Machine learning algorithms predicting cancer associated with diabetes and hypertension: NHANES 2021 to 2023

Name: Chieh H Chang

Question

- What are the relationships between diabetes, hypertension, and cancer?

- Predict the development of cancer using 4 Machine learning algorithms.

Background and Significance

Cancer, diabetes, and hypertension are major health issues in the U.S. Research shows that type 2 diabetes and hypertension are linked and may increase cancer risk (Xu & Huang, 2024).

Early cancer prediction in individuals with diabetes and hypertension is vital. However, current machine learning risk assessments using national data are insufficient (Vangeepuram, Liu, Chiu, et al. (2021).

This study utilizes the NHANES dataset from 2021-2023, and machine learning to assess cancer risk in a high-risk group, aiming to enhance early detection and public health initiatives (Olshvang et al,2024).

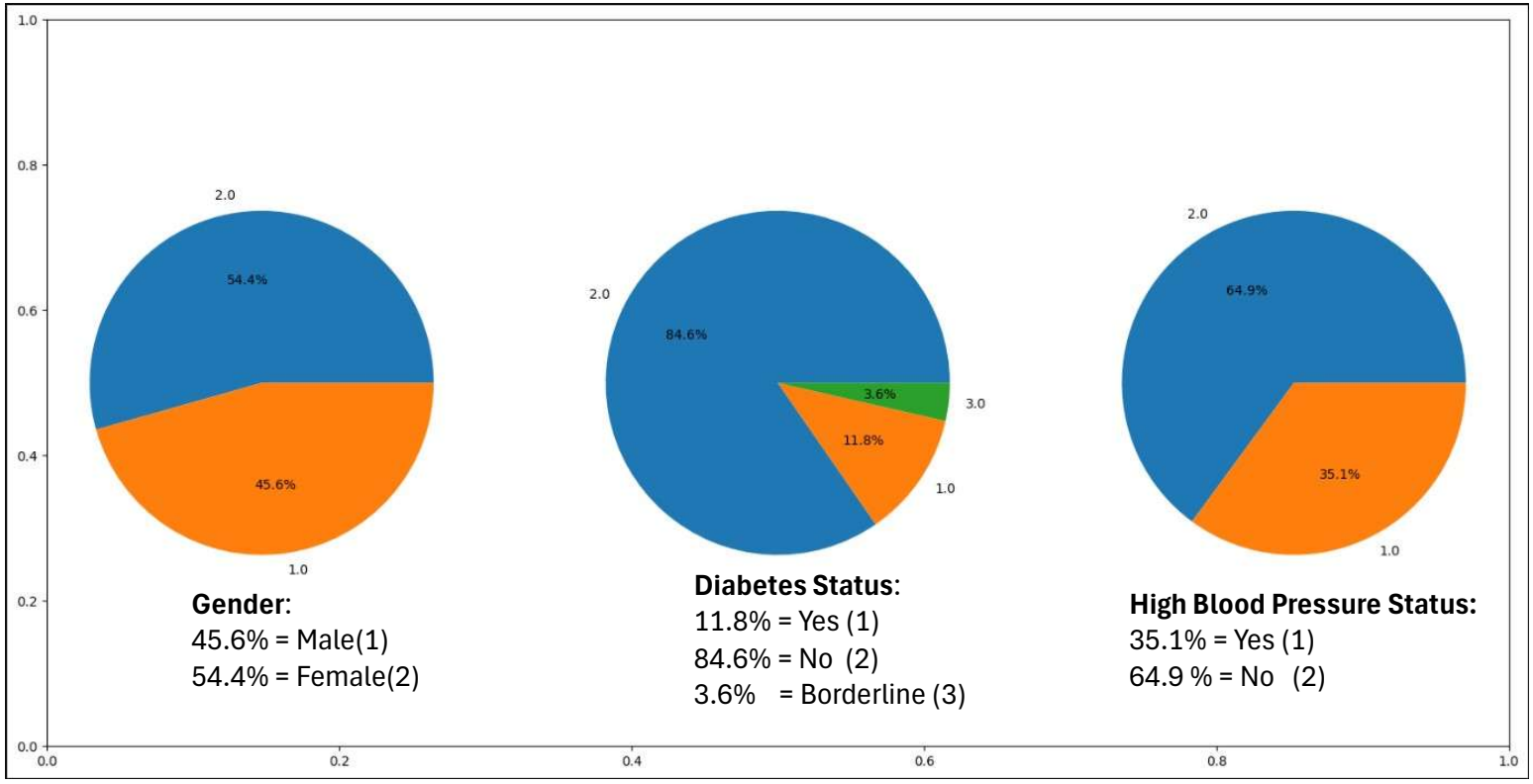
Approach

- What are the relationships between diabetes, hypertension, and cancer?
  - ➢using multivariate logistic regression.
- Can we predict the development of cancer?
  - ➢using 4 Machine learning algorithms.

Dataset

Key Variables by Diagnosis(MCQ220), N=2,135							
Diabetes Status(DIQ010)	n(%)	0 = has no cancer	1= has cancer	Ratio of Income Poverty (INDFMPIR)	n(%)	0 = has no cancer	1= has cancer
Yes	252(11.8)	195(9.1)	57(2.7)	Equal the poverty (RIP ≤ 1)	95(10.8)	86(9.9)	9(0.9)
No	1807(84.6)	1560(73.1)	247(11.6)	Above the poverty (RIP > 1)	713(89.2)	598(75.2)	115(14.0)
Borderline	76(3.6)	62(2.9)	14(0.7)	Smoked(SMQ020)			
Hypertension Status(BPQ020)				Yes	814(38.1)	678(31.8)	136(6.4)
Yes	749(35.1)	593(27.8)	156(7.3)	No	1321(61.9)	1139(53.3)	182(8.5)
No	1386(64.9)	1224(57.3)	162(7.6)	Physical Activities(PAD790Q)			
Age Group (RIDAGEYR)				Yes	1609(75.4)	1378(64.5)	231(10.8)
20-40 years	546(25.6)	539(25.2)	8(0.3)	No	526(24.6)	439(20.6)	87(4.1)
40-49 years	301(14.1)	238(13.5)	13(0.6)	Body Mass Index kg/m**2 (BMXBMI)			
50-59 years	328(15.4)	291(13.6)	37(1.7)	Underweight (≤ 18.5)	23(1.1)	2(0.1)	23(1.1)
60-69 years	545(25.5)	431(20.2)	114(5.3)	Normal Weight (18.5–24.9)	590(27.6)	499(23.4)	91(4.3)
70-75 years	224(10.5)	156(7.3)	68(3.2)	Overweight (25-29.9)	710(33.3)	606(28.4)	104(4.9)
>75 years	191(8.9)	112(5.2)	79(3.7)	Obesity (≥ 30)	812(38.0)	691(32.4)	121(5.7)
Gender(RIAGENDR)				Fasting Glucose mg/dL (LBXGLU)			
Male	974(45.6)	823(38.5)	151(7.1)	Normal (≤ 100)	971(45.5)	865(40.5)	106(5)
Female	1161(54.4)	994(46.6)	167(7.8)	Prediabetes (100-125)	920(43.1)	758(35.5)	162(7.6)
Ethnicity (RIDRETH3)				Diabetes (≥ 126)	244(11.4)	194(9.1)	50(2.3)
Mexican American	150(7.0)	147(6.9)	3(0.1)	Total Cholesterol mg/dL (LBXTC)			
Other Hispanic	176(8.2)	163(7.6)	13(0.6)	Normal (< 200)	1310(61.4)	1106(51.8)	204(9.6)
Non-Hispanic White	1345(63.0)	1083(50.7)	262(12.3)	Borderline High (200-239)	586(27.4)	508(23.8)	78(3.7)
Non-Hispanic Black	208(9.7)	196(9.2)	12(0.6)	High (≥ 240)	239(11.2)	203(9.5)	36(1.7)
Non-Hispanic Asian	121(5.7)	112(5.2)	9(0.4)	HS C-Reactive Protein mg/L(LBXHSCRP)			
Other Race	135(6.3)	116(5.4)	19(0.9)	Normal (< 3)	1411(66.1)	1191(55.8)	220(10.3)
Education (DMDEDUC2)				High (≥ 3)	724(33.9)	626(29.3)	98(4.6)
Less than 9th grade	68(3.2)	60(2.8)	8(0.4)				
9-11th grade	116(5.4)	104(4.9)	12(0.6)				
High school graduate/GED	388(18.2)	334(15.6)	54(2.5)				
Some college or AA degree	636(29.8)	559(26.2)	77(3.6)				
College graduate or above	927(43.4)	760(35.6)	167(7.8)				

Exploratory Data Analysis (EDA)



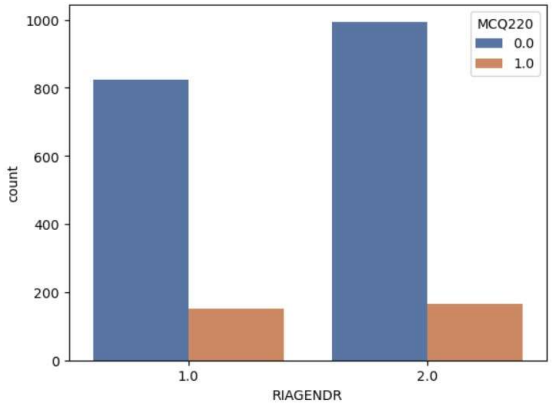
Gender:
45.6% = Male(1)
54.4% = Female(2)

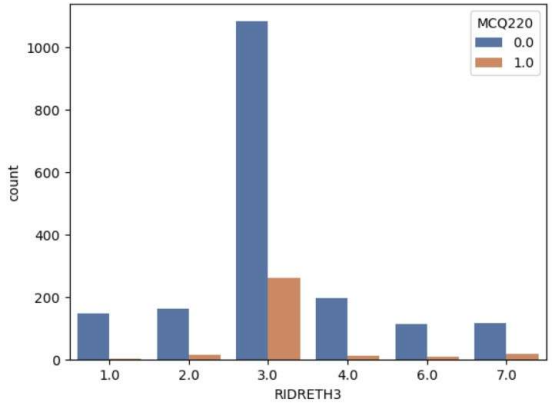
Diabetes Status:
11.8% = Yes (1)
84.6% = No (2)
3.6% = Borderline (3)

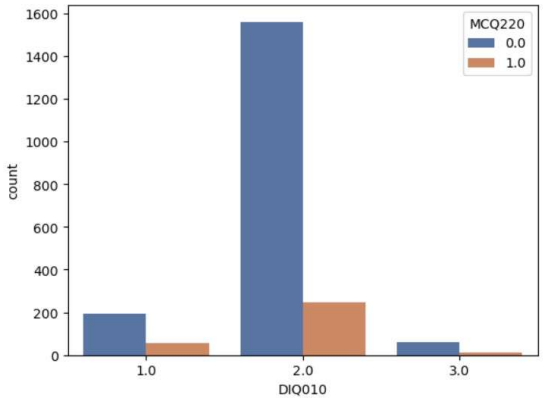
High Blood Pressure Status:
35.1% = Yes (1)
64.9 % = No (2)

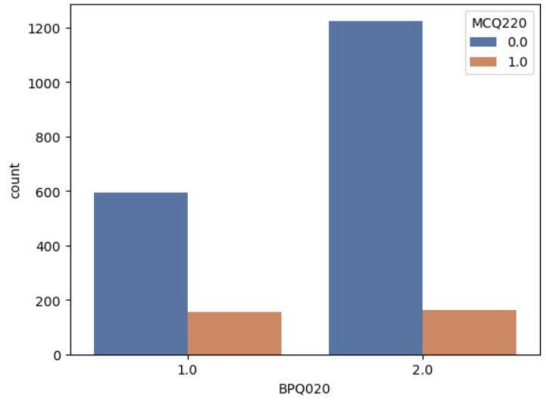
The dataset has a balanced gender distribution, which is beneficial for analysis. Many participants have no diabetes, affecting diabetes-related health outcomes

Exploratory Data Analysis (EDA)









Chi-Square Test

	Variable	Chi-square Statistic	P-value
2	RIDAGEYR	340.516719	7.830944e-41
4	RIDRETH3	68.850655	1.777188e-13
1	BPQ020	32.040688	1.509771e-08
0	DIQ010	14.748389	6.272318e-04
5	DMDEDUC2	13.732096	8.201059e-03
10	LBXGLU	192.888797	3.899405e-02
7	SMQ020	3.411381	6.474822e-02
11	LBXTC	247.868502	7.374837e-02
8	PAD790Q	19.063021	3.249268e-01
3	RIAGENDR	0.523132	4.695085e-01
6	INDFMPIR	338.722535	8.981408e-01
12	LBXHSCRP	712.289840	9.213791e-01
9	BMXBMI	262.292344	9.949314e-01

RIDAGEYR, RIDRETH3, BPQ020, DIQ010, DMDEDUC2, and LBXGLU: All six variables have p-values well below 0.05, indicating significant associations with the outcome.

SMQ020, LBXTC, PAD790Q, RIAGENDR, INDFMPIR LBXHSCRP, and BMXBMI: All seven variables have p-values greater than 0.05, indicating no significant associations with the outcome.

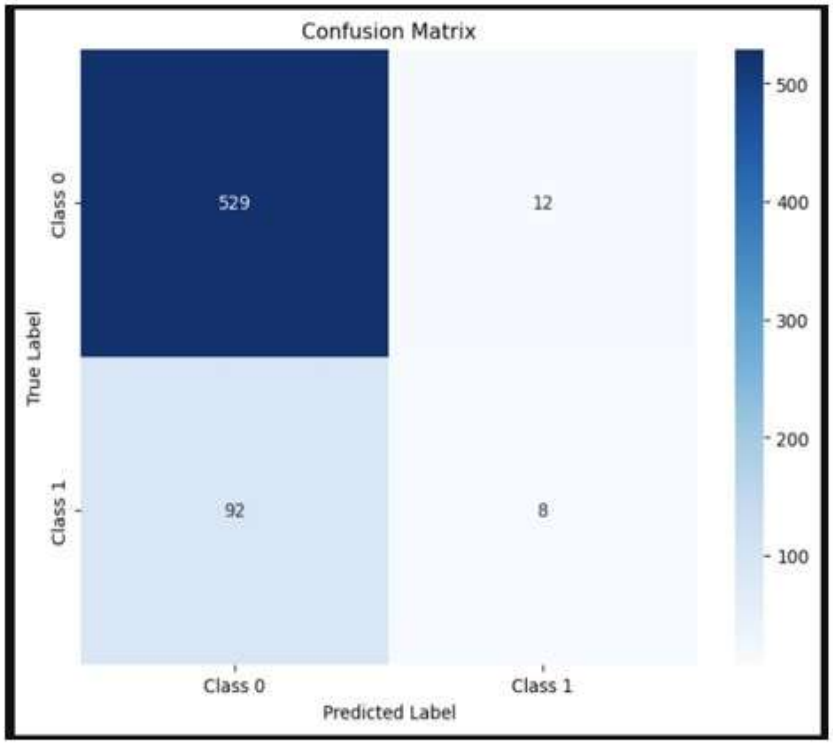
Logistic Regression P-Value

```
Optimization terminated successfully.
         Current function value: 0.350414
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                  MCQ220   No. Observations:                 1494
Model:                           Logit   Df Residuals:                     1481
Method:                            MLE   Df Model:                           12
Date:                 Sun, 20 Apr 2025   Pseudo R-squ.:                  0.1568
Time:                         19:44:06   Log-Likelihood:                -523.52
converged:                       True   LL-Null:                       -620.85
Covariance Type:            nonrobust   LLR p-value:                 4.120e-35
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
DIQ010        -0.9689      0.198     -4.895      0.000      -1.357      -0.581
BPQ020        -0.5221      0.164     -3.186      0.001      -0.843      -0.201
RIDAGEYR       0.0631      0.006     10.562      0.000       0.051       0.075
RIAGENDR       0.0498      0.164      0.303      0.762      -0.272       0.372
RIDRETH3      -0.1141      0.065     -1.764      0.078      -0.241       0.013
DMDEDUC2       0.1135      0.086      1.314      0.189      -0.056       0.283
INDFMPIR       0.0186      0.059      0.316      0.752      -0.097       0.134
SMQ020        -0.2340      0.164     -1.430      0.153      -0.555       0.087
PAD790Q        0.0098      0.032      0.303      0.762      -0.054       0.073
BMXBMI        -0.0450      0.013     -3.350      0.001      -0.071      -0.019
LBXGLU        -0.0103      0.003     -3.322      0.001      -0.016      -0.004
LBXTC         -0.0016      0.002     -0.859      0.391      -0.005       0.002
LBXHSCRP      -0.0021      0.012     -0.185      0.853      -0.025       0.020
==============================================================================
```

Significant Predictors (<0.001): DIQ010, BPQ020, RIDAGEYR, BMXBMI, and LBXGLU.

Non- Significant Predictors(>0.05) : RIAGENDR, RIDRETH3, DMDEDUC2, INDFMPIR, SMQ020, PAD790Q, LBXTC and LBXGLU.

Logistic Regression Confusion Matrix




```
confusion_matrix(y_test,y_pred)
```

```
array([[529,  12],
       [ 92,   8]], dtype=int64)
```

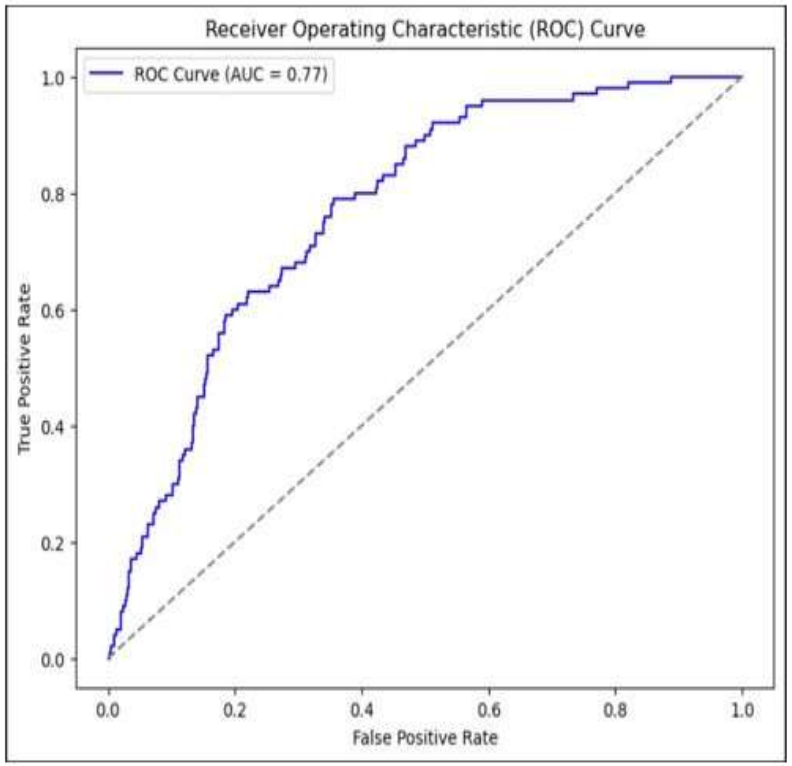
True Positives (TP): 8 individuals accurately diagnosed with cancer.

False Positives (FP): 12 individuals incorrectly identified as having cancer when they do not.

True Negatives (TN): 529 individuals accurately diagnosed with cancer.

False Negatives (FN): 92 individuals incorrectly identified as not having cancer when they actually do.

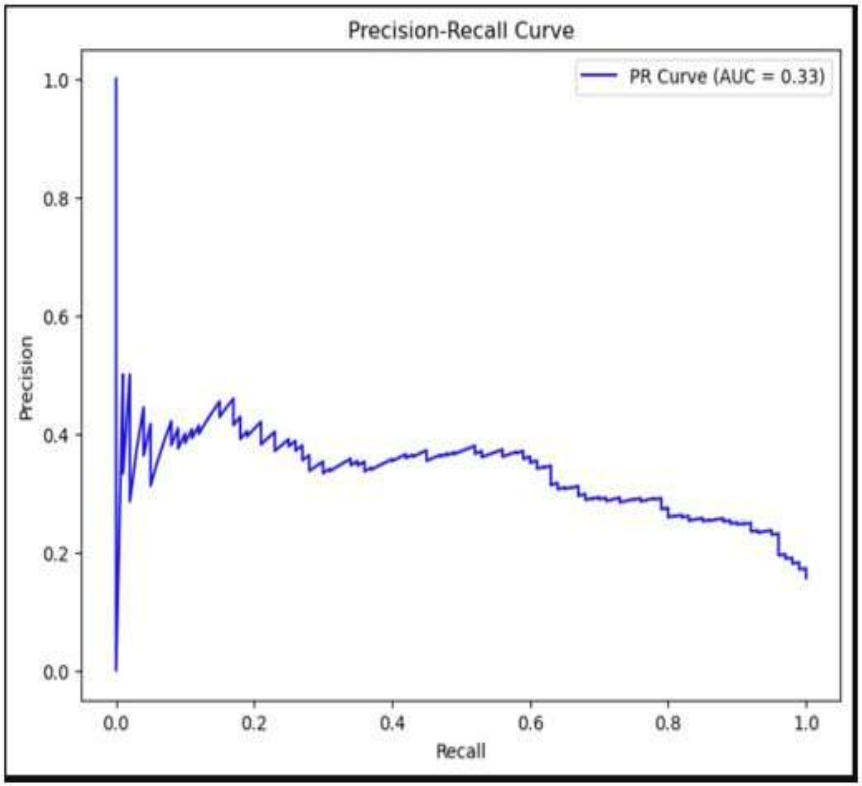
Logistic Regression
Receiver Operating Characteristic (ROC)



Acceptable Diagnostic Ability: An AUC of 0.77 signifies that the model has good ability to distinguish between the two classes (cancer vs. no cancer).

high false negatives and low false positives: The model Less effectively identifies true positives (individuals with cancer) while maintaining a low rate of false positives (individuals incorrectly identified as having cancer).

Logistic Regression
Precision-Recall (PR)



Poor Model Performance: An AUC of 0.33 suggests that the model has poor performance in distinguishing between the positive (cancer) and negative (no cancer) classes.

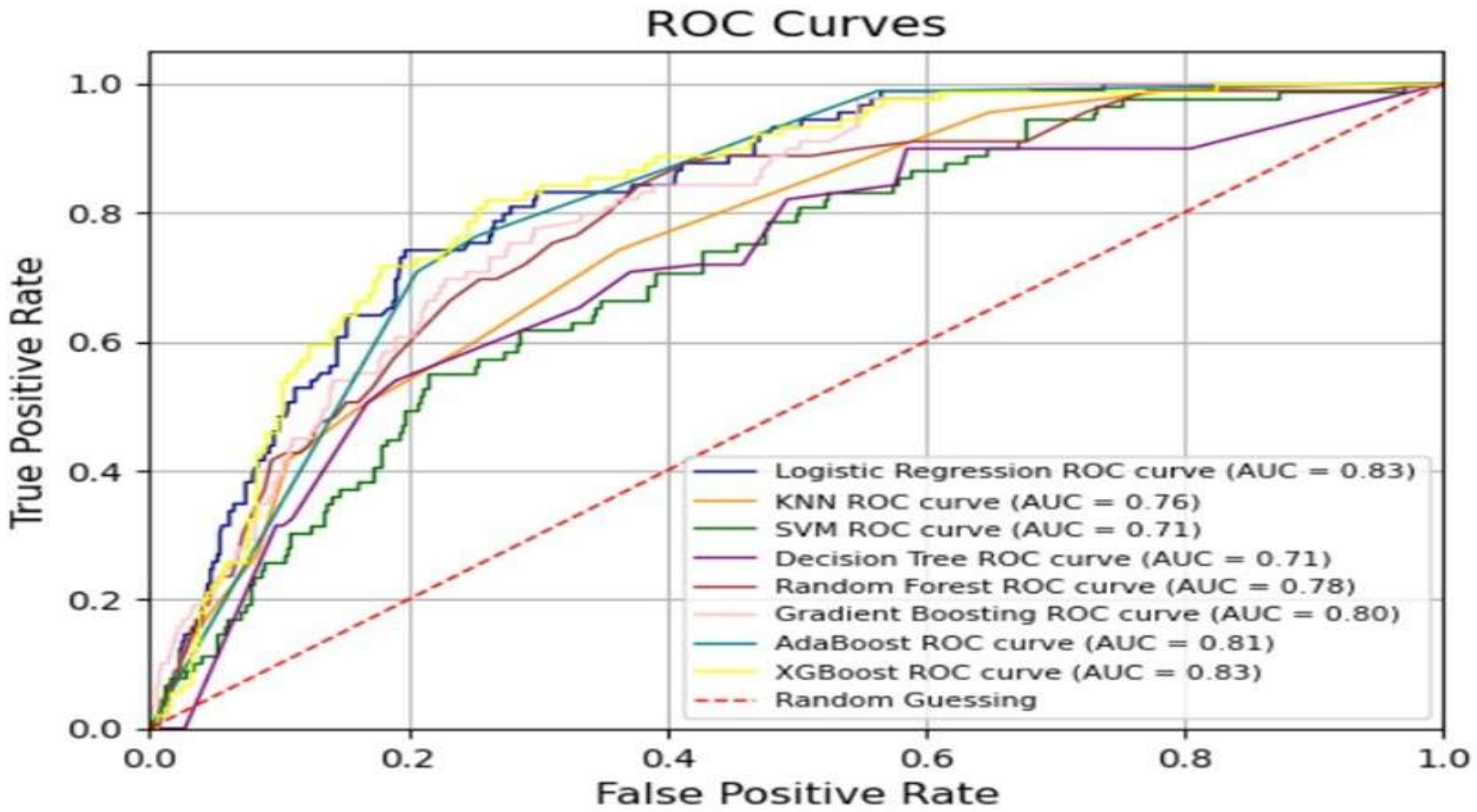
Balance between Precision and Recall: The very low AUC indicates that the model maintains a poor balance between precision (correctly predicting positives) and recall (correctly identifying actual positives).

Comparative Chart For All Models

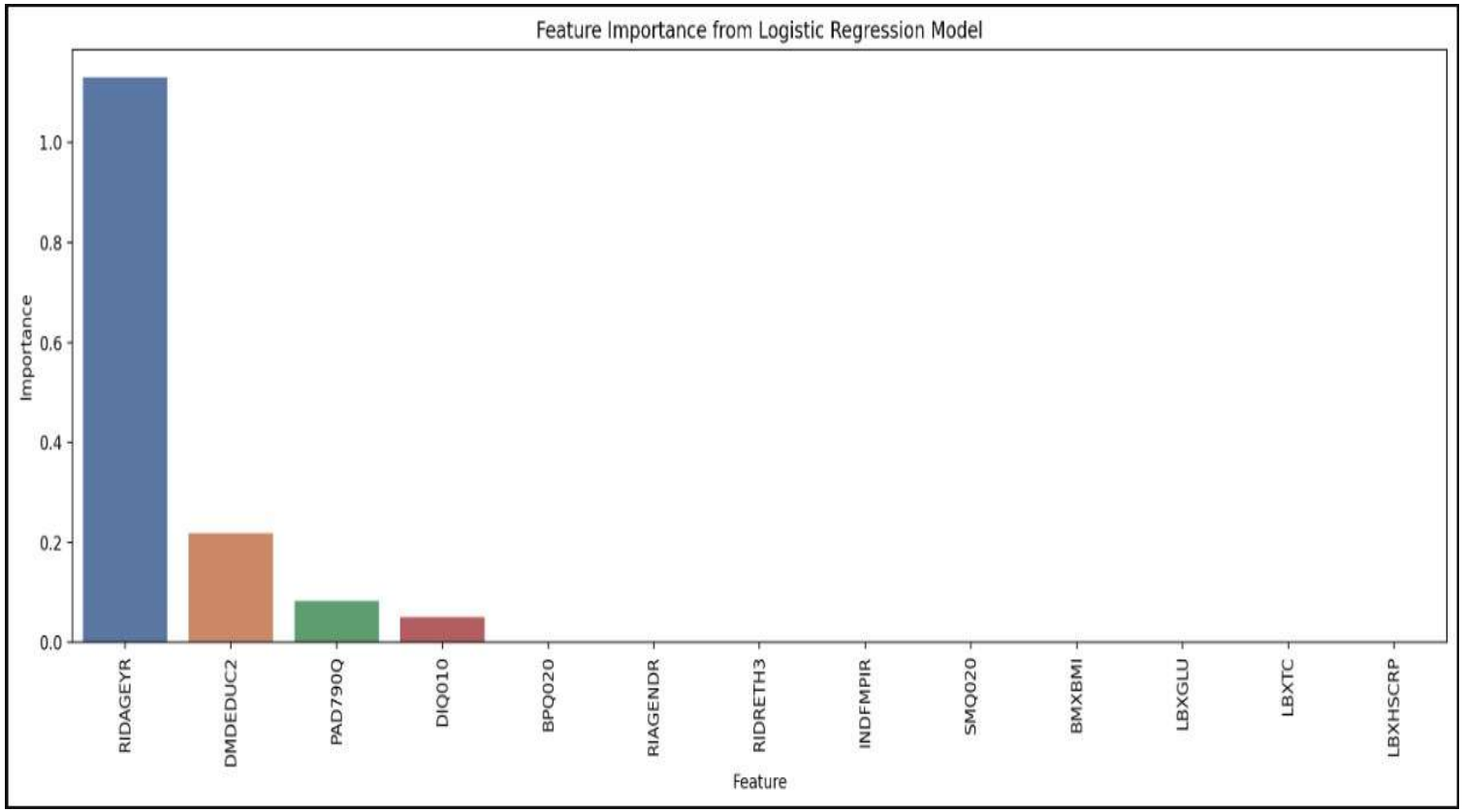
Model	Logistic Regression	KNN	SVM	Decision Tree	Random Forest	AdaBoost	Gradient Boosting	XGBoost
Precision (0)	0.85	0.86	0.84	0.84	0.85	0.85	0.85	0.86
Precision (1)	0.45	0.30	0.00	0.00	0.46	0.38	0.35	0.42
Recall (0)	0.98	0.93	1.00	1.00	0.99	0.97	0.97	0.95
Recall (1)	0.09	0.16	0.00	0.00	0.06	0.10	0.09	0.20
F1-Score (0)	0.91	0.89	0.91	0.91	0.91	0.91	0.91	0.90
F1- Score (1)	0.15	0.21	0.00	0.00	0.11	0.16	0.14	0.27
Accuracy	0.84	0.81	0.84	0.84	0.84	0.83	0.83	0.83
AUC	0.82	0.57	0.66	0.60	0.71	0.76	0.78	0.81

Logistic Regression, Random Forest, AdaBoost, Gradient Boosting, and XGBoost appear to be the more balanced performers overall.

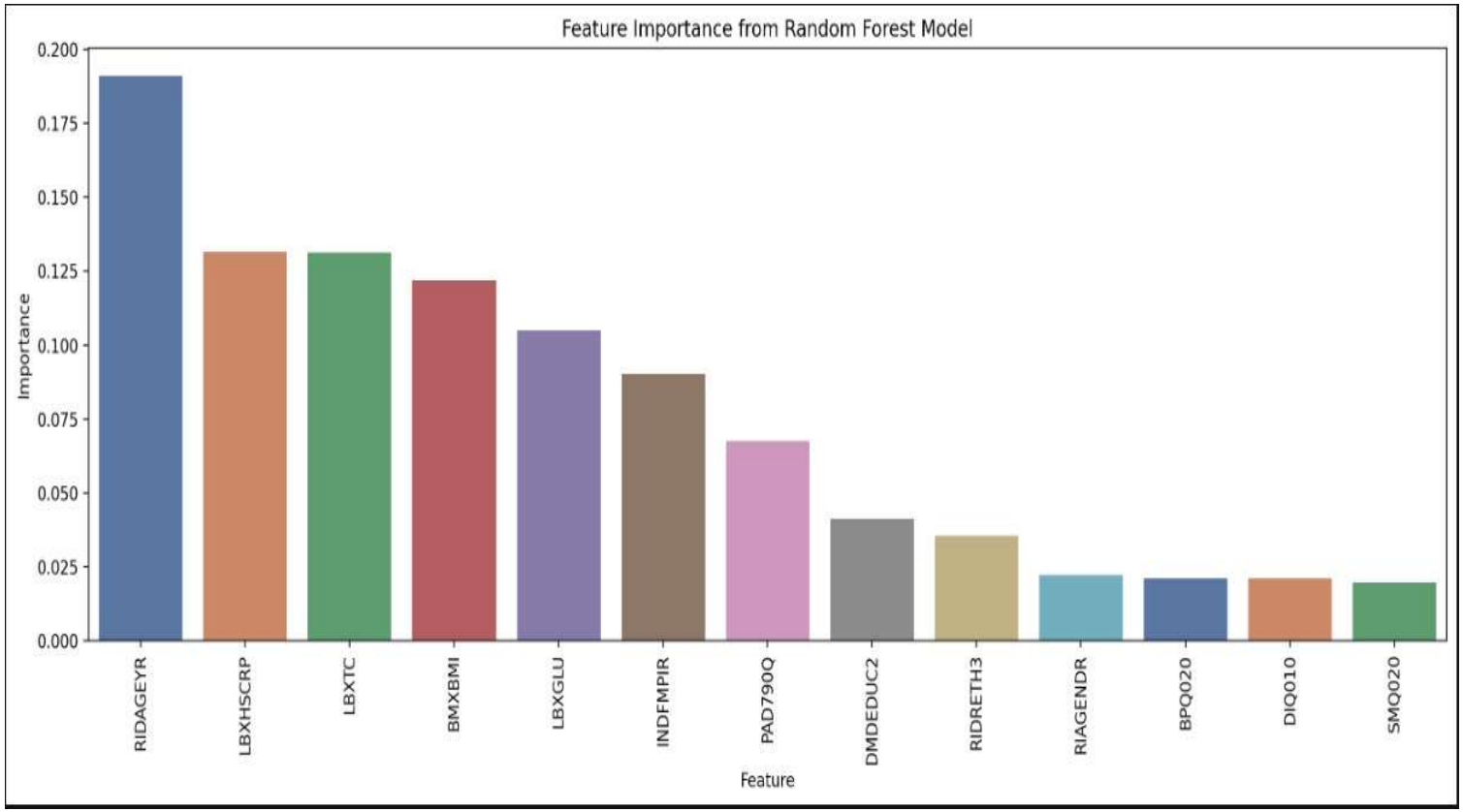
Comparative ROC For All Models



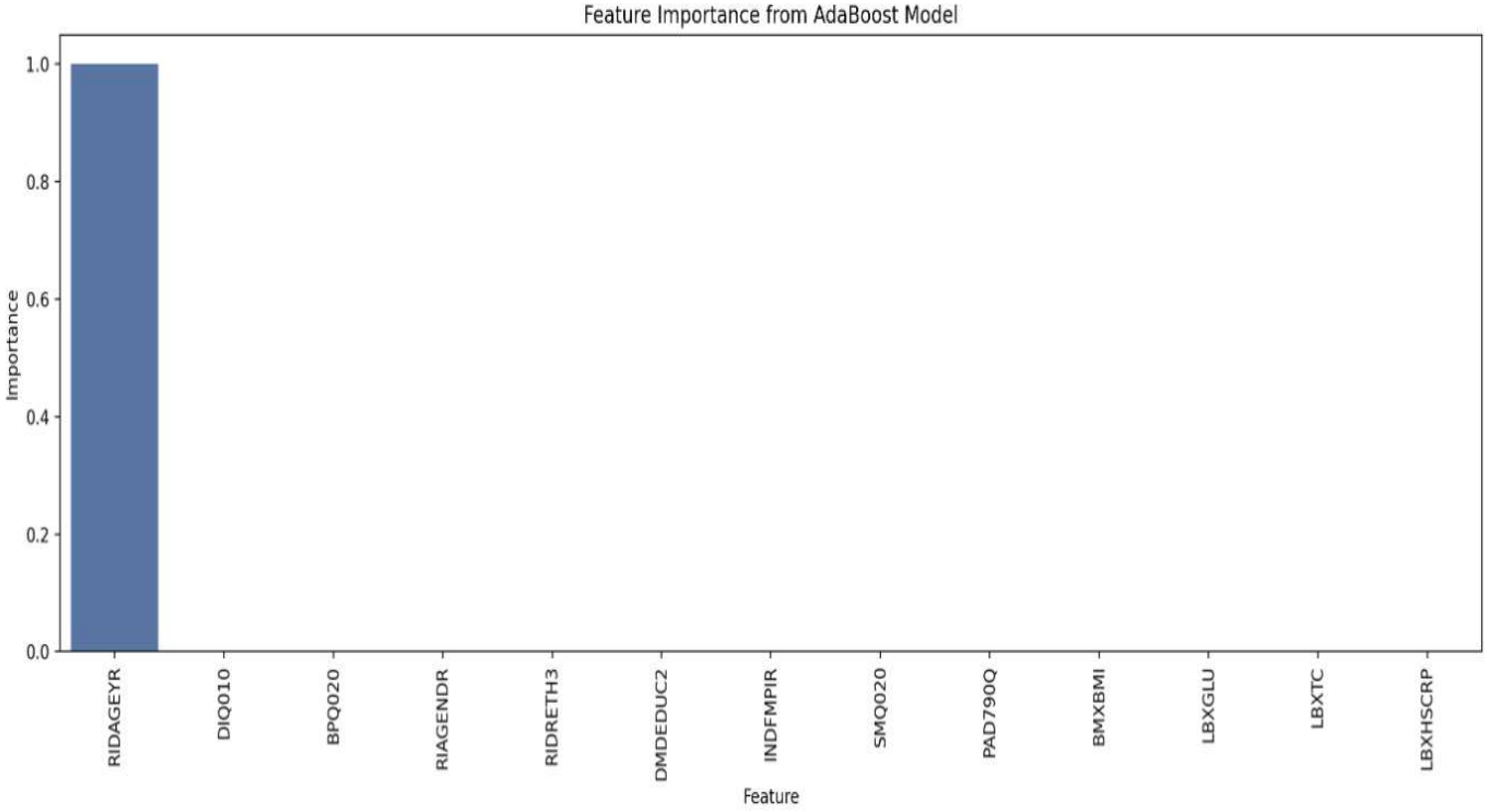
Feature Importance from Logistic Regression



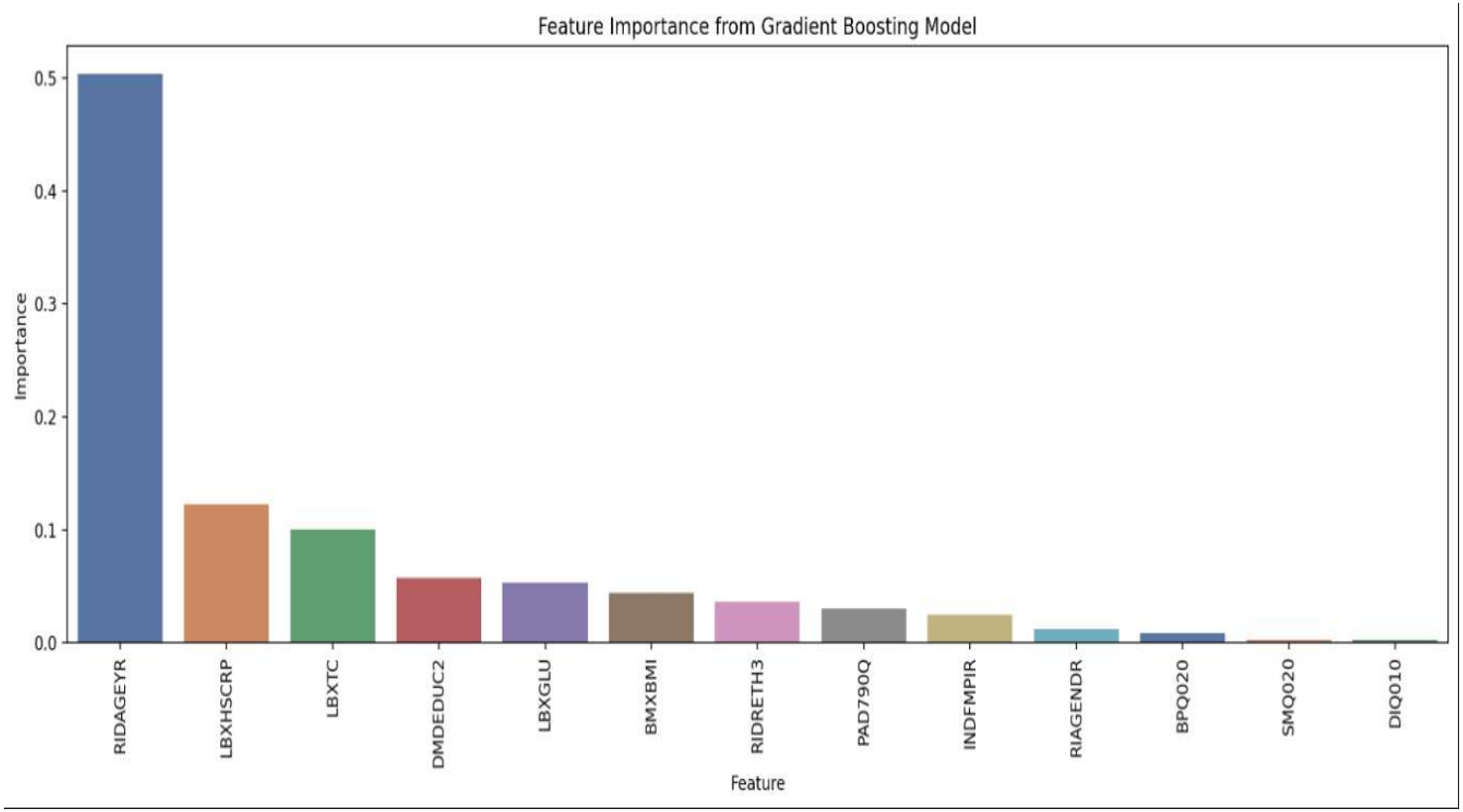
Feature Importance from Random Forest



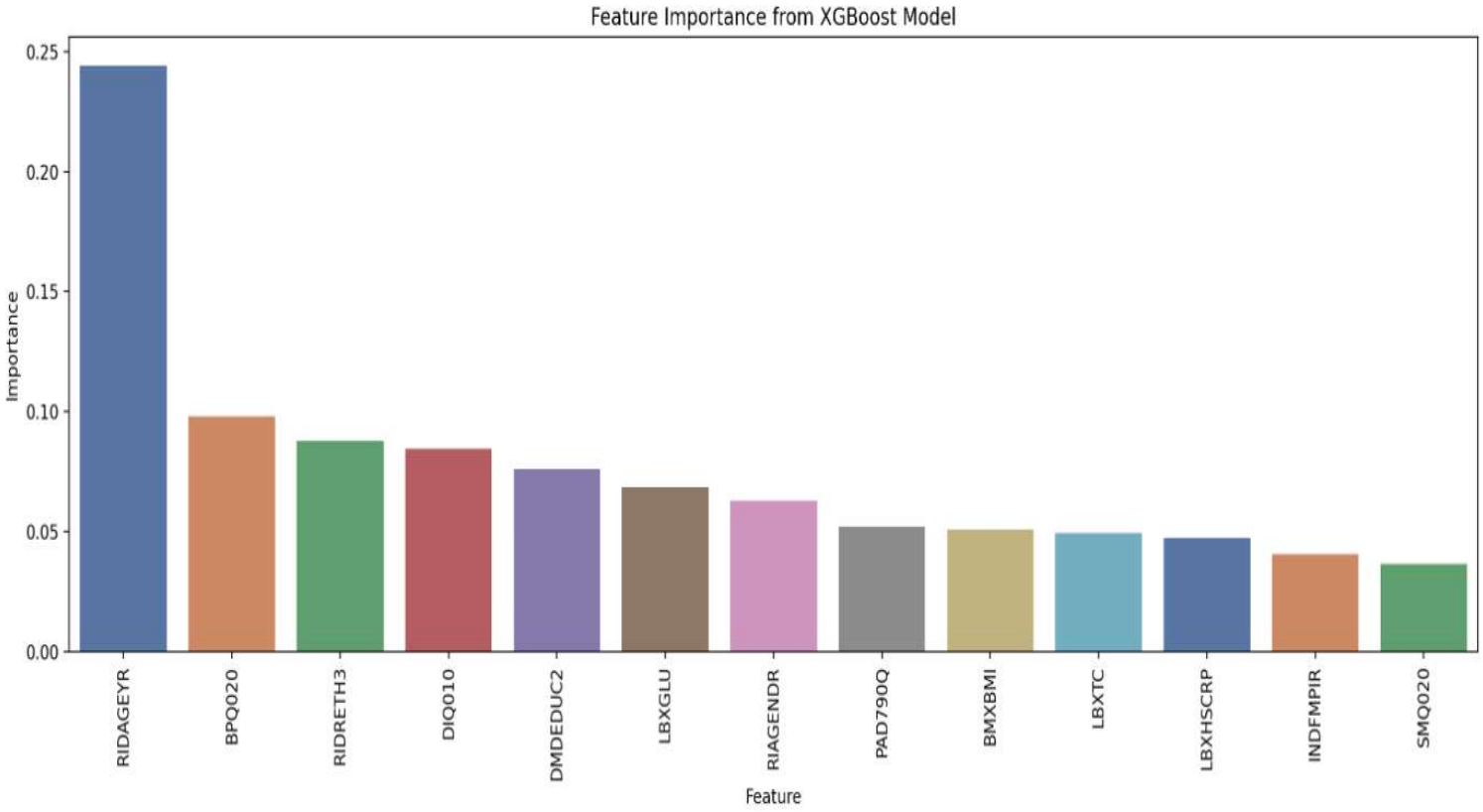
Feature Importance from AdaBoost



Feature Importance from Gradient Boosting



Feature Importance from XGBoost



Model Prediction

```
Logistic Regression Model's prediction for the 25th sample: 1.0
Actual prediction for the 25th sample: 1.0

KNN Model's prediction for the 25th sample: 0.0
Actual prediction for the 25h sample: 1.0

SVM Model's prediction for the 25th sample: 0.0
Actual prediction for the 25th sample: 1.0

Decision Tree Model's prediction for the 25th sample: 1.0
Actual prediction for the 25th sample: 1.0

Random Forest Model's prediction for the 25th sample: 0.0
Actual prediction for the 25th sample: 1.0

Gradient Boosting Model's prediction for the 25th sample: 1.0
Actual prediction for the 25th sample: 1.0

AdaBoost Model's prediction for the 25th sample: 0.0
Actual prediction for the 25th sample: 1.0

XGBoost Model's prediction for the 25th sample: 0
Actual prediction for the 25th sample: 1.0
```

Predictions for all models

```
from pickle import dump
dump(finalmodel,open('insurancemodelf.pkl','wb'))
```

```
new_data=pd.DataFrame({'DIQ010':2, 'BPQ020': 2,'RIDAGEYR':50, 'RIAGENDR':1, 'RIDRETH3':3, 'DMDEDUC2':3, 'INDFMPIR':3,'SMQ020':1, 'PAD790Q':20, 'BMXBMI':20
new_data['SMQ020']=new_data['SMQ020'].map({'1':1,'2':2})
finalmodel.predict(new_data)
```

```
array([0.44])
```

```
new_data=pd.DataFrame({'DIQ010':1, 'BPQ020': 1,'RIDAGEYR':50, 'RIAGENDR':1, 'RIDRETH3':3, 'DMDEDUC2':3, 'INDFMPIR':3,'SMQ020':1, 'PAD790Q':20, 'BMXBMI':20
new_data['SMQ020']=new_data['SMQ020'].map({'1':1,'2':2})
finalmodel.predict(new_data)
```

```
array([0.56])
```

```
new_data=pd.DataFrame({'DIQ010':1, 'BPQ020': 1,'RIDAGEYR':30, 'RIAGENDR':1, 'RIDRETH3':3, 'DMDEDUC2':3, 'INDFMPIR':3,'SMQ020':1, 'PAD790Q':20, 'BMXBMI':20
new_data['SMQ020']=new_data['SMQ020'].map({'1':1,'2':2})
finalmodel.predict(new_data)
```

```
array([0.14])
```

A simple cancer incidence prediction using the Random Forest model

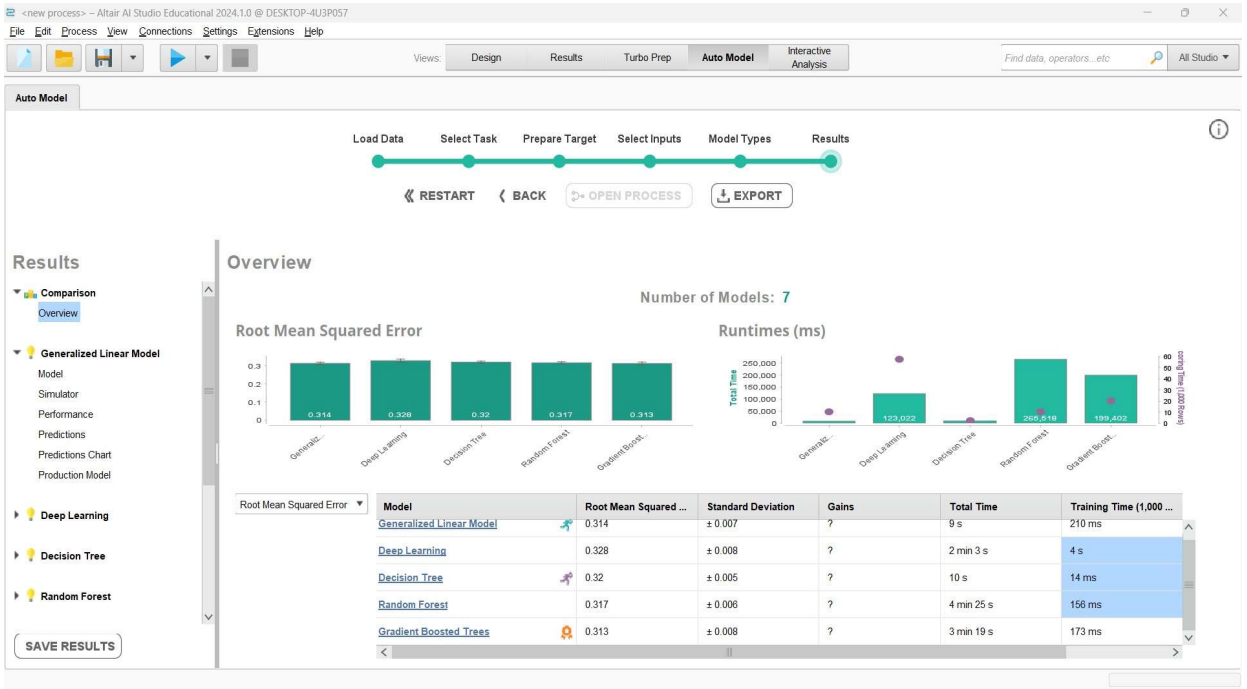
Potential Challenges

- Class Imbalance: When Class 1 (had cancer) is much less frequent than Class 0 (had no cancer)in the training data. This can lead to models favoring the majority class (Class 0), resulting in strong performance for Class 0 but poor performance for Class 1.

- Class 1 distinction difficulty: Class 1 distinction difficulty: Features may lack clarity or sufficient quality, making it hard for models to distinguish between Class 0 and Class 1 due to their similarities.

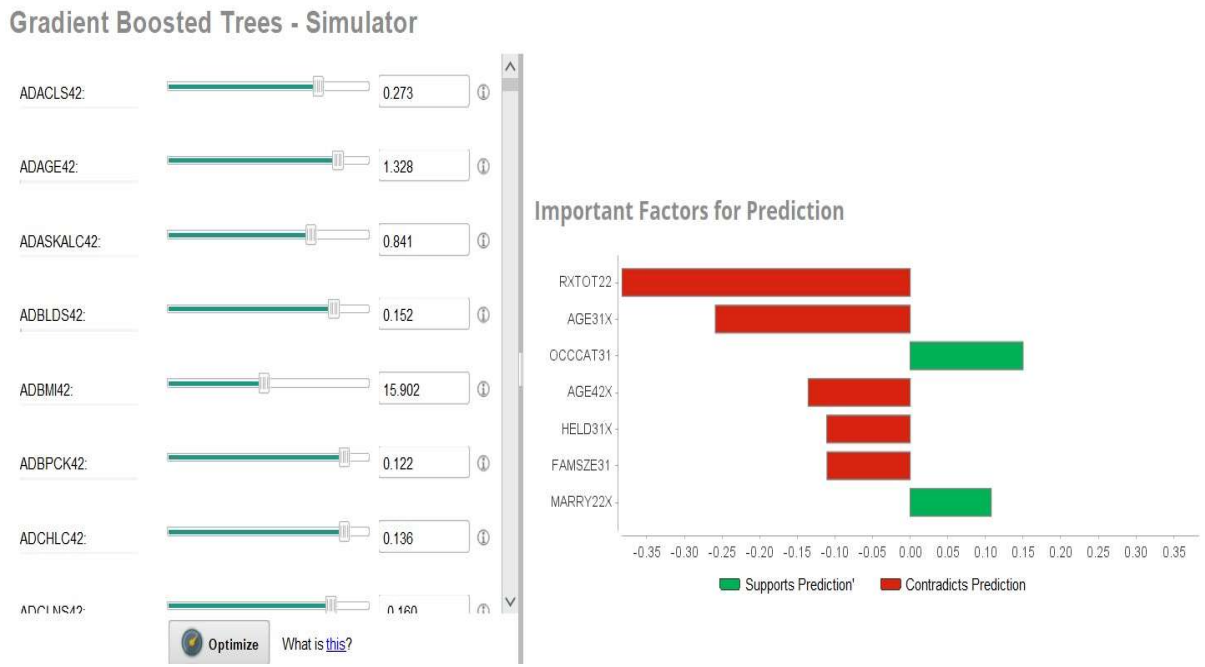
AI Studio: Comparative For All Models

Medical Expenditure Panel Survey (MEPS) HC-243: 2022 Full Year Consolidated Data File



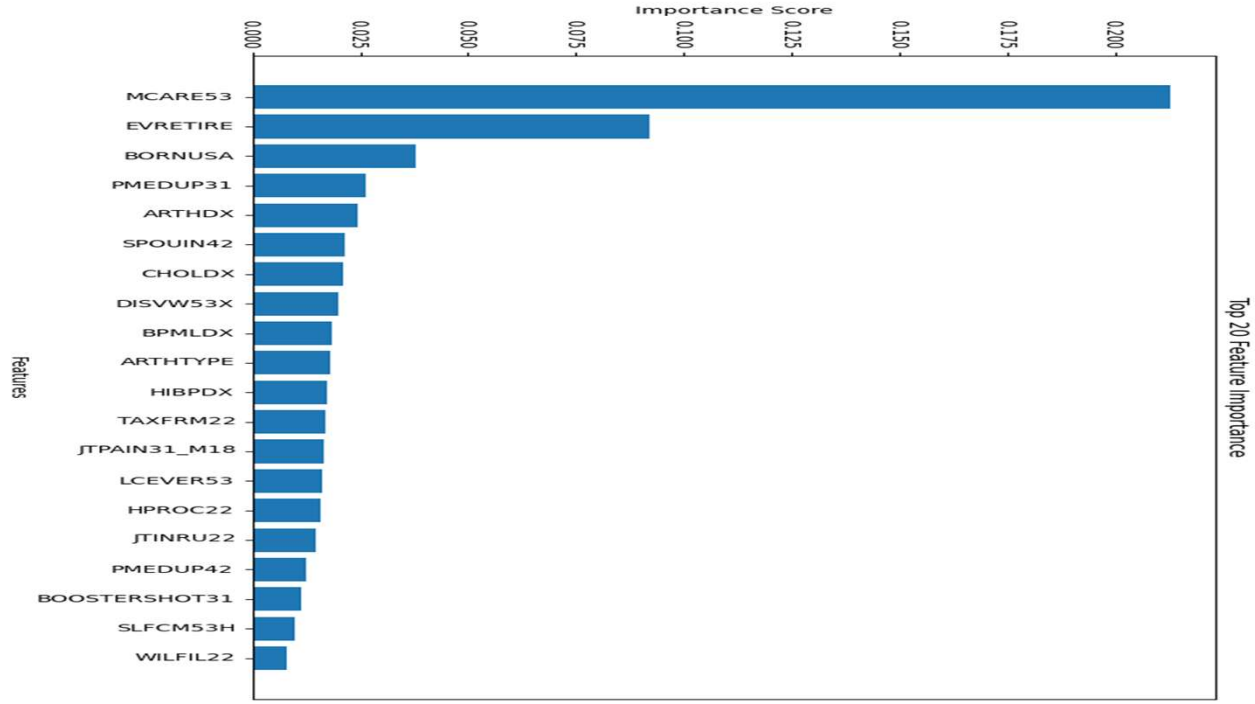
AI Studio:
Feature Importance from Gradient Boosting

Medical Expenditure Panel Survey (MEPS) HC-243: 2022 Full Year Consolidated Data File



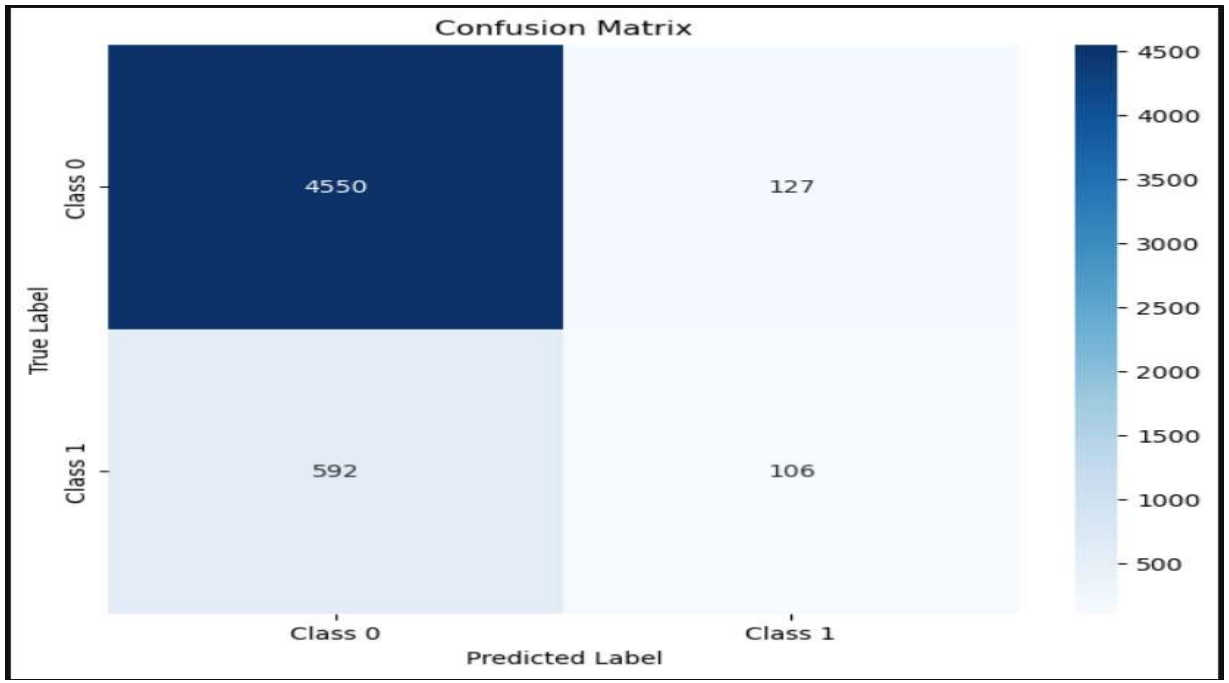
Feature Importance Top 20

Medical Expenditure Panel Survey (MEPS) HC-243: 2022 Full Year Consolidated Data File



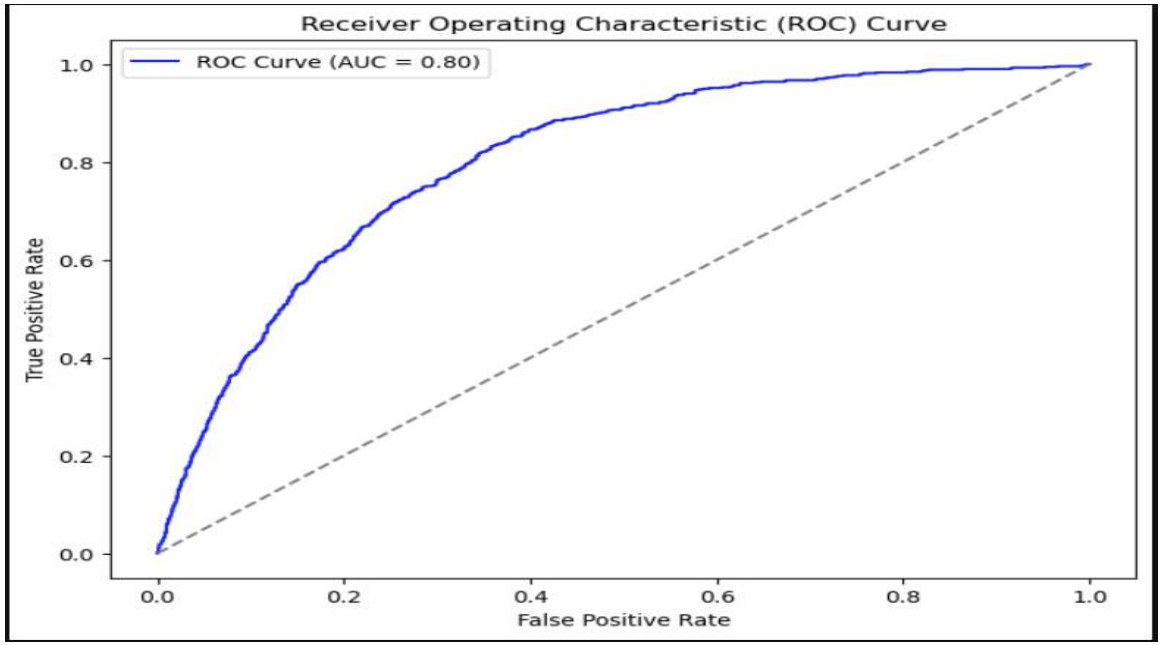
Logistic Regression Confusion Matrix

Medical Expenditure Panel Survey (MEPS) HC-243: 2022 Full Year Consolidated Data File



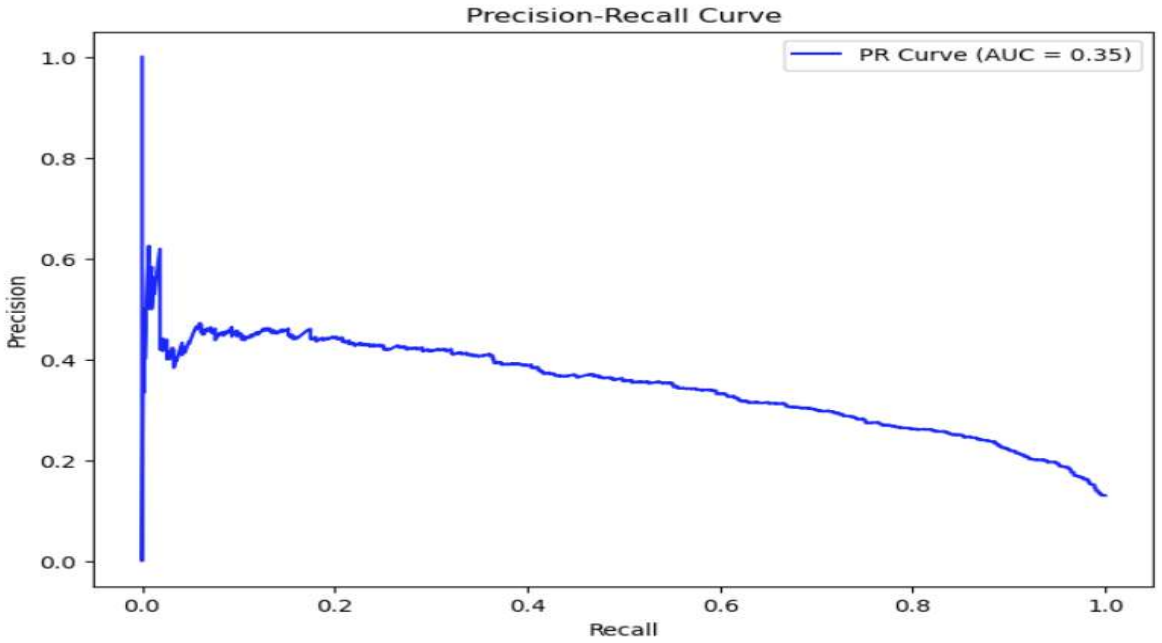
Logistic Regression
Receiver Operating Characteristic (ROC)

Medical Expenditure Panel Survey (MEPS) HC-243: 2022 Full Year Consolidated Data File



Logistic Regression
Receiver Operating Characteristic (ROC)

Medical Expenditure Panel Survey (MEPS) HC-243: 2022 Full Year Consolidated Data File



Logistic Regression Classification Report

Medical Expenditure Panel Survey (MEPS) HC-243: 2022 Full Year Consolidated Data File

	precision	recall	f1-score	support
0	0.88	0.97	0.93	4677
1	0.45	0.15	0.23	698
accuracy			0.87	5375
macro avg	0.67	0.56	0.58	5375
weighted avg	0.83	0.87	0.84	5375

References

Xu, Siying MDa; Huang, Jing MDa,*. Machine learning algorithms predicting bladder cancer associated with diabetes and hypertension: NHANES 2009 to 2018. Medicine 103(4):p e36587, January 26, 2024. | DOI: 10.1097/MD.0000000000036587

Vangeepuram, N., Liu, B., Chiu, Ph. et al. Predicting youth diabetes risk using NHANES data and machine learning. Sci Rep 11, 11212 (2021). https://doi.org/10.1038/s41598-021-90406-0

Olshvang D, Harris C, Chellappa R, Santhanam P (2024) Predictive modeling of lean body mass, appendicular lean mass, and appendicular skeletal muscle mass using machine learning techniques: A comprehensive analysis utilizing NHANES data and the Look AHEAD study. PLOS ONE 19(9): e0309830. https://doi.org/10.1371/journal.pone.0309830

Yang, F., Zhang, J., Chen, W. et al. DeepMPM: a mortality risk prediction model using longitudinal EHR data. BMC Bioinformatics 23, 423 (2022). https://doi.org/10.1186/s12859-022-04975-6

Centers for Disease Control and Prevention. (2023, Aug 16). Impact of Racism on our Nation's Health. Retrieved from Minority Health: https://www.cdc.gov/minorityhealth/racism-disparities/impact-of-racism.html

Curtis, K., & Cheng, S. (2022, April 21). How to Combat Disparities in Healthcare for Minority Populations. Retrieved from EDUMED: https://www.edumed.org/resources/combating-disparities-in-healthcare/

McCullom, R., & Holder, N. L. (2022, October 3). What science tells us about structural racism's health impact. Retrieved from Harvard Public Health: https://harvardpublichealth.org/equity/what-science-tells-us-about-structural-racisms-health-impact/

Radley, D. C., Baumgartner, J. C., Collins, S. R., Zephyrin, L. C., & Schneider, E. C. (2021, NOVEMBER 18). Achieving Racial and Ethnic Equity in U.S. Health Care: A Scorecard of State Performance. Retrieved from Commonwealth Fund: https://doi.org/10.26099/ggmq-mm33