**Diabetes Prediction Model by Utilizing the Health Indicator Diabetes Datasets**

COURSE NAME

Under the guidance of

Professor NAME

Project Group

Team Members:

* Praduvi

Table of Contents

[Acknowledgments 5](#_Toc181902835)

[Abstract 6](#_Toc181902836)

[Introduction 7](#_Toc181902837)

[Research Questions: 8](#_Toc181902838)

[Tools: 8](#_Toc181902839)

[Literature Review 9](#_Toc181902840)

[Research Methods 13](#_Toc181902841)

[Algorithms follow to build the models: 13](#_Toc181902842)

[ Clean the datasets for NA/NULL Values 13](#_Toc181902843)

[ Perform the statistical analysis on each health marker vs Diabetes 13](#_Toc181902844)

[ Remove the Outliers from the datasets 13](#_Toc181902845)

[ Perform the features selection (PCA/Feature Importance) 14](#_Toc181902846)

[ Build the Model, measure converge time, and performance (accuracy) 15](#_Toc181902847)

[Supervised Machine Learning Model: 15](#_Toc181902848)

[ Logistic Regression: 15](#_Toc181902849)

[ Gaussian Naïve Bayes 15](#_Toc181902850)

[ Decision Tree Classifier: 16](#_Toc181902851)

[ Random Forest Classifier: 16](#_Toc181902852)

[ XGBoost Classifier: 17](#_Toc181902853)

[Data Analysis 18](#_Toc181902854)

[Discussion 22](#_Toc181902855)

[Conclusion 26](#_Toc181902856)

[List of References 29](#_Toc181902857)

Table of Figures

[**Figure 1 : Random Forest Classifier used for Feature Selection** 14](#_Toc181485467)

[**Figure 2 : Bias Vs Variance for building ML Model** 19](file:////Users/sanjayk/Dropbox/Work/Careerera/CareereraData/ML/HealthIndicatorDataset/FinalReport.docx#_Toc181485468)

[**Figure 3: High Correlated Health Indicator Marker towards Diabetes** 23](#_Toc181485469)

[**Figure 4: Random Forest Classifier is used for Feature Selection** 24](#_Toc181485470)

[**Figure 5: Model Comparison of Accuracy Vs Time Taken** 24](#_Toc181485471)

Table of Tables

[**Table 1 : Demographics Information about Dataset** 20](#_Toc131972049)

# Acknowledgments

# 

# Abstract

Diabetes is a chronic disease that affects how the body processes glucose, which is the primary source of energy for the cells. There are two main types of diabetes: type 1 diabetes, which is caused by the immune system attacking the cells in the pancreas that produce insulin, and type 2 diabetes, which is caused by the body becoming resistant to insulin or the pancreas not producing enough insulin.

In this project, we have selected the Health Indicator Diabetes Dataset from Kaggle. The dataset allows us to illustrate which indicator which is relevant to cause diabetes in the common people with different economic, education, and other kinds of day-to-day habit which may cause or impact the body to trigger to our immune system to stop release of the insulin which cause us to become Diabetes.

Our focus on this project to build the Machine Learning based system which can detect the early stage diabetes. For this purpose, from the dataset we have consider the three samples, which was given to us. We have selected only one sample which has around 70,692 patients, as it is balanced dataset. After performing cleaning, analyzing the dataset and we have built different model and evaluate their performance in terms of accuracy and time taken for building. We have evaluated and discuss each model with their pros and cons.

In conclusion, we have discussed the models which we built for early detection of diabetes in everyone life. Outcomes of our model to show a person is becoming diabetes if the given indicator is no change in our day-to-day lifestyle.

# Introduction

Diabetes is a chronic disease that affects how the body processes glucose, which is the primary source of energy for the cells. There are two main types of diabetes: type 1 diabetes, which is caused by the immune system attacking the cells in the pancreas that produce insulin, and type 2 diabetes, which is caused by the body becoming resistant to insulin or the pancreas not producing enough insulin.

Diabetes has become increasingly common in recent years, and it is estimated that millions of people worldwide are affected by this disease. In the United States, according to the Centers for Disease Control and Prevention (CDC), as of 2018, 34.2 million Americans have diabetes, and 88 million have prediabetes.

However, the CDC estimates that many people with diabetes or prediabetes are unaware of their risk. Specifically, they estimate that 1 in 5 people with diabetes and roughly 8 in 10 people with prediabetes are unaware of their condition.

There are several reasons why people may be unaware of their risk for diabetes or prediabetes. One reason is that the early symptoms of diabetes can be mild or even absent, which can make it difficult to detect the disease in its early stages. Some of the early symptoms of diabetes include increased thirst, frequent urination, fatigue, blurred vision, and slow wound healing.

Another reason why people may be unaware of their risk for diabetes or prediabetes is that they may not undergo regular health checkups or screenings. Regular health checkups and screenings are important for detecting diabetes and other health conditions in their early stages when they are more treatable.

In conclusion, diabetes is a chronic disease that affects millions of people worldwide, and its prevalence is increasing. Many people with diabetes or prediabetes are unaware of their condition, which can lead to serious health complications. Regular health checkups and screenings can help in detecting diabetes and other health conditions in their early stages, which is crucial for effective treatment and prevention of complications.

## Research Questions:

1. Which indicators are more correlated towards Diabetes?
2. Which indicator variables has more importance on Diabetes while performing predictive analysis?
3. Which predictive model is the best for your prediction analysis, in terms of time and accuracy?
4. What are the applications to utilize develop this predictive model?

## Tools:

* Language: Python
* Module: pandas, NumPy, scikit-learn, matplotlib, seaborn
* Editor: python-notebook

# Literature Review

In the paper, “*Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study*” by Liying Zhang has discussed about Diabetes Mellitus, is a metabolic disorder, which is prevalently increasing worldwide including China. The Henan Rural Cohort Study reported the rural population of Henan province highly affected. It’s difficult to screen thousands of patients with high risk conditions. Therefore, it’s a call for analytical techniques to mass screen number of patients. Various risk score based statistical tool developed, based on assumptions that each risk factor was linear with outcome. Machine learning can analyzed nonlinear interaction with large data. Various machine learning approaches such as ANN, novel joint clustering and classification (JCC) method, neural network, decision tree, and random forest to predict diabetes mellitus with 14 attributes etc. were used. Here, the study focused on performance comparison of prediction techniques with fixed number of variables, and on a small population sample. Therefore, the purpose of this study was to evaluate an array of machine learning algorithms for predicting the risk of T2DM in a rural Chinese population, identify the important variables, and reveal the performance of each model with number of variables. The study comprise logistic regression, artificial neural networks, classification and regression tree, support vector machine, and ensemble learning (random forest and gradient boosting machine) to build the risk assessment mode. Statistically, research involved factors such as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under precision recall curve (AUPR) and area under curve (AUC) to evaluate discrimination. In this study, high predictive performance was obtained in all the models with AUC value (0.811to 0.872).The GBM model perform best with AUC value (0.872) compare to other models. The New Chinese Diabetes Risk Score included sex, age, family history of diabetes, waist circumference, BMI, SBP were important risk factors. Obesity is the major cause of T2D according to study. The boosting method was outstanding performer in T2D risk assessment.

In the paper, “*An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study*” by Umm e Laila stated that Diabetes is a chronic condition, increase high blood glucose level. World Health Organization (WHO) conducted a research reveal increase in the number of diabetes cases. Ignorance in diagnosis and treatment of Diabetes cause neuropathy, retinopathy, nephropathy and muscle dystrophy. The prognosis of diabetes and treatment is very important to prevent health risk. Numerous machine learning techniques used for disease prediction but attain accuracy more than 80%. This study aims to increase the accuracy of machine learning ensemble standard algorithms (including AdaBoost, Bagging, and Random Forest) by analyzing the UCI diabetes dataset and comparing their performances. The diabetes data with 17 attributes were collected from the UCI repository which contains different datasets. The dataset utilized 17 attributes reflecting patient and hospital outcomes. The 517 instances including a class attribute used to predict the negative and positive possibility of having diabetes or not. When the ensemble standard algorithms compared during cross validation and found that Random Forest gives the best accuracy, precision, recall and F-measure compared to other ML techniques. The Chi-Square attributes selection approach used to calculate score of individual approach. The attribute polyuria obtained high score (208) and age; itching got lowest attribute (0). The study will be helpful in healthcare system and for people to managing disease. It is necessary to develop advance algorithm methods, gather data, Improve data quality etc. in future.

In the paper, “*Development of Various Diabetes Prediction Models (DM) Using Machine Learning Techniques*” by Juyoung Shin focus on to develop valid and applicable DM prediction models using an electronic medical record database. The data were extracted from electronic medical records of the Health Promotion Center of Seoul St. Mary’s Hospital between 2009 and 2018. Four prediction model develop in this study to ensure accuracy. Model-1 and Model-2 were predicting the Diabetes development after 2 and 1 years, respectively (normal or prediabetic). Model-3 was a model for predicting the development of DM after 1 year in prediabetic subjects. Model-4 was a model for predicting the development of DM after 1 year in prediabetic subjects after who knows the difference between one and 2 years before diabetes diagnosis. Gradient boosting algorithms were used in Model-1, Model-2, and Model-3, and random forest algorithms were used in Model-4. The non-diabetic and diabetic subjects adjusted with same number in each model. Two set of variables used to develop prediction model. One set comprises 62 easily access variables from hospital and other model developed by 27 variables out of 62 from national health checkup records. The variables were age, sex, medication use, underlying diseases, family history, physical examinations etc. The information collected from medical records and found 3952 individuals were diagnose with diabetes. The test included self-reported information, physical examinations, and laboratory tests, pulmonary functional test, fasting glucose, HbA1c levels. These parameters make model more effective with AUC value approximately 0.9. This model help to health providers, clinicians to encourage people to follow healthy lifestyle.

In the paper, “*Diabetes Prediction Using Machine Learning*” by Jyotirani has discussed the usage of Machine Learning for prediction of early diagnosis of Diabetes which will be helpful for the management and prevention of disease. The focus of study to predict early diagnosis of diabetes with higher accuracy by combining different machine learning techniques. These algorithms were used in in the study i.e. k-nearest neighbour, Logistic Regression, Random forest, Support Vector Machine (SVM) and Decision tree. The algorithm was calculated to check the model accuracy. Diabetes dataset contain 2000 individuals collected from <https://www.kaggle.com/johndasilva/diabetes>. The outcome of data was based on the patient condition either diabetic or nondiabetic with denotation 0 and 1. Conclusively, the decision tree algorithm achieved 99% accuracy for the prognosis of diabetes.

In the paper, “*Quality indicators in type 2 diabetes patient care: analysis per care-complexity level*” by Josiane Schneiders focus on the study, that was to evaluate care quality indicator in diabetes patients at primary and tertiary health care level in Brazilian health care system. The analysis was evaluated based on the comparison of these healthcare system. The study was based on the assessment of quality care for the diabetes in these healthcare centres. According to ADA, the diabetic patient with control glycemic value should have at least two HbA1c tests and the patient with no well control glycemic should have at least 4 HbA1c tests. Only 50% of the people tested in primary health care and one fifth patient in tertiary health care. The diabetic neuropathy feet assessment test rate was very low especially in primary care. The study also found that lower percentage of patients had undergone retinopathy test. The reason might be lack of facilities in health care center to examine the problem and due to presence of lack of professionals. The reason is expected because these centers refer to the patients to the ophthalmologist. Nephropathy is the most common quality indicator and easily assessed by creatinine test for glomerular filtration test included in the study. The quality care indicator assessment was at lower rate in health care centre according to this study. If the findings of the study considered, it could increase the prevalence of quality care indicator for the patients best possible care.

# Research Methods

## Algorithms follow to build the models:

### Clean the datasets for NA/NULL Values

Many machines learning classifier doesn’t work on the dataset which has null values in their samples as it produces error or generate inaccurate prediction, So It’s better to clean the dataset before performing any step towards model building. We have applied NumPy library to check and remove the sample consist the null values.

### Perform the statistical analysis on each health marker vs Diabetes

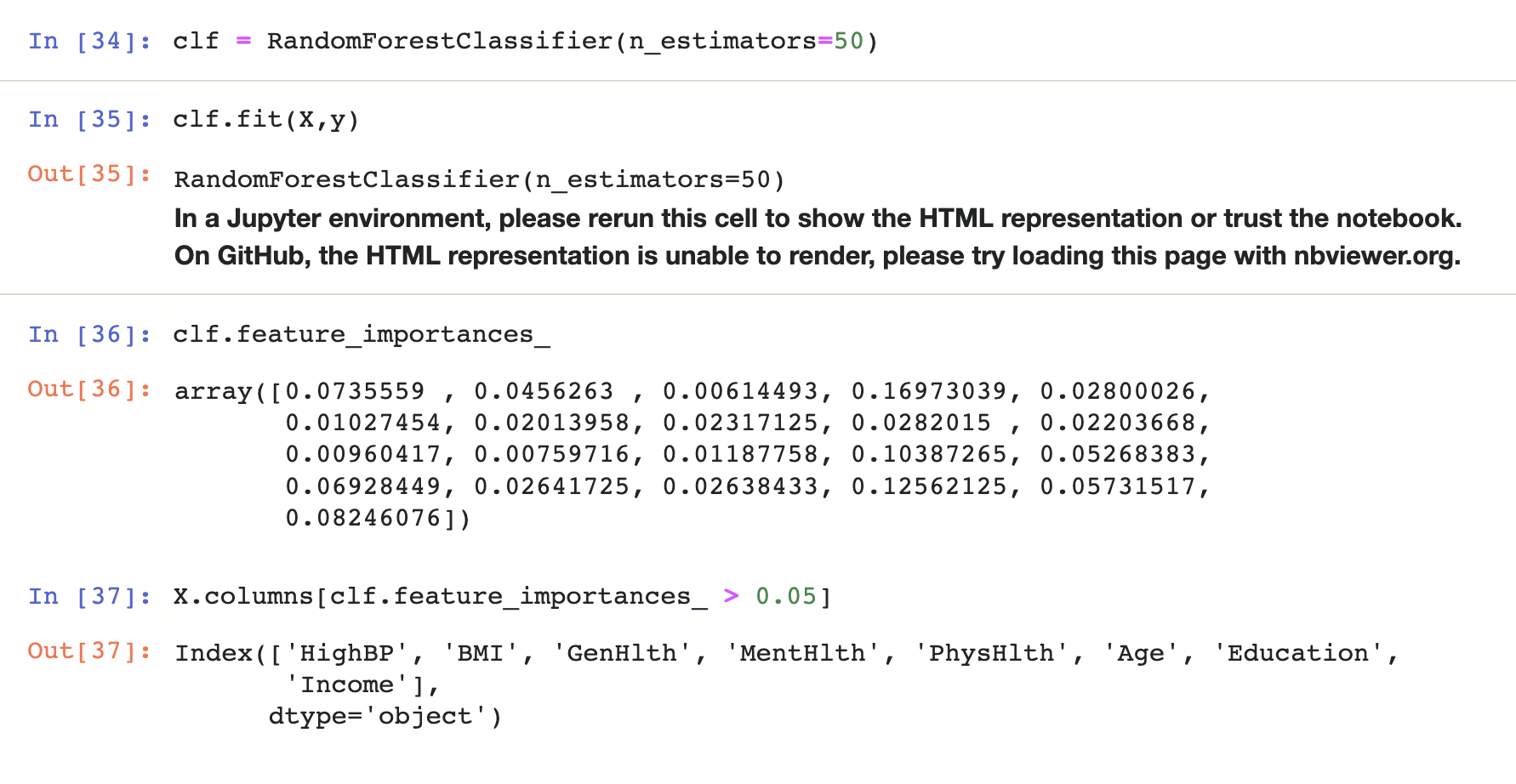
We have utilized the Chi-Square test to perform the statistical analysis of each marker vs Diabetes to prove Null Hypothesis False. Null Hypothesis states that any relevance feature vs Diabetes is just luck by chance, there is no direct relevance between under the significance value.

### Remove the Outliers from the datasets

We have utilized the box-plot to visualize the outlier of non-binary columns i.e., BMI, Education, MenHlths, Income. Removal of these outlier was performed by utilizing the inter-quartile-range (IQR) to remove the outliers’ samples from the datasets. This may change our dataset percentage of Diabetes Vs Non-Diabetes little off balance, but It was the necessary to get the highest accuracy.

### Perform the features selection (PCA/Feature Importance)

In our datasets, we have total 21 health markers, which is not directly connect with diabetes, but still they have relation which can be seen in heatmap plot of the datasets. For reducing the dimension of the dataset for better prediction accuracy. We have tried many methods i.e. PCA/ Random Forest Feature Elimination / Extra Tree Classifier Feature Elimination / Recursive Feature Elimination. In most of the methods, we need to specify the dimension which we want to reduce to find the appropriate feature for building the model which can improve the classifier performance.



**Figure 1 : Random Forest Classifier used for Feature Selection**

We have used the Random Forest to build each health marker, feature importance. Which is shown in figure 2. In the next step we have check with many values of threshold for removing the, but if we choose the threshold value below 0.05, it was selecting almost all the features. So we have taken threshold for 0.05, when we did that, It almost selected all the non-binary column (health markers) and few binary column (health markers).

### Build the Model, measure converge time, and performance (accuracy)

In machine learning model building, we used to normalize the dataset when the features have many diverse values. In our dataset, all the values categorical values except few such as BMI. Rest of the other non-binary column data is categorical as bin data i.e. Education, MentHlth, GenHlth, Income, PhyHlth, Age. So didn’t perform the scaling in our datasets.

## Supervised Machine Learning Model:

We have selected fives machine learning classifier i.e. Logistics Regression, Random Forest, Decision Tree Classifier, XGBoost Classifier, Naive Bayes Classifier to build our machine learning model.

### Logistic Regression:

Logistic Regression is a statistical Machine Learning algorithm that is used for classification problems. It is based on the concept of probability. It is used when the dependent variable (target) is categorical. It is widely used when the classification problem at hand is binary; true or false, yes or no, etc. Logistics regression uses the sigmoid function to return the probability of a label.

### Gaussian Naïve Bayes

Naïve Bayes classifier used the Bayes Theorem for prediction of samples. The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category.

Picture 5

* P(c|x) - Posterior Probability of the response (target) variable given the training data inputs
* P(c) - Prior probability of the class (target)
* P(x|c) - Probability of the predictor (x) given the class/target (c)
* P(x) - Prior probability of the predictor (x).

### Decision Tree Classifier:

Decision Tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions. Decision trees can perform both classification and regression tasks. Decision is kind of an umbrella term, the intuition behind Decision Trees is that you use the dataset features to create yes/noquestions and continually split the dataset until you isolate all data points belonging to each class. With this process you’re organizing the data in a tree structure. Every time you *ask a* question*,* you’re adding a node to the tree. And the first node is called the root node. The result of *asking a question* splits the dataset based on the value of a feature, and creates new nodes. If you decide to stop the process after a split, the last nodes created are called leaf nodes.

### Random Forest Classifier:

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning). Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model’s prediction.

### XGBoost Classifier:

XGBoost is an ensemble learning algorithm meaning that it combines the results of many models, called base learners to make a prediction. Just like in Random Forests, XGBoost uses Decision Trees as base learners. XGBoost Library is parallelizable which means the core algorithm can run on clusters of GPUs or even across a network of computers. This makes it feasible to solve ML tasks by training on hundreds of millions of training examples with high performance. Due to Its speed and performance are unparalleled and it consistently outperforms any other algorithms aimed at supervised learning tasks.

The results of these model will be discussed in the Conclusion section.

# Data Analysis

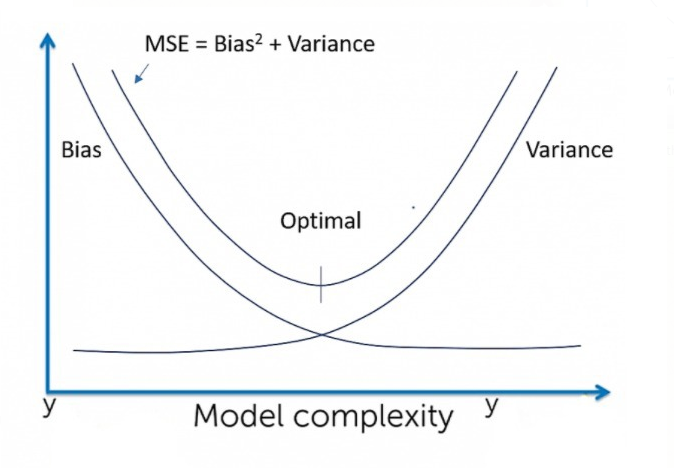
For understanding the diabetes, I have browsed through many data maintaining website i.e. data-world, Kaggle, few hospital website as well which keeps the information of diabetes patients who are facing diabetes as challenges in their life.  We have selected the datasets, “Diabetes Health Indicators dataset”, was downloaded from Kaggle. This dataset was generated by health-survey named “The Behavioral Risk Factor Surveillance System” ([BRFSS](https://rpubs.com/dMero/dim_reduction)) that is collected in 2015 by the “Central of Disease Control” (CDC). Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. This original [dataset](https://rpubs.com/dMero/dim_reduction) contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

 This dataset contains 3 files:

1. diabetes\_012\_health\_indicators\_BRFSS2015.csv: is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes\_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. This dataset has 21 feature variables and is imbalanced.
2. diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv: is a clean dataset of 70,692 survey responses to the CDC's BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable diabetes binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced. (Selected for Building Model)
3. diabetes\_binary\_health\_indicators\_BRFSS2015.csv: is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable diabetes binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced.

For the selected dataset, we have the written the following important facts:

1. It has many independent variables (20 feature) and 1 dependent variable (diabetes binary)
2. It has many numbers of samples in each kind of dataset files 70692 (Selected CSV File). It is good for predictive model to have sufficient number of samples for training and testing. But
3. It has the features i.e. Diabetes\_binary, HighBP, HighChol, CholCheck, BMI, Smokes, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education and Income.

****

**Figure 2 : Bias Vs Variance for building ML Model**

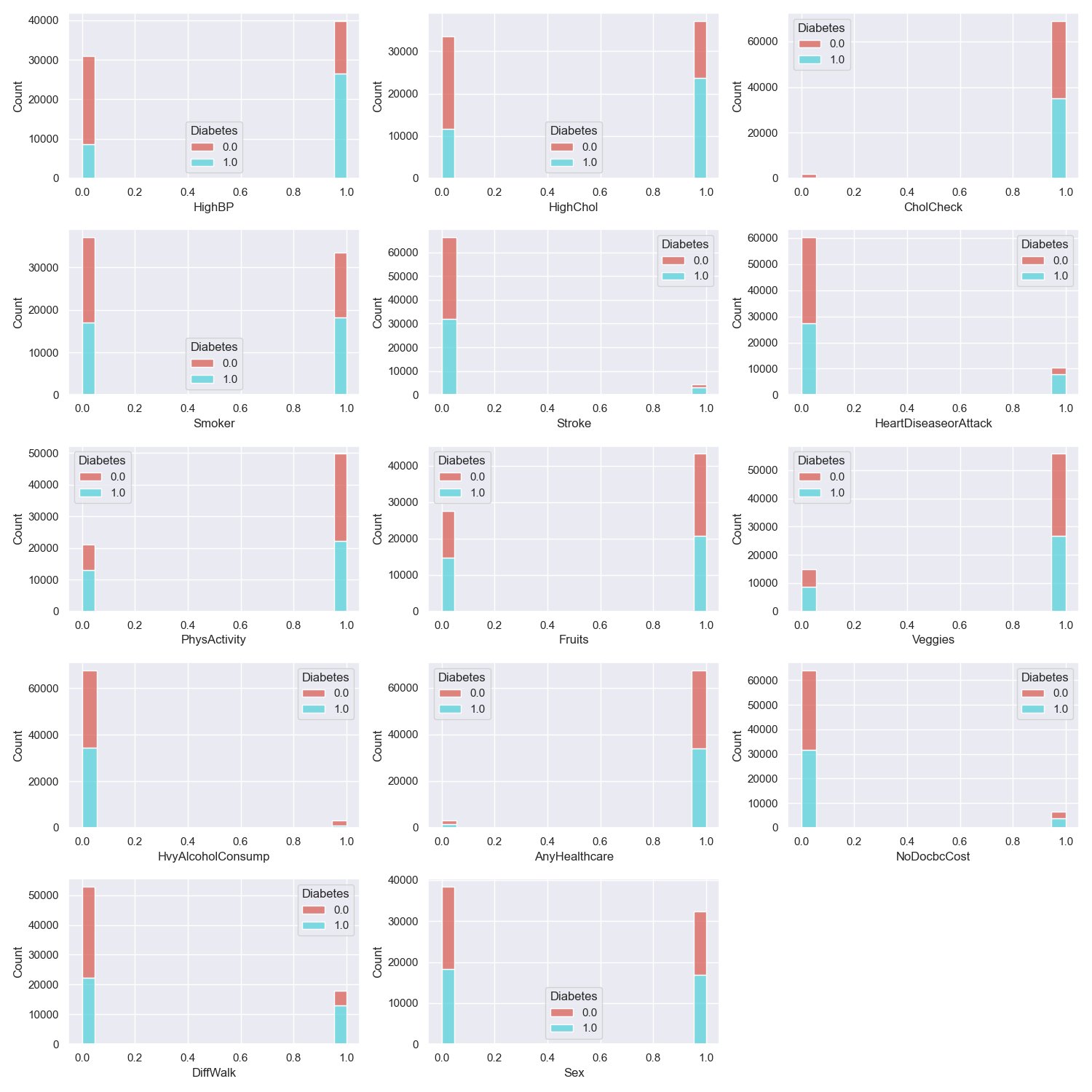
**Table 1 : Demographics Information about Dataset**

|  |  |
| --- | --- |
| **Characteristics** | **Values** |
| Age  Mean + Std (Range)  Divided into 13 age categories  1=18-24, 2=25-29, 3=30-34,4=35-39,5=40-44,6=45-49,7=50-54,8=55-59,9=60-64,10=65-69,11=70-74,12=75-79,13=80 or older | 8.58 + 2.85 (1-13) |
| **AnyHealthcare**  Binary (1-0)  1 – Yes  0 – No/ Don’t know/Not sure/Refused | 67508 (95.49%)  3184 (4.50%) |
| **BMI** Mean + Std (Range) | 29.86 + 7.11 (12-98) |
| **CholCheck**  Binary (1-0)  1 – Yes, Cholesterol check in 5 years  0 – No Cholesterol check in 5 years | 68943 ( 97.52%)  1749 (2.47%) |
| **Diabetes**  Binary (1-0)  1 – Yes, Diabetes  0 – No, Non-Diabetes | 35346 (50%)  35346 (50%) |
| **DiffWalk**  Binary (1-0) :difficulty walking or climbing stairs ?  1 – Yes,  0 – No, | 35346 (74.72%)  35346 (25.27%) |
| **Education**  1 – Never attended School  2 – Grades 1 through 8 (Elementary)  3 – Grades 9 to 11 (some high school)  4 – Grades 12 or GED ( High School graduate)  5 – College 1 year to 3 years (some college )  6 – College 4 year or more (college graduate) | 75  (0.10%)  1647 ( 2.32%)  3447 (4.87%)  19473 ( 27.54%)  20030 ( 28.33%)  26020 ( 36.80%) |
| **Fruits**  Binary (1-0)  1 – Yes, Take Fruits regularly  0 – No, Didn’t eat fruits regularly | 27443 (38.82%)  43249 (61.17%) |
| **GenHealth**  1 – Excellent  2 – Very good  3 – Good  4 – Fair 12 or GED ( High School graduate)  5 – Poor / Not sure/ Refused | 8282 (11.71%)  19872 ( 28.11%)  23427 (33.13%)  13303 ( 18.81%)  5808 ( 8.21%) |

|  |  |
| --- | --- |
| **Characteristics** | **Values** |
| **HeartDiseaseorAttack**  Binary (1-0)  1 – Yes,  0 – No, | 10449 (14.78%)  60243 (85.21%) |
| **HighBP**  Binary (1-0)  1 – Yes,  0 – No, | 39832 (56.34%)  30860 (43.65%) |
| **HighChol**  Binary (1-0)  1 – Yes,  0 – No, | 37163 (52.57%)  33529 (48.42%) |
| **HvyAlcholConsump**  Binary (1-0)  1 – Yes,  0 – No, | 3020 (4.27%)  67672 (95.72%) |
| **Income**  1 – Less than $10, 000  (0 to 10000)  2 – Less than $15,000   (10001 to 15000)  3 – Less than $20, 000  (15001 to 20000)  4 – Less than $25,000   (20001 to 25000)  5 – Less than $35, 000  (25001 to 35000)  6 – Less than $50,000   (35001 to 50000)  7 – Less than $75,000   (50000 to 75000)  8 – more than $75,000   (75000 to more) | 3611  (5.10%)  4498 ( 6.36%)  5557 (7.86%)  6658 (9.41%)  8010 ( 11.33%)  10287 ( 14.55%)  11425 (16.16%)  20646 (29.20%) |
| **MentHlth**  Mean + Std (Range)  Number of day of poor mental health scale | 3.75+8.16 (0-30) |
| **PhysActivity**  Binary (1-0)  1 – Yes,  0 – No, | 49699 (70.30%)  20993 (29.63%) |
| **NoDocbcCost**  Binary (1-0)  1 – Yes,  0 – No, | 6639 (9.39%)  64053 (90.60%) |
| **Sex**  Binary (1-0)  1 – Yes,  0 – No, | 32306 (45.69%)  38386 (54.30%) |
| **PhysHlth**  Mean + Std (Range)  Number of  day of poor mental health scale | 5.81 + 10.06 (0-30) |

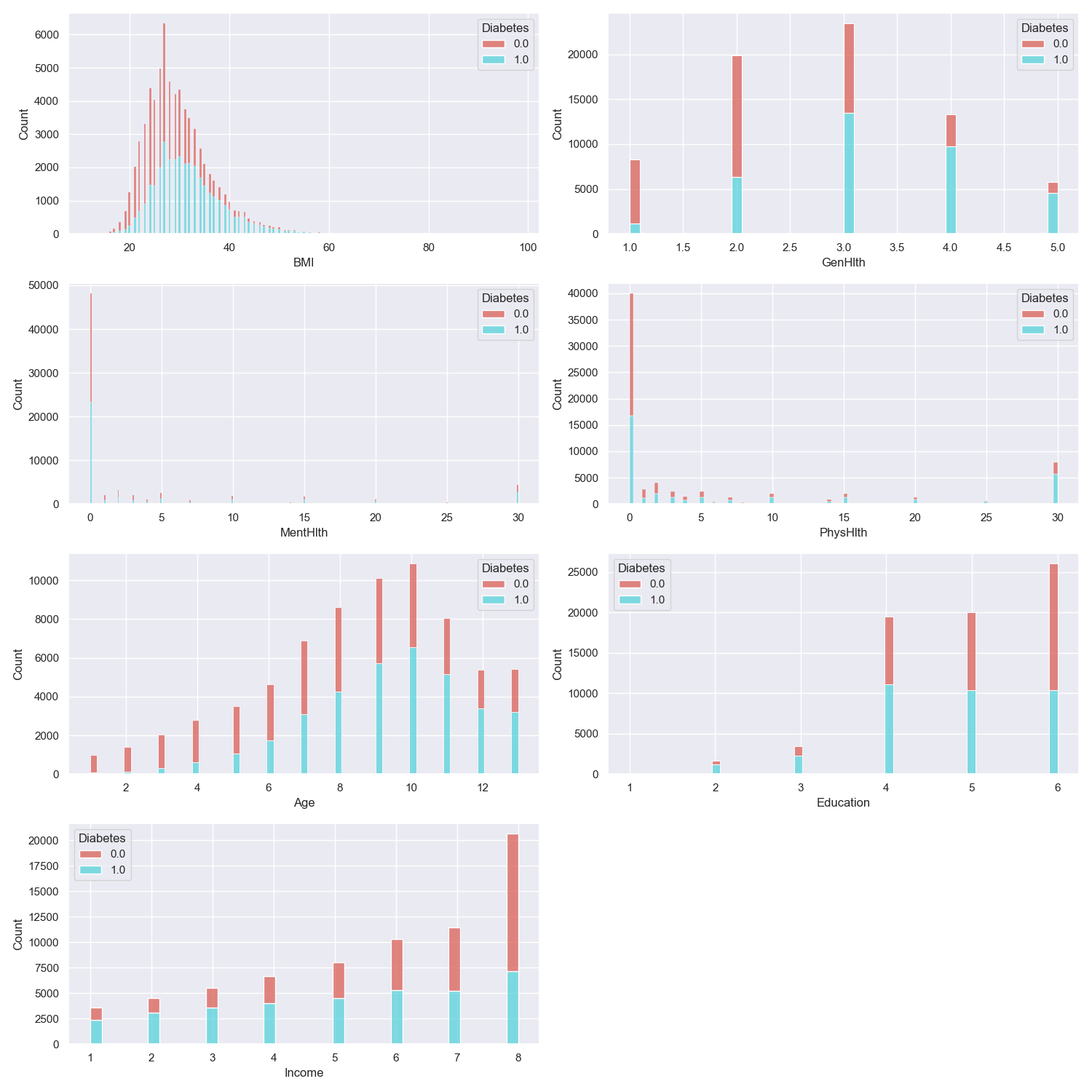
# Discussion

For building any model, we need to understand how independent feature variables (Age, Income, Education, etc.) are correlated with dependent variable (Diabetes). We have divided the independent features in to two category (Numeric and Binary) and plotted them explicitly.



**Figure 3:Plots the Binary variable w.r.to. Diabetes**

From the figure 3, we can observe that patients who don’t perform CholCheck, are very few who are diabetic. The patients who are HbyAlcoholConsump, very few are diabetic.



**Figure 4: Plot of Numeric Variables w.r.to Diabetes**

From all the numeric feature variables, Higher BMI patients has more diabetic than lower BMI patients. Higher the Age, number of patients are more diabetic.

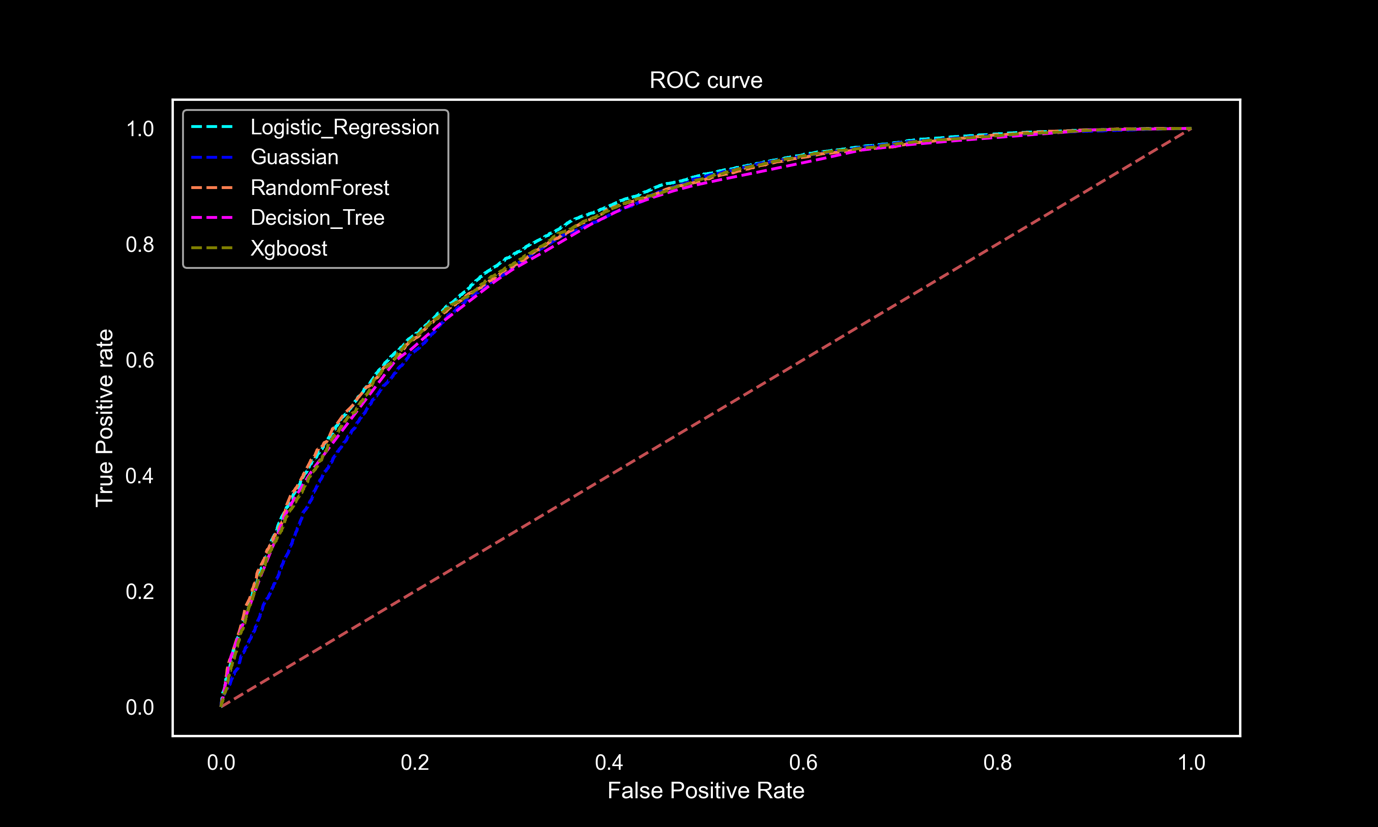
As the income of the patients increases, people become more careful towards their health (diabetic) control. There is not much correlation of Education vs Diabetes.

In Diabetes Health Indicators Dataset analysis, there were many questions raise in terms of analysis and choice of algorithms while performing feature/indicator selections. I have built the model by consider the 70% dataset for Training, and 30% dataset for Testing.

For relative importance of health marker, we have performed the feature importance by utilizing the Random Forest Classifier, which provides the feature importance of each health markers. It allows us the select the appropriate number of features by passing the threshold for feature importance. We have Display each health marker importance score in Figure 7.

If two models (model1, model2) have the similar performance even a little less model2 has the little less performance accuracy. I would definitely go the model which has the less converge time with the similar accuracy, as it is advisable for anyone who has less efficient computer, still he/she can run that classifier and get the similar performance comparison to someone who has powerful computer and getting the similar performance by choosing the other model.

From the Figure 5, It is shown that Area Under Curve (AUC) of all the model built have similar values. But Logistic Regression has the highest AUC value in ROC curve which is plotted between True Positive Rate vs False Positive Rate. For an ideal model should have AUC equals to 1.0. In our case, I got the highest AUC value as 0.81 for Logistic Regression and other models are lower AUC than Logistic Regression.

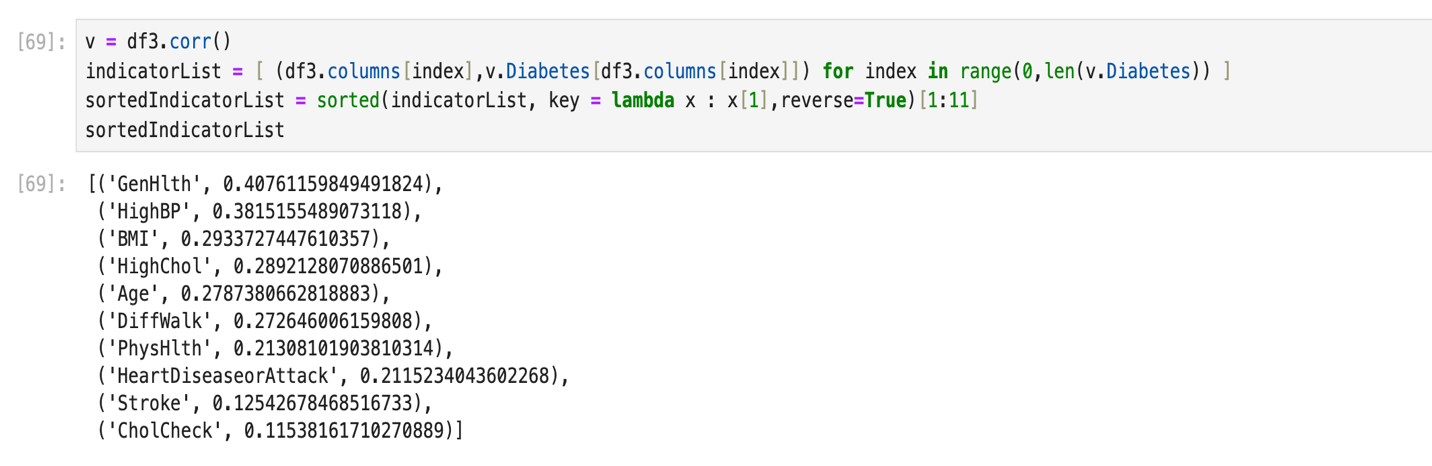


**Figure 5: ROC curve for different Models**

# Conclusion

In this project, we have explored the health marker which impacts a human life to trigger his system to become a Diabetes Person. We have developed diabetes prediction models using machine learning kernels and different dataset cases. The studies found that Age, BMI, family history of diabetes, physical activity, smoking status, waist circumference, systolic blood pressure, fasting plasma glucose, glycated hemoglobin (HbA1c), triglycerides, and total cholesterol were significant predictors of diabetes.

The very high correlated health marker towards a person to become Diabetes are listed in Out[22] in Python notebook.



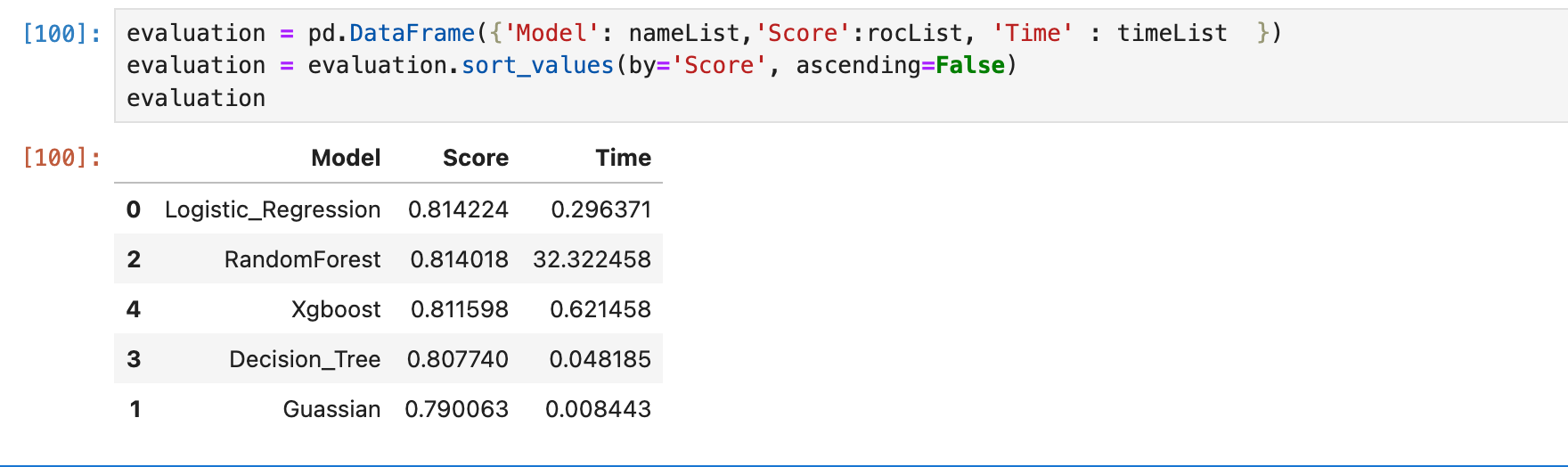
**Figure 6: High Correlated Health Indicator Marker towards Diabetes**

In the project, we found these factors (HighBp, BMI, GenHlth, MentHlth, PhysHlth, Age, Education and Income) are among important for a person becomes Diabetes ( found by Random Forest Classifier by using feature importance)



**Figure 7: Random Forest Classifier is used for Feature Selection**

The models developed had good discriminatory power, accuracy ranging from 79% to 81%. These findings demonstrate the potential of machine learning algorithms in predicting the risk of diabetes and improving early detection and prevention strategies.



**Figure 8: Model Comparison of Accuracy Vs Time Taken**

Finally In the last, once we have selected the appropriate features, we build our classifier. In our case, we have taken 5 classifiers to build for classification Logistic Regression, Random Forest Classifier, Decision Tree Classifier, XGBoost Classifier, Naive Bayes Classifier. In the Figure 5, we have displayed the classifier results, according to their classification accuracy and Converse Time (Fitting Time taken by classifier).

As you observe, the Logistic Regression Classifier has similar accuracy as Random Forest and XGBoost Classifier. But if you can observe, that it has the minimum converse time comparison to other two classifier. It is 10x time faster than XGBoost classifier and 39x faster than Random Forest Kernel where number of Decision tree taken in Random Forest was 100 trees.

# List of References

1. Zhang, Liying, et al. "Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study." Scientific reports 10.1 (2020): 4406
2. Laila, Umm E., et al. "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study." Sensors 22.14 (2022): 5247.
3. Shin, Juyoung, et al. "Development of various diabetes prediction models using machine learning techniques." Diabetes & Metabolism Journal 46.4 (2022): 650-657.
4. Rani, KM Jyoti. "Diabetes prediction using machine learning." International Journal of Scientific Research in Computer Science Engineering and Information Technology 6 (2020): 294-305
5. Schneiders, Josiane, et al. "Quality indicators in type 2 diabetes patient care: analysis per care-complexity level." Diabetology & metabolic syndrome 11 (2019): 1-9.
6. Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.
7. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Emerging Risk Factors Collaboration
8. Website : [TowardDataScience](https://towardsdatascience.com/hypothesis-testing-for-data-scientists-everything-you-need-to-know-8c36ddde4cd2) <https://towardsdatascience.com/hypothesis-testing-for-data-scientists-everything-you-need-to-know-8c36ddde4cd2>
9. Kaggle Dataset link : <https://towardsdatascience.com/hypothesis-testing-for-data-scientists-everything-you-need-to-know-8c36ddde4cd2>