Part 1 : Exploring the Dataset

General Information about the Dataset

- Dataset consists of (70692 rows/observations and 22 columns) and has no missing values.
- Dataset contains (21 independent variables) and (1 dependent variable)

Summary and Description of (Independent/Input Variables)

Categorical Variables:

- **Diabetes_binary (**Presence of diabetes or pre-diabetes**)**
  Values: 0 (no diabetes), 1 (pre-diabetes or diabetes)
- **HighBP (**Presence of high blood pressure**)**
  Values: 0 (no high blood pressure), 1 (high blood pressure)
- **HighChol (**Presence of high cholesterol**)**
  Values: 0 (no high cholesterol), 1 (high cholesterol)
- **CholCheck (**Whether individual had a cholesterol check in the last 5 years**)**
  Values: 0 (no cholesterol check in 5 years), 1 (cholesterol check in 5 years)
- **Smoker (**Smoking history (at least 100 cigarettes in life)**)**
  Values: 0 (no), 1 (yes)
- **Stroke (**History of being told they had a stroke**)**
  Values: 0 (No), 1 (Yes)
- **HeartDiseaseorAttack (**History of coronary heart disease or myocardial infarction**)**
  Values: 0 (no), 1 (yes)
- **DiffWalk (**Serious difficulty walking or climbing stairs**)**
  Values: 0 (no), 1 (yes)
- **PhysActivity (**Engagement in physical activity in past 30 days (excluding job-related activities)**)**
  Values: 0 (no), 1 (yes)
- **Fruits (**Consumption of fruits one/more times per day**)**
  Values: 0 (no), 1 (yes)
- **Veggies (**Consumption of vegetables one/more times per day**)**
  Values: 0 (no), 1 (yes)
- **HvyAlcoholConsump (**Heavy alcohol consumption (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)**)**
  Values: 0 (no), 1 (yes)
- **AnyHealthcare (**Availability of healthcare coverage (e.g. health insurance)**)**
  Values: 0 (no), 1 (yes)
- **NoDocbcCost (**Difficulty in seeing a doctor due to cost in past 12 months**)**
  Values: 0 (no), 1 (yes)
- **Sex (**Gender**)**
  Values: 0 (female), 1 (male)

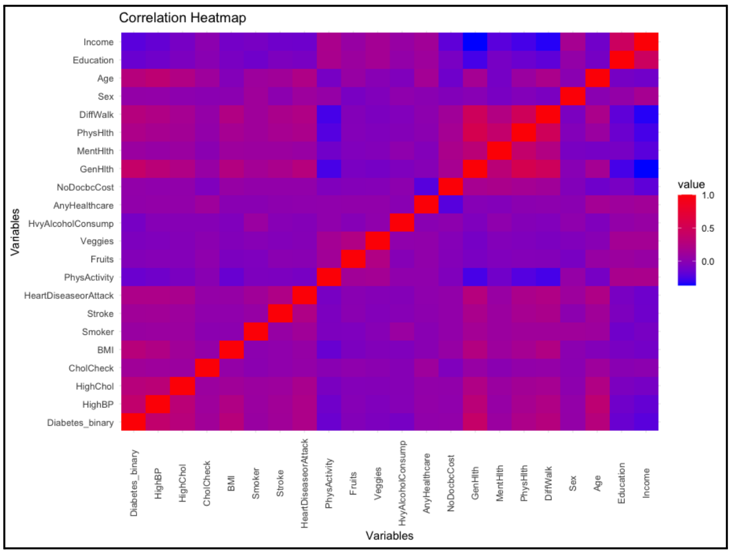
Ordinal Variables:

- **GenHlth (**General health perception**)**
  Values: 1 (excellent), 2 (very good), 3 (good), 4 (fair), 5 (poor)
- **Age (**Age categories**)**
  Values: 1 (age from 18 to 24), ..., 9 (age 60 to 64), 13 (age 80 or above)
- **Education (**Education level scale**)**
  Values: 1 (never attended school or only kindergarten), 2 (elementary), ...
- **Income (**Income scale**)**
  Values: 1 (less than 10k), 5 (less than 35k), 8 (75k or more)

Numeric Variables:

- **BMI (**Body mass index (a continuous numeric variable)**)**
  Values: Continuous numerical value (e.g., 25.4).
- **MentHlth (**number of days in past 30 days when mental health was not good**)**
  Values: Numerical count (continuous value).
- **PhysHlth (**number of days in the past 30 days when physical health was not good**)**
  Values: Numerical count (continuous value).

Summary and Description of (Dependent/Target Variable)



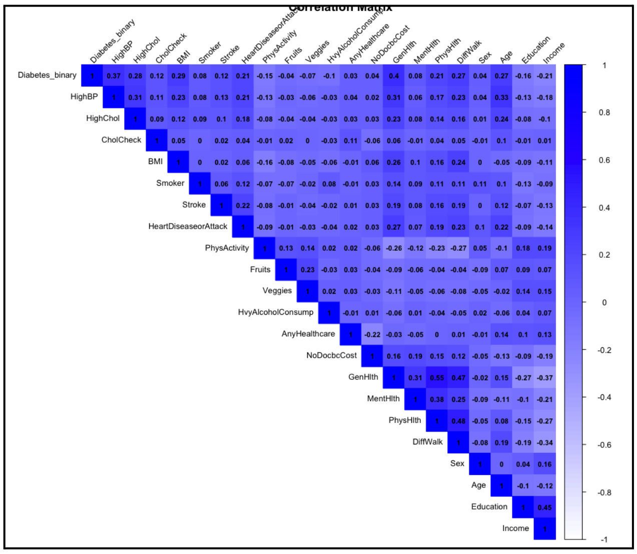


Figure 1 : Correlation Heatmap of all Columns *Figure 2 : Correlation Matrix of all Columns*

Through analysis of figure 1 (heatmap) and figure 2 (correlation matrix), I was able to identify that:

- Variables (GenHlth and PhysHlth) are highly correlated with each other (positive relation)
- Variables (GenHlth and Income) are highly correlated with each other (negative relation)

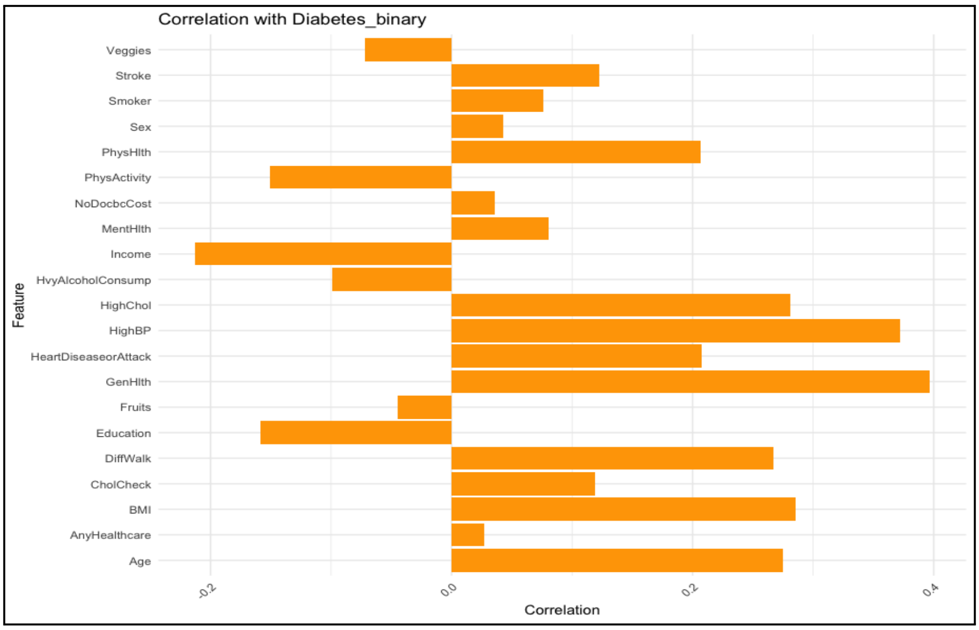
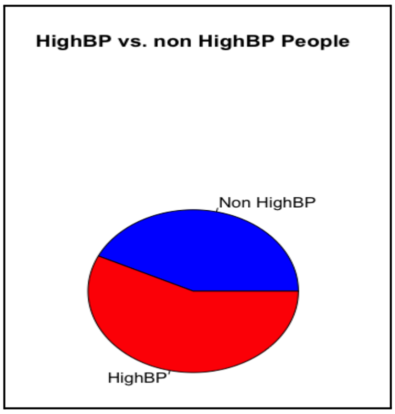


Figure 3 : Bar Graph of Correlation of Diabetes_binary with all the independent variables of the dataset

Through the analysis of the bar graph in Figure 2 I've identified that:

- Variables that least correlated with **Diabetes_binary** were:
  (Fruits | AnyHealthcare | NoDocbccost | Sex | Veggies | CholCheck)
- Variables that have significant correlation with **Diabetes_binary** are:
  (HighBP | HighChol | BMI | Smoker | Stroke | HeartDiseaseorAttack | PhysActivity | HvyAlcoholconsump | GenHlth | PhysHlth | Age | Education | Income | DiffWalk)

Specific Analysis of relationship between (HighBP) and (Diabetes_binary) show:



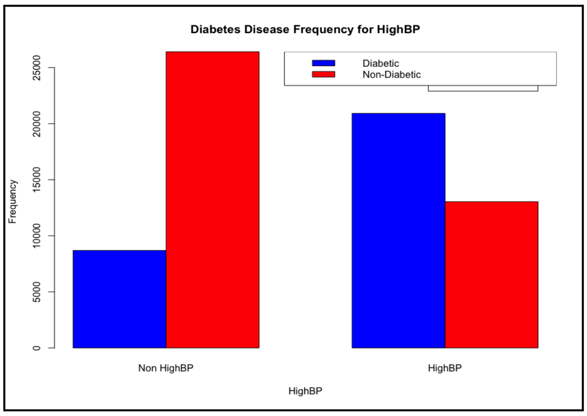


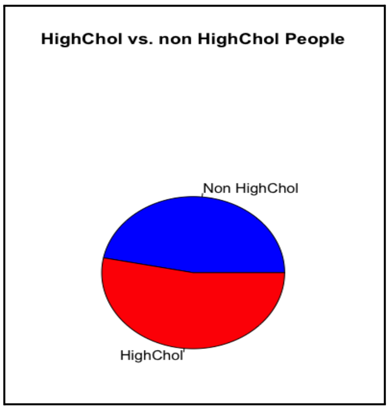
Figure 4 : Pie chart of HighBP Values *Figure 5 : Bar Charts of Diabetes Frequency for HighBP Values*

	Diabetic	Non-Diabetic
High-BP	26405 (75.23435%)	13042 (38.40400%)
Non High-BP	8692 (24.76565%)	20918 (61.59600%)

Table 1 : Contingency Table in Values and Proportions of Relationship between HighBP and Diabetes

Conclusion: Based on the above figures (3, 4) and tables (1) I’ve concluded that HighBP (elevated blood pressure) plays a significant role in the development of diabetes. As the prevalence of high blood pressure rises, so does the incidence of diabetes.

Specific Analysis of relationship between (HighChol) and (Diabetes_binary) show:



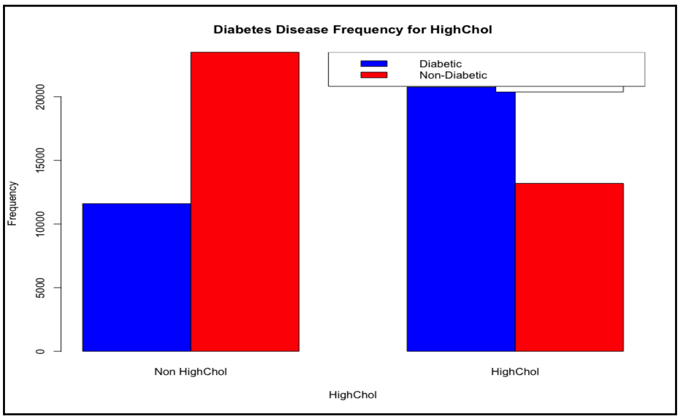


Figure 6 : Pie chart of HighChol Values *Figure 7 : Bar Charts of Diabetes Frequency for HighChol Values*

	Diabetic	Non-Diabetic
High-Chol	23496 (66.94589%)	13196 (38.85748%)
Non High-Chol	11601 (33.05411%)	20764 (61.14252%)

Table 2 : Contingency Table in Values and Proportions of Relationship between HighChol and Diabetes

Conclusion: Based on the above figures (5, 6) and tables (3, 4) I’ve concluded that elevated cholesterol levels play a significant role in the development of diabetes. As instances of high cholesterol rise, so do occurrences of diabetes.

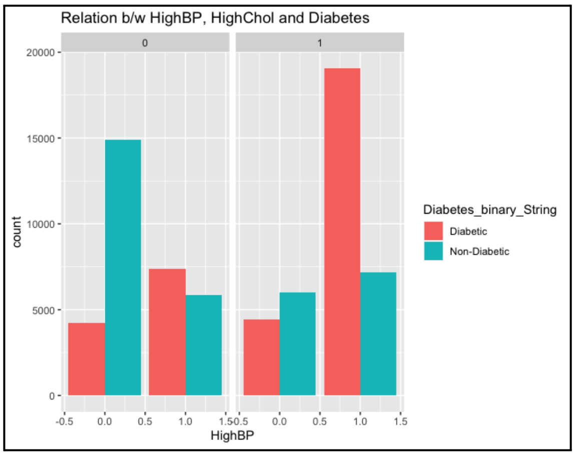
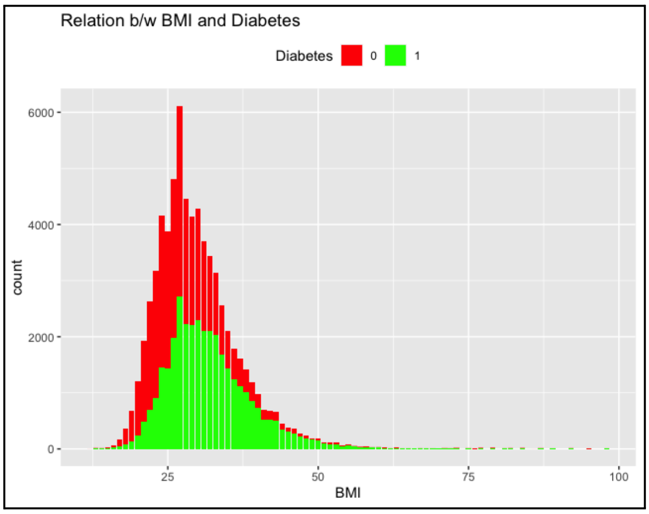


Figure 8 : Countplot of the BMI Variable faceted by Diabetes_binary

Analysis of combined effect of (HighChol & High_Chold) on (Diabetes_binary) show:

Based on the table (5) and figure (7), we see that the combination of having both HighBP and HighChol simultaneously elevates the likelihood of developing diabetes. Specific Analysis of (BMI) on (Diabetes_binary) show:



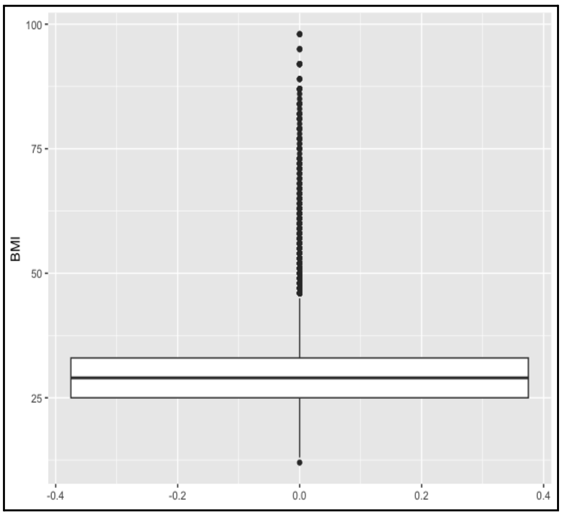


Figure 9 : Countplot of BMI Variable faceted by Diabetes   *Figure 10 : Boxplot of the BMI Variable*

Category 1 (0 < BMI <= 20)		*Category 2 (20 < BMI <= 50)*		*Category 3 (50 < BMI <= 100)*	
Diabetes	Non-Diabetes	Diabetes	Non-Diabetes	Diabetes	Non-Diabetes
545	1990	33849	31750	703	220

Table 3 : Contingency Table of the Categorised BMI Variable by Diabetes

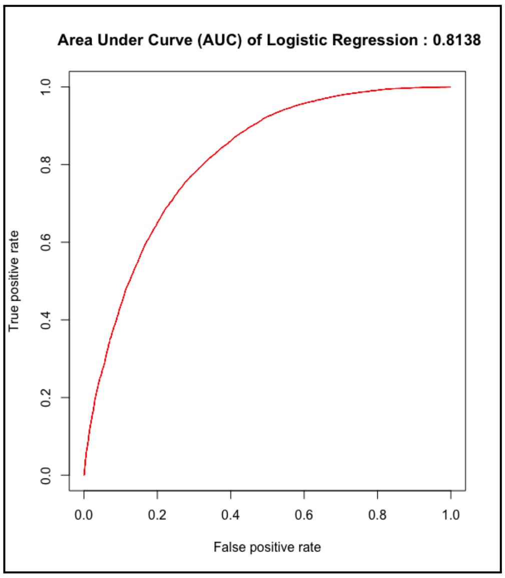
Countplot in figure (7), boxplot in figure (8) and categorised contingency tables of BMI in tables (5, 6, 7) shows that:

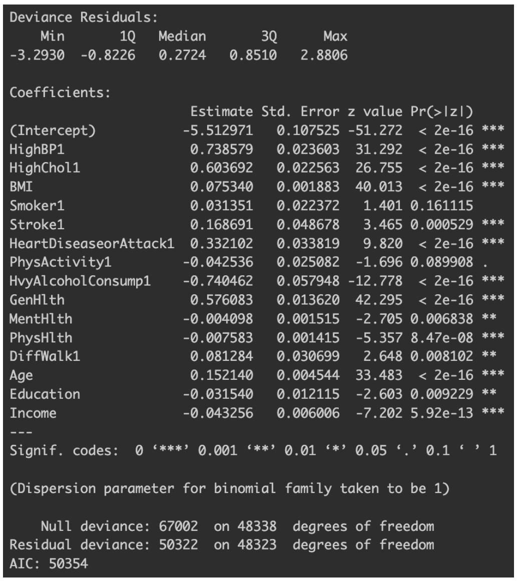
- Boxplot indicates the presence of outliers in the BMI variable.
- Countplot reveals that the majority of individuals with diabetes have a BMI between 20 & 50
- Out of 70,692 individuals, 65,599 (93%) fall within the BMI range of 20 to 50.
- Among 35,097 people with diabetes, 33,849 (96%) have a BMI between 20 and 50.
- Hence, it can be concluded that the second BMI group (20 < BMI <= 50) significantly influences diabetes.

Part 2 : Building the Classification Model

We assume that our priority is to correctly identify individuals who actually have diabetes and we are less worried about falsely predicting a diabetes diagnosis when the person actually does not have it. Therefore, correctly identifying diabetes cases is of greater importance than minimising false positives.

Classification Model 1 : (Logistic Regression) Model



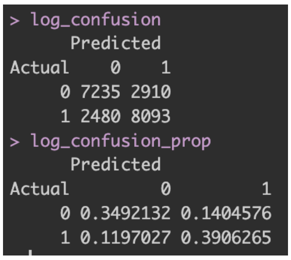


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2930  -0.8226   0.2724   0.8510   2.8806

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -5.512971   0.107525 -51.272  < 2e-16 ***
HighBP1                0.738579   0.023603  31.292  < 2e-16 ***
HighChol1              0.603692   0.022563  26.755  < 2e-16 ***
BMI                    0.075340   0.001883  40.013  < 2e-16 ***
Smoker1                0.031351   0.022372   1.401 0.161115
Stroke1                0.168691   0.048678   3.465 0.000529 ***
HeartDiseaseorAttack1  0.332102   0.033819   9.820  < 2e-16 ***
PhysActivity1         -0.042536   0.025082  -1.696 0.089908 .
HvyAlcoholConsump1    -0.740462   0.057948 -12.778  < 2e-16 ***
GenHlth                0.576083   0.013620  42.295  < 2e-16 ***
MentHlth              -0.004098   0.001515  -2.705 0.006838 **
PhysHlth              -0.007583   0.001415  -5.357 8.47e-08 ***
DiffWalk1              0.081284   0.030699   2.648 0.008102 **
Age                    0.152140   0.004544  33.483  < 2e-16 ***
Education             -0.031540   0.012115  -2.603 0.009229 **
Income                -0.043256   0.006006  -7.202 5.92e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67002  on 48338  degrees of freedom
Residual deviance: 50322  on 48323  degrees of freedom
AIC: 50354
```

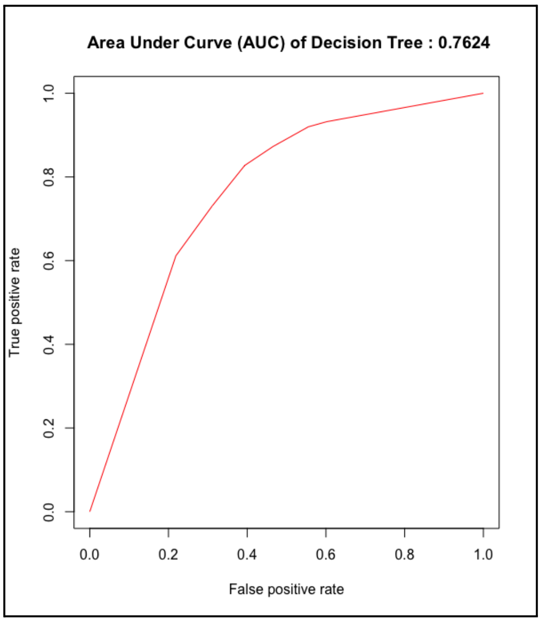


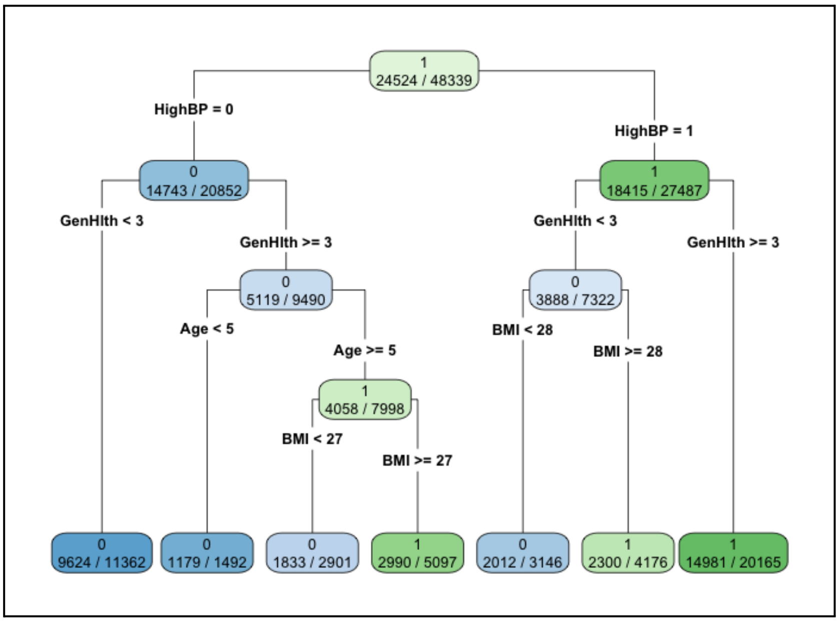
```
> log_confusion
      Predicted
Actual    0    1
     0 7235 2910
     1 2480 8093
> log_confusion_prop
      Predicted
Actual         0         1
     0 0.3492132 0.1404576
     1 0.1197027 0.3906265
```

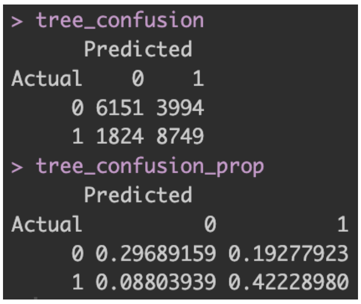
Accuracy=74.28% | Precision=73.52% | Recall=76.52% | Specificity=71.35%

Based on the above summary of the fitted logistic regression model, it appears to display quite a robust goodness of fit for predicting diabetes given this particular dataset. Firstly, the deviance residuals, which indicate the ability of the model to predict the dependent variable, range from -3.293 to 2.8806 and have a median of 0.2724, which suggests that the deviance residuals are relatively small in magnitude, scattered around zero and with a range not excessively large, therefore, indicating a good fit and that model's predictions closely align with the actual outcomes. In terms of coefficients, the significant positive coefficients such as HighBP, HighChol, BMI, Stroke, HeartDiseaseorAttack, GenHlth, DiffWalk, and Age, suggest that an increase in these variables is associated with an increased likelihood of having diabetes. In contrast, the significant negative coefficients such as HvyAlcoholConsump, MentHlth, PhysHlth, PhysActivity, Education, and Income, suggest that an increase in these variables is associated with a decreased likelihood of having diabetes. The fact that the p-values of most coefficients are well below 0.05 suggest that these variables are statistically significant in predicting diabetes. Largest coefficients correspond to variables BMI, Age, and GenHlth, suggesting that they will have substantial impact on the log-odds of diabetes. The fact that we have variables with both positive and negative coefficients, reflects the complexity of predicting diabetes and so having both positive and negative coefficient variables will contribute to a more comprehensive understanding of factors that influence diabetes in people. Lastly, the logistic regression model has a high AUC (Area Under the Curve) value of 0.8138, further indicating its effective ability to correctly distinguish between individuals with and without diabetes. The accuracy of 74.28% suggests a reasonably good overall correctness in classifying individuals as diabetic or non-diabetic. The precision of 73.52% shows the model's accuracy when predicting diabetes, aligning with the priority of correctly identifying individuals with the diabetes. In addition, the recall of 76.52% signifies the model's effectiveness in capturing a significant proportion of actual diabetes cases, minimising the risk of false negatives. Lastly, the specificity of 71.35% showcases the model's ability to accurately identify non-diabetic individuals. In conclusion, the logistic regression model demonstrates quite a good fit for the dataset, particularly in prioritising the correct identification of individuals with diabetes.

Classification Model 2 : (Decision Tree) Model





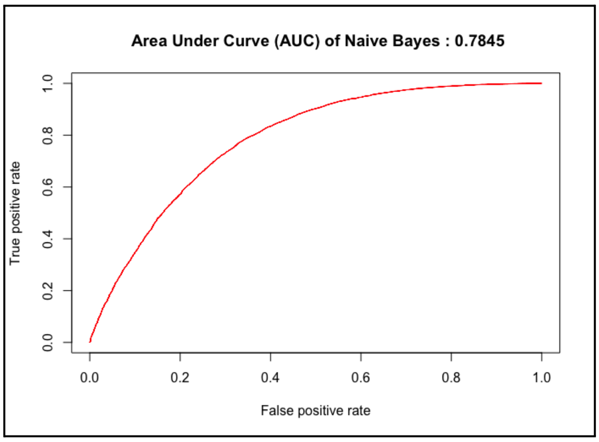


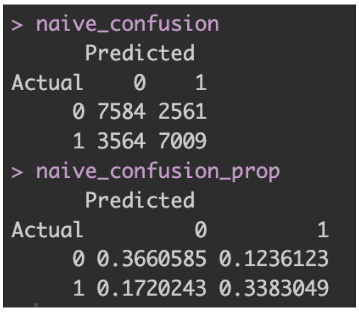
```
> tree_confusion
      Predicted
Actual    0    1
     0 6151 3994
     1 1824 8749
> tree_confusion_prop
      Predicted
Actual          0          1
     0 0.29689159 0.19277923
     1 0.08803939 0.42228980
```

Accuracy=67.13%
Precision=68.66%
Recall=82.73%
Specificity=60.63%

Based on the above summary of the fitted decision tree model, it appears to display a moderate level of goodness of fit, as evidenced by an AUC of 0.7624. Through analysis of the confusion matrix, it is seen that the model correctly predicts 67.13% of instances with. Furthermore, the precision of 68.66% suggests that when it predicts an individual has diabetes, it is correct 68.66% of the time. So, while not extremely high, the decision tree is able to provide a reasonable level of confidence in positive predictions. In addition, the recall of 82.73%, emphasises its effective ability to identify a significant portion of actual positive instances of diabetes, this is particularly important in our medical context of diabetes, where missing positive cases (false positives) could have significant consequences, therefore, this relatively high recall value aligns with our priority to correctly identifying individuals with diabetes as it shows its high ability to capture true positive instances. Lastly the specificity of 60.63%, displays its accuracy in predicting non-diabetic cases, while it is not as high as recall, it still indicates a reasonable ability to avoid false positives in non-diabetic predictions. In conclusion, the decision tree model demonstrates a satisfactory balance between precision and recall, taking into account that our focus is on correctly identifying individuals with diabetes, although the decision tree still loses to logistic regression overall in AUC metric.

Classification Model 3 : (Naive Bayes) Model



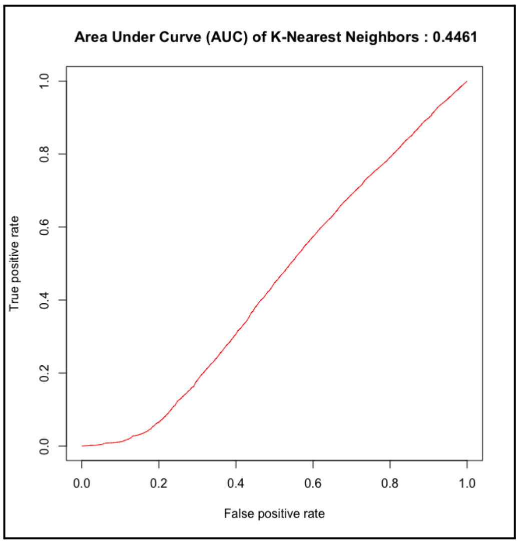


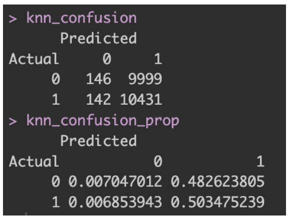
```
> naive_confusion
      Predicted
Actual    0    1
     0 7584 2561
     1 3564 7009
> naive_confusion_prop
      Predicted
Actual         0         1
     0 0.3660585 0.1236123
     1 0.1720243 0.3383049
```

Accuracy=70.44%
Precision=66.29%
Recall=73.24%
Specificity=68.00%

Based on the above summary of the fitted naive bayes model, it appears to demonstrate a reasonably good fit. With an AUC of 0.7845 the naive bayes model displays its ability to differentiate reasonably well between diabetic and non-diabetic cases. The confusion matrix reveals that the naive bayes model has an accuracy of 70.44%, indicating that over 70% of predictions are correct. Furthermore, the precision of 66.29% signifies that when the model predicts diabetes, it is correct approximately ⅔ of the time. Importantly, the recall of 73.24% suggests that the model effectively captures over 73% of the actual diabetes cases, which is a crucial aspect for our medical problem, as we prioritise correctly identifying actual diabetes cases over the concern for falsely identifying non-diabetes cases. Specificity of 68% depicts that 68% of non-diabetes cases get identified correctly.

Classification Model 4 : (KNN) Model





```
> knn_confusion
      Predicted
Actual     0     1
     0   146  9999
     1   142 10431
> knn_confusion_prop
      Predicted
Actual           0           1
     0 0.007047012 0.482623805
     1 0.006853943 0.503475239
```

Accuracy=51.05%
Precision=51.06%
Recall=98.66%
Specificity=50.69%

Based on the above summary of the fitted KNN model, we see that the low AUC value of 0.4461 suggests a modest discriminatory power of this KNN model for this dataset. However, a deeper analysis through the confusion matrix and the relevant metrics actually provide a more detailed understanding of its performance. What immediately stands out is the recall of 98.66% which indicates that the KNN model is able to effectively identify a large majority of individuals with actual diabetes correctly, which perfectly aligns with our priority of correctly identifying cases of actual diabetes, which is often most crucial in healthcare applications. However, the precision of 51.06% is relatively lower, indicating a notable number of false positives. Next, the specificity of 50.69% indicates that the KNN model has a tendency to incorrectly classify some non-diabetic individuals as diabetic. Finally, the overall accuracy of 51.05% suggests that the model's performance is only slightly better than random guessing. In conclusion, the overall performance of KNN model with this dataset is quite poor as it only excels in identifying individuals with diabetes and it comes at a big cost of a notable number of false positives.

Part 3 : Conclusion

As previously mentioned, we assumed that due to the medical context of the problem, our priority is to correctly identify individuals with actual diabetes. Therefore, when evaluating the performance of all the four classification machine learning models on predicting diabetes, the logistic regression model stood out as the most suitable choice. The reason for that is because the logistic regression model achieved an accuracy of 74.28%, with a precision of 73.52%, recall of 76.52%, and specificity of 71.35%. Furthermore, the deviance residuals indicate a good model fit and the significant coefficients of the variables provide valuable insights into the factors that influence diabetes. In addition, the AUC value of 0.8138 further supports the model's high effectiveness in distinguishing between individuals with and without diabetes. In contrast, the decision tree model, while displaying a reasonable balance between precision and recall, it still falls short of the logistic regression model with an AUC of 0.7624. The naive bayes model also performs reasonably well but does not surpass the logistic regression in terms of accuracy and AUC. The KNN model, despite having an impressive recall of 98.66%, suffers severely from low precision and low overall accuracy, making it less suitable for our priority of correctly identifying diabetes cases. In summary, the logistic regression model stands out as the most reliable choice for this problem with the provided dataset, because it offers a robust goodness of fit, significant insights into predictive independent variables and a high AUC value. Furthermore, its balanced performance across various metrics, particularly recall and specificity, aligns with our prioritisation on correctly identifying individuals with diabetes while minimising false positives. When considering the unique advantages of the models, the logistic regression seems particularly effective in this application due to its ability to consider the varying importance of different variables through their varying coefficients and its ability to effectively handle a mix of categorical and numerical variables for more accurate predictions. Lastly, another important advantage of using a logistic regression model in this context is its interpretability of coefficients which provides people the ability to interpret how changes in each of the independent variables influence the likelihood of having diabetes. This interpretability can be crucial in a medical context, for doctors and scientists to be able to understand the impact of different factors on the risk of their patient having diabetes.