83      0.86      0.83     50667
```

Decision Tree Model

```
Best hyperparameters: {'max_depth': 10, 'min_samples_leaf': 10, 'min_samples_split': 2}
The tuned decision tree has 825 splits (internal nodes).
The tuned decision tree has 826 leaves.
```

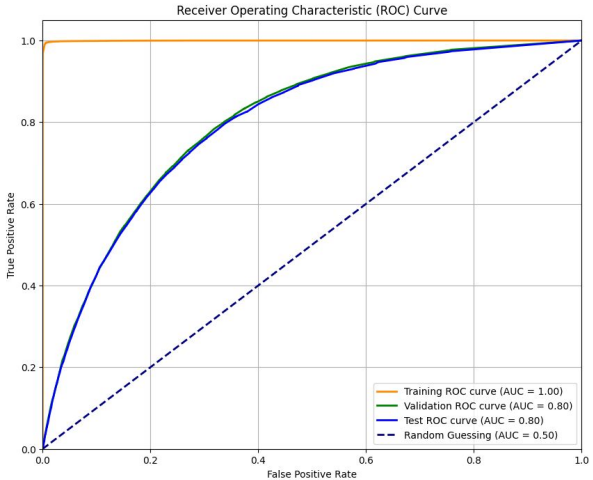



```
Confusion Matrix - Tuned Training Set:
[[128626   2282]
 [ 17581   3510]]

Confusion Matrix - Tuned Validation Set:
[[42766   859]
 [ 6034  1008]]

Confusion Matrix - Tuned Test Set:
[[42622   878]
 [ 6194   973]]
```

```
Classification Report - Training Set:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93    130908
         1.0       0.61      0.17      0.26     21091

    accuracy                           0.87    151999
   macro avg       0.74      0.57      0.59    151999
weighted avg       0.84      0.87      0.84    151999


Classification Report - Validation Set:
              precision    recall  f1-score   support

         0.0       0.88      0.98      0.93     43625
         1.0       0.54      0.14      0.23      7042

    accuracy                           0.86     50667
   macro avg       0.71      0.56      0.58     50667
weighted avg       0.83      0.86      0.83     50667


Classification Report - Test Set:
              precision    recall  f1-score   support

         0.0       0.87      0.98      0.92     43500
         1.0       0.53      0.14      0.22      7167

    accuracy                           0.86     50667
   macro avg       0.70      0.56      0.57     50667
weighted avg       0.82      0.86      0.82     50667
```

Bootstrap Forest Model

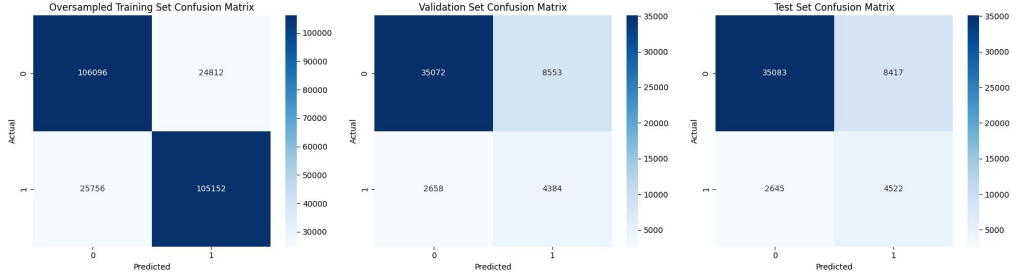
```
rf_tuned_model = RandomForestClassifier(
    n_estimators=300,
    max_depth=12,
    min_samples_split=5,
    min_samples_leaf=3,
    random_state=42
)
```

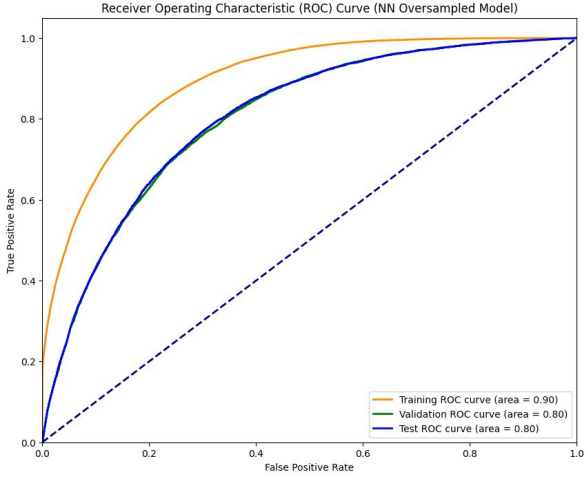



```
Confusion Matrix - Training Set:
[[129907   1001]
 [ 17982   3109]]

Confusion Matrix - Validation Set:
[[43191   434]
 [ 6259   783]]

Confusion Matrix - Test Set:
[[43007   493]
 [ 6437   730]]
```

```
Classification Report - Training Set:
              precision    recall  f1-score   support

         0.0       0.88      0.99      0.93    130908
         1.0       0.76      0.15      0.25     21091

    accuracy                           0.88    151999
   macro avg       0.82      0.57      0.59    151999
weighted avg       0.86      0.88      0.84    151999


Classification Report - Validation Set:
              precision    recall  f1-score   support

         0.0       0.87      0.99      0.93     43625
         1.0       0.64      0.11      0.19      7042

    accuracy                           0.87     50667
   macro avg       0.76      0.55      0.56     50667
weighted avg       0.84      0.87      0.83     50667


Classification Report - Test Set:
              precision    recall  f1-score   support

         0.0       0.87      0.99      0.93     43500
         1.0       0.60      0.10      0.17      7167

    accuracy                           0.86     50667
   macro avg       0.73      0.55      0.55     50667
weighted avg       0.83      0.86      0.82     50667
```

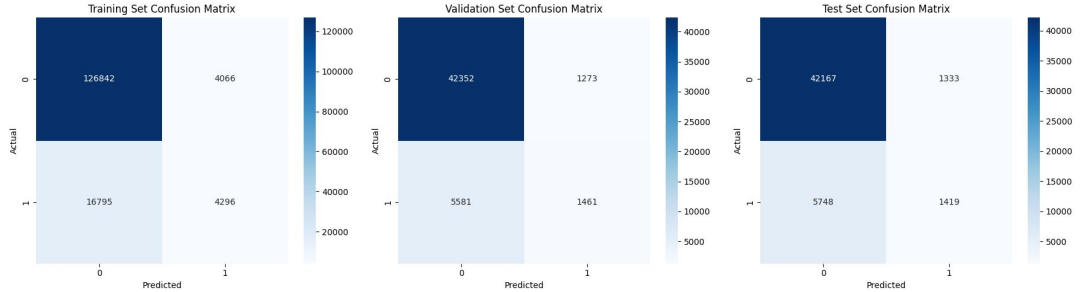
Neural Network Oversampled Model

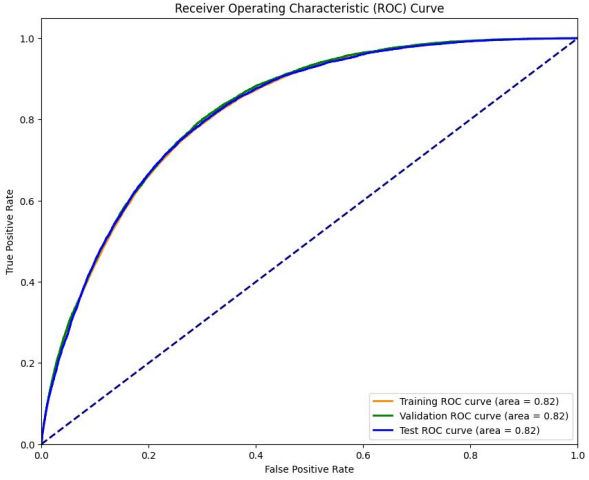





```
Oversampled Training Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.80      0.81      0.81    130908
         1.0       0.81      0.80      0.81    130908

    accuracy                           0.81    261816
   macro avg       0.81      0.81      0.81    261816
weighted avg       0.81      0.81      0.81    261816


Validation Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.93      0.80      0.86     43625
         1.0       0.34      0.62      0.44      7042

    accuracy                           0.78     50667
   macro avg       0.63      0.71      0.65     50667
weighted avg       0.85      0.78      0.80     50667


Test Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.93      0.81      0.86     43500
         1.0       0.35      0.63      0.45      7167

    accuracy                           0.78     50667
   macro avg       0.64      0.72      0.66     50667
weighted avg       0.85      0.78      0.81     50667
```

Discriminant Analysis Model



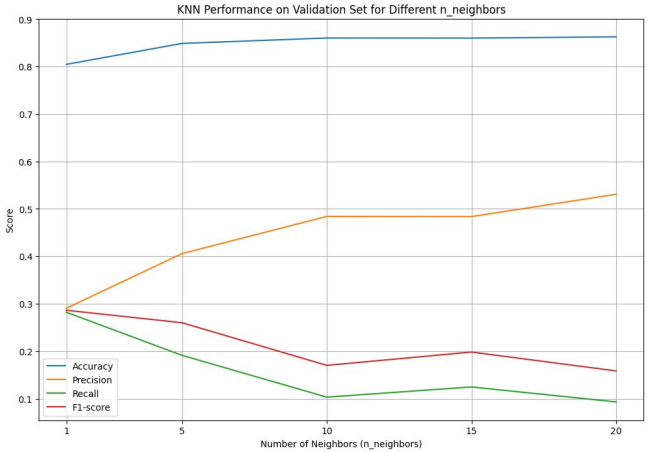



```
Training Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.88      0.97      0.92    130908
         1.0       0.51      0.20      0.29     21091

    accuracy                           0.86    151999
   macro avg       0.70      0.59      0.61    151999
weighted avg       0.83      0.86      0.84    151999


Validation Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.88      0.97      0.93     43625
         1.0       0.53      0.21      0.30      7042

    accuracy                           0.86     50667
   macro avg       0.71      0.59      0.61     50667
weighted avg       0.84      0.86      0.84     50667


Test Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.88      0.97      0.92     43500
         1.0       0.52      0.20      0.29      7167

    accuracy                           0.86     50667
   macro avg       0.70      0.58      0.60     50667
weighted avg       0.83      0.86      0.83     50667
```

```
Discriminant Function Coefficients:
HighBP                  0.731441
HighChol                0.550650
CholCheck               0.429144
BMI                     0.076142
Smoker                 -0.058603
Stroke                  0.425958
HeartDiseaseorAttack    0.681965
PhysActivity           -0.072947
Fruits                  0.011419
Veggies                -0.047441
HvyAlcoholConsump      -0.523783
AnyHealthcare           0.113335
NoDocbcCost            -0.024537
GenHlth                 0.464343
MentHlth               -0.006179
PhysHlth                0.001301
DiffWalk                0.375048
Sex                     0.152794
Age                     0.074460
Education              -0.035209
Income                 -0.059036
dtype: float64

Intercept: -6.916556620445842
```

KNN Model

Optimal K = 15



Training Set Classification Report:

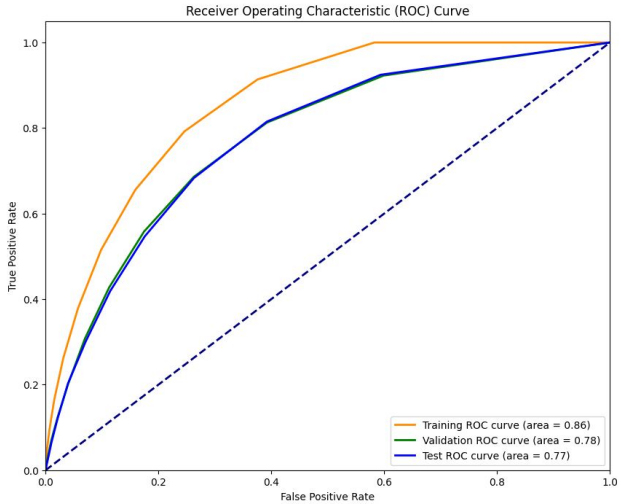
	precision	recall	f1-score	support
0.0	0.88	0.98	0.93	130908
1.0	0.63	0.17	0.26	21091
accuracy			0.87	151999
macro avg	0.76	0.57	0.60	151999
weighted avg	0.85	0.87	0.84	151999

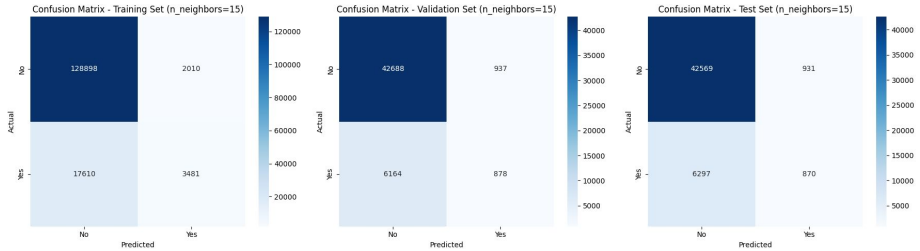
Validation Set Classification Report:

	precision	recall	f1-score	support
0.0	0.87	0.98	0.92	43625
1.0	0.48	0.12	0.20	7042
accuracy			0.86	50667
macro avg	0.68	0.55	0.56	50667
weighted avg	0.82	0.86	0.82	50667

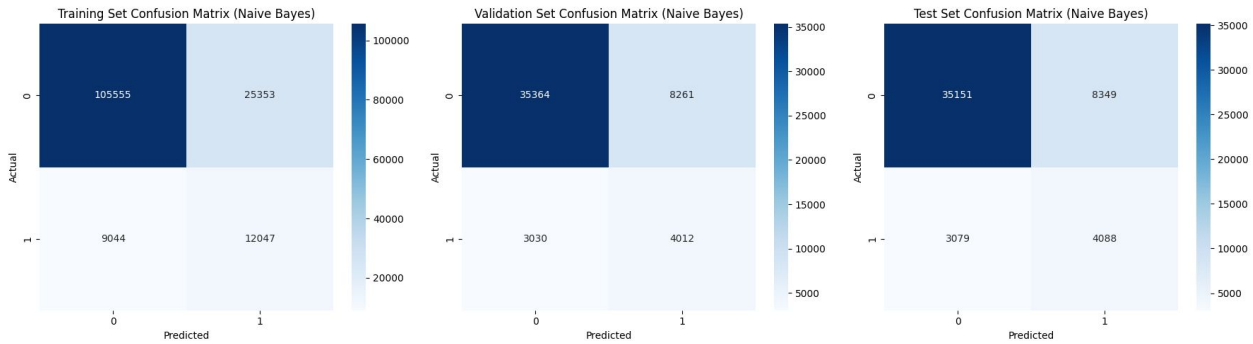
Test Set Classification Report:

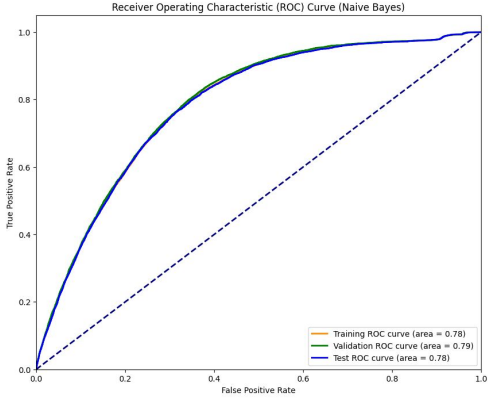
	precision	recall	f1-score	support
0.0	0.87	0.98	0.92	43500
1.0	0.48	0.12	0.19	7167
accuracy			0.86	50667
macro avg	0.68	0.55	0.56	50667
weighted avg	0.82	0.86	0.82	50667





Naive Bayes Model





Training Set Classification Report (Naive Bayes):

	precision	recall	f1-score	support
0.0	0.92	0.81	0.86	130908
1.0	0.32	0.57	0.41	21091
accuracy			0.77	151999
macro avg	0.62	0.69	0.64	151999
weighted avg	0.84	0.77	0.80	151999

Validation Set Classification Report (Naive Bayes):

	precision	recall	f1-score	support
0.0	0.92	0.81	0.86	43625
1.0	0.33	0.57	0.42	7042
accuracy			0.78	50667
macro avg	0.62	0.69	0.64	50667
weighted avg	0.84	0.78	0.80	50667

Test Set Classification Report (Naive Bayes):

	precision	recall	f1-score	support
0.0	0.92	0.81	0.86	43500
1.0	0.33	0.57	0.42	7167
accuracy			0.77	50667
macro avg	0.62	0.69	0.64	50667
weighted avg	0.84	0.77	0.80	50667

References

Diabetes Kaggle Dataset

Gemini - Code Assistance

Timothy M. Dall, Wenya Yang, Karin Gillespie, Michelle Mocarski, Erin Byrne, Inna Cintina, Kaleigh Beronja, April P. Semilla, William Iacobucci, Paul F. Hogan; The Economic Burden of Elevated Blood Glucose Levels in 2017: Diagnosed and Undiagnosed Diabetes, Gestational Diabetes Mellitus, and Prediabetes. Diabetes Care 1 September 2019; 42 (9): 1661–1668. https://doi.org/10.2337/dc18-1226

Kuo S, Ye W, Wang D, McEwen LN, Villatoro Santos C, Herman WH. Cost-effectiveness of the National Diabetes Prevention Program: A Real-world, 2-Year Prospective Study. Diabetes Care. 2025 Jul 1;48(7):1180-1188. doi: 10.2337/dc24-1110. PMID: 39565893; PMCID: PMC12178621.

Park J, Bigman E, Zhang P. Productivity Loss and Medical Costs Associated With Type 2 Diabetes Among Employees Aged 18-64 Years With Large Employer-Sponsored Insurance. Diabetes Care. 2022 Nov 1;45(11):2553-2560. doi: 10.2337/dc22-0445. PMID: 36048852; PMCID: PMC9633402.

A Report Card: Diabetes in the United States Infographic: https://www.cdc.gov/diabetes/communication-resources/diabetes-statistics.html

Shannon Arens in Benefits. Addressing Rising Medical Costs: Strategies for Startups. July 29, 2024: https://www.sequoia.com/2024/07/addressing-rising-medical-costs-strategies-for-startups/

Cathy DeWitt Dunn. Rising Healthcare Costs Draining Retirement Savings: https://www.annuitywatchusa.com/rising-healthcare-costs-draining-retirement-savings/