

Project Summary

Batch details	DSE-MAY22-G2
Team members	Akshay Jadhav Arun Cibi Chellasamy A Varshiney M Sabarivasan R
Domain of Project	Healthcare
Proposed project title	Risk factors Prediction Model for Diabetes
Group Number	GROUP 2
Team Leader	Sabarivasan R
Mentor Name	Mr. Pratik Sonar

Date: 09-09-2022

Mr. Pratik Sonar

Sabarivasan R

Signature of the Mentor

Signature of the Team Leader

Table of Contents

Sr. No.	Topic	Page No
1	Introduction	3
2	Data Set and Domain	4
3	Exploratory Data Analysis (EDA)	8
4	Statistical Test	20
5	Models Building	23
6	ML Model Result Comparison	37
7	Business Interpretations	39
8	References	40

1. INTRODUCTION

Diabetes is among the most prevalent chronic diseases in the world, impacting millions of people each year and exerting a significant financial burden on the economy. Diabetes is a serious chronic disease where the individual's pancreas loses its ability to effectively secrete insulin which thereby regulates level of glucose in the blood, and can lead to reduced quality of life and life expectancy. Diabetes is generally characterized by either the body not making enough insulin or being unable to use the insulin that is made as effectively as needed.

Complications like heart disease, vision loss, lower-limb amputation, and kidney disease are associated with chronically high levels of sugar remaining in the bloodstream for those with diabetes. While there is no cure for diabetes, strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harms of this disease in many patients. Early diagnosis can lead to lifestyle changes and more effective treatment, making machine learning models to predict diabetes risk an important tool for public and public health officials

Problem Statement

The scale of this problem is also important to recognize. The Centers for Disease Control and Prevention has indicated that 37.3 million Americans (11.3% of the US population) have diabetes and 96 million have prediabetes (38.0% of the adult US population). Furthermore, the CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 pre diabetics are unaware of their risk. While there are different types of diabetes, type II diabetes is the most common form and its prevalence varies by age, education, income, Blood Pressure, Cholesterol and other social determinants of health.

Business Objective

The goal is to predict the Diabetes risk of an individual from the data collected via simple survey questions which will result in early diagnosis of the disease which can lead to lifestyle changes and more effective treatments. Medical Insurance and Fitness companies can also make targeted sales by this predictive model.

2. DATA SET AND DOMAIN

Data Dictionary

HighBP	Blood Pressure
HighChol	Cholesterol Level
CholCheck	Did cholesterol check in last 5 years
BMI	Body Mass Index
Smoker	Whether or not has smoked in his life
Stroke	Whether or not had a stroke
HeartDiseaseorAttack	Whether had Heart Disease
PhysActivity	Physical activity in past 30 days
Fruits	Consume Fruit on a regular interval
Veggies	Consume Vegetables on a regular interval
HvyAlcoholConsump	Consumption level of Alcohol
AnyHealthcare	Have any kind of health care coverage
NoDocbcCost	Needed to see doctor but could not because of cost
GenHlth	General health rating
MentHlth	Mental health on 1 to 5 scale
PhysHlth	Physical health on 1 to 5 scale
DiffWalk	Have difficulty in walking
Sex	Gender
Age	Age of the person
Education	Education level
Income	Annual Income

Data Types

```

RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                                Non-Null Count  Dtype
---  -
 0   Diabetes_binary                      253680 non-null float64
 1   HighBP                              253680 non-null float64
 2   HighChol                            253680 non-null float64
 3   CholCheck                           253680 non-null float64
 4   BMI                                  253680 non-null float64
 5   Smoker                              253680 non-null float64
 6   Stroke                              253680 non-null float64
 7   HeartDiseaseorAttack                253680 non-null float64
 8   PhysActivity                        253680 non-null float64
 9   Fruits                              253680 non-null float64
10  Veggies                              253680 non-null float64
11  HvyAlcoholConsump                   253680 non-null float64
12  AnyHealthcare                       253680 non-null float64
13  NoDocbcCost                         253680 non-null float64
14  GenHlth                             253680 non-null float64
15  MentHlth                            253680 non-null float64
16  PhysHlth                            253680 non-null float64
17  DiffWalk                            253680 non-null float64
18  Sex                                  253680 non-null float64
19  Age                                  253680 non-null float64
20  Education                           253680 non-null float64
21  Income                              253680 non-null float64
dtypes: float64(22)

```

Variable Categorization

Numerical Variables

	PhysHlth	MentHlth	BMI
mean	4.242081	3.184772	28.382364
std	8.717951	7.412847	6.608694
min	0.000000	0.000000	12.000000
25%	0.000000	0.000000	24.000000
50%	0.000000	0.000000	27.000000
75%	3.000000	2.000000	31.000000
max	30.000000	30.000000	98.000000

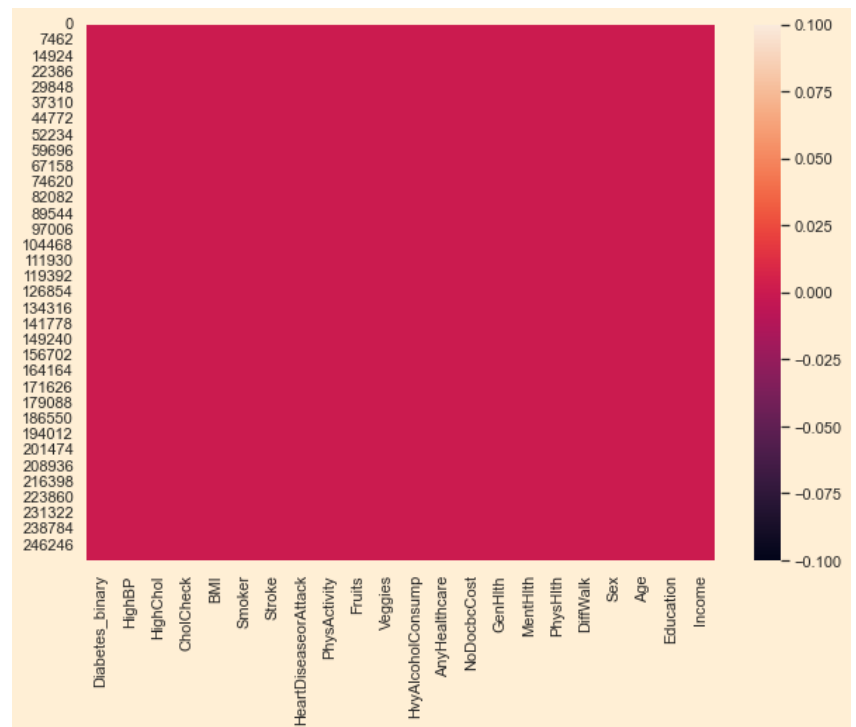
Categorical Variables

columns	
0	Diabetes_binary
1	HighBP
2	HighChol
3	CholCheck
4	Smoker
5	Stroke
6	HeartDiseaseorAttack
7	PhysActivity
8	Fruits
9	Veggies
10	HvyAlcoholConsump
11	AnyHealthcare
12	NoDocbcCost
13	GenHlth
14	DiffWalk
15	Sex
16	Age
17	Education
18	Income

Preprocessing Data Analysis

Count of missing/null values

1	dia.isnull().sum()
Diabetes_binary	0
HighBP	0
HighChol	0
CholCheck	0
BMI	0
Smoker	0
Stroke	0
HeartDiseaseorAttack	0
PhysActivity	0
Fruits	0
Veggies	0
HvyAlcoholConsump	0
AnyHealthcare	0
NoDocbcCost	0
GenHlth	0
MentHlth	0
PhysHlth	0
DiffWalk	0
Sex	0
Age	0
Education	0
Income	0
dtype:	int64



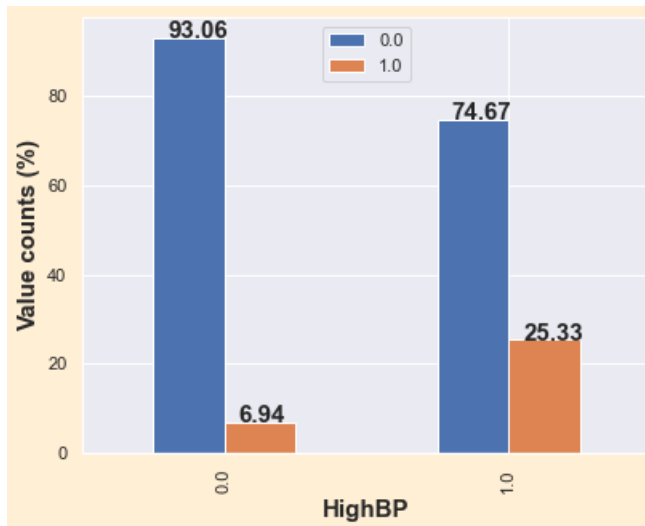
There are no missing values present in the dataset.

Duplicate Values:

On checking the whole data set there were 24206 duplicate observations found. Those were removed from the dataset using the `.drop_duplicates()` method, keeping first values.

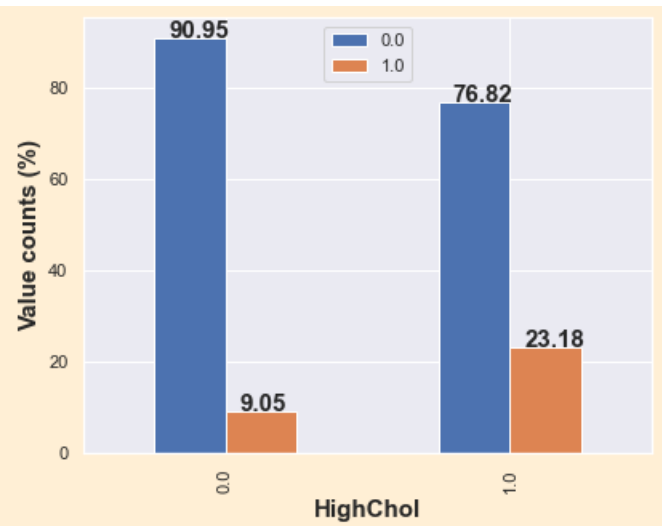
3. DATA EXPLORATION (EDA)

High Blood Pressure vs Diabetes Risk



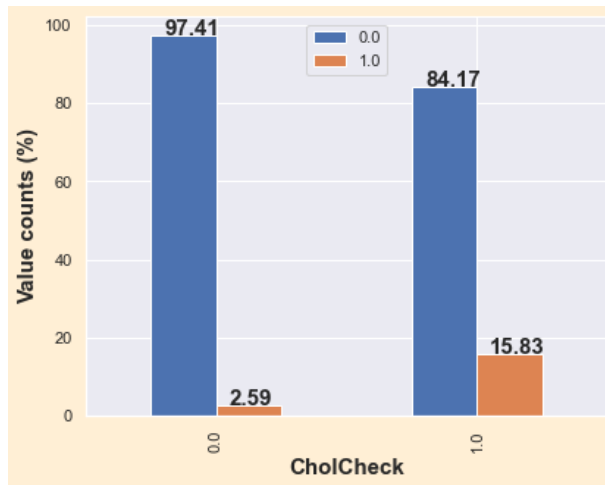
High Blood pressure indicates high vulnerability towards Diabetes compared to Low BP

High Cholesterol vs Diabetes Risk



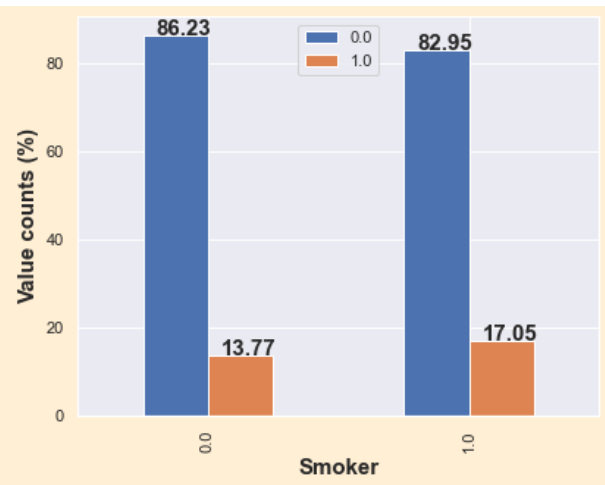
Individuals with High Cholesterol are at a high risk of Diabetes compared to the individuals with low cholesterol

Cholesterol Check vs Diabetes Risk



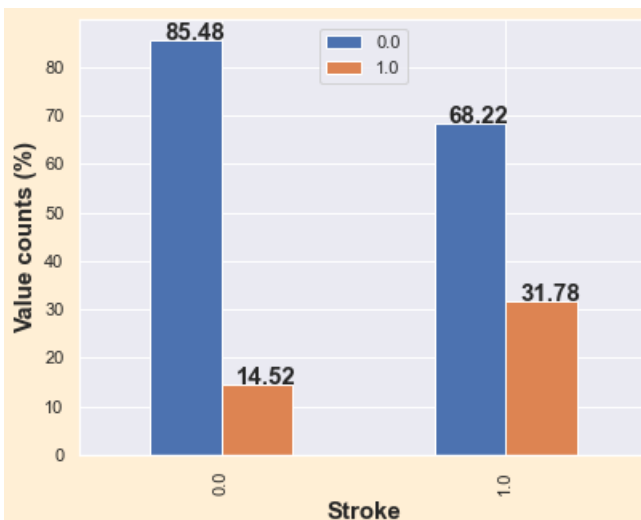
Data represent patients who are having Cholesterol check up are having high risk compare to the patients who are not having cholesterol check

Smoking Habit vs Diabetes Risk



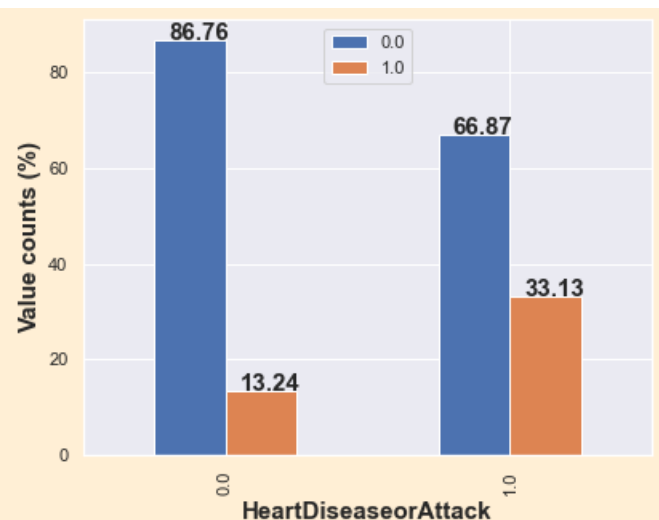
There is no marginal difference of Diabetes risk for a Active Smoker and non smoker

Stroke vs Diabetes Risk



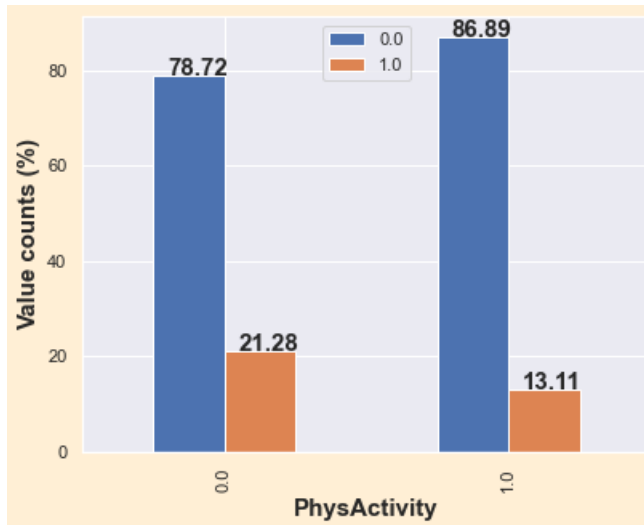
Presence of Stroke resembles to high risk of Diabetes, while risk is less for not having a stroke

Heart Attack vs Diabetes Risk



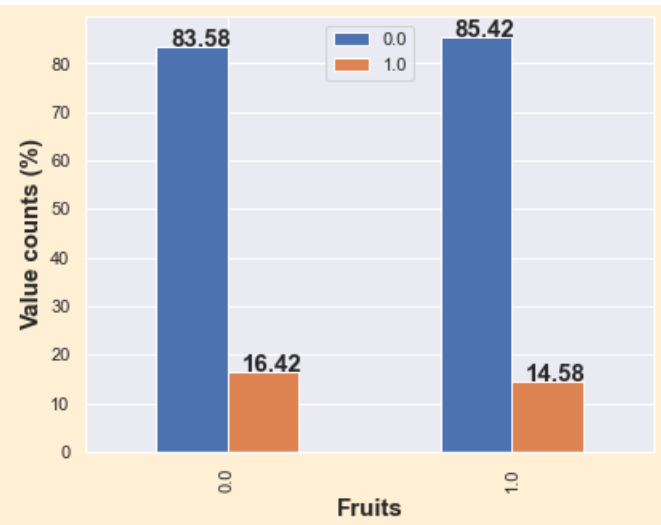
High likelihood is observed for an individual having an history of Heart attack compared to others

Physical Activity vs Diabetes Risk



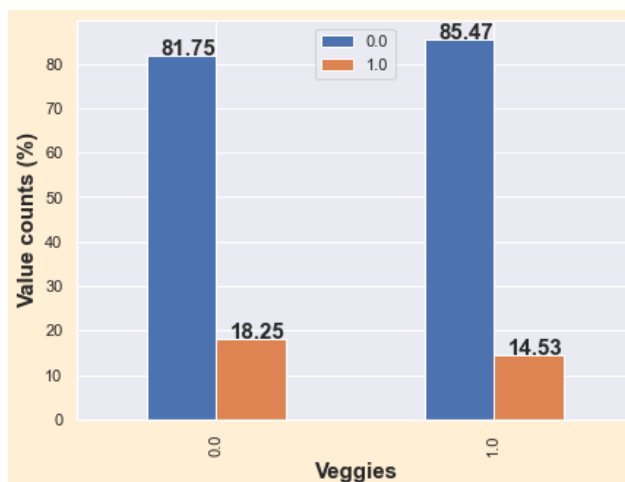
People who are not Physically active are more feasible for the Diabetes that the people who are Physically active

Fruits Consumption vs Diabetes Risk



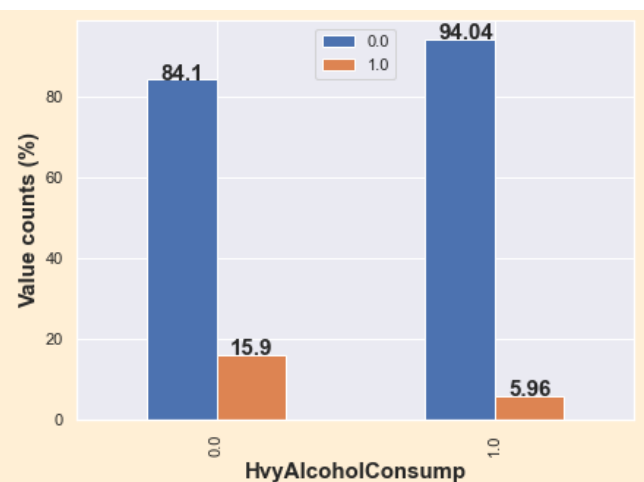
Deviation in risk factor is not much high for an individual having fruits in or not in the diet

Veggies Intake vs Diabetes Risk

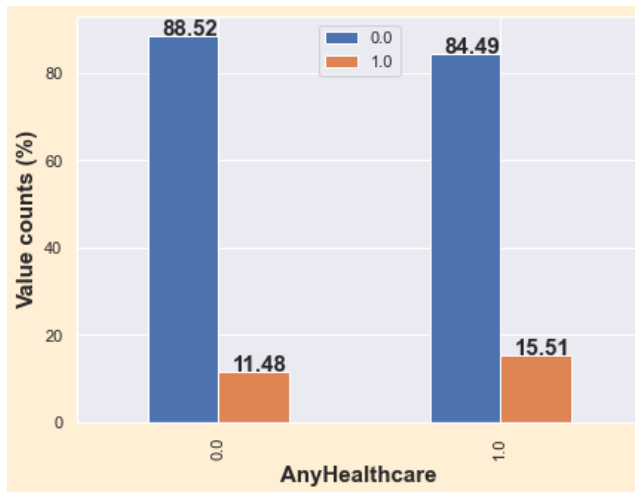


Diabetes has less attentiveness to the individuals who are consuming vegetables than the non consumers of vegetables

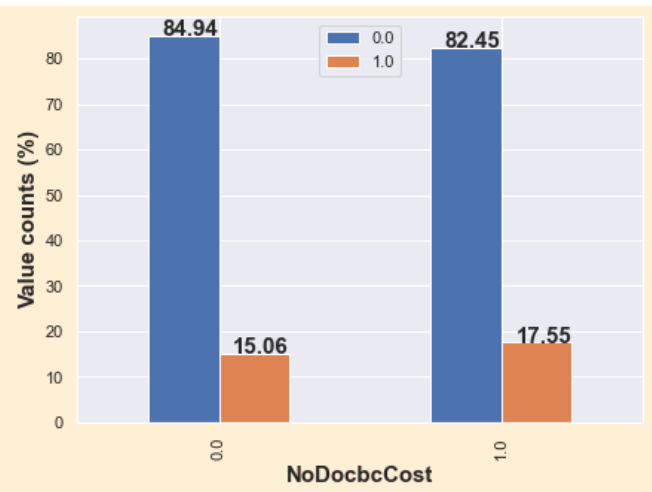
Heavy Alcohol vs Diabetes Risk



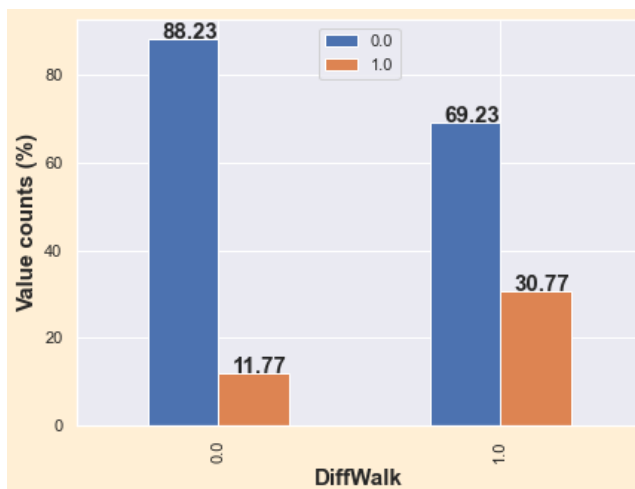
Data shows the Heavy Alcohol Drinker having less affection towards the Diabetes than remainings

Any Healthcare vs Diabetes Risk

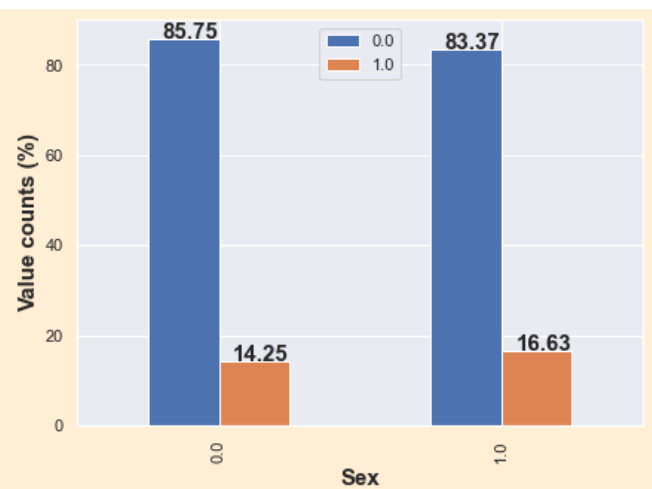
Data represents individuals who are having any Healthcare persons in the Family are less vulnerable to the diabetes in contrast to an individual not having the same

NoDocbcCost vs Diabetes Risk

No Doctor visit due to Medical Cost discovers the potential risk of having an Diabetes

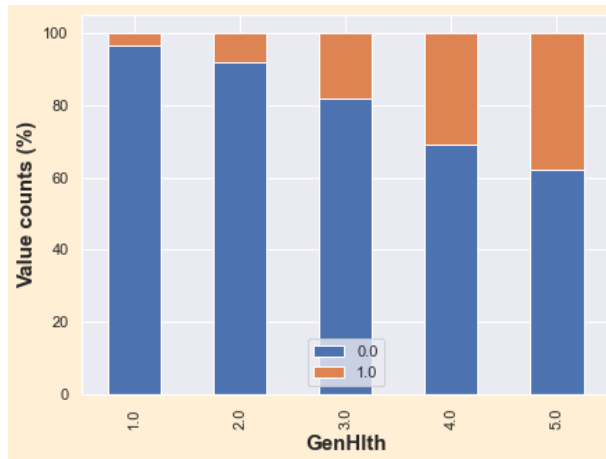
DiffWalk vs Diabetes Risk

Data speaks personnel having serious difficulty in walking or climbing stairs are more prone to have Diabetes compare to the rest

Gender vs Diabetes Risk

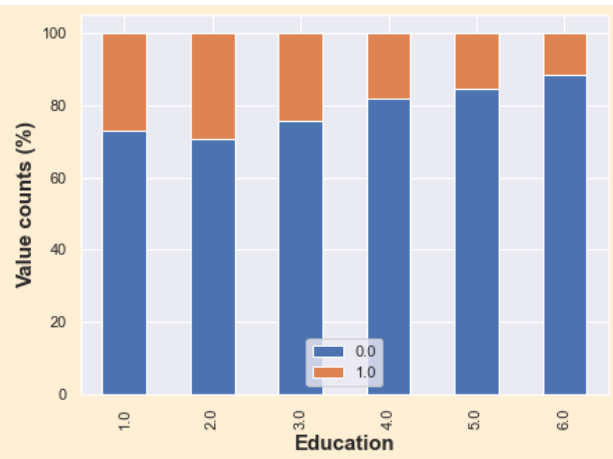
Diabetes follows feminism & not discriminant towards any gender significantly.

General Health vs Diabetes Risk



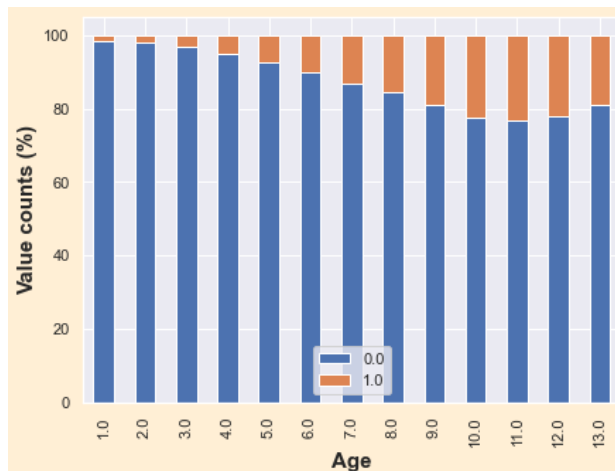
Data speaks personnel having poor general health are more prone to diabetes than those who have good general health.

Education vs Diabetes Risk



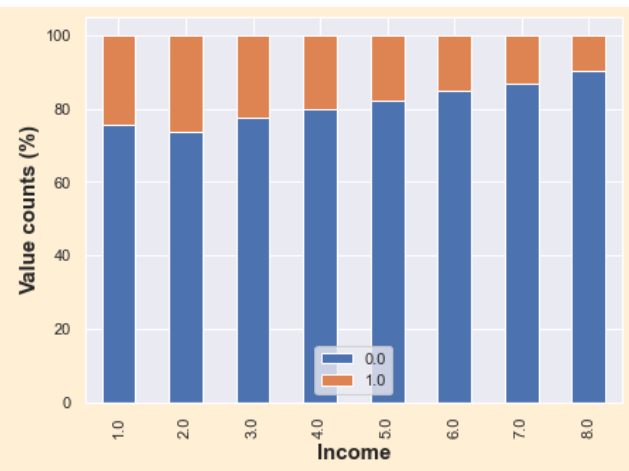
Those who are well educated have awareness about diabetes and are less prone to diabetes.

Age vs Diabetes Risk



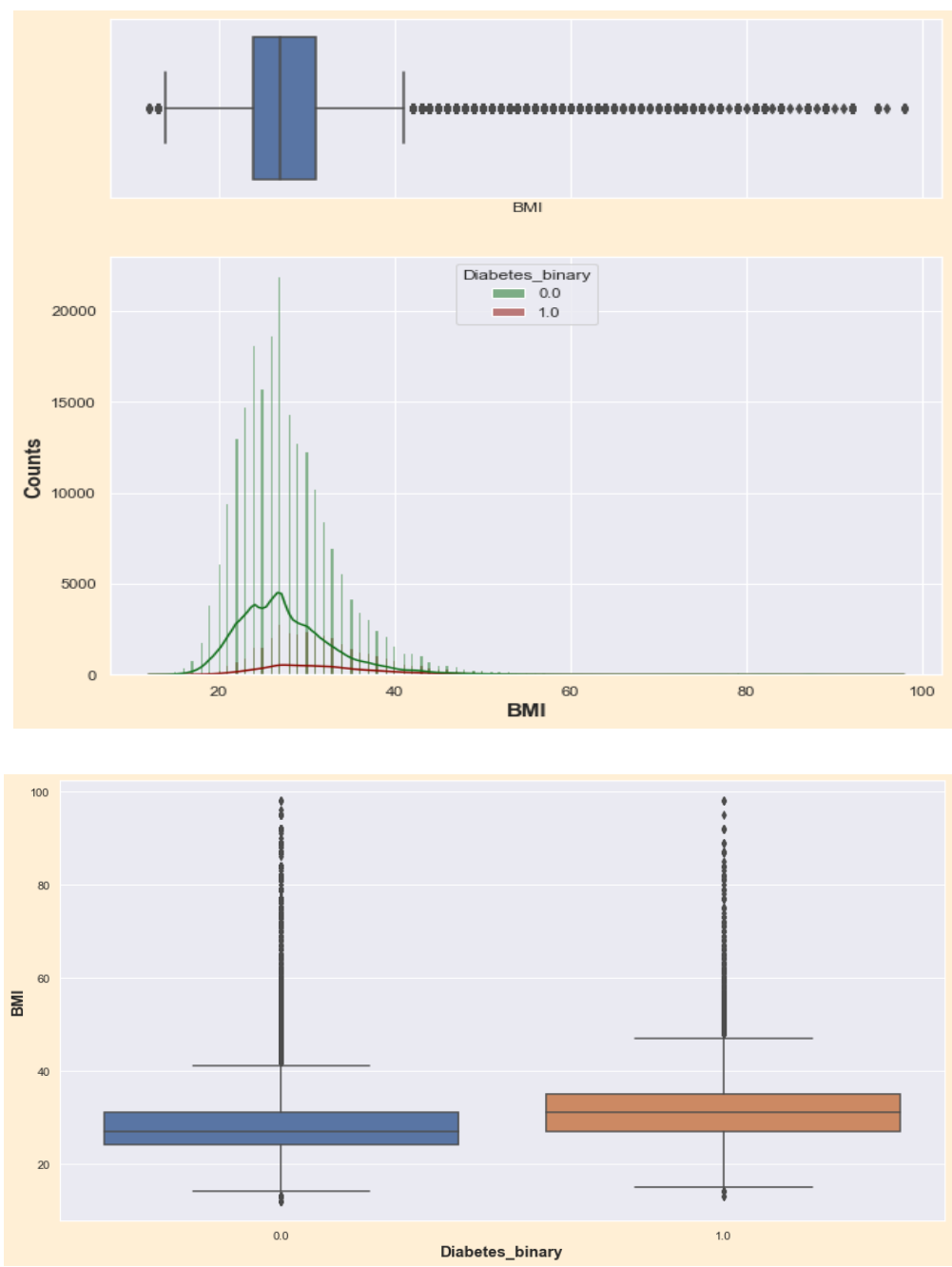
Population below the age of 40 years is less susceptible towards diabetes compared to the population above 40 years

Income vs Diabetes Risk



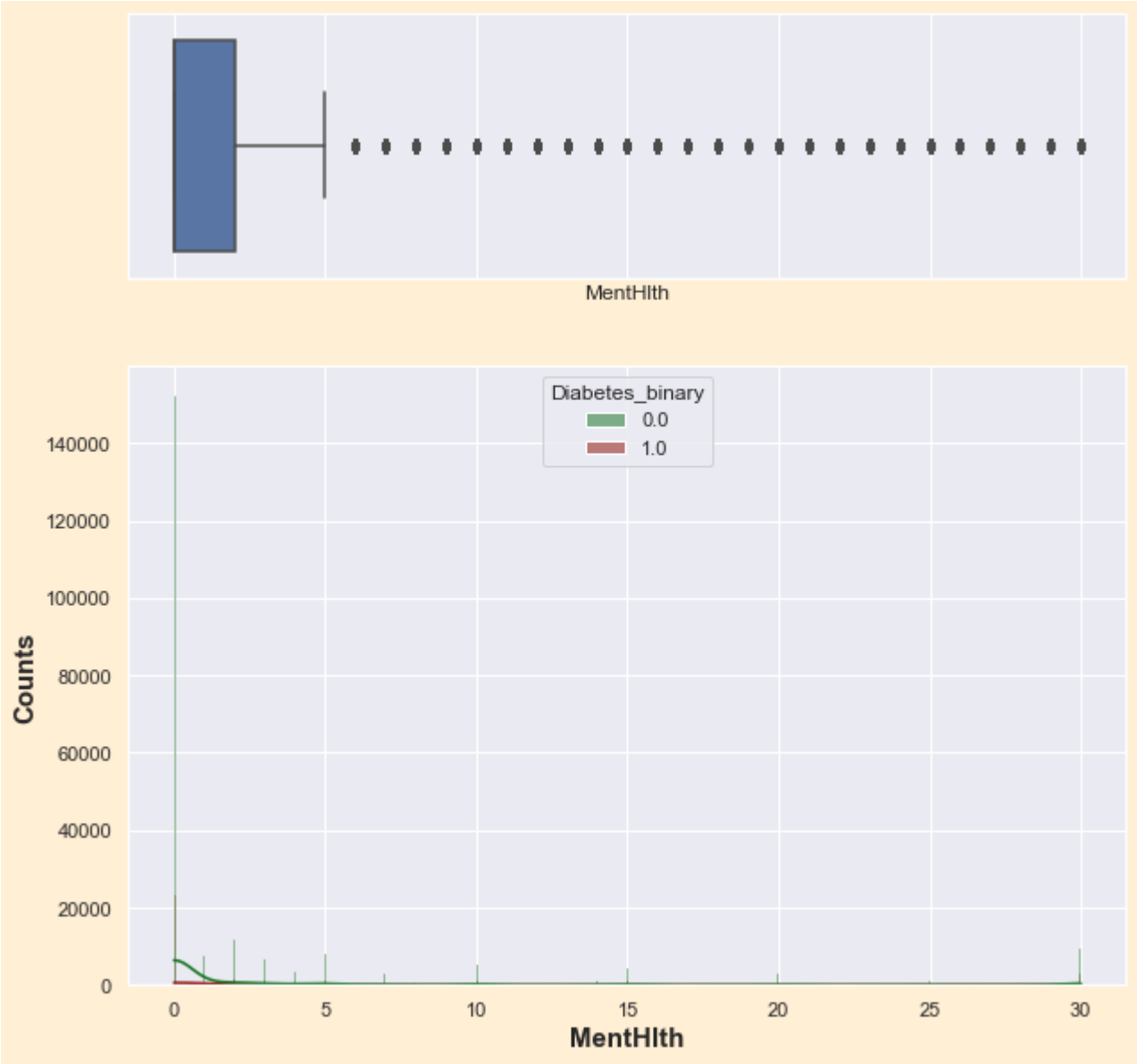
Group of Individuals which belongs to Low income group are more sensitive to have diabetes compared to Higher income groups

Distribution of BMI Feature



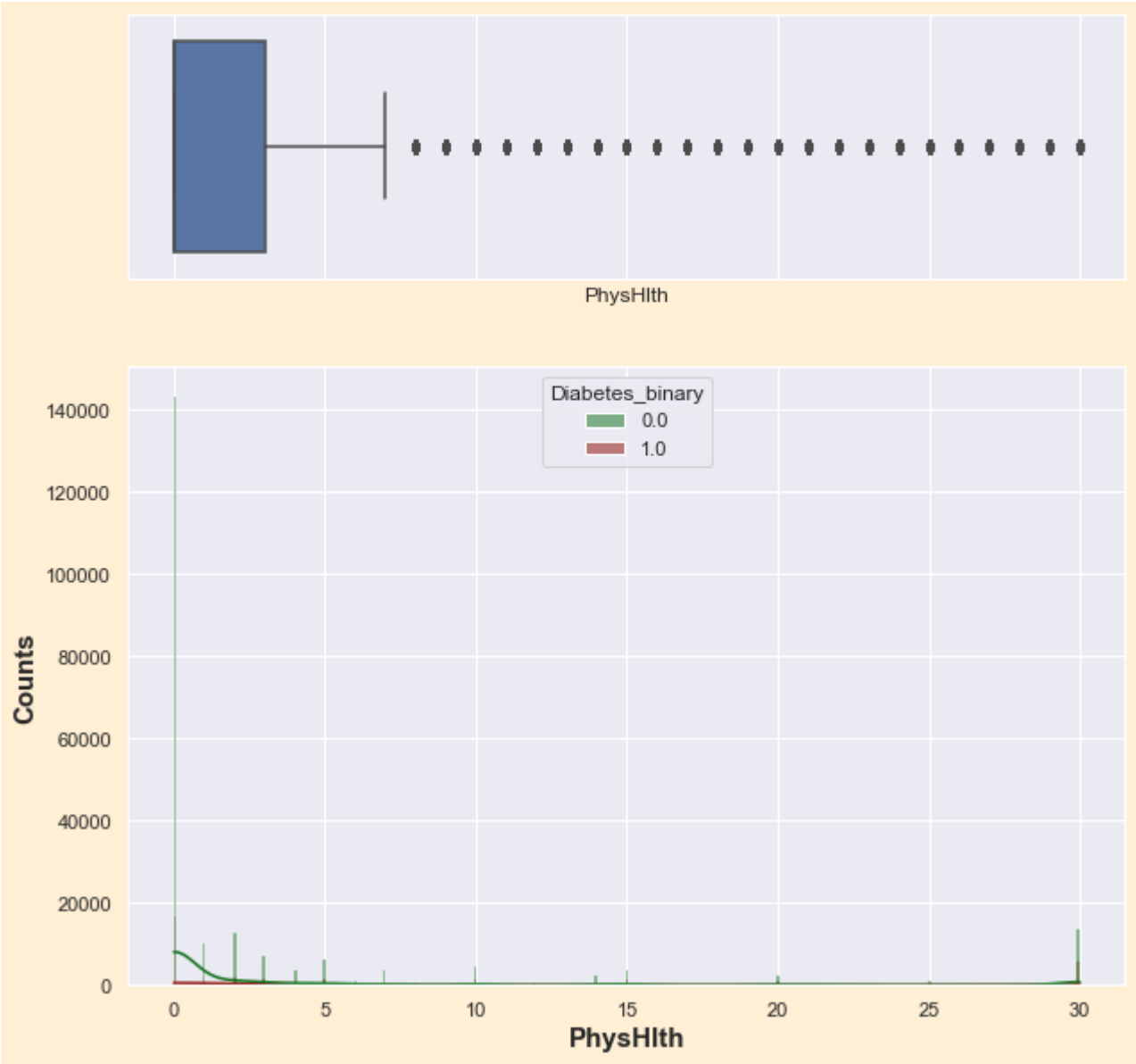
The general population has BMI value in the range of 25 - 30 and people with Diabetes on Average has higher BMI (or) obese in BMI scale

Distribution of Mental Health Feature

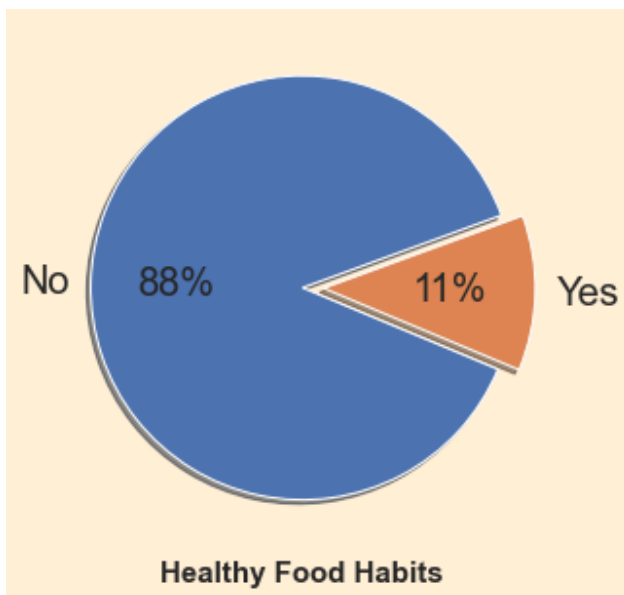


On an Average Most of the population is suffer 3 days of stress, depression, and emotional problems

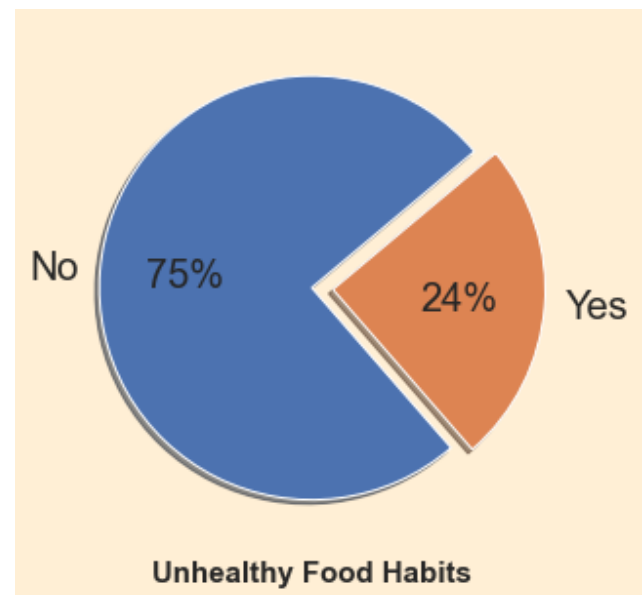
Distribution of Physical Health Feature



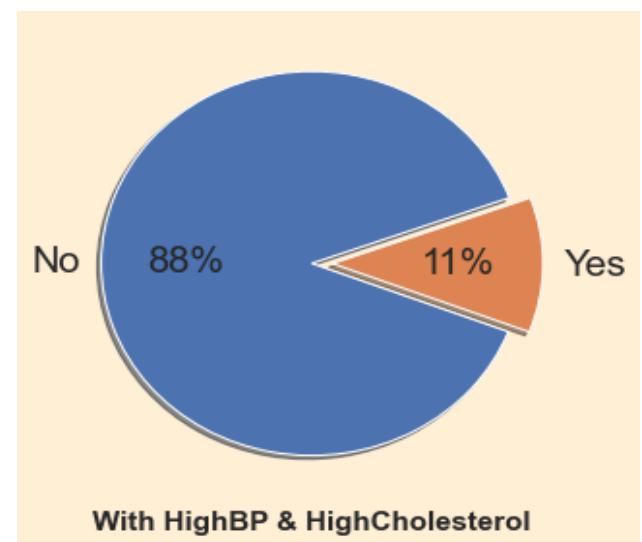
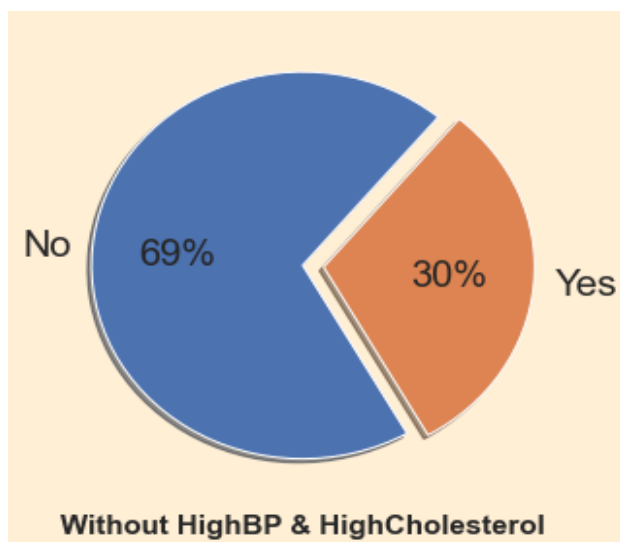
As we can see from the plot 75% of the population is Physically ill for 5 days

Healthy Habits vs Diabetes Risk

Population having healthy food habits such as consuming Fruits, Veggies and having Physical Activities are Less endangered to have diabetes

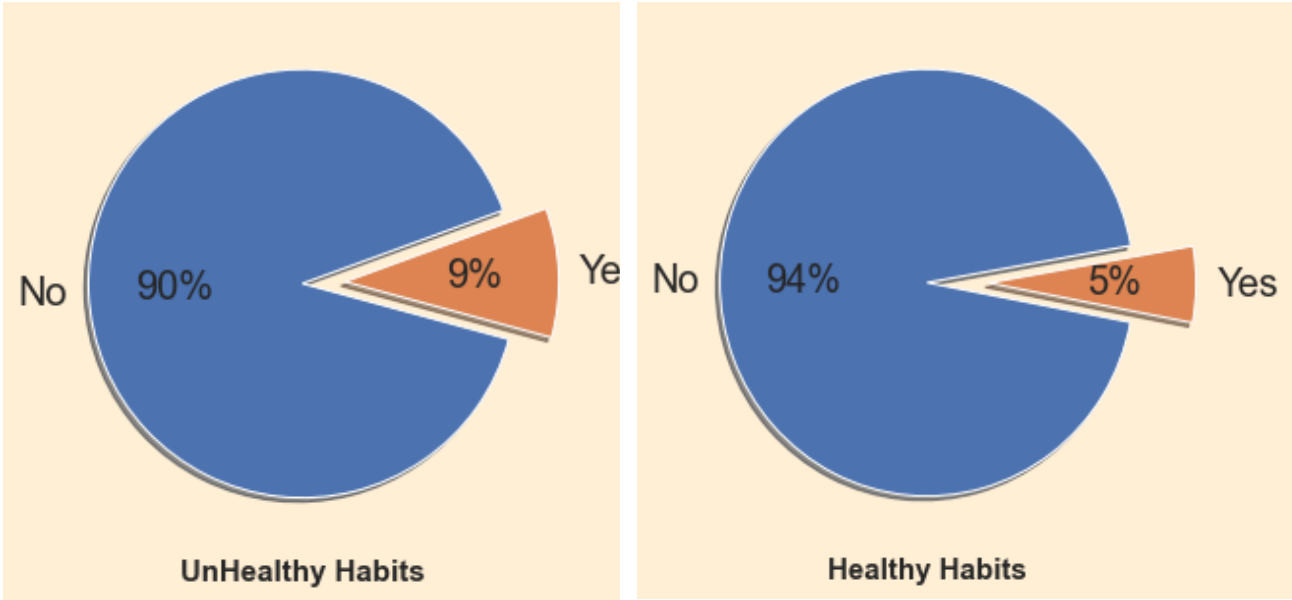
Unhealthy Habits vs Diabetes Risk

Individuals which are not consuming vegetables, fruits and are not much Physically active are having high risk to have diabetes (24%)

Relation Between HighBP, High Cholesterol & Diabetes Risk

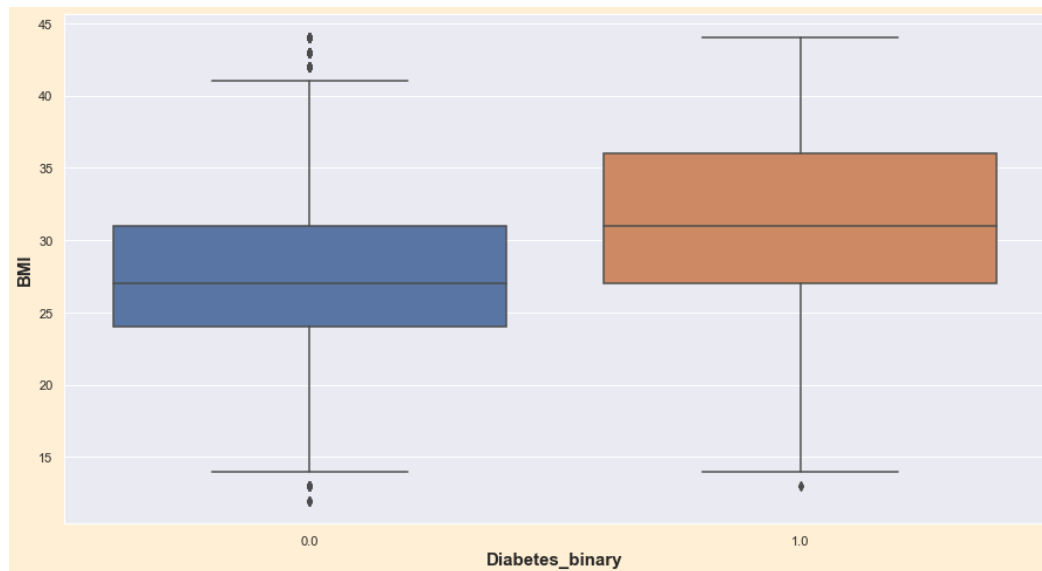
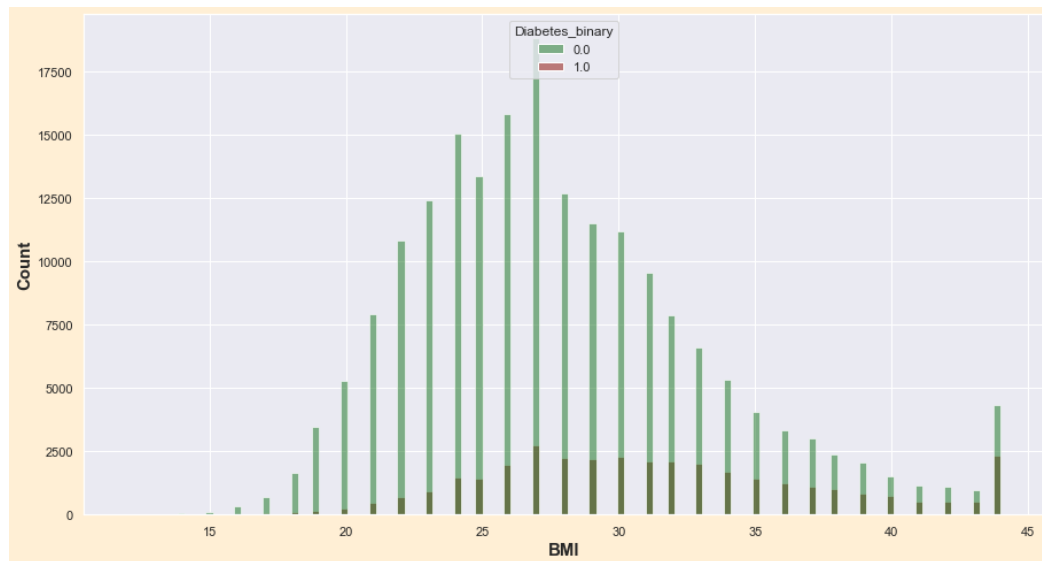
Individuals with High Blood Pressure and High Cholesterol are more vulnerable to diabetes than the general population.

Relation between Individual Habits & Diabetes Risk



Data resembles persons who are Heavy Alcohol consumers, Smokers and who are not Physically active are unshielded towards diabetes compare to the individuals whose alcohol consumption & Smoking is less & being physically active

Presence of outliers and its treatment



Overweight – BMI greater than or equal to 25 to 29.9 kg/m²

Obesity – BMI greater than or equal to 30 kg/m²

Obesity class I – BMI 30 to 34.9 kg/m²

Obesity class II – BMI 35 to 39.9 kg/m²

Obesity class III – BMI greater than or equal to 40 kg/m² (also referred to as severe, extreme, or massive obesity)

As the highest recorded BMI in history is 104 we come to the conclusion of outliers are due to manual error we choose to cap it

As an Individual begins to overweight (BMI greater than or equal to 25 to 29.9 kg/m²) the chances of getting an Diabetes increases

Scaling

```
1 dia.reset_index(drop=True,inplace=True)
2 Y = dia[['Diabetes_binary']]
3 X = dia.drop('Diabetes_binary',axis=1)
4 xtrain,xtest,ytrain,ytest = train_test_split(X,Y,test_size=0.3,random_state=10,stratify=Y,shuffle=True)
```

```
1 xtr_num = xtrain[['MentHlth','PhysHlth','BMI']].copy()
2 xtr_ob = xtrain.drop(['PhysHlth','MentHlth','BMI'],axis=1)
3 xts_num = xtest[['MentHlth','PhysHlth','BMI']].copy()
4 xts_ob = xtest.drop(['PhysHlth','MentHlth','BMI'],axis=1)
```

```
1 ssl=StandardScaler()
2 xtr_num = DF(ssl.fit_transform(xtr_num),columns=xtr_num.columns,index=xtrain.index)
```

```
1 xtrain_s = pd.concat([xtr_ob,xtr_num],axis=1)
2 print(xtrain.shape)
3 xtrain_s
```

	HighBP	HighChol	CholCheck	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocvisits
198405	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	
84698	1.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	
90877	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	
69519	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	
94634	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	
...
41660	1.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	
184288	0.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	
71464	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	
51551	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	
188656	1.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	

The numerical variables are scaled using Standard Scaler after Train Test Split to avoid data leakage.

5. Statistical Tests

Following Statistical tests has been used for

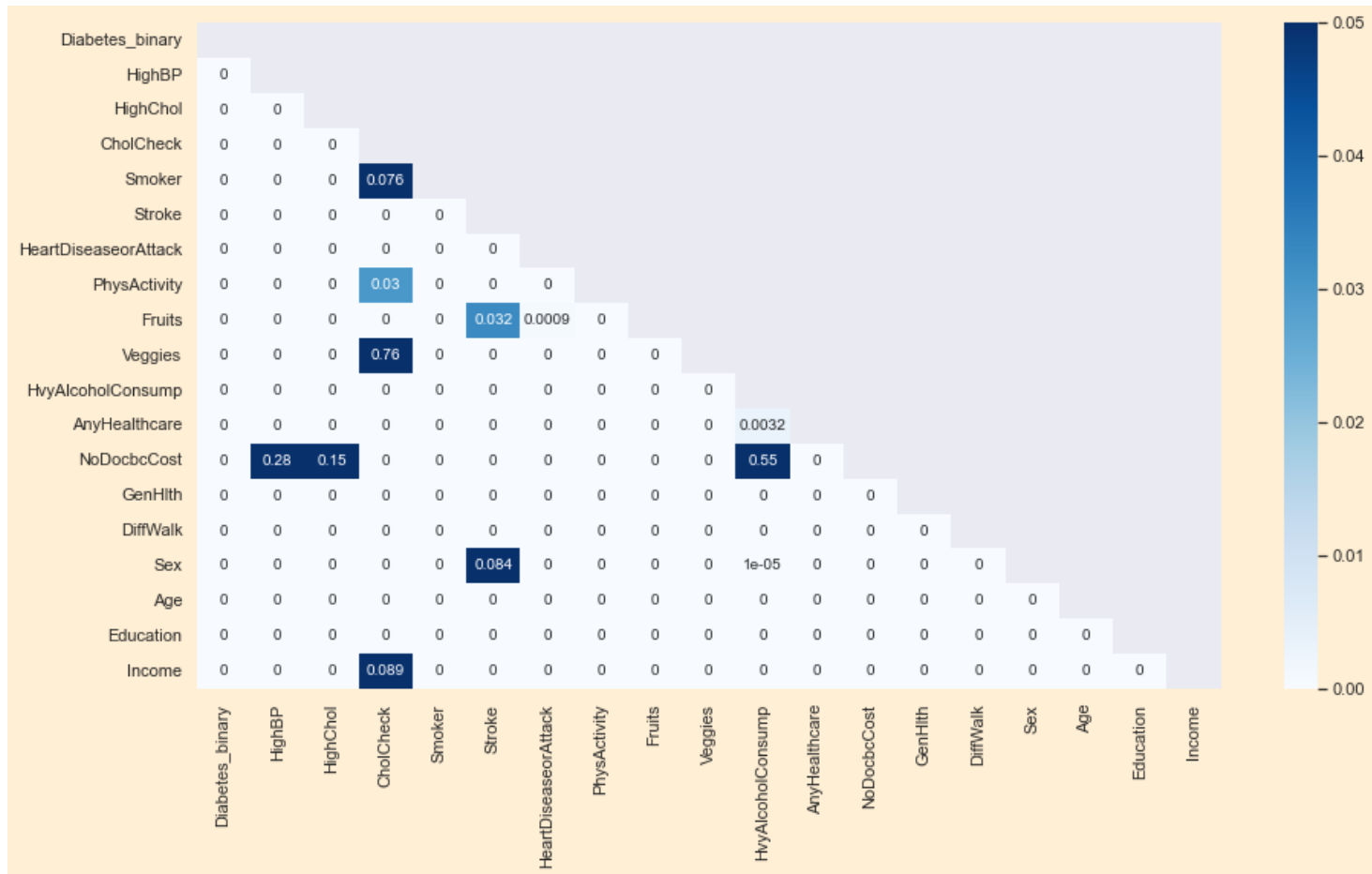
1. Anova Test for Numerical Variables
2. Chi Square Contingency for Categorical Variables
3. Cramer's V Score for Categorical Variables

Anova Table for Numerical Values:

	sum_sq	df	F	PR(>F)
HighBP	1.713796e+05	1.0	5944.544061	0.0000
HighChol	4.407740e+03	1.0	152.888682	0.0000
CholCheck	3.791613e+03	1.0	131.517449	0.0000
Smoker	1.849631e+04	1.0	641.570721	0.0000
Stroke	6.455626e+03	1.0	223.922486	0.0000
HeartDiseaseorAttack	9.073301e+02	1.0	31.472023	0.0000
PhysActivity	3.103872e+04	1.0	1076.621583	0.0000
Fruits	4.320391e+03	1.0	149.858838	0.0000
Veggies	6.311447e-01	1.0	0.021892	0.8824
HvyAlcoholConsump	1.442944e+04	1.0	500.505296	0.0000
AnyHealthcare	2.906971e+02	1.0	10.083237	0.0015
NoDocbcCost	8.603444e+01	1.0	2.984226	0.0841
DiffWalk	9.306345e+04	1.0	3228.036964	0.0000
Sex	1.187156e+04	1.0	411.781932	0.0000
Age	3.185693e+05	12.0	920.835306	0.0000
Income	5.522912e+03	7.0	27.367138	0.0000
Education	1.330971e+04	5.0	92.333213	0.0000
GenHlth	1.105392e+05	4.0	958.551637	0.0000
Diabetes_binary	1.448264e+05	1.0	5023.508369	0.0000
Residual	6.614406e+06	229430.0	NaN	NaN

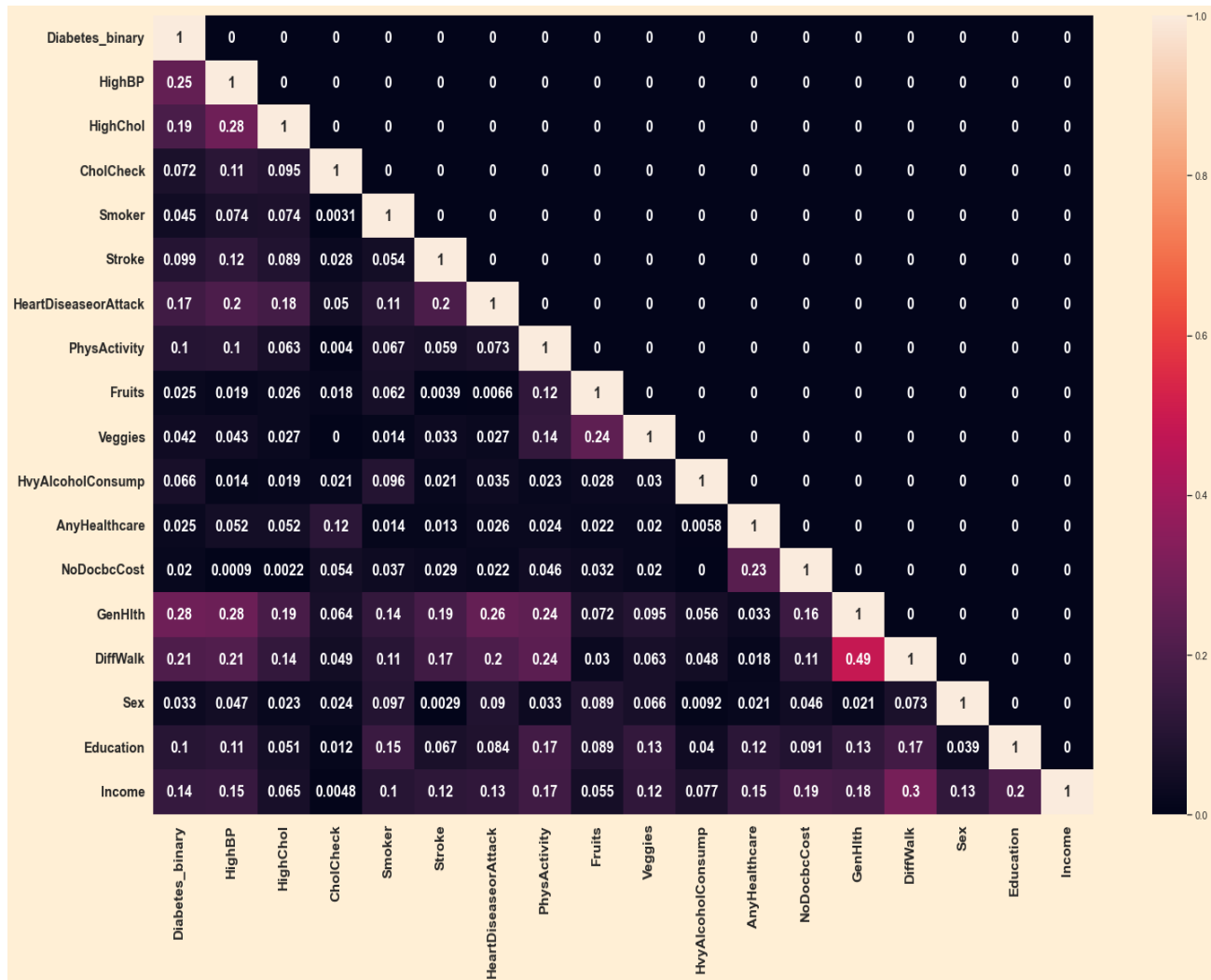
Where all p_value of less than 0.05 so by rejecting null hypothesis we can say that mean of each sub groups of all categorical features are different from each other

Chi Square P-Value test plot for Correlation between categorical columns



chi2_contingency tells just two variables are dependent are not but doesn't tell us about the strength of association so we can use Carmer's V

Cramer's V score of Categorical Features



As we can see from the above plot, there is moderate Positive Association between Difficulty to Walk and General Health.

Target Variable has high Association with GenHlth, Blood Pressure, High Cholesterol.

6. MODEL BUILDING

LOGISTIC REGRESSION

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

Logit Regression Results						
Dep. Variable:	Diabetes_binary	No. Observations:	160631			
Model:	Logit	Df Residuals:	160609			
Method:	MLE	Df Model:	21			
Date:	Wed, 12 Oct 2022	Pseudo R-squ.:	0.1999			
Time:	12:59:49	Log-Likelihood:	-54978.			
converged:	True	LL-Null:	-68716.			
Covariance Type:	nonrobust	LLR p-value:	0.000			
	coef	std err	z	P> z	[0.025	0.975]
const	-6.1453	0.104	-59.015	0.000	-6.349	-5.941
HighBP	0.6856	0.018	38.623	0.000	0.651	0.720
HighChol	0.5663	0.016	34.620	0.000	0.534	0.598
CholCheck	1.2190	0.080	15.162	0.000	1.061	1.377
Smoker	-0.0088	0.016	-0.556	0.578	-0.040	0.022
Stroke	0.1663	0.030	5.539	0.000	0.107	0.225
HeartDiseaseorAttack	0.2357	0.021	11.052	0.000	0.194	0.277
PhysActivity	-0.0098	0.017	-0.570	0.569	-0.044	0.024
Fruits	-0.0300	0.016	-1.825	0.068	-0.062	0.002
Veggies	-0.0152	0.019	-0.798	0.425	-0.052	0.022
HvyAlcoholConsump	-0.7622	0.046	-16.672	0.000	-0.852	-0.673
AnyHealthcare	0.0710	0.039	1.800	0.072	-0.006	0.148
NoDocbcCost	0.0139	0.027	0.510	0.610	-0.040	0.067
GenHlth	0.4973	0.010	50.778	0.000	0.478	0.516
DiffWalk	0.1108	0.020	5.470	0.000	0.071	0.150
Sex	0.2626	0.016	16.166	0.000	0.231	0.294
Age	0.1330	0.003	39.337	0.000	0.126	0.140
Education	-0.0167	0.008	-2.011	0.044	-0.033	-0.000
Income	-0.0489	0.004	-11.449	0.000	-0.057	-0.041
MentHlth	-0.0188	0.008	-2.405	0.016	-0.034	-0.003
PhysHlth	-0.0685	0.008	-8.094	0.000	-0.085	-0.052
BMI	0.4878	0.008	62.133	0.000	0.472	0.503

Here we have built a base Logistic regression model on the entire dataset after doing all the proper encoding and transformation processes by using the Statsmodels Logit method.

From the model summary we can see that the Pseudo R^2 value is fairly low but the LLR- PValue is very low at 0.00 which means that further analysis needs to be done before using this data set for target variable prediction.

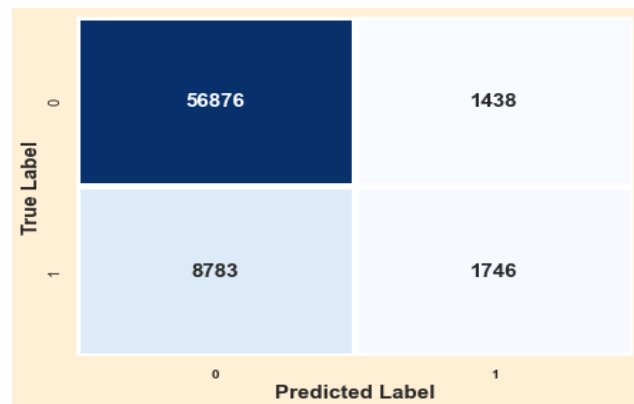
Log Odds Coefficient

	LogOdds
HighBP	1.985037
HighChol	1.761789
GenHlth	1.644275
BMI	1.628803
Sex	1.300358
HeartDiseaseorAttack	1.265767
Stroke	1.180918
Age	1.142275
DiffWalk	1.117117
AnyHealthcare	1.073607
NoDocbcCost	1.014027
Smoker	0.991199
PhysActivity	0.990221
Veggies	0.984945
Education	0.983407
MentHlth	0.981353
Fruits	0.970464
Income	0.952295
PhysHlth	0.933786
HvyAlcoholConsump	0.466641
const	0.002144

With the summary and odds ratio of the Logistic Regress Base Model we can see that the HighBP variable is highly contributing to the target variable, i.e if You have BP it increases the Log of odds of having Diabetes by 1.98 times.

Followed by HighChol showing Log of odds of having Diabetes is 1.76 times high if an Individual have high cholesterol

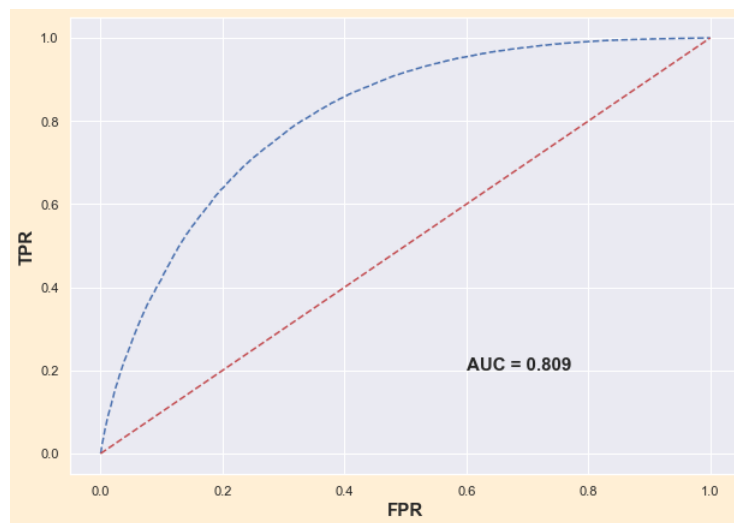
Confusion Matrix



Classification Report

	precision	recall	f1-score	support
0.0	0.88	0.83	0.86	58314
1.0	0.29	0.39	0.34	10529
accuracy			0.76	68843
macro avg	0.59	0.61	0.60	68843
weighted avg	0.79	0.76	0.78	68843

- **ROC Curve**



As we can see the Area Under Curve is 81% for the Base Model, which needs to be increased as 81% AUC is not enough in Medical Domain.

Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules. Below diagram illustrate the basic flow of decision tree for decision making with labels (Rain(Yes), No Rain(No))

Decision tree is one of the predictive modeling approaches used in statistics, data mining and machine learning.

Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks.

Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification And Regression Tree (CART) is the general term for this.

- **Confusion Matrix**

True Label	0	1
	56708	1606
1	8840	1689
Predicted Label		

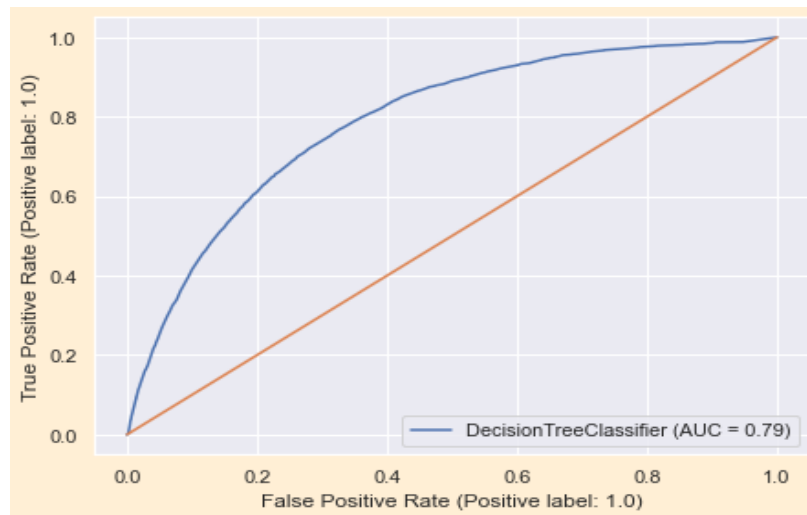
Percentage of ***correctly classified*** observation :
69.59%

Percentage of ***misclassified*** observation :
30.40%

- **Classification Report**

	precision	recall	f1-score	support
0.0	0.94	0.68	0.79	58314
1.0	0.30	0.76	0.43	10529
accuracy			0.70	68843
macro avg	0.62	0.72	0.61	68843
weighted avg	0.84	0.70	0.74	68843

- **ROC Curve**



For the Decision Tree Base Model, Area Under Curve is 79% for the Base Model, which is less compared to the Logistic Regression base model.

Also the Precision Score (0.69), F1-Score (0.58) & Kappa Score (0.19) for the Decision Tree model is less compared to the Logistic Regression base model

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is that it can be used for both classification and regression problems.

Random Forest has nearly the same hyper parameters as a decision tree or a bagging classifier.

• Classification Report :

	precision	recall	f1-score	support
0.0	0.87	0.96	0.91	58314
1.0	0.47	0.17	0.25	10529
accuracy			0.84	68843
macro avg	0.67	0.57	0.58	68843
weighted avg	0.81	0.84	0.81	68843

• Confusion Matrix :

True Label	0	56267	2047
	1	8702	1827
		0	1
		Predicted Label	

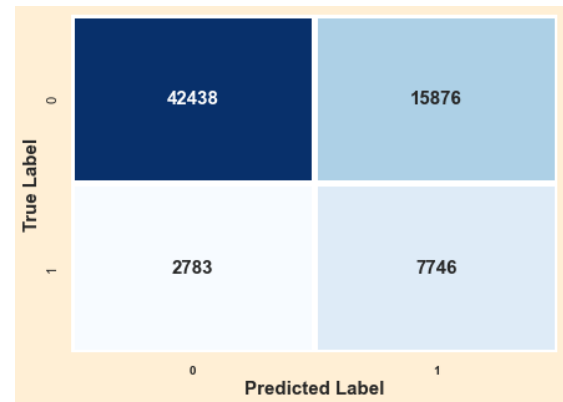
Random Forest (after tuning)

The tuned parameters that were used in Random Forest are *n_estimators* as **250**, *min_samples_split* as **5**, *min_samples_leaf* as **9**, *max_depth* as **5**

• Classification Report :

	precision	recall	f1-score	support
0.0	0.94	0.73	0.82	58314
1.0	0.33	0.74	0.45	10529
accuracy			0.73	68843
macro avg	0.63	0.73	0.64	68843
weighted avg	0.85	0.73	0.76	68843

• Confusion Matrix :



After model tuning, recall of the Random Forest Model has been increased to 0.73 from 0.57 as we are using the Weighted_loss Function along with the GridSearchCV.

Also the model evaluation metrics Kappa Score (0.19 → 0.31) & F1 Score (0.58 → 0.64) increased significantly after model tuning.

Naive Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

The fundamental Naive Bayes assumption is that each feature makes an independent equal contribution to the outcome.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

Bernoulli Naive Bayes

BernoulliNB implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable. Therefore, this class requires samples to be represented as binary-valued feature vectors

The decision rule for Bernoulli naive Bayes is based on

$$P(x_i | y) = P(x_i = 1 | y)x_i + (1 - P(x_i = 1 | y))(1 - x_i)$$

• Classification Report :

	precision	recall	f1-score	support
0.0	0.89	0.91	0.90	58314
1.0	0.40	0.35	0.37	10529
accuracy			0.82	68843
macro avg	0.64	0.63	0.63	68843
weighted avg	0.81	0.82	0.82	68843

• Confusion Matrix :

True Label	0	52830	5484
	1	6863	3666
		0	1
		Predicted Label	

Recall value by the BernoulliNB model is 0.63 & the rest metrics evaluation parameters such as F1 Score = 0.63 and Kappa Score is 0.27

Sampling of Data:

Imbalanced classification involves developing predictive models on classification datasets that have a severe class imbalance.

The challenge of working with imbalanced datasets is that most machine learning techniques will ignore, and in turn have poor performance on, the minority class, although typically it is performance on the minority class that is most important.

One approach to addressing imbalanced datasets is to oversample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples don't add any new information to the model. Instead, new examples can be synthesized from the existing examples.

For data made of only categorical data, one can use the **SMOTEN**. The algorithm changes in two ways:

- The nearest neighbors search does not rely on the Euclidean distance. Indeed, the value difference metric (VDM) also implemented in the class **ValueDifferenceMetric** is used.
- A new sample is generated where each feature value corresponds to the most common category seen in the neighbors samples belonging to the same class.

Model built on SMOTE Data:

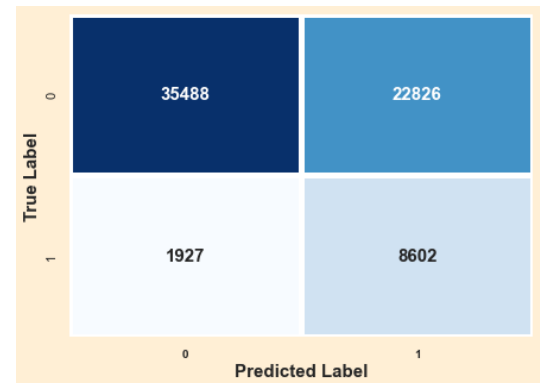
1. Random Forest Model using SMOTE Data
2. XGBoost Model using SMOTE Data
3. Stacked Model using SMOTE Data

Random Forest on SmoteData

• Classification Report :

	precision	recall	f1-score	support
0.0	0.95	0.61	0.74	58314
1.0	0.27	0.82	0.41	10529
accuracy			0.64	68843
macro avg	0.61	0.71	0.58	68843
weighted avg	0.85	0.64	0.69	68843

• Confusion Matrix :

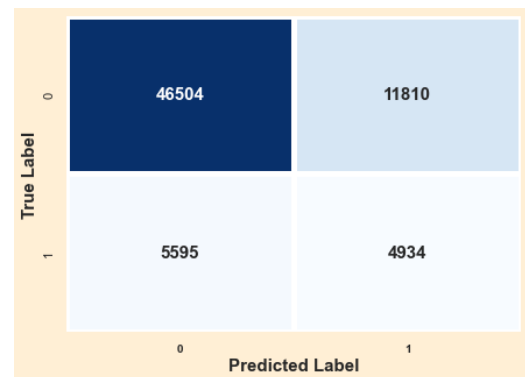


XGBoost on SMOTE Data

• Classification Report :

	precision	recall	f1-score	support
0.0	0.89	0.80	0.84	58314
1.0	0.29	0.47	0.36	10529
accuracy			0.75	68843
macro avg	0.59	0.63	0.60	68843
weighted avg	0.80	0.75	0.77	68843

• Confusion Matrix :



The recall by Random Forest on SMOTE data (0.71) is more compared to the XGBoost on SMOTE data (0.63). Also the F1 Score for the models are 0.58 and 0.6 respectively whereas the Kappa Score is 0.23 and 0.21 respectively.

Boosting Models

Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. First, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

AdaBoostclassifier

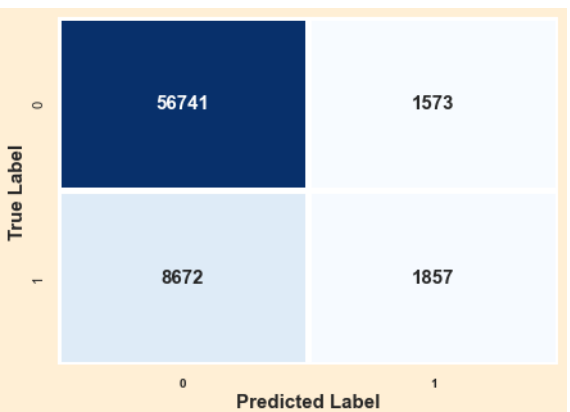
AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps.

This algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a low error is received.

• Classification Report :

	precision	recall	f1-score	support
0.0	0.87	0.97	0.92	58314
1.0	0.54	0.18	0.27	10529
accuracy			0.85	68843
macro avg	0.70	0.57	0.59	68843
weighted avg	0.82	0.85	0.82	68843

• Confusion Matrix :



True Label	0	1
	56741	1573
1	8672	1857
	0	1
Predicted Label		

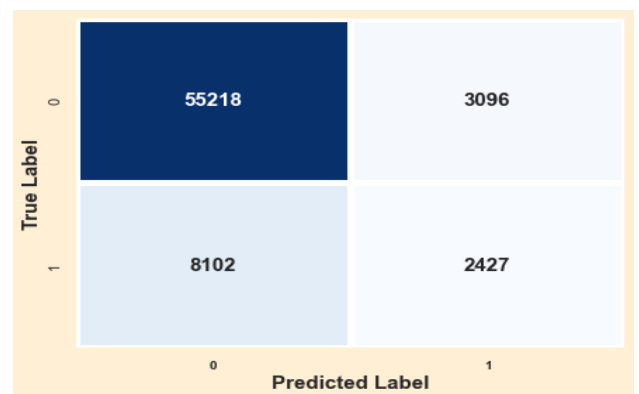
XGBoost

In XGBoost, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

• Classification Report :

	precision	recall	f1-score	support
0.0	0.87	0.95	0.91	58314
1.0	0.44	0.23	0.30	10529
accuracy			0.84	68843
macro avg	0.66	0.59	0.61	68843
weighted avg	0.81	0.84	0.82	68843

• Confusion Matrix :



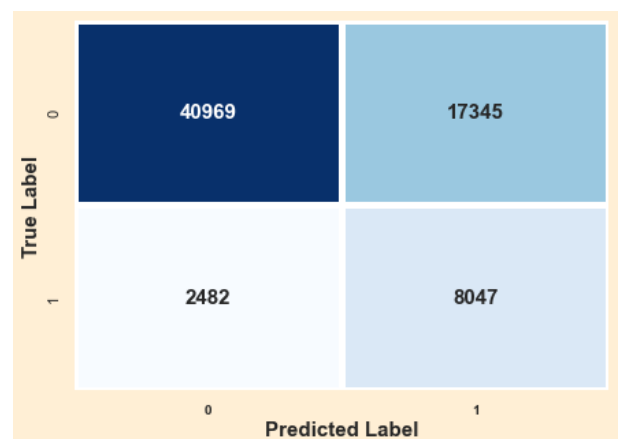
XGBoost (after tuning)

The tuned parameters that were used in XG Boost are *subsample* as **0.7**, *n_estimators* as **150**, *max_depth* as **5**, *max_delta_step* as **2**, *learning_rate* as **0.2**

Classification Report :

	precision	recall	f1-score	support
0.0	0.94	0.70	0.81	58314
1.0	0.32	0.76	0.45	10529
accuracy			0.71	68843
macro avg	0.63	0.73	0.63	68843
weighted avg	0.85	0.71	0.75	68843

Confusion Matrix :



Comparing all Boosting Models (Adaboost, XGBoost & Tuned XGBoost) we find that Tuned XGBoost is giving the Maximum Recall score as 0.73

F1 Score observed for AdaBoost (0.59), XGBoost (0.60) & Tuned XGBoost is (0.63), where as the Kappa Score is AdaBoost (0.21), XGBoost (0.22) & Tuned XGBoost is (0.29)

Stacked Model

Stacked Generalization is an ensemble machine learning algorithm. It involves combining the predictions from multiple machine learning models on the same dataset, like bagging and boosting.

The architecture of a stacking model involves two or more base models, often referred to as level-0 models, and a meta-model that combines the predictions of the base models, referred to as a level-1 model.

- **Level-0 Models (*Base-Models*):** Models fit on the training data and whose predictions are compiled.
- **Level-1 Model (*Meta-Model*):** Model that learns how to best combine the predictions of the base models.

The meta-model is trained on the predictions made by base models on out-of-sample data. That is, data not used to train the base models is fed to the base models, predictions are made, and these predictions, along with the expected outputs, provide the input and output pairs of the training dataset used to fit the meta-model. The outputs from the base models used as input to the meta-model may be real value in the case of regression, and probability values, probability like values, or class labels in the case of classification

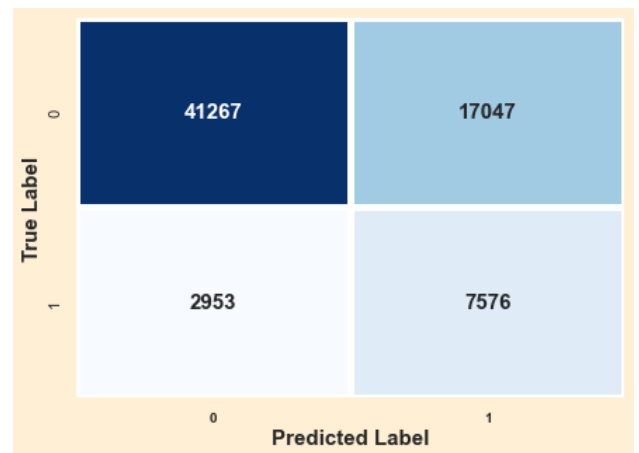
Stacked Model

- Base Estimators : DecisionTreeClassifier, RandomForestClassifier, BernoulliNB, XGBClassifier
- Final Estimators : LogisticRegression

Classification Report :

	precision	recall	f1-score	support
0.0	0.94	0.70	0.80	58314
1.0	0.31	0.76	0.44	10529
accuracy			0.71	68843
macro avg	0.63	0.73	0.62	68843
weighted avg	0.84	0.71	0.75	68843

Confusion Matrix :



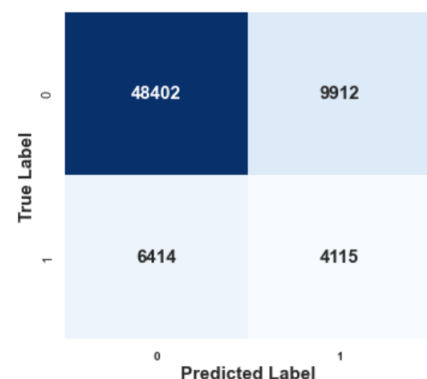
Stacked Model on SMOTE Data:

- Base Estimators : DecisionTreeClassifier, RandomForestClassifier, BernoulliNB, XGBClassifier
- Final Estimators : LogisticRegression

Classification Report :

	precision	recall	f1-score	support
0.0	0.88	0.83	0.86	58314
1.0	0.29	0.39	0.34	10529
accuracy			0.76	68843
macro avg	0.59	0.61	0.60	68843
weighted avg	0.79	0.76	0.78	68843

Confusion Matrix :

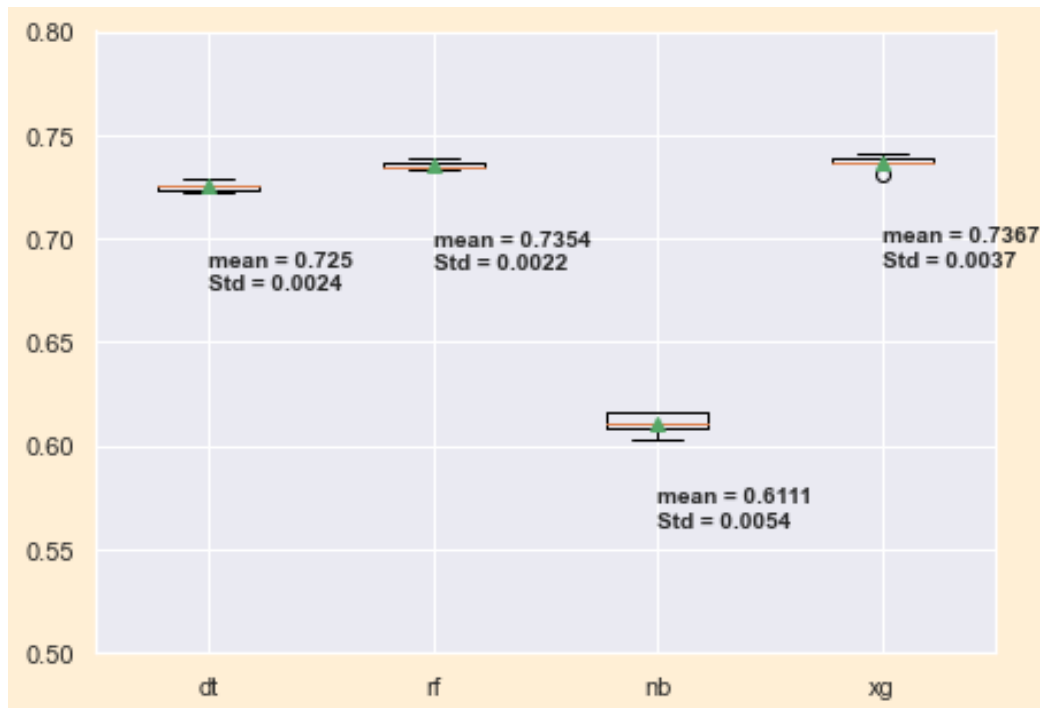


7. ML Model Result Comparison:

	Model	Recall	Precision	f1-score	Accuracy	cohen_kappa_score
0	XGBClassifier(tunned)	0.73	0.63	0.63	0.71	0.29
1	RandomForestClassifier(tunned)	0.73	0.63	0.64	0.73	0.31
2	DecisionTreeClassifier(tunned)	0.72	0.62	0.61	0.70	0.28
3	StackedClassifier	0.73	0.63	0.62	0.71	0.29
4	StackedClassifier(Smote)	0.61	0.59	0.60	0.76	0.19
5	XGBClassifier(Smote)	0.63	0.59	0.60	0.75	0.21
6	RandomForestClassifier(Smote)	0.71	0.61	0.58	0.64	0.23
7	XGBClassifier	0.59	0.66	0.60	0.84	0.22
8	RandomForestClassifier	0.57	0.67	0.58	0.84	0.19
9	AdaBoostClassifier	0.57	0.70	0.59	0.85	0.21
10	BernoulliNB	0.63	0.64	0.63	0.82	0.27
11	DecisionTreeClassifier(Base)	0.57	0.69	0.58	0.85	0.19
12	LogisticRegression(Base)	0.57	0.71	0.59	0.85	0.20

From the above we can find that XGBoost (Tunned) and Random Forest (using randomizer search cv) perform well. Where final performance of ML model on entire data can be used as a deciding parameter in choosing the final model.

ML Model Result Comparison on Entire dataset

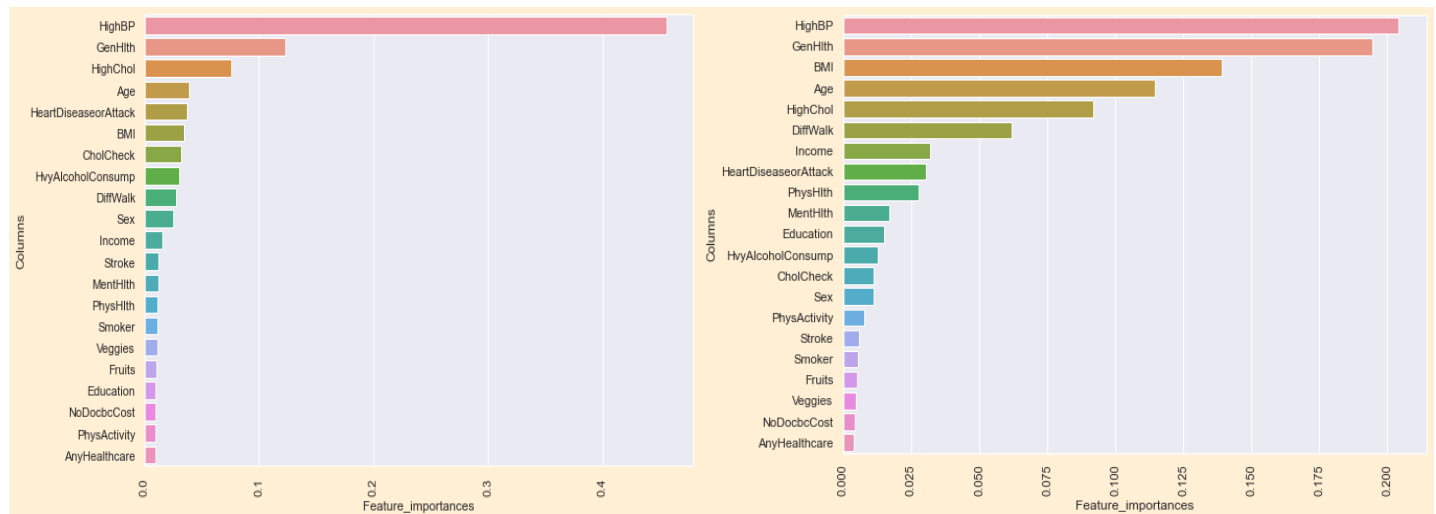


Conclusion for the model built

Out of the built model, Random Forest with hyperparameter tuning and XGB Classifier(tuned) performed well for the train and test scores with less variance in scores. The final model trained on entire data with 5 fold cross validation method XGB Classifier(tuned) has higher average macro recall with low variance.

Early detection of diabetes is one of the significant challenges in the healthcare industry. In our research, we designed a system, which can predict diabetes with high accuracy

10. SUGGESTIONS FOR BUSINESS



Suggestions to stay healthy

- Reducing stress can help in keeping the blood pressure level optimum and put one at a lower risk of diabetes
- Physical exercise, a healthy sleep cycle, good food habits makes one General health better .
- Proper balanced diet helps you control your blood sugar (glucose), manage your weight and control heart disease risk factors, such as high blood pressure and high blood fats.
- As higher bmi indicates poor health and less physical activity .
- Age is a important factor in diabetes as age increase so does the risk of being diabetic

REFERENCES

[1] [Data Scaling](#)

Collected from Analytics India Magazine. Retrieved 17th May 2022.

[2] [Feature Selection](#)

Collected from JavaTPoint. Retrieved 17th May 2022.

[3] [Dimensionality Reduction](#)

Collected from JavaTPoint. Retrieved 17th May 2022.

[4] [Assumptions of Logistic Regression](#)

Collected from Stratology.Org

[5] [Data Set Location](#)

Data set collected from Kaggle.com