APPLIED MACHINE LEARNING PROJECT

Diabetes (Risk) Prediction

Harish Bhupalam
Sai Phani Sudheer

Introduction:

Our decision to focus on the Diabetes Health Indicators Dataset is driven by the profound and widespread impact of diabetes, a chronic disease that not only compromises the well-being and life expectancy of millions of Americans but also imposes substantial health and financial burden on the nation as a whole. The dataset, derived from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey conducted by the Centers for Disease Control and Prevention (CDC), serves as a valuable resource for comprehending the intricate dynamics of diabetes and formulating strategies to address its multifaceted challenges.

Context:

Diabetes, characterized by the ineffective regulation of blood glucose levels, stems from either insufficient insulin production or the inefficient utilization of produced insulin. The repercussions are severe, leading to complications such as heart disease, vision loss, lower-limb amputation, and kidney disease. While a cure remains elusive, proactive measures such as weight management, healthy eating, physical activity, and medical treatments can mitigate its impact. The dataset underscores the critical importance of early diagnosis, accentuating the need for predictive models to identify individuals at risk and intervene promptly.

The scale of the Problem:

As of 2018, the Centers for Disease Control and Prevention revealed alarming statistics – 34.2 million Americans diagnosed with diabetes and an additional 88 million in the prediabetic stage. What intensifies the challenge is the substantial portion of those affected who are unaware of their risk. Type II diabetes, the most prevalent form, exhibits varying prevalence influenced by factors like age, education, income, location, race, and other social determinants of health. Notably, the burden disproportionately falls on those with lower socioeconomic status. The economic impact is staggering, with diagnosed diabetes costs reaching approximately $327 billion annually, and the overall financial burden, including undiagnosed diabetes and prediabetes, nearing an astronomical $400 billion.

The preceding paragraphs illuminate the multifaceted nature of the diabetes epidemic, emphasizing the need for proactive measures, early intervention, and strategic predictive models to navigate this public health challenge. The ensuing exploration of the diabetes health indicators dataset seeks to unravel deeper insights into the dynamics of diabetes, contributing to a more informed and effective approach in the ongoing battle against this pervasive chronic condition.

About the Dataset : Decoding Health Patterns from BRFSS 2015

Dataset Overview: The Behavioral Risk Factor Surveillance System (BRFSS) annually captures the health-related insights of over 400,000 Americans through a telephone survey conducted by the Centers for Disease Control and Prevention (CDC). This comprehensive dataset, spanning the year 2015, encompasses responses from 441,455 individuals, unveiling a tapestry of health-related risk behaviours, chronic conditions, and preventative service utilization. The dataset, available on Kaggle, is presented in three distinct files, each shedding light on diabetes-related health indicators.

1. **Diabetes_012 Health Indicators (Imbalanced):**
- File: **diabetes_012_health_indicators_BRFSS2015.csv**
- Clean dataset: 253,680 survey responses
- Target Variable: Diabetes_012 (3 classes - 0: no diabetes, 1: prediabetes, 2: diabetes)
- Class imbalance noted
- Features: 21 variables capturing diverse health aspects
2. **Diabetes Binary 50-50 Split Health Indicators (Balanced):**
- File: **diabetes_binary_5050split_health_indicators_BRFSS2015.csv**
- Clean dataset: 70,692 survey responses
- Target Variable: Diabetes_binary (2 classes - 0: no diabetes, 1: prediabetes or diabetes)
- Balanced dataset with equal representation
- Features: 21 variables providing a holistic view of health characteristics
3. **Diabetes Binary Health Indicators (Imbalanced):**
- File: **diabetes_binary_health_indicators_BRFSS2015.csv**
- Clean dataset: 253,680 survey responses
- Target Variable: Diabetes_binary (2 classes - 0: no diabetes, 1: prediabetes or diabetes)
- Imbalanced class distribution
- Features: 21 variables capturing nuanced health factors

Research Questions: Exploring Diabetes Predictors: The exploration of this dataset is anchored in several pertinent research questions aimed at unravelling the intricate landscape of diabetes risk factors and prediction.

1. **Survey Predictive Accuracy:**
- Can survey questions from the BRFSS accurately predict an individual's diabetes status?
2. **Identifying Key Risk Factors:**
- What risk factors emerge as the most predictive indicators of diabetes risk?

3. **Subset Predictive Power:**
- Can a subset of risk factors be isolated to accurately predict an individual's diabetes status?

4. **Creating a Concise Predictive Form:**
- Is it possible to distil a shorter set of questions from the BRFSS using feature selection, yet maintain accuracy in predicting diabetes or high-risk status?

Data Dictionary for our binary class Diabetes dataset

Diabetes_012 0 = no diabetes 1 = pre-diabetes 2 = diabetes

HighBP 0 = no high BP 1 = high BP

HighChol 0 = no high cholesterol 1 = high cholesterol

CholCheck 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years

BMI Body Mass Index

Smoker Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes

Stroke (Ever told) you had a stroke. 0 = no 1 = yes

HeartDiseaseorAttack coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes

PhysActivity physical activity in past 30 days - not including job 0 = no 1 = yes

Fruits Consume Fruit 1 or more times per day 0 = no 1 = yes

Veggies Consume Vegetables 1 or more times per day 0 = no 1 = yes

HvyAlcoholConsump Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes

AnyHealthcare Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = 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 = no 1 = yes

GenHlth Would you say that in general, your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor

MentHlth Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days

PhysHlth Now thinking 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 = no 1 = yes

Age 13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older

Education Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)

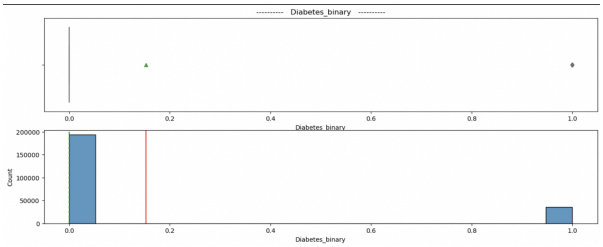
Income Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

Exploratory Data Analysis and Pre-Processing

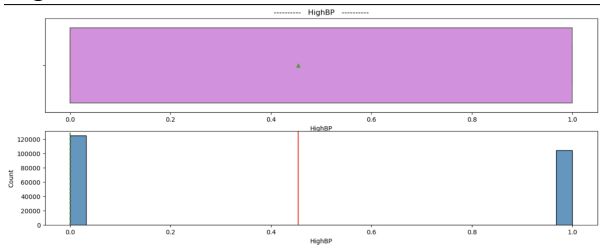
Our initial steps were to look for duplicate observations in the dataset and remove them. There were 24,206 duplicate records which were eliminated from the dataset which resulted in the final version of the data having 229,474 records. We also checked for features having no variance throughout all the observations. Fortunately for us, there were no such features. There were also no missing values for any of the features or for any observations in the entire data.

Our next step was to look at the univariate distributions of all the features which are displayed in the table below:

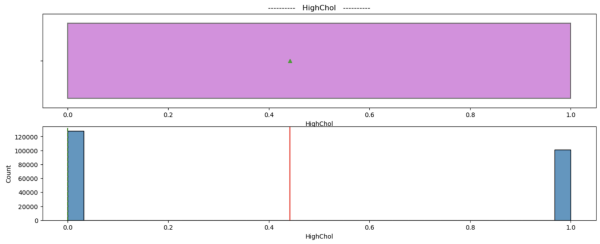
Diabetes Binary (Target Variable)



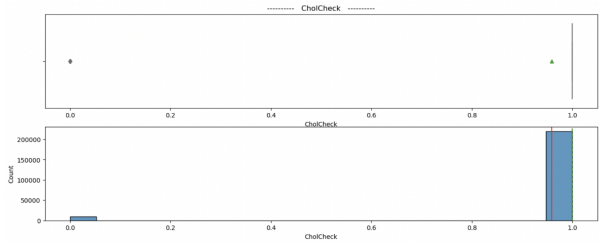
High BP



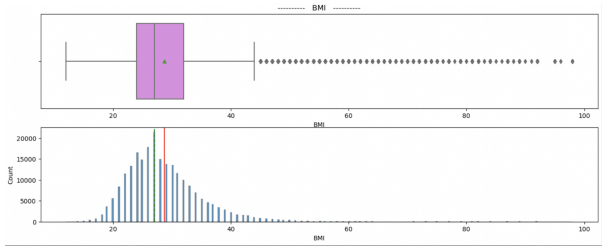
High Cholesterol



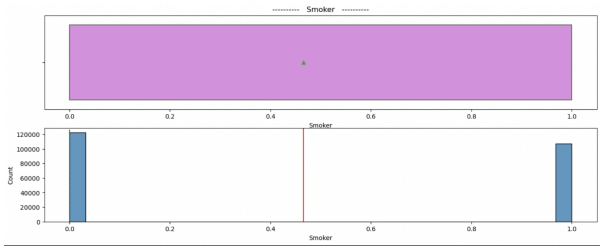
Cholesterol Checkups



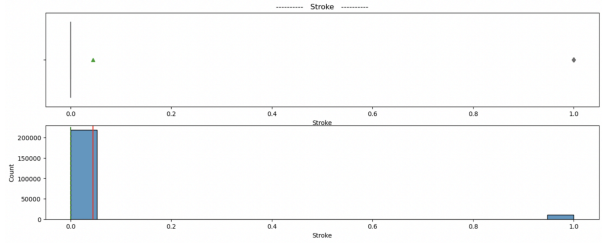
BMI



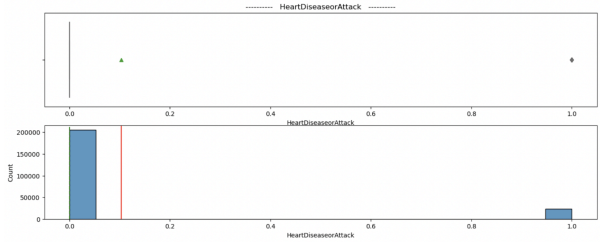
Smoker



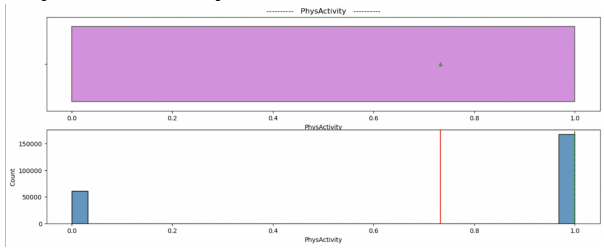
Stroke



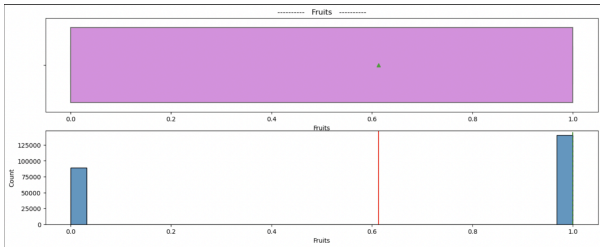
Heart Disease or Attack



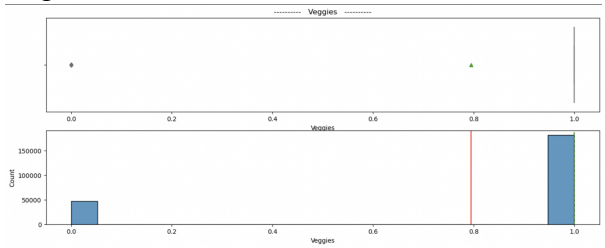
Physical Activity



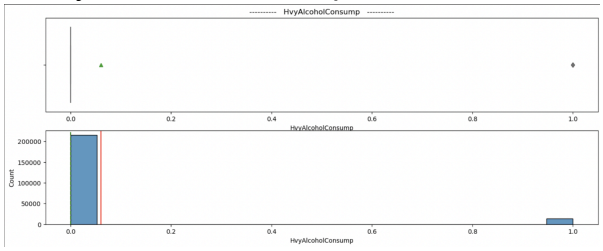
Fruits



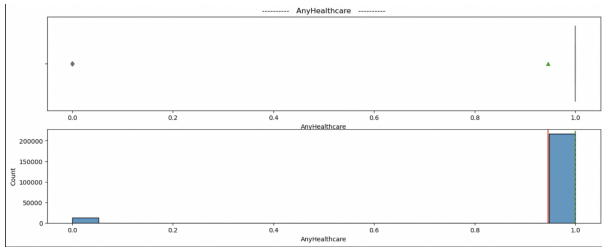
Vegetables



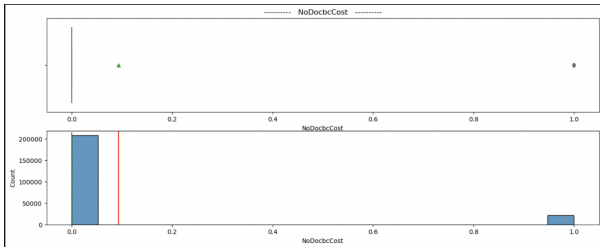
Heavy Alcohol Consumption



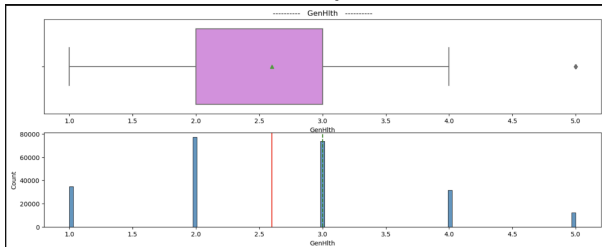
Any Health Care Checkup



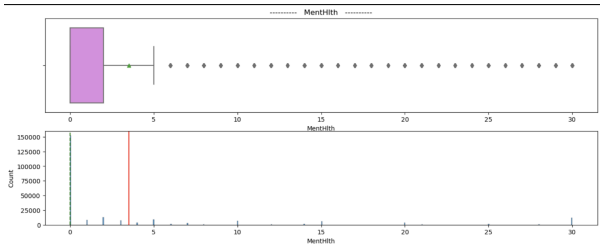
No Doc due to Cost



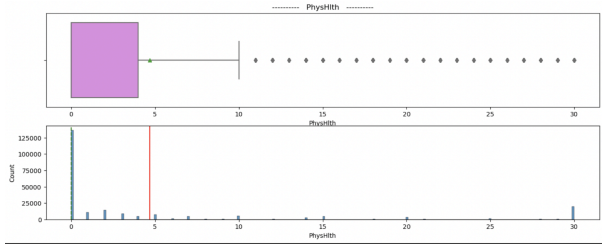
General Health Checkup



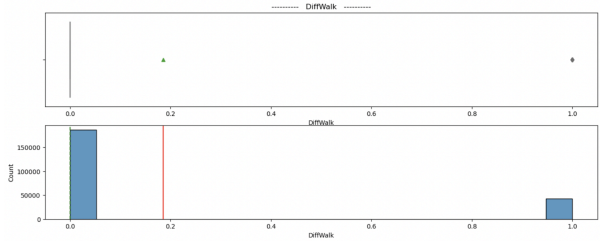
Mental Health



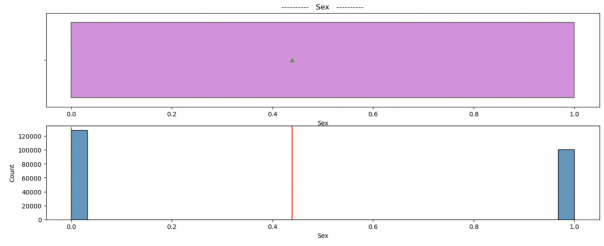
Physical Health



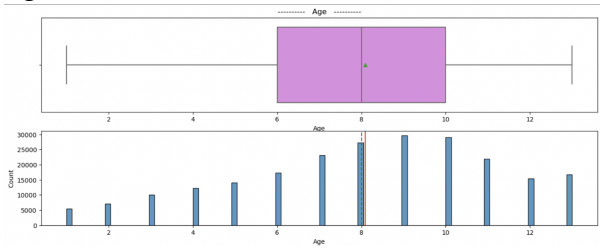
Difficulty Walking

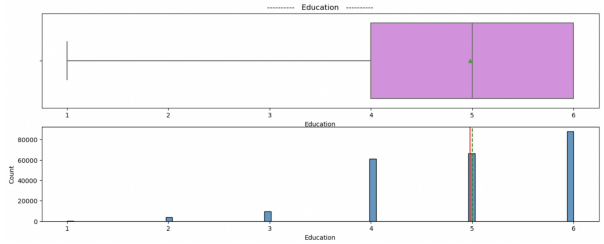
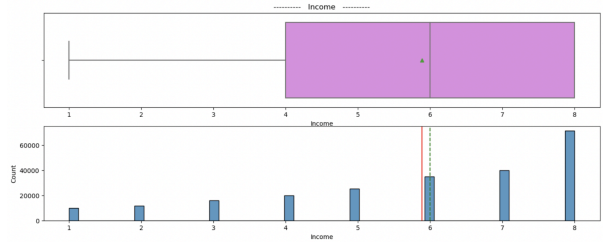


Sex



Age



Education	Income
	

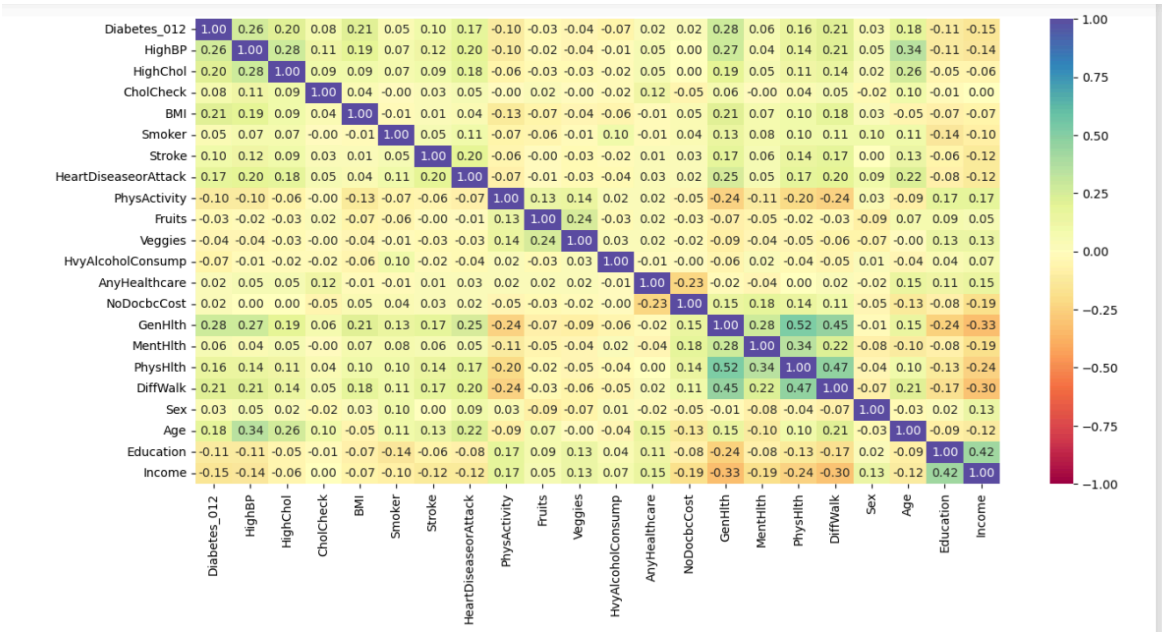
Observations:

Diabetes Binary : The target variable labelling a patient as Diabetic (1) or Not-Diabetic (0) was imbalanced with only about 15% of the observations being diabetic. The majority (85%) of the dataset has a value of 0, indicating no diabetes.

HighBP and HighChol: Both HighBP and HighChol have a mean close to 0.5, indicating a roughly equal distribution between the presence and absence of these conditions. Further investigation can explore how these conditions relate to diabetes status.

BMI: Body Mass Index, a continuous variable. The mean BMI is 28.69, with a standard deviation of 6.79 and slightly skewed towards right.

Correlations between the Features:



Observations:
From the heatmap, we can see that there are very few dark spots. This means that there is no high correlation between any of these features. The highest positive correlation is 0.52 between Physical Health and General Health Check-ups, followed by 0.47 between Physical Health and Difficulty Walking. These ranges of correlations are safely handled by all the models that we use and therefore we chose not to do any treatment for this.

Feature Correlation with Target Variable:

The below-listed features are the ones having highest correlation with the target variable.

Feature Name	Correlation Value
GenHlth	0.28
HighBP	0.25
DiffWalk	0.21
BMI	0.21
HighChol	0.19
Age	0.18
HeartDiseaseorAttack	0.17
PhysHlth	0.16
Stroke	0.10
CholCheck	0.07

Positive Correlations:

GenHlth (0.2769):
A positive correlation suggests that individuals who rate their general health higher are more likely to have diabetes.

Age (0.1773):
Positive correlation indicates that as age increases, the likelihood of having diabetes also increases.

PhysHlth (0.1562):
Individuals with more physical health issues in the past 30 days tend to have a higher correlation with diabetes.

HeartDiseaseorAttack (0.1682):
Positive correlation implies a connection between a history of heart disease or attack and the likelihood of diabetes.

BMI (0.2051):
Positive correlation indicates that higher BMI is associated with a higher likelihood of diabetes.

DiffWalk (0.2053):
The positive correlation suggests that individuals with difficulty walking or climbing stairs are more likely to have diabetes.

Negative Correlations:

PhysActivity (-0.1004):
A negative correlation implies that individuals who engaged in physical activity in the past 30 days are less likely to have diabetes.

Education (-0.1027):
A negative correlation suggests a potential trend where higher education is associated with a lower likelihood of diabetes.

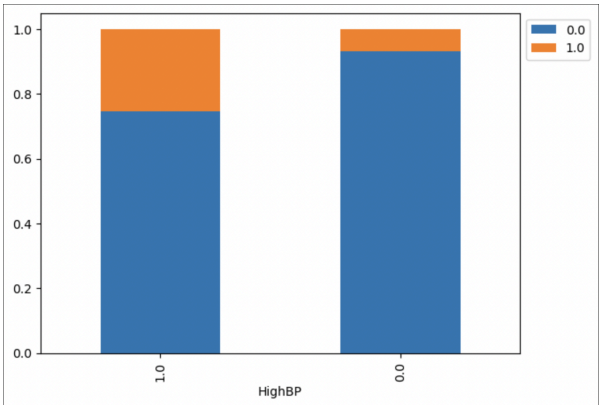
Income (-0.1407):
A negative correlation indicates that individuals with higher income levels may have a lower likelihood of diabetes.

HvyAlcoholConsump (-0.0660):
A negative correlation suggests that heavy alcohol consumption is associated with a lower likelihood of diabetes.

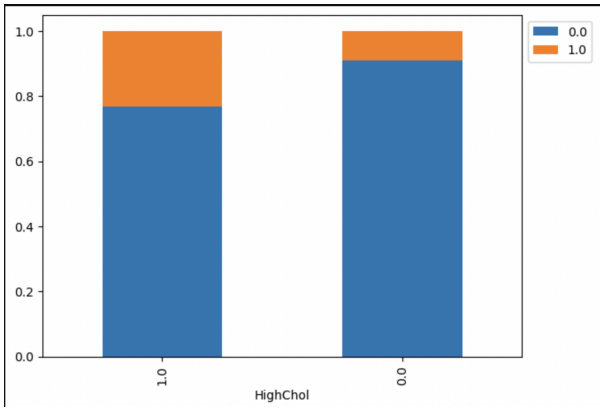
Fruits (-0.0248) and **Veggies** (-0.0417):
Interestingly there is a slight negative correlation between the consumption of fruits and vegetables and the likelihood of diabetes.

Feature Correlation with Target Variable:

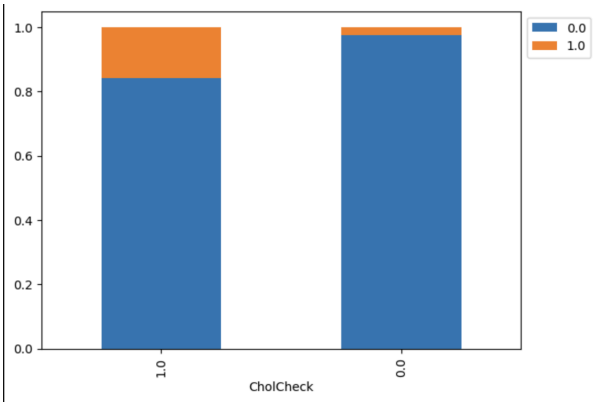
In the table below the orange portion in the stacked bar charts represents the proportion of diabetic patients among that particular class of the variable being observed, while the blue portion represents the non-diabetic population's



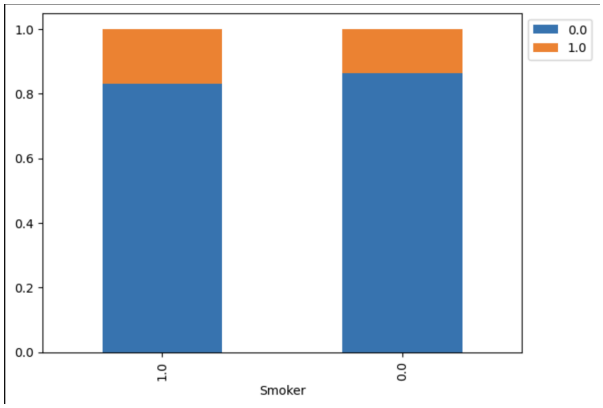
Higher proportion of diabetic patients in people with high Blood Pressure



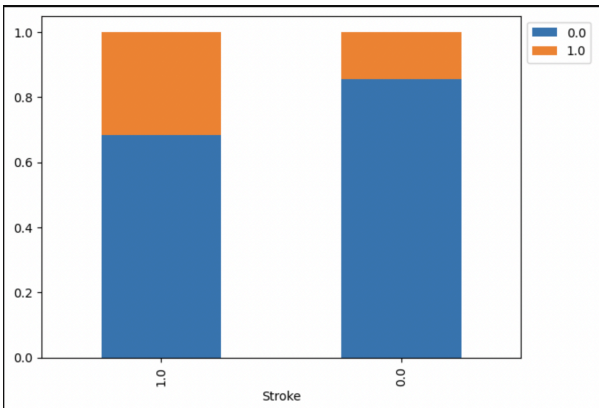
Higher proportion of diabetic patients in people with high Cholesterol



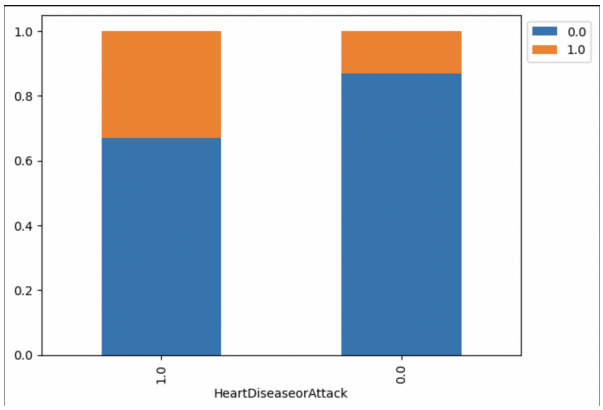
Higher Proportion of diabetic patients among those who underwent Cholesterol check in last 5 years



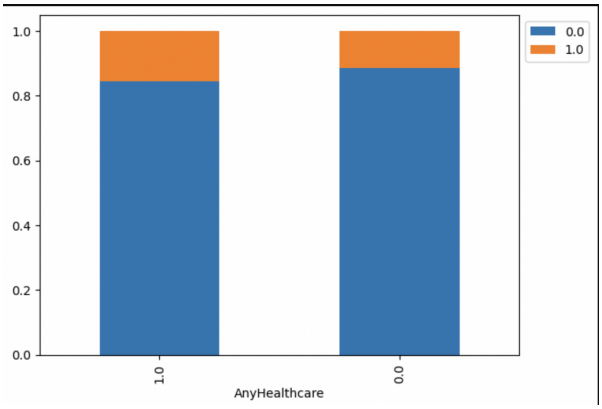
Smokers have marginally higher proportion of diabetic patients.

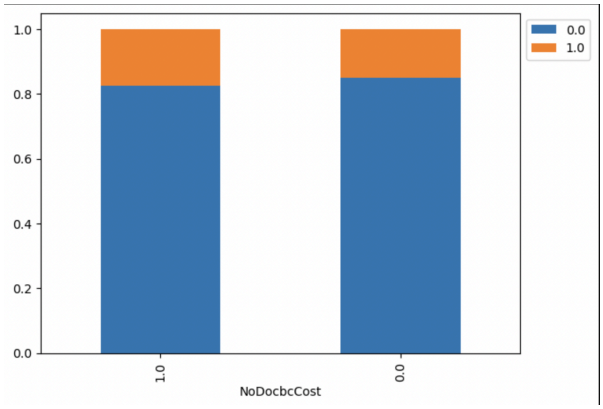


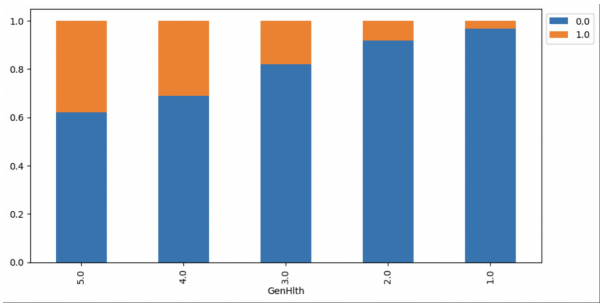
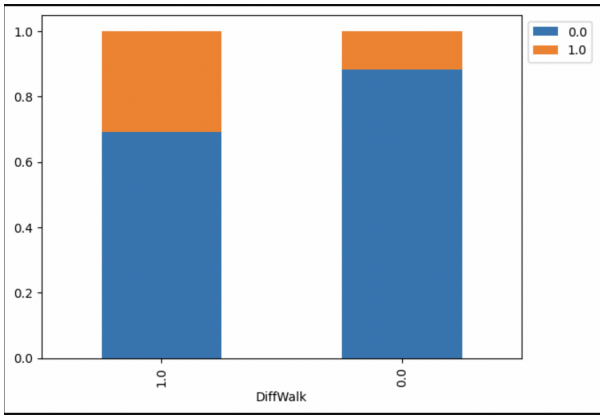
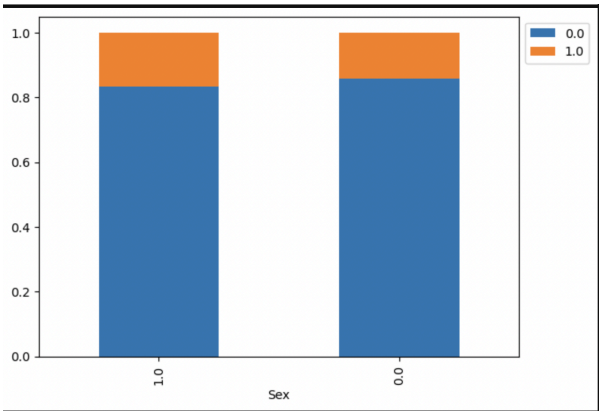
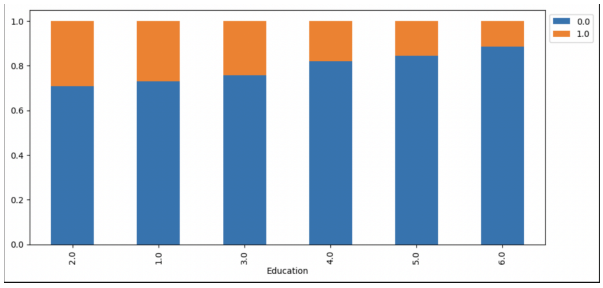
People who experienced a stroke have a higher proportion of diabetic individuals

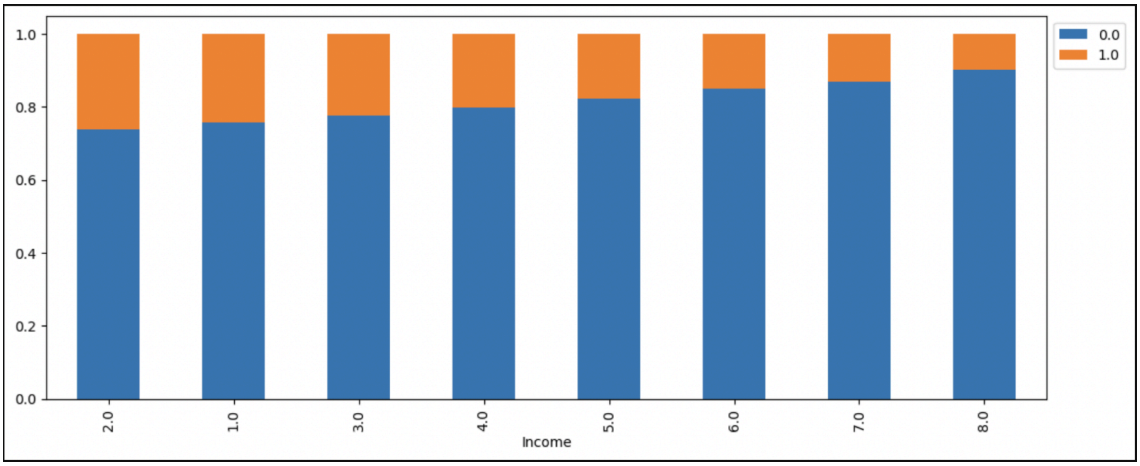


People who experienced a heart attack or heart disease have higher proportion of diabetic individuals

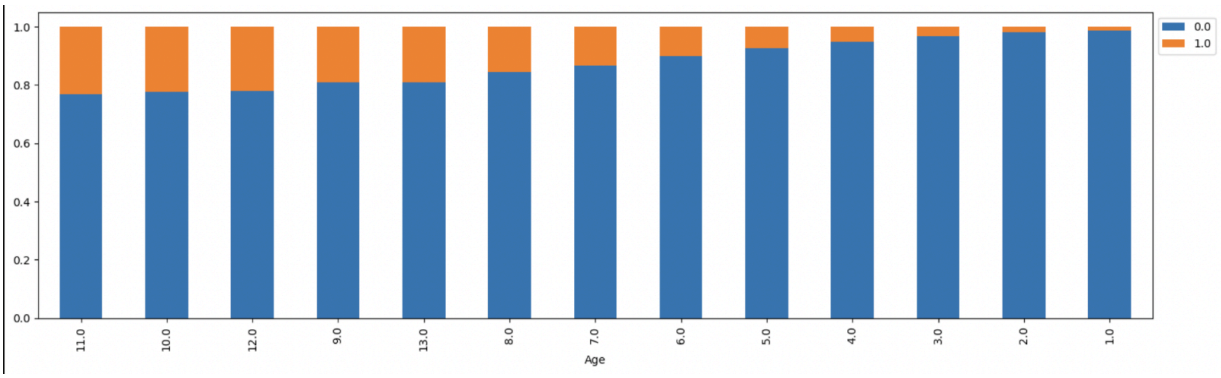




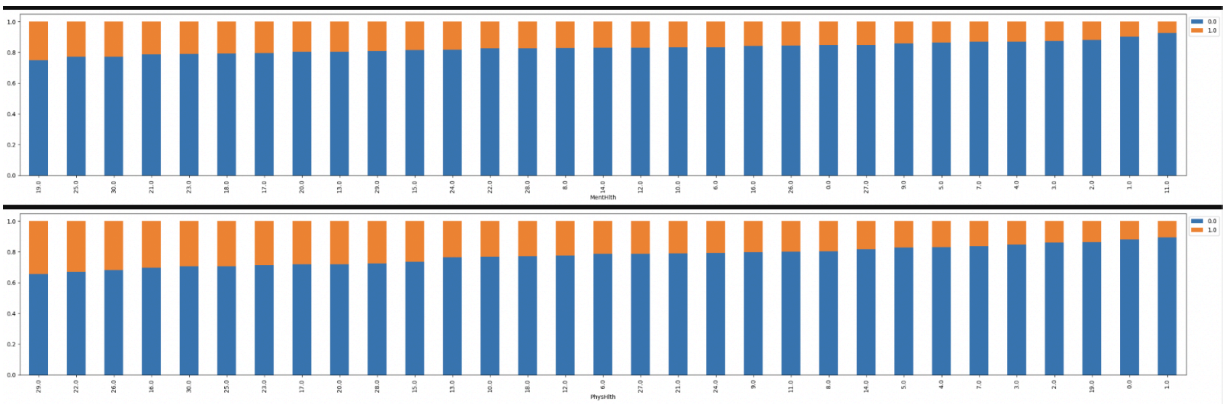
People who underwent any kind of healthcare test have a slightly higher proportion of diabetic diagnoses	Inability to visit a doctor due to costs does not seem to have an effect on the prevalence of diabetes
 Increase in the number of General Health check-ups have an increasing trend of diagnosis of diabetes. Perhaps a possibility of selection bias.	 Higher proportion of diabetic patients in people who faced difficulty in walking
 Marginally higher prevalence of diabetes seen in males compared to females	 Generally increasing trend of prevalence of diabetes with the level of education



Higher prevalence of diabetes seen among those having lower levels of income



Generally increasing trend of prevalence of diabetes with age



Those who experienced Physical or Mental Health issues for a higher number of days in a month, generally, had a higher proportion of diabetic patients

As final steps in our Pre-Processing, We partitioned 20% of the data for Testing and used the remaining 80% to train our models. We used Stratified partitioning of the data, given the imbalance in the distribution of target variable. Stratification ensured that both training and testing partitions have about 15% of the positive classes which is the proportion in the original dataset.

Before building the models we also Standardized the features using StandardScaler to bring all these columns with variable ranges of distributions into a comparable scale. This is important since many of the models we use are sensitive to the varying scales of the features, and in order for the models to work effectively, we need to bring them onto a common scale.

Building our Machine Learning Models:

We started off with a few traditional Machine Learning Models for the prediction of diabetes using the features. We used Logistic Regression, Gaussian Naive Bayes and Decision tree models. Since our objective was to detect the prevalence of diabetes among the individuals we chose to use Recall as the primary metric for our model comparisons while also keeping an eye out for Precision in order not to make the costs of predicting a positive class too prohibitive.

1. Logistic Regression

First up, we ran a Logistic regression on the scaled dataset. The initial model produced the following output metrics:

Evaluation metrics on the training dataset

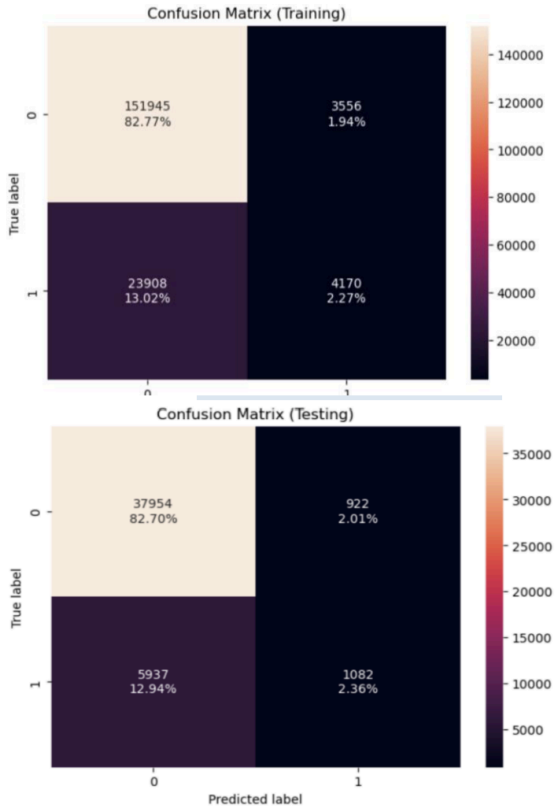
	Accuracy	Recall	Precision	F1
0	0.850397	0.148515	0.539736	0.232935

Evaluation metrics on the testing dataset

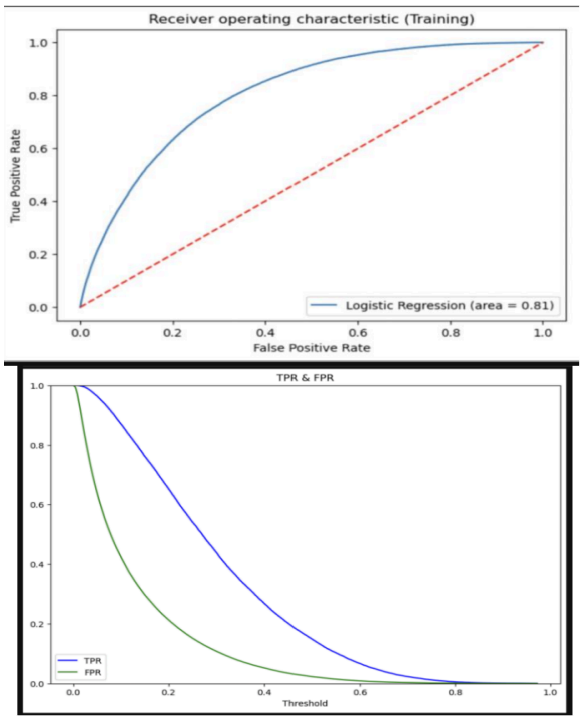
	Accuracy	Recall	Precision	F1
0	0.85055	0.154153	0.53992	0.239832

The model produced a decent enough accuracy score of 85% on both training and testing partitions. However, the model performed very poorly on Recall which is the metric of our interest. A somewhat better Precision of about 53% also lifted the F1-Score a bit higher than that of Recall.

The results were visualised through Confusion Matrices for Training and Testing Partitions which is shown below:



To improve our model we tried optimizing for Threshold based on the ROC area under the curve.



The optimal classification threshold based on the ROC Curve was obtained as 0.14 and the performance metrics were as follows:

Evaluation metrics on the training dataset

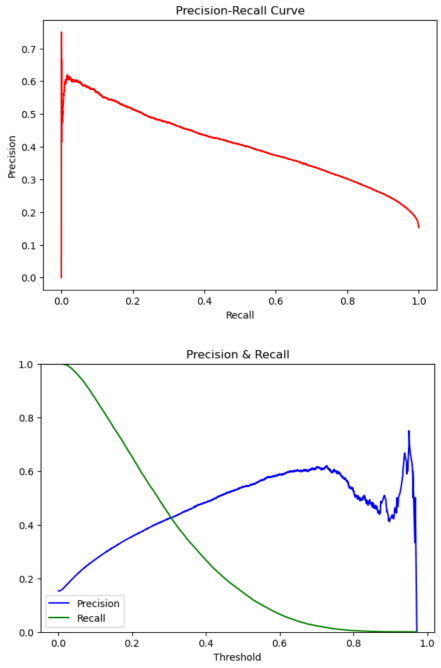
	Accuracy	Recall	Precision	F1
0	0.701344	0.779863	0.310406	0.444064

Evaluation metrics on the testing dataset

	Accuracy	Recall	Precision	F1
0	0.698137	0.785155	0.308618	0.443078

It can be seen that the Recall value, which is the metric of our interest improved significantly to about 78% on both training and testing partitions. There was a slight drop in Precision and Recall measures, but that is an acceptable trade-off considering our objective.

Next, We continued our search for an optimal threshold by optimising it using the Precision-Recall curve which gave us a threshold of about 0.3.



However, the performance metrics were worse off compared to even the default Logistic Regression Model. We therefore chose to ignore the results of this and retained the threshold of 0.14 for Logistic Regression and for further comparisons with other models.

The table below provides the model interpretation in terms of Feature Coefficients and how the odds of having diabetes change with the various features

	Feature_Names	Coefficient	odds	percent_change_odds
0	Intercept	-2.230149	0.107512	-89.248754
1	HighBP	0.366906	1.443262	44.326218
2	HighChol	0.279440	1.322389	32.238851
3	CholCheck	0.251201	1.285568	28.556819
4	BMI	0.396702	1.486913	48.691254
5	Smoker	-0.013207	0.986880	-1.311977
6	Stroke	0.025968	1.026308	2.630775
7	HeartDiseaseorAttack	0.065302	1.067481	6.748101
8	PhysActivity	-0.013582	0.986510	-1.349030
9	Fruits	-0.012916	0.987167	-1.283321
10	Veggies	-0.004034	0.995974	-0.402567
11	HvyAlcoholConsump	-0.190941	0.826181	-17.381861
12	AnyHealthcare	0.021527	1.021761	2.176086
13	NoDocbcCost	-0.005299	0.994715	-0.528484
14	GenHlth	0.532962	1.703972	70.397197
15	MentHlth	-0.033577	0.966980	-3.301981
16	PhysHlth	-0.060369	0.941417	-5.858317
17	DiffWalk	0.050980	1.052302	5.230203
18	Sex	0.130137	1.138984	13.898389
19	Age	0.373001	1.452086	45.208565
20	Education	-0.019556	0.980634	-1.936638
21	Income	-0.092953	0.911236	-8.876406

2. Gaussian Naive Bayes

The next model we tried was the Gaussian Naive Bayes which is based on Bayes theorem for Probabilities and assumes conditional independence of the features with respect to prediction class..

The default model provided us with the following evaluation metrics:

Evaluation metrics on the training dataset

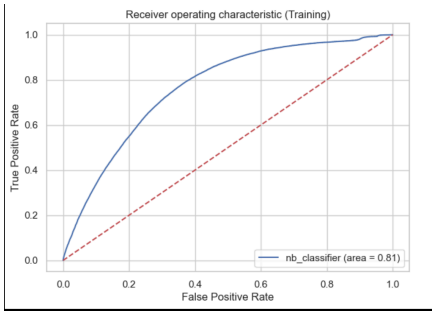
	Accuracy	Recall	Precision	F1
0	0.7555	0.569984	0.32785	0.416267

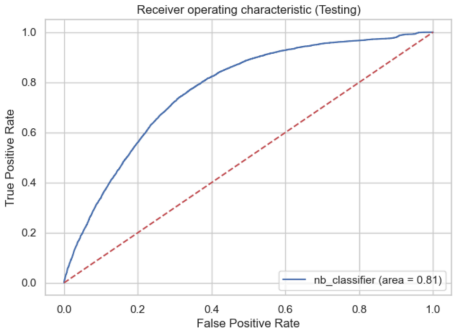
Evaluation metrics on the testing dataset

	Accuracy	Recall	Precision	F1
0	0.756727	0.583416	0.331955	0.423146

The Recall value was in the middling range of 0.56 to 0.58. Compared to the Recall we obtained in the Logistic Regression model after optimising for the threshold, this does not look good.

To improve the recall value we found the optimal threshold value based on ROC-AUC which gave us 0.08





Evaluation metrics on the training dataset

	Accuracy	Recall	Precision	F1
0	0.658387	0.783959	0.279841	0.412453

Evaluation metrics on the testing dataset

	Accuracy	Recall	Precision	F1
0	0.659157	0.791708	0.281538	0.415368

The Recall value improved significantly here too while the Precision dropped by a few points. However, the threshold value of 0.08 seemed to be too low and problematic. It is virtually 0 which means we have to catch every person in a given population to identify about 79% of diabetic patients.

3. Decision Tree Classifier

Next, we ran a Decision tree model without any pruning. The default Decision Tree grew to a depth of 42 and obtained the following performance metrics:

Evaluation metrics on the training dataset

	Accuracy	Recall	Precision	F1
0	0.994591	0.964634	1.0	0.981999

Evaluation metrics on the testing dataset

	Accuracy	Recall	Precision	F1
0	0.774137	0.320416	0.286679	0.30261

The default decision tree achieved high accuracy and Recall scores on the training set but exhibited signs of overfitting, as evidenced by a substantial drop in accuracy on the test set. We performed hyper-parameter tuning in search of a Decision-Tree model that gives us the best F1-score (which is a weighted average of Recall and Precision). The tuned/pruned decision tree had a max depth of 32 with the following metrics:

Evaluation metrics on the training dataset

	Accuracy	Recall	Precision	F1
0	0.993752	0.960289	0.998815	0.979173

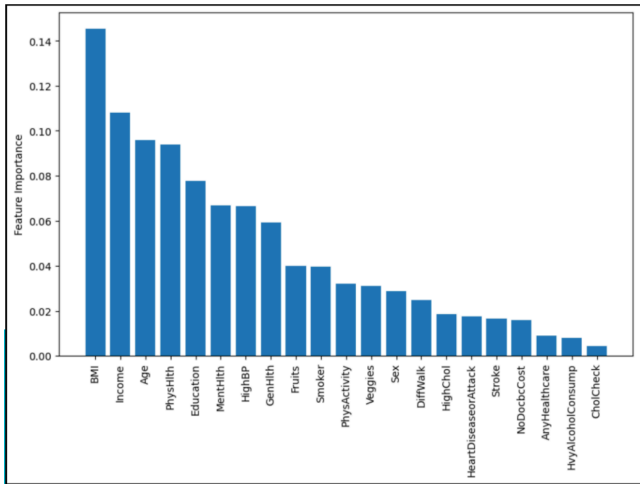
Evaluation metrics on the testing dataset

	Accuracy	Recall	Precision	F1
0	0.772197	0.314717	0.281258	0.297048

As we can see from the Performance metrics, pruning the tree showed some improvement in generalization ability, however, the decision tree was still largely overfitting. The precision for predicting diabetes (class 1) remains low, suggesting that the model struggles with correctly identifying positive cases.

The recall for diabetes prediction is also modest, indicating that the model may miss a significant number of actual positive cases. The F1-score, a balance between precision and recall, is relatively low for the positive class

Feature Importance for Decision-Tree Classifier:



ENSEMBLE MODELS

4. After trying these individual Machine-Learning models, we wanted to combine multiple models (wisdom of the crowd) to improve our prediction performance metrics. Even in choosing Ensemble methods, the idea was to have variety. We therefore chose the Random-Forest which works on Bagging and Booststrapped Sampling, followed by the LightGBM model which is based on Gradient Boosting. We followed these models with ADA Boost which uses weights to boost the performance and then the XG Boost model based on Newton method. We finally tried to combine the models as well as ensembles using a Meta-Classifier or Stacking Classifier to create an "Ensemble of Ensembles"**Random Forest Classifier**

Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions to obtain more accurate and stable results.

The model shows signs of overfitting, as evidenced by the significant drop in recall from the training set (73.70%) to the test set (20.42%).

	precision	recall	f1-score	support
0.0	0.87	0.96	0.91	38876
1.0	0.47	0.17	0.25	7019
accuracy			0.84	45895
macro avg	0.67	0.57	0.58	45895
weighted avg	0.80	0.84	0.81	45895

The model exhibits high precision and recall for predicting no diabetes (Class 0), suggesting it is robust in identifying individuals without the condition.

For diabetes prediction (Class 1), the model struggles, highlighting the challenge of identifying positive cases accurately

Tuned Random Forest model for better Recall metric:

Optimal Hyperparameters:

n_estimators: 20
max_samples: 0.9
max_features: 0.9
max_depth: 20

	precision	recall	f1-score	support
0.0	0.87	0.96	0.91	38876
1.0	0.47	0.20	0.28	7019
accuracy			0.84	45895
macro avg	0.67	0.58	0.60	45895
weighted avg	0.81	0.84	0.82	45895

The tuned Random Forest model achieves comparable accuracy to the default model, suggesting that hyperparameter tuning has a modest impact on overall accuracy.

The precision for predicting diabetes (Class 1) improved only marginally from 47% to 47.14%

Recall for diabetes prediction (Class 1) increased from 17% to 20.42%, implying a slight improvement in capturing actual positive cases. However, like the Decision Trees, We can see that there is a case of excessive overfitting in the model based on the gap between Training and Testing Performances.

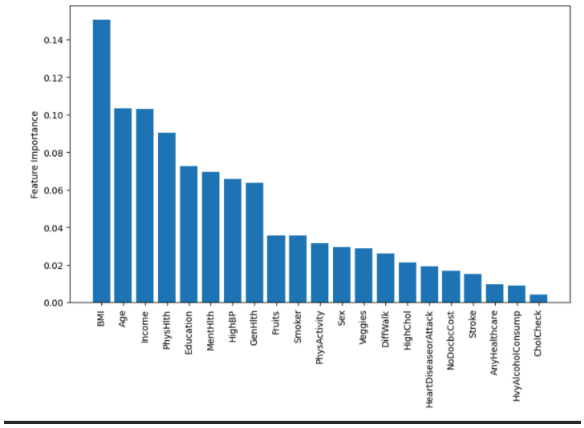
Evaluation metrics on the training dataset

	Accuracy	Recall	Precision	F1
0	0.958585	0.737018	0.989528	0.844808

Evaluation metrics on the testing dataset

	Accuracy	Recall	Precision	F1
0	0.843273	0.20416	0.471382	0.284919

The most important features of Random Forest can be seen below:



5. Light GBM

LightGBM is a gradient-boosting framework that uses tree-based learning algorithms. It's particularly efficient for large datasets and handles class imbalances well. In the context of diabetic disease detection, achieving a balance between precision and recall is crucial, especially when dealing with a low threshold for class imbalance.

	precision	recall	f1-score	support
0.0	0.87	0.98	0.92	38876
1.0	0.61	0.17	0.26	7019
accuracy			0.86	45895
macro avg	0.74	0.57	0.59	45895
weighted avg	0.83	0.86	0.82	45895

Optimal Hyperparameters:

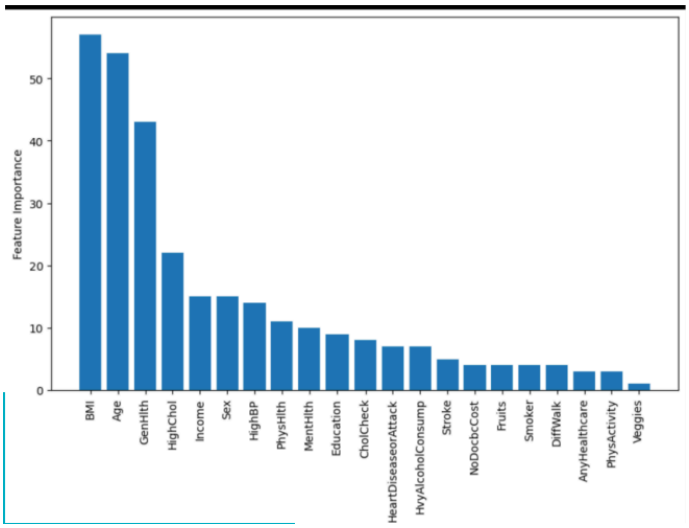
n_estimators: 100
max_depth: 2
learning_rate: 0.5

	precision	recall	f1-score	support
0.0	0.87	0.98	0.92	38876
1.0	0.59	0.17	0.27	7019
accuracy			0.86	45895
macro avg	0.73	0.58	0.59	45895
weighted avg	0.83	0.86	0.82	45895

The tuned LightGBM model shows promising improvements in precision and recall for predicting diabetes. However, achieving a balance between precision and recall

remains a challenge, and further refinement may be necessary for practical application. The model's robustness in predicting no diabetes highlights its potential for identifying individuals without the condition accurately.

However, LightGBM provides valuable insights into feature importance to further understand which features contribute most to the model's predictions.



6. Ada-Boost

AdaBoost is an ensemble learning technique that combines the predictions of weak learners, typically decision trees, to create a strong classifier. In the context of diabetic disease detection, AdaBoost can provide valuable insights into how the model balances precision and recall, especially with a low threshold for class imbalance.

For the default Ada-boost model:

	precision	recall	f1-score	support
0.0	0.87	0.97	0.92	38876
1.0	0.57	0.19	0.28	7019
accuracy			0.85	45895
macro avg	0.72	0.58	0.60	45895
weighted avg	0.82	0.85	0.82	45895

The optimal hyperparameters used for the model are

optimal hyper-parameters:

n_estimators: 200
learning_rate: 0.5

	precision	recall	f1-score	support
0.0	0.87	0.97	0.92	38876
1.0	0.57	0.19	0.28	7019
accuracy			0.85	45895
macro avg	0.72	0.58	0.60	45895
weighted avg	0.82	0.85	0.82	45895

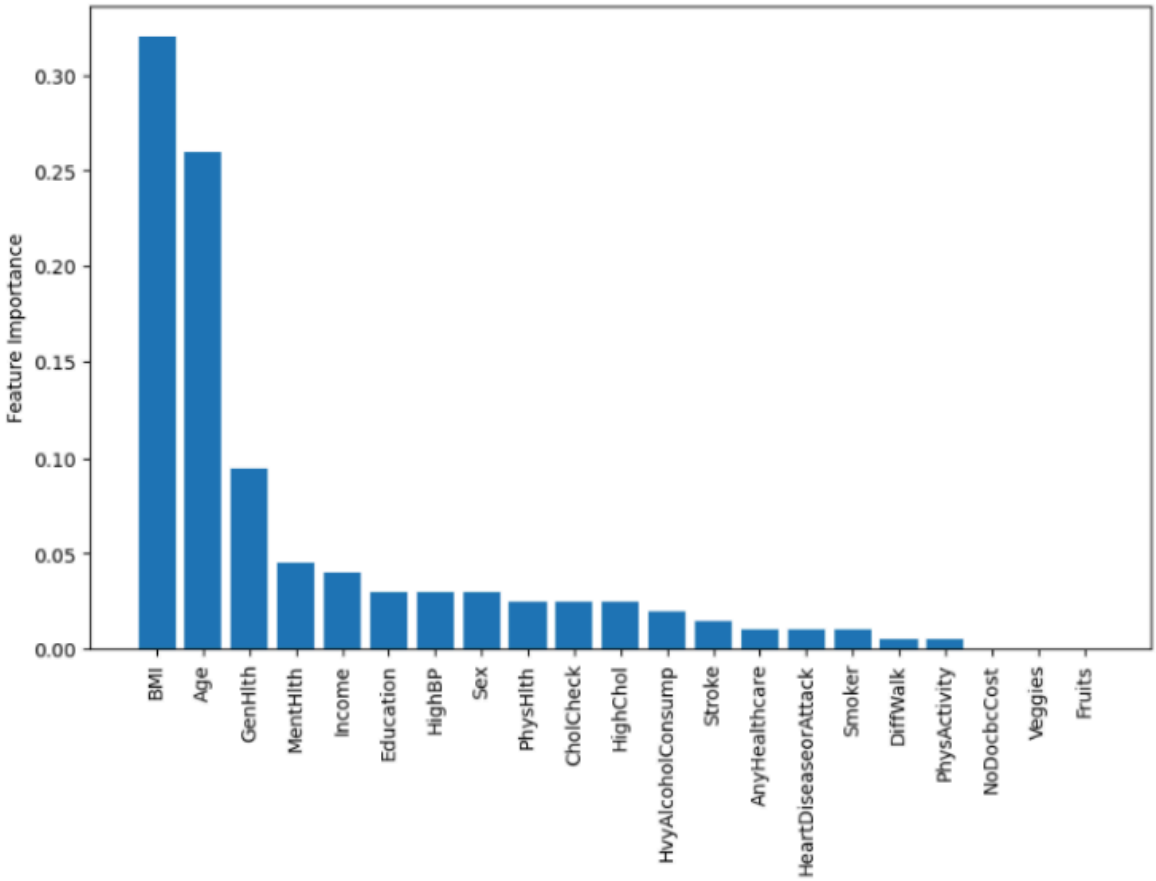
The tuned AdaBoost model maintained an accuracy of 85%, with a precision of 57% and recall of 19% for predicting diabetes (Class 1).

The F1-score for Class 1 improved slightly from 28% to 28.14%.

The model's performance on the testing dataset was consistent with that on the training dataset.

The AdaBoost model, even with hyperparameter tuning, faces challenges in achieving a high recall for predicting diabetes. Balancing precision and recall remains crucial in the context of diabetic disease detection, and further refinements may be needed to enhance the model's ability to identify positive cases while maintaining a low false positive rate.

Feature importance based on Adaboost model:



7. XG-Boost

Xtreme gradient boosting algorithm widely used for classification tasks. In the context of diabetic disease detection, XGBoost can provide insights into how well the model balances precision and recall, particularly with a low threshold for class imbalance.

	precision	recall	f1-score	support
0.0	0.87	0.98	0.92	38876
1.0	0.57	0.18	0.27	7019
accuracy			0.85	45895
macro avg	0.72	0.58	0.59	45895
weighted avg	0.82	0.85	0.82	45895

The default XGBoost model achieved an accuracy of 87.3%, with a precision of 57% and recall of 18% for predicting diabetes (Class 1).

The F1-score, representing the balance between precision and recall, was 27% for Class 1.

The model demonstrates a relatively high precision for predicting individuals without diabetes (Class 0), indicating a low false positive rate.

Through hyperparameter tuning, the XGBoost model was fine-tuned with the following optimal parameters:

Optimal parameters:
subsample: 0.7
scale_pos_weight: 5
n_estimators: 100
max_depth: 2
learning_rate: 0.05
colsample_bytree: 0.9
colsample_bylevel: 0.5

	precision	recall	f1-score	support
0.0	0.94	0.72	0.82	38876
1.0	0.33	0.76	0.46	7019
accuracy			0.73	45895
macro avg	0.64	0.74	0.64	45895
weighted avg	0.85	0.73	0.76	45895

The tuned XGBoost model shows promise in improving recall for predicting diabetes for both the classes and is an ideal candidate for our final predictor

We then moved on to the Meta-Learner method

8. Meta Learner method (Stacking Classifier)

A meta-learner, represented by a stacking classifier, combines the predictions of multiple base models to enhance the overall performance. In the context of diabetic disease detection, the stacking classifier utilizes Adaboost, LightGBM, and XGBoost as base models.

	precision	recall	f1-score	support
0.0	0.94	0.71	0.81	38876
1.0	0.32	0.77	0.46	7019
accuracy			0.72	45895
macro avg	0.63	0.74	0.63	45895
weighted avg	0.85	0.72	0.76	45895

The default stacking classifier achieved an accuracy of 72%, with a precision of 32% and recall of 77% for predicting diabetes (Class 1).

The F1-score, representing the balance between precision and recall, was 45.6% for Class 1.

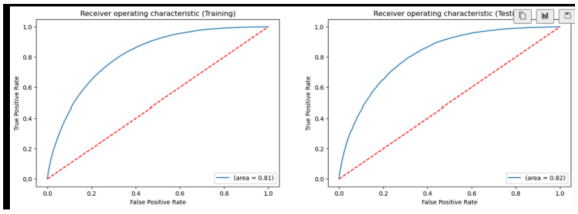
The model demonstrates a trade-off between precision and recall, indicating sensitivity to positive cases at the expense of an increased false positive rate.

Optimal Classification Threshold:

An optimal classification threshold, based on the ROC Curve, was determined to be approximately 0.465.

Adjusting the threshold resulted in updated performance metrics, with a

slight decrease in overall accuracy but an improvement in recall.



With a slight decrease in the accuracy, but improved recall after the optimal threshold adjustment:

Evaluation metrics on the training dataset

	Accuracy	Recall	Precision	F1
0	0.693783	0.803049	0.307894	0.445124

Evaluation metrics on the testing dataset

	Accuracy	Recall	Precision	F1
0	0.692123	0.808947	0.307467	0.445578

With the adjusted threshold, the stacking classifier achieved an accuracy of 69.2%, with a recall of 80.9% for predicting diabetes (Class 1).

The F1-score for Class 1 remained at 44.6%, suggesting that the model maintains a balance between precision and recall with the new threshold.

The stacking classifier, as a meta-learner, exhibited the best performance in terms of Recall Value for Diabetes detection while having a reasonably higher Precision in comparison to other models. We therefore choose this as our final model.

Model Comparisons

Training performance comparison:

	Accuracy	Recall	Precision	F1
Logistic Regression	0.850397	0.148515	0.539736	0.232935
Logistic Regression (changed threshold)	0.701344	0.779863	0.310406	0.444064
Naive Bayes	0.755500	0.569984	0.327850	0.416267
Naive Bayes (changed threshold)	0.658387	0.783959	0.279841	0.412453
Decision Tree	0.993752	0.960289	0.998815	0.979173
Random Forest	0.958585	0.737018	0.989528	0.844808
Light GBM	0.853785	0.166287	0.576277	0.258098
ADA Boost	0.852380	0.175155	0.555204	0.266298
XG Boost	0.726951	0.744106	0.327302	0.454631
Meta-Learner	0.720927	0.762875	0.324575	0.455396
Meta-Learner (changed threshold)	0.693783	0.803049	0.307894	0.445124

Testing performance comparison:

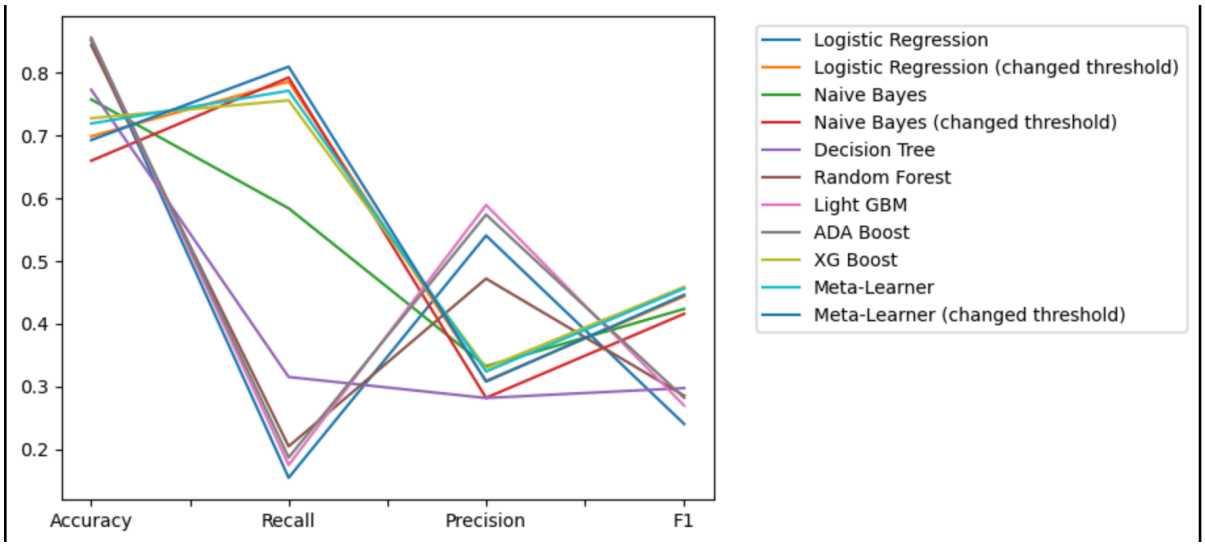
	Accuracy	Recall	Precision	F1
Logistic Regression	0.850550	0.154153	0.539920	0.239832
Logistic Regression (changed threshold)	0.698137	0.785155	0.308618	0.443078
Naive Bayes	0.756727	0.583416	0.331955	0.423146
Naive Bayes (changed threshold)	0.659157	0.791708	0.281538	0.415368
Decision Tree	0.772197	0.314717	0.281258	0.297048
Random Forest	0.843273	0.204160	0.471382	0.284919
Light GBM	0.855104	0.174526	0.588659	0.269231
ADA Boost	0.854363	0.186494	0.573368	0.281445
XG Boost	0.726855	0.755378	0.328888	0.458254
Meta-Learner	0.718379	0.770623	0.323427	0.455629
Meta-Learner (changed threshold)	0.692123	0.808947	0.307467	0.445578

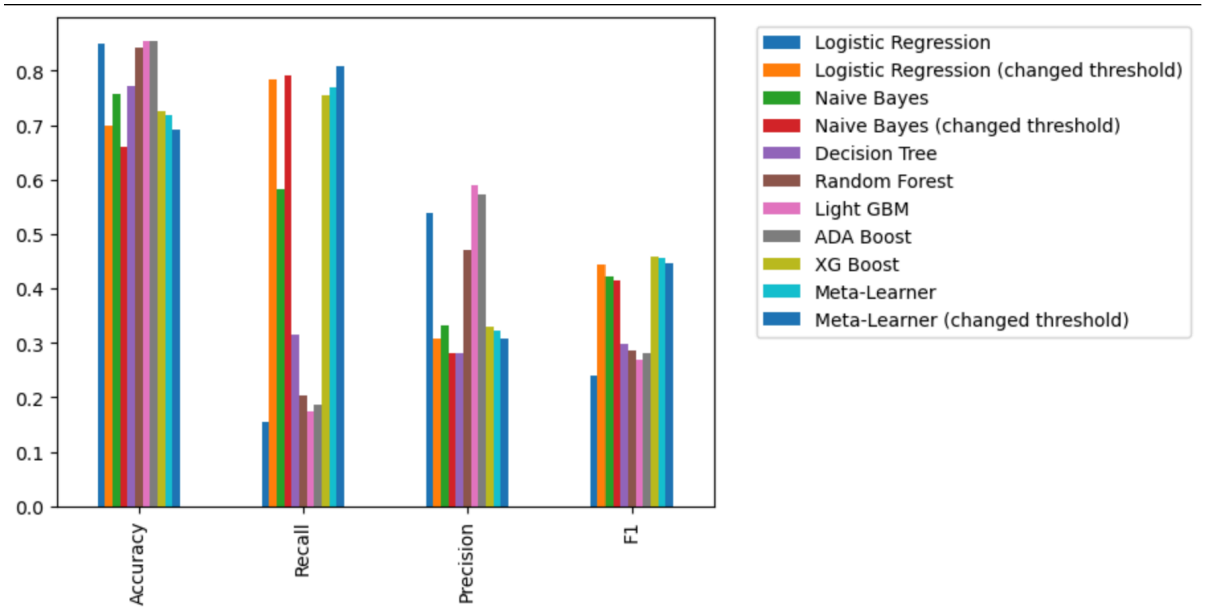
Decision Tree and Random Forest: These models exhibit high accuracy and F1-scores training sets but the performance drops significantly on testing sets, indicating a high degree of overfitting.

Logistic Regression: The model, even with a threshold adjustment, demonstrates moderate performance, emphasizing the need for further tuning or consideration of ensemble methods.

Light GBM, ADA Boost, and XG Boost: These models show varying degrees of success in balancing precision and recall but perform better than Decision Trees and Random Forest Models

Meta-Learner: While the stacking classifier combines various models, it exhibits the best performance on Recall, particularly after optimising for threshold.





Concluding Remarks:

On comparing the models based on our chosen metrics of Recall and Precision, We propose the Stacking Classifier / Meta-learner model for predicting the prevalence of diabetes. It has to be noted that the predictions of this model are heavily influenced by the predictions of the XG-Boost model more than others. Another recommendation that we have is that the logistic Regression Model also fares comparably well albeit with a slightly reduced performance on the metrics. If the model interpretation is important, then we would recommend using the Logistic Regression model given its robustness in feature interpretability despite the low threshold at which it provides this performance