



Diabetes Prediction Model using Survey Data



ISE 599 – INTRODUCTION TO HEALTH ANALYTICS



HEALTH PROBLEM

The rising prevalence of diabetes and its associated complications pose significant challenges to public health systems globally. Despite its prevalence, many individuals remain undiagnosed, highlighting the need for effective predictive tools.



MOTIVATION

Addressing diabetes is crucial due to its detrimental impact on individual health outcomes, healthcare costs, and overall population well-being. By implementing effective preventive measures and treatments, we aim to mitigate the burden of diabetes-related complications, improve the quality of life for affected individuals and also reduce the healthcare costs.



OBJECTIVE

Our goal is to assist clinicians in enhancing diabetes detection and treatment outcomes. To achieve this, we aim to develop a predictive model capable of accurately identifying individuals at risk of developing diabetes or experiencing diabetes-related complications. By harnessing data obtained from survey-related questions, we seek to enable early intervention and personalized healthcare strategies by clinicians. This proactive approach aims to prevent or effectively manage diabetes, thereby improving patient outcomes and reducing the burden on healthcare systems.



DATA SOURCE



The dataset utilized in this study is the Diabetes Health Indicators dataset obtained from the UCI Machine Learning Repository. While the exact source of the data is not stated explicitly, the description says that the data is taken from the Behavioural Risk Factor Surveillance System (BRFSS) surveys of 2015, which was subsequently modified for analysis purposes.

The curated version of the dataset used in our project comprises responses from -

- 253680 observations
- 22 features
 - 21 predictor variables (survey-type questions)
 - 1 response variable - Diab_Binary



DATA DICTIONARY

Variable	Description
High BP	Individual suffers from High BP (0 if no high BP, 1 if high BP)
HighChol	Individual suffers from High Chol (0 if no high Chol, 1 if high Chol)
CholCheck	Has had Cholesterol Check in 5 years (0 if no cholesterol check in 5 years, 1 if yes cholesterol check in 5 years)
BMI	Body Mass Index: weight (kg) / height (m)^2
Smoker	Have you smoked atleast 100 cigarretes (5 packs) in your entire life? (0 if no, 1 if yes)
Stroke	Ever told you had a stroke? (0 if no, 1 if yes)
HeartDiseaseorAttack	Ever had Coronary Heart Disease (CHD) or Myocardial Infarction(MI)? (0 if no, 1 if yes)
PhysActivity	Physical Activity in Past 30 days- not including job (0 if no, 1 if yes)
Fruits	Consume fruits one or more times a day (0 if no, 1 if yes)
Veggies	Consume Vegetables one or more times a day (0 if no, 1 if yes)
HvyAlcoholConsump	Heavy drinkers: adult men/ adult women having more than 14/ 7 drinks per week respectively (0 if no, 1 if yes)
AnyHealthCare	Have any kind of heathcare coverage, including health insurance, prepaid plans such as HMO, etc? (0 if no, 1 if yes)
NoDocbcCost	Was there a time in the past 12 months, when you needed to see a doctor but could not because of cost? (0 if no, 1 is yes)
GenHlth	Would you say that in General your Health is : scale 1-5 (1 if excellent, 2 if very good, 3 if good, 4 if fair, 5 if poor)
MentHlth	Now think about your mental Health which includes stress, depression, problems with emotions, for how many days during the past 30 days was your mental health not good? (scale: 1-30 days)
PhyScHlth	Now think about your physical Health which includes physical illness, and injury, for how many days during the past 30 days was your physical health not good? (scale: 1-30 days)
DiffWalk	Do you have serious difficulty walking or climbing stairs? (0 if no, 1 if yes)
Sex	Gender of individual (0 if female, 1 if male)
Age	13 level Age Category (_AGEGSYR see codebook) 1=18-24...,9=60-64...,13= 80 or older.
Education	Education Level (EDUCA) scale1-6 (1= Never attended school or only kindergarten, 2= Grades 1 through 8(Elementary), 3= Grades 9 through 11 (high school), 4=Grade 12 (high school graduate), 5=College 1 year to 3 years (some college or technical school), 6=College 4 years or more (College graduate))
Income	Income Scale 1-8 (1= less than \$10000..., 5 = less than \$35000...,8 = \$75000 or more)
Diabetes_binary	1 if individual has diabetics or is prediabetic, 0 otherwise (no diabetics)



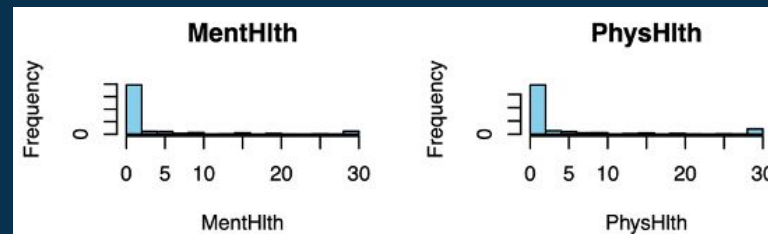
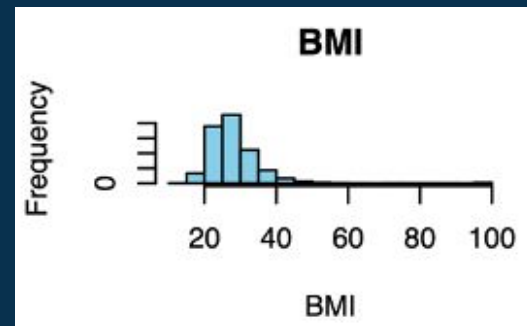
EXPLORATORY DATA ANALYSIS



VARIABLE DISTRIBUTIONS

Characteristic	0, N = 194,377	1, N = 35,097
HighBP		
0	116,522 (60%)	8,692 (25%)
1	77,855 (40%)	26,405 (75%)
HighChol		
0	116,528 (60%)	11,601 (33%)
1	77,849 (40%)	23,496 (67%)
CholCheck		
0	9,057 (4.7%)	241 (0.7%)
1	185,320 (95%)	34,856 (99%)
Smoker		
0	105,711 (54%)	16,874 (48%)
1	88,666 (46%)	18,223 (52%)
Stroke		
0	187,361 (96%)	31,829 (91%)
1	7,016 (3.6%)	3,268 (9.3%)
HeartDiseaseorAttack		
0	178,520 (92%)	27,241 (78%)
1	15,857 (8.2%)	7,856 (22%)
PhysActivity		
0	48,222 (25%)	13,038 (37%)
1	146,155 (75%)	22,059 (63%)
Fruits		
0	74,289 (38%)	14,592 (42%)
1	120,088 (62%)	20,505 (58%)
Veggies		
0	38,535 (20%)	8,602 (25%)
1	155,842 (80%)	26,495 (75%)
HvyAlcoholConsump		
0	181,259 (93%)	34,265 (98%)
1	13,118 (6.7%)	832 (2.4%)

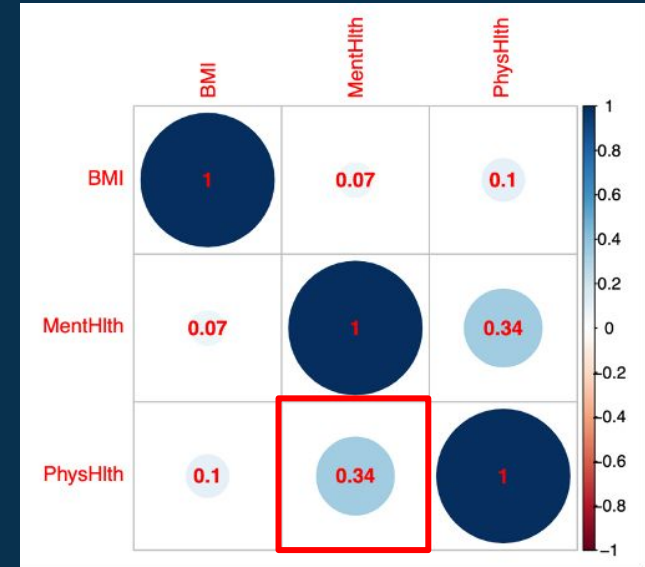
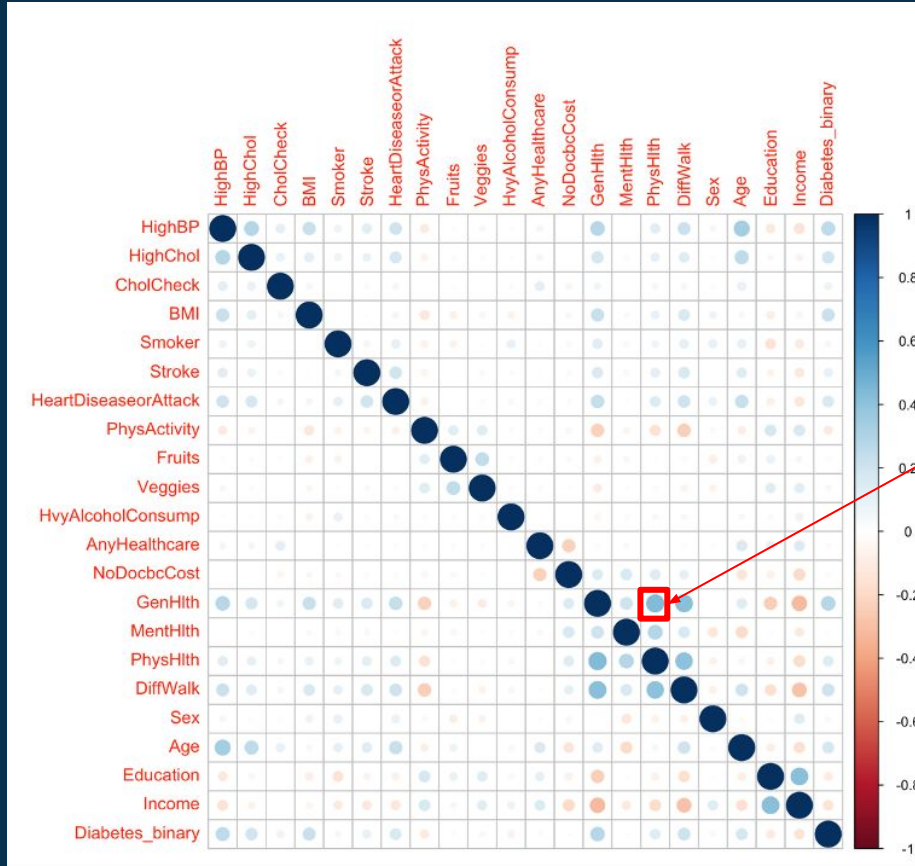
AnyHealthcare		
0	10,967 (5.6%)	1,422 (4.1%)
1	183,410 (94%)	33,675 (96%)
NoDocbcCost		
0	176,796 (91%)	31,355 (89%)
1	17,581 (9.0%)	3,742 (11%)
GenHlth		
1	33,719 (17%)	1,135 (3.2%)
2	71,085 (37%)	6,280 (18%)
3	60,308 (31%)	13,324 (38%)
4	21,764 (11%)	9,781 (28%)
5	7,501 (3.9%)	4,577 (13%)
DiffWalk		
0	164,866 (85%)	21,983 (63%)
1	29,511 (15%)	13,114 (37%)
Sex		
0	110,370 (57%)	18,345 (52%)
1	84,007 (43%)	16,752 (48%)
Education		
1	127 (<0.1%)	47 (0.1%)
2	2,857 (1.5%)	1,183 (3.4%)
3	7,171 (3.7%)	2,296 (6.5%)
4	50,092 (26%)	11,032 (31%)
5	56,133 (29%)	10,311 (29%)
6	77,997 (40%)	10,228 (29%)
Income		
1	7,408 (3.8%)	2,383 (6.8%)
2	8,670 (4.5%)	3,086 (8.8%)
3	12,356 (6.4%)	3,564 (10%)
4	15,906 (8.2%)	4,047 (12%)
5	20,837 (11%)	4,489 (13%)
6	29,697 (15%)	5,260 (15%)
7	34,905 (18%)	5,226 (15%)
8	64,598 (33%)	7,042 (20%)



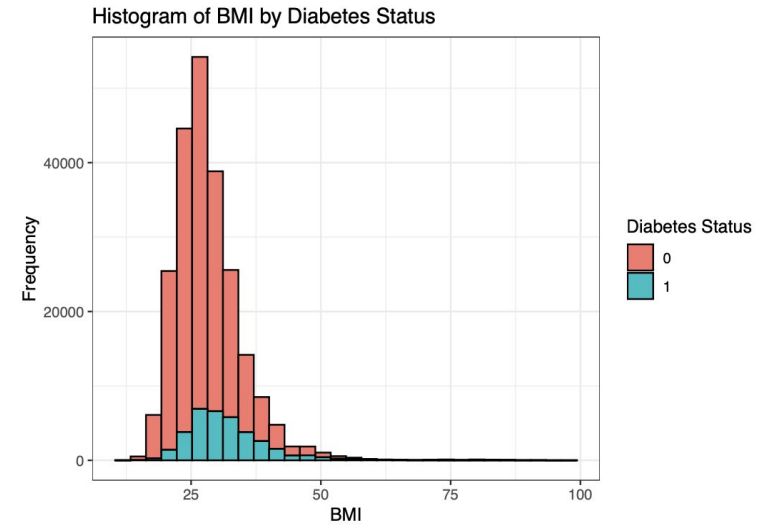
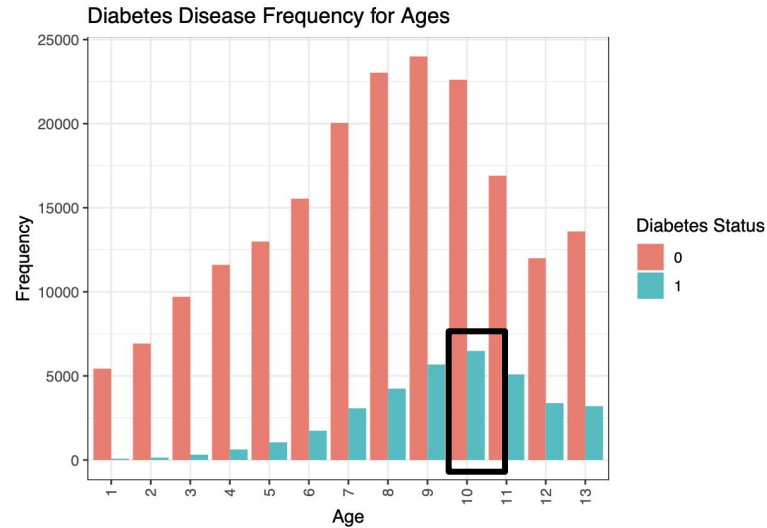
Categorical variable distributions wrt the Target

Histograms for numeric variables

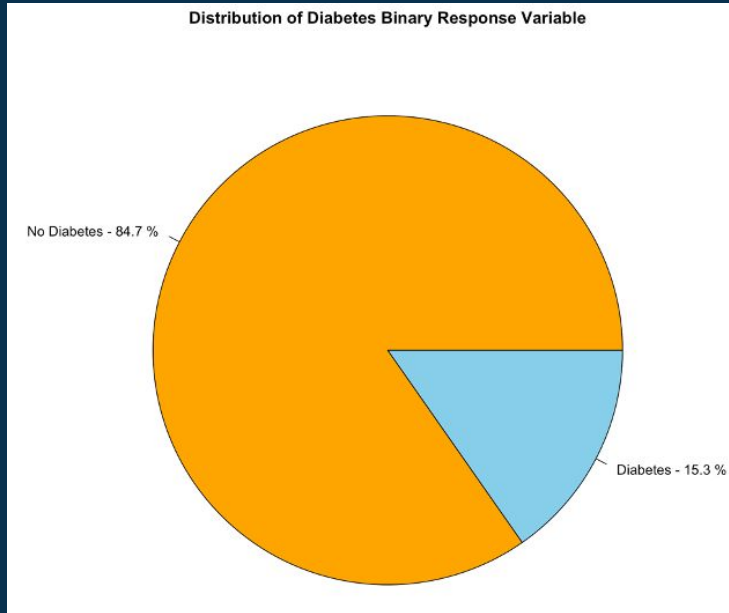
CORRELATION ANALYSIS



VARIABLE DISTRIBUTION ANALYSIS



HIGH IMBALANCE - TARGET VARIABLE



Techniques used-

- Stratified Sampling
- Random Over Sampling
- SMOTE - Synthetic Minority Oversampling Technique



ADDRESSING THE PROBLEM

Binary Classification Problem

- Logistic Regression-Baseline Approach
- Other Machine Learning Classifiers (KNN, Decision-Tree, Random Forest, XGBoost..)
- Hyperparameter Tuning - optimize model performance
- Model evaluation and best model selection
- Cost Based Analysis
- Feature Importance - Explainable AI

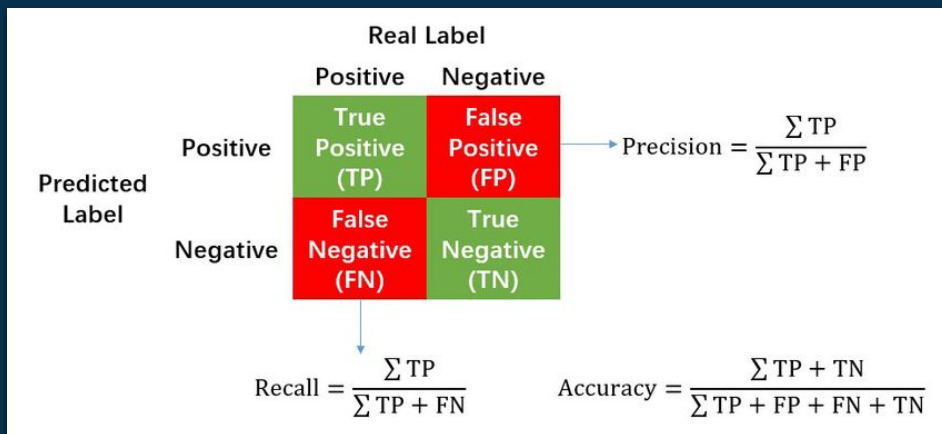


BASELINE MODEL – LOGISTIC REGRESSION

Baseline Model	Threshold	Precision	Recall	F1-Score	AUC-ROC	Accuracy
Full Logistic Regression	Default = 0.5	0.5494	0.1487	0.234	0.811	0.8511
Full Logistic Regression at best threshold	0.2 by optimizing the Recall	0.3579	0.6687	0.4662	0.811	0.7658
Logistic Regression with Stepwise Backward Elimination	0.2	0.3579	0.6693	0.4664	0.8110	0.7658
Lasso CV Logistic Regression	0.2	0.3604	0.6614	0.4665	0.8095	0.7687
Ridge Regression CV Logistic Regression	0.2	0.3609	0.6578	0.4661	0.8091	0.7695



HYPERPARAMETER TUNING



- Grid Search Cross Validation
- Optimized for Higher Recall
→ Capture higher “True Positives”
Positive cases are critical
- Impact on Model Performance
→ Trade-off with Precision -
Compromise on Precision(willingness to accept more false positives)



COMPARING DIFFERENT ML MODELS

Model	Hyperparameters	Precision	Recall	F1-Score	AUC-ROC	Accuracy
Logistic Regression (Stepwise Backward Elimination)	best threshold = 0.2, (selected by optimizing the recall)	0.36	0.67	0.47	0.81	0.76
KNN Classifier	n_neighbors = 5	0.25	0.62	0.36	0.71	0.69
Decision Tree Classifier	min_samples_split = 2, min_samples_leaf = 1, max_tree_depth = 12	0.31	0.74	0.44	0.79	0.71
Random Forest Classifier	max_tree_depth = 12, n_estimators = 10	0.30	0.74	0.42	0.80	0.72
XGBoost Classifier	learning_rate = 0.1, max_depth = 5, min_child_weight = 3	0.33	0.79	0.46	0.82	0.72
Perceptron Classifier	Penalty = l2 regularization, Max_iter = 1000	0.29	0.85	0.43	0.80	0.65



ANALYSIS USING COST MATRIX

<p>True Negatives (Actual – Non Diabetic, Predicted – Non Diabetic) NO COSTS</p>	<p>False Positives (Actual – Non Diabetic Predicted – Diabetic) COSTS - \$7 / glucose test</p>
<p>False Negatives (Actual – Diabetic Predicted – Non Diabetic) COSTS - \$50000 / on average per annum due to further complications</p>	<p>True Positives (Actual – Diabetic Predicted – Diabetic) COSTS - \$7 / glucose test + \$12022 / medical expenses per annum</p>

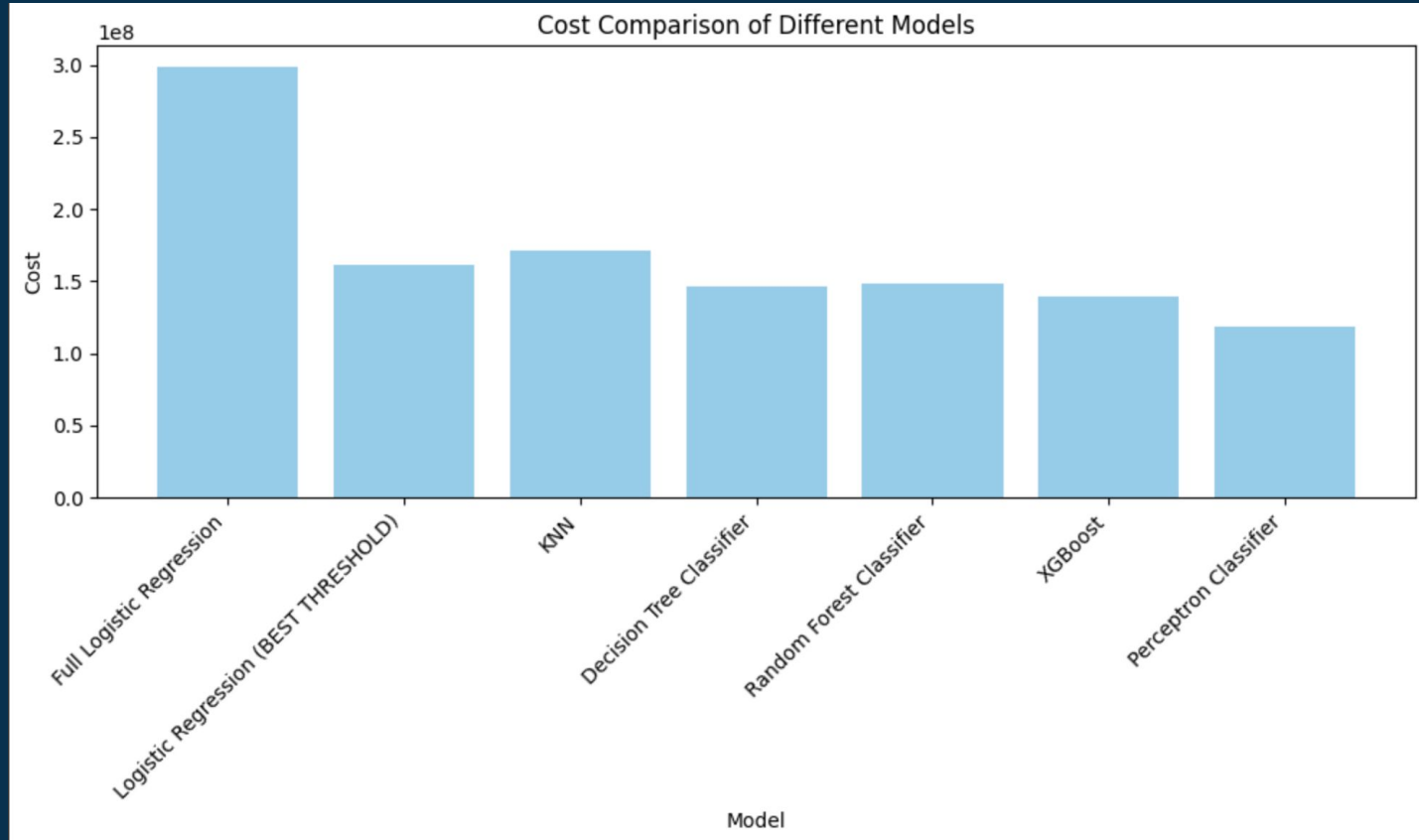


MODEL COMPARISON BASED ON COST MATRIX

Model	TP	FP	TN	FN	Total Cost Approximations (in USD)
Full Logistic Regression	997	908	36464	5728	298399269
Logistic Regression at Best Threshold	4070	8596	28776	2255	161768202
KNN	4360	12970	30769	2637	170887230
Decision Tree Classifier	5165	11777	25471	1684	146412224
Random Forest Classifier	5118	10521	26727	1731	148188069
XGBoost	5355	11006	26242	1494	139192337
Perceptron Classifier	5907	14862	22386	942	118259337

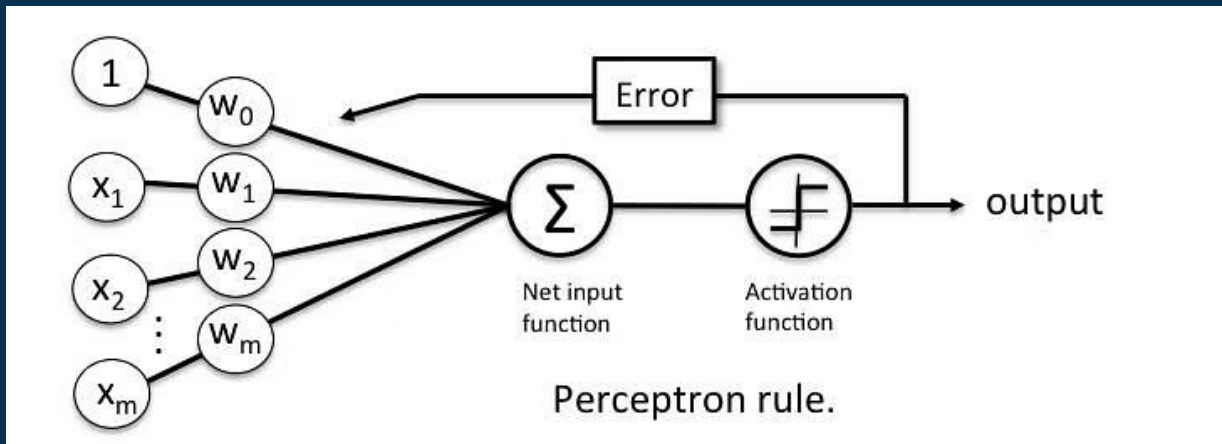


COSTS OF DIFFERENT MODELS

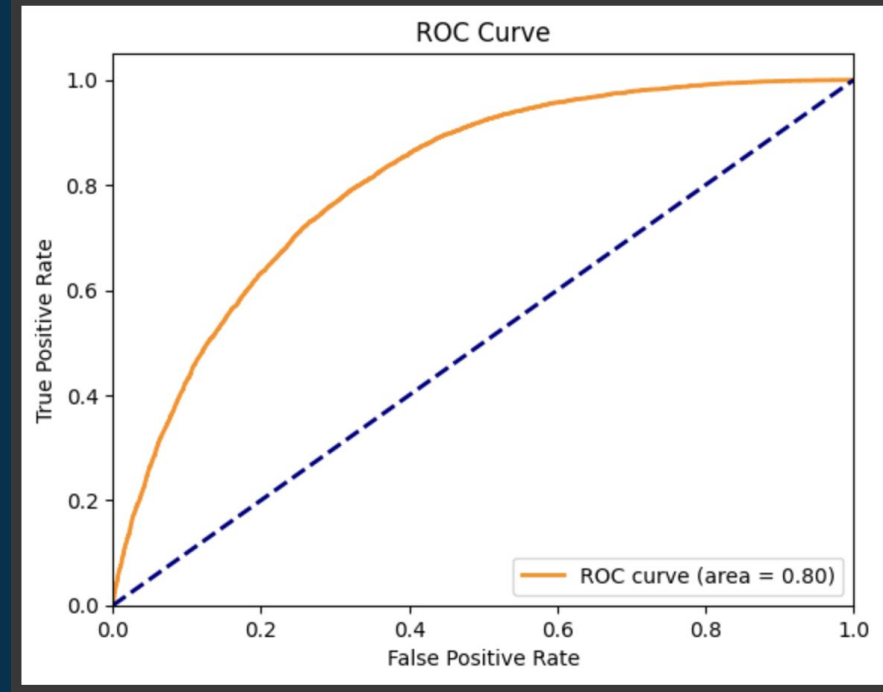
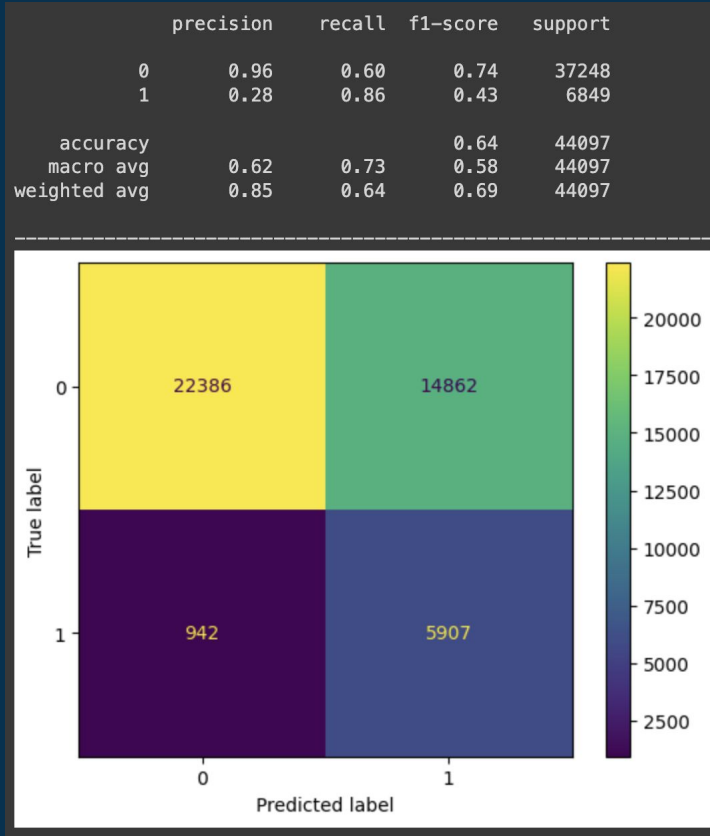


PERCEPTRON CLASSIFIER

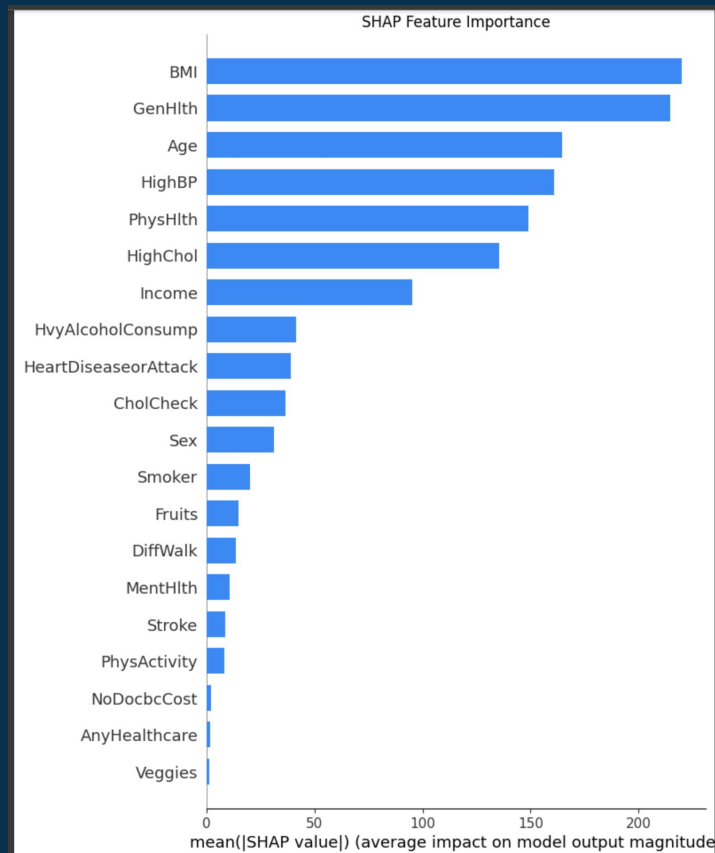
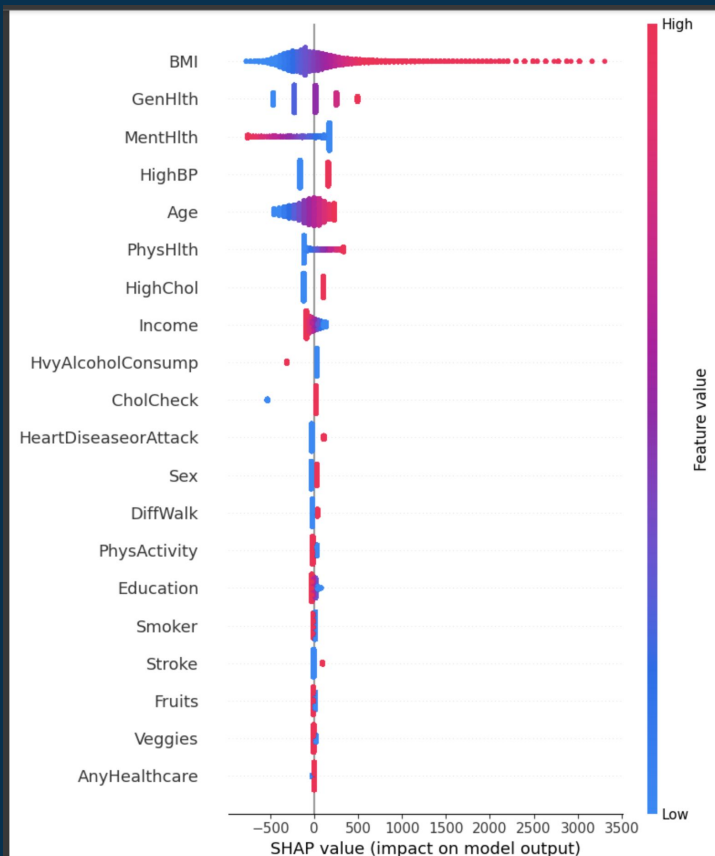
- Perceptron was introduced by Frank Rosenblatt
- Simplest form of Neural Network Classifier
- Algorithm for supervised learning of binary classifiers.
- Enables neurons to learn and processes elements in the training set one at a time.



CLASSIFICATION REPORT - PERCEPTRON CLASSIFIER



FEATURE IMPORTANCE (Perceptron) - SHAP



LOGISTIC REGRESSION -IMPORTANT FEATURES

```
X2_train_resampled.head()
```

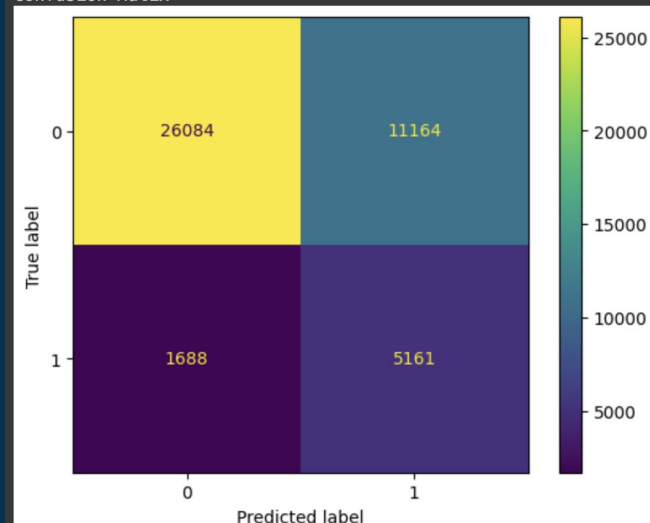
	BMI	GenHlth	Age	HighBP	PhysHlth	HighChol
0	27	3	13	1	0	1
1	33	3	9	1	3	1
2	27	4	4	0	1	0
3	20	2	7	0	0	0
4	27	2	8	1	0	1

```
y2_train_resampled.head()
```

	Diabetes_binary
0	1
1	0
2	0
3	0
4	0

	precision	recall	f1-score	support
0	0.94	0.70	0.80	37248
1	0.32	0.75	0.45	6849
accuracy			0.71	44097
macro avg	0.63	0.73	0.62	44097
weighted avg	0.84	0.71	0.75	44097

Confusion Matrix



KEY FINDINGS

Predictive Performance-

- **Perceptron Classifier** demonstrated superior predictive performance compared to traditional machine learning models in our trade-off case.
- 85% of individuals with diabetes are correctly classified.



Feature Importance Analysis -

All models showed relatively similar features as important in predicting diabetes. The most important demographic and clinical factors significantly influencing diabetes risk are-

1. **BMI**
2. **GenHlth**
3. **Age**
4. **HighBP**
5. **PhysHlth**
6. **HighChol**



RECOMMENDATIONS

- 
1. **Screening and Diagnosis:** Prioritize further screenings and tests for individuals classified as predicted positives to confirm diabetes diagnosis. Tailor treatment plans based on the type of diabetes identified.
 2. **Preventive Measures:** Refer individuals identified as at-risk for diabetes to dieticians for personalized preventive strategies, focusing on lifestyle modifications such as maintaining a healthy weight, balanced diet, and regular physical activity.
 3. **Regular Health Check-ups:** Incorporate regular screenings for diabetes risk factors, including BMI, blood pressure, cholesterol levels, and age, into routine health check-ups for all patients.
 4. **Personalized Healthcare:** Implement personalized healthcare strategies based on individual risk profiles identified by predictive models. Offer tailored interventions and treatments to manage diabetes effectively and prevent complications.
- 
-

FUTURE WORK

1. **Deep Learning Integration:** Implementation of deep learning models to enhance predictive performance beyond traditional machine learning techniques. Deep learning architectures such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) could offer improved robustness in diabetes prediction tasks.
2. **Data Augmentation:** Increase the diversity and volume of data by integrating additional sources such as electronic health records (EHRs), wearable devices, genetic data, and dietary logs. This enriched dataset can provide deeper insights into diabetes risk factors and enable more accurate predictive modeling.
3. **Clinical Validation:** Validate the predictive models in clinical settings to assess their real-world performance and utility. Collaborate with healthcare professionals to evaluate model accuracy, usability, and impact on patient outcomes, facilitating seamless integration into clinical practice.
4. **Fuzzy Optimization:** Implement fuzzy optimization techniques to fine-tune model parameters and enhance precision. By optimizing sensitivity and precision simultaneously, we can achieve a balanced approach that improves model performance across various evaluation metrics.





Thank you!

