BRAC
UNIVERSITY

Inspiring Excellence

# CSE422 - ARTIFICIAL INTELLIGENCE

## *Project on Diabetes Prediction using Machine Learning*

**Group Information ( Section - 1 )**

Akib Zabed Ifti
ID: 20101113

MD.Sazidur Rahim
ID: 17301048

Riana Islam Shawon
ID: 20101374

Md. Anonto Shuvo
ID: 20301301

## INTRODUCTION

Diabetes is a chronic illness that affects how your body converts food into energy. Untreated Diabetes may cause some major issues in a person like: heart related problems, kidney problems, blood pressure, eye damage and it can also affect other organs of the human body. If we can identify diabetes earlier, we can prevent further health problems that can be caused by Diabetes. We will perform early diabetes prediction in a human body or patient using a range of machine learning approaches for a greater level of accuracy. ways to use machine learning We can better predict outcomes by creating models from patient data sets accurately. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which are Naive Bayes (NB), Logistic Regression (LR) and Random Forest (RF). When compared to other models, each model's accuracy varies. The project work reveals that the model is capable of accurately predicting diabetes with an accuracy of 95% or higher. Our findings demonstrate that Random Forest outperformed other machine learning methods in terms of accuracy.

## METHODOLOGY

**Dataset Description :** Our dataset 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.

### Libraries used:

The project would use the following libraries:

● Pandas - for reading the data file

● Numpy - for arrays

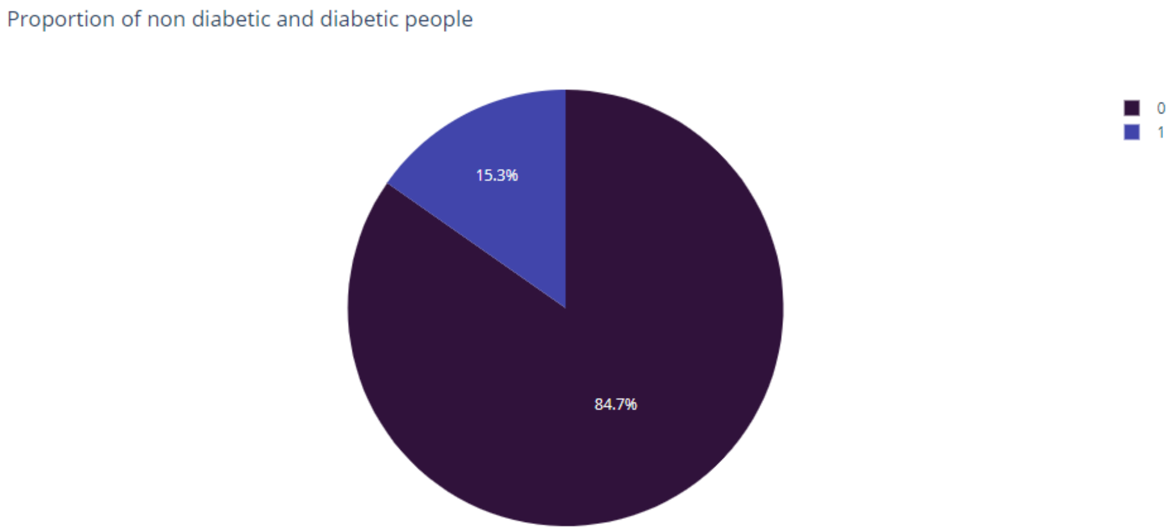
● Tensorflow - for building the model

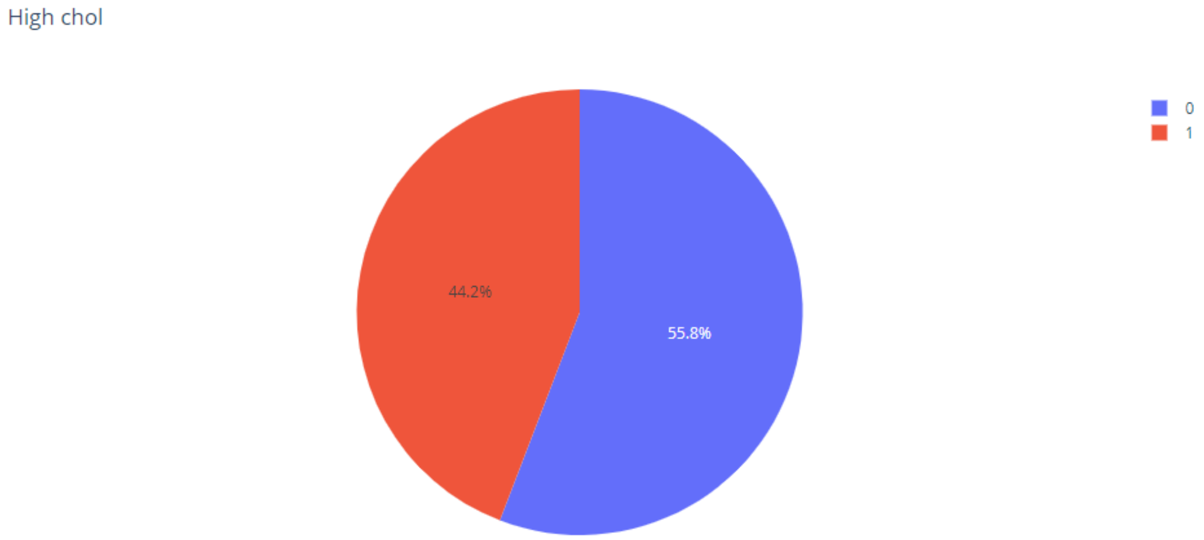
● Sklearn - for splitting the dataset

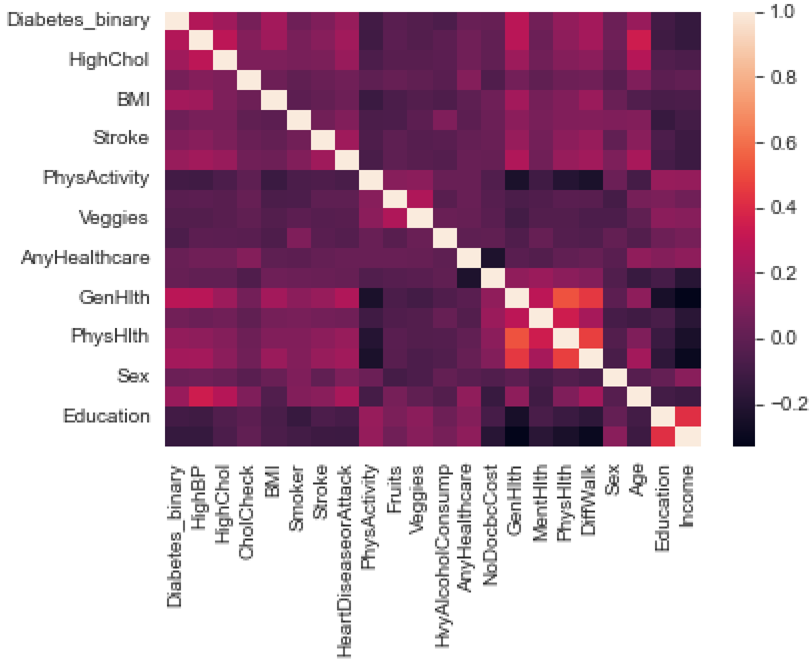
● Matplotlib - for plotting graphs

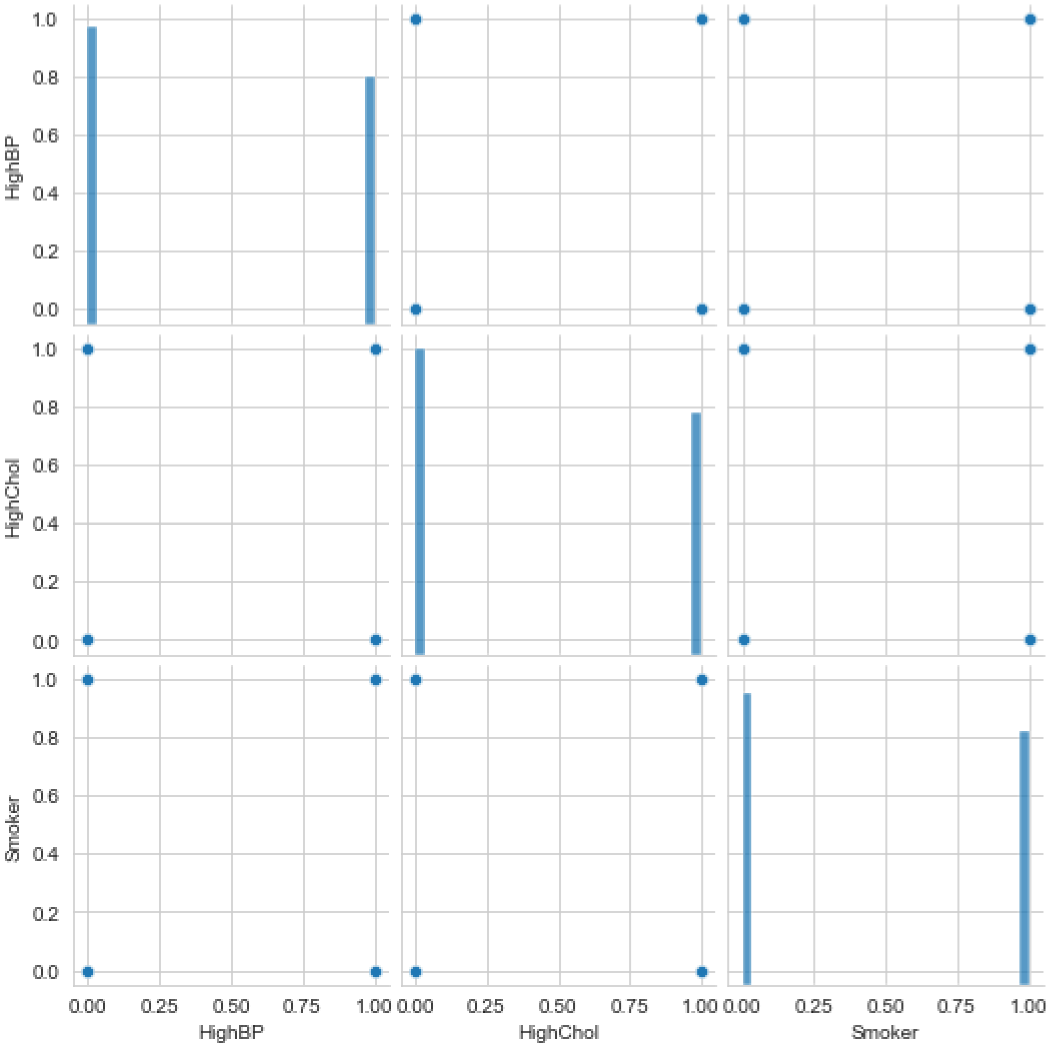
**Project architecture:** Using the dataset, we trained our model using different algorithms.. First of all, we initialized our preprocessing in order to split our data into training and testing datasets. Then we explored and analyzed our data. After splitting, we compiled and fit the model with a train dataset to build our model. Furthermore, we extracted the model after compilation and fitting. Using the model, we ran our test set and got the desired results or outcomes.

**Exploratory Data analysis :** In this part we tried to analyze our data with some pie-plot, pair-plot and a correlation matrix.




High chol
0
1
44.2%
55.8%


Diabetes_binary
HighChol
BMI
Stroke
PhysActivity
Veggies
AnyHealthcare
GenHlth
PhysHlth
Sex
Education
Diabetes_binary
HighBP
HighChol
CholCheck
BMI
Smoker
Stroke
HeartDiseaseorAttack
PhysActivity
Fruits
Veggies
HvyAlcoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



**Data Pre-Processing :** As we don't have any null values in our data we already were free from further work. We have 253680 rows and 22 columns in our dataset where the equivalence of our target variable is missing. We have way more 0 values than 1 in our diabetes column.  So, we oversampled our data and made them equal After oversampling the number of rows for 0’s and 1’s are 194377. Then the Standard scalar method was applied as if we had non-even values in different columns we would lead us to chaotic model sampling because the models tend to prioritize the columns with higher values. So, we scaled all of our values between 0 to 1.

## Applied Models :

### 1. Naive Bayes

For our project, we have used the Naive Bayes Classifier model to predict the possibility of a person having diabetics using the dataset. The reason for using this model is that it's one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier i.e it predicts on the basis of the probability of an object. It converts the given dataset into frequency tables, generates likelihood tables by finding the probabilities of given features and uses Bayes theorem to calculate the posterior probability. For our project we specifically used the Gaussian Naive Bayes Model because it assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

### 2. Logistic Regression

Logistic regression is a statistical analysis method to predict a binary outcome. Logistic regression is used in various fields, including machine learning. Many medical scales used to assess severity of a patient have been developed using logistic regression. Logistic regression may be used to predict the risk of developing a given disease based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests etc.) The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product.

In our project, we have used the logistic regression algorithm to predict whether a person has diabetes or not. We have used this algorithm because a logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

### 3. Random forest:

Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. In machine learning, it can be applied to both classification and regression issues. It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance. Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead of depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. There are various escorts where random forests can be used. Positive aspects of Random Forest can be that both classification and regression tasks can be handled by Random Forest. To add on, It is able to handle big datasets with lots of dimensions. Lastly, It improves the model's accuracy and avoids the overfitting problem. However, Random forest can be used for both classification and regression tasks, but regression tasks are not better suited for it. This method can be used to identify illness patterns and risk factors.

In our project, we have used the Random forest algorithm to predict whether a person has diabetes or not. We have used this algorithm because this model can handle big datasets with lots of dimensions.

## RESULTS:

**Prediction** :

We have taken random data as input except the “Diabetes Binary” which is the result of our dataset. Then we have changed the input arrays to numpy arrays because it is faster and more compact than Python lists. An array consumes less memory and is convenient to use whereas numpy uses much less memory to store data and it provides a mechanism of specifying the data types. Thus allows the code to be optimized even further. We have reshaped the input as we are only working with only one row of data rather than using the entire data. After that, we have standardized the input data as there is variation in the values of the features. So by standardizing, we have kept the values from 0 to 1. Finally, we used predictions on the standardized data and found the result. If it's 0, then the person is not diabetic and if it's 1, then The person is diabetic.

**For Random Forest:**

**Accuracy = 90%**

```
 precision    recall  f1-score   support



        0.0       0.95      0.85      0.90      38705
        1.0       0.86      0.96      0.91      39046
```

**For Logistic Regression :**

**Accuracy : 73%**

```
precision    recall  f1-score   support

        0.0       0.74      0.71      0.73      38705
        1.0       0.72      0.75      0.74      39046
```

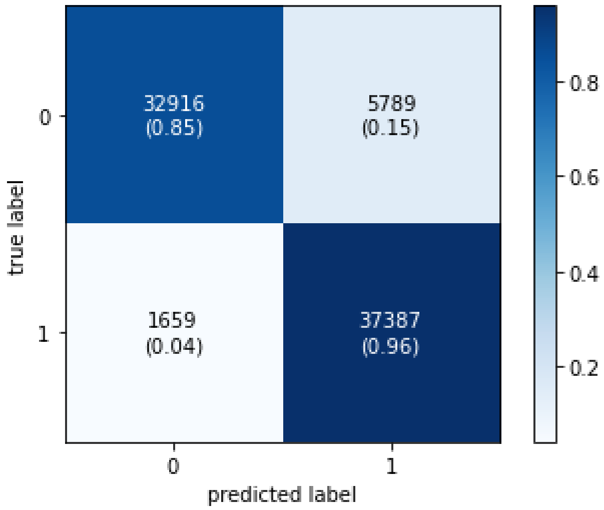
**For Naive Bayes :**

**Accuracy : 70%**

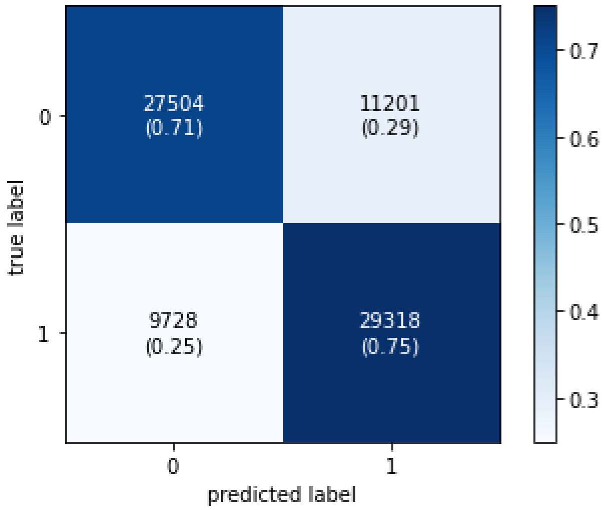
```
precision    recall  f1-score   support

        0.0       0.71      0.69      0.70      38705
        1.0       0.70      0.72      0.71      39046
```

**The confusion Matrix for the algorithms are :**

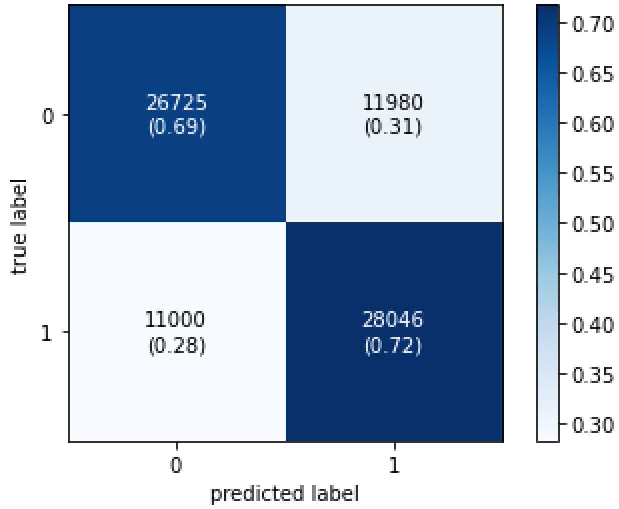
## Random Forest :

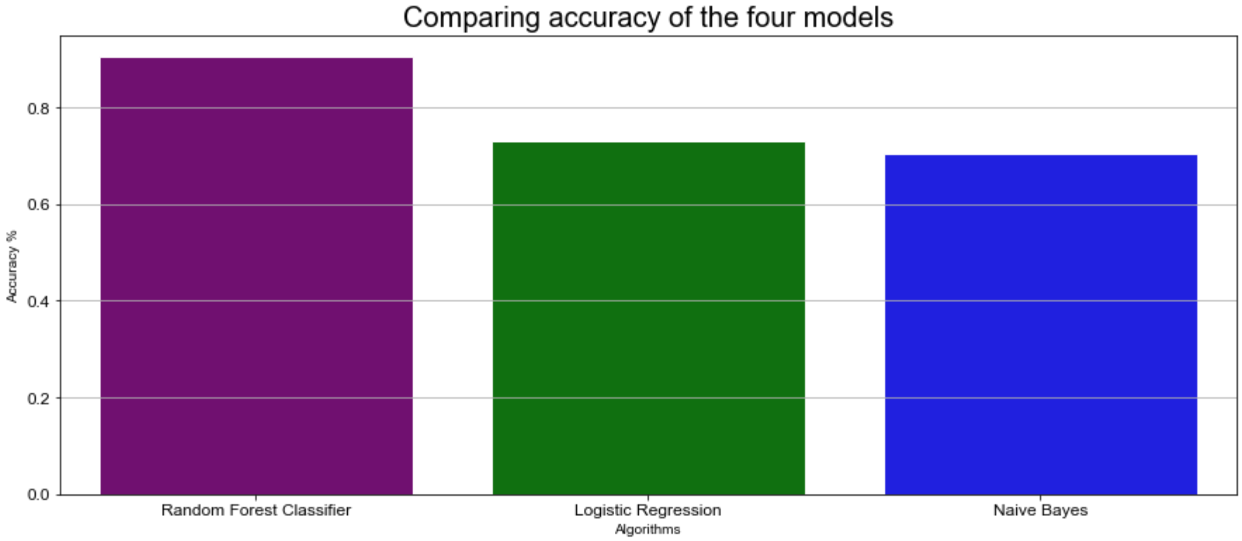


## Logistic Regression :



### Naive Bayes :


26725
(0.69)
11980
(0.31)
11000
(0.28)
28046
(0.72)
true label
0
1
predicted label
0
1
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30


Comparing accuracy of the four models
Accuracy %
0.0
0.2
0.4
0.6
0.8
Random Forest Classifier
Logistic Regression
Naive Bayes
Algorithms

## CONCLUSION

Finally after implementing all 3 algorithms we can conclude that Random Forest is performing better than the other Models and it has an accuracy score of 90%. Based on the accuracy of the Random Forest classifier, a prediction test to see How well the model can predict for our randomly given data which is from our dataset.

## REFERENCES

1. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset