DEVELOPMENT OF A SMOKING CESSATION SUPPORT SYSTEM

MILESTONE: Project Report

GROUP 90

Sai Srinivas Madamraju

Sri Mahalakshmi Harika Punati

Contact number

(609)-803-1239

(510)-460-8178

Email

madamraju.s@northeastern.edu

punati.s@northeastern.edu

Percentage of effort contribution by Sai Srinivas Madamraju: 50%

Percentage of effort contribution by Sri Mahalakshmi Harika Punati: 50%

Signature of student 1: Sai Srinivas Madamraju

Signature of student 2: Sri Mahalakshmi Harika Punati

Submission Date: 04-21-2023

Table of Contents

Problem Setting:

Smoking causes an estimated around 7 million deaths annually, which is a serious public health hazard. It is a major contributor to avoidable deaths and is linked to a variety of illnesses, such as heart disease, chronic obstructive pulmonary disease (COPD), and lung cancer. Additionally, smoking not only harms the smoker, but it also affects people nearby who are exposed to second-hand smoke. The financial burden of smoking on society is also significant, with billions of dollars spent each year on healthcare costs related to smoking-related illnesses. Despite the well-known risks of smoking and the availability of several initiatives to help people quit, many people still smoke. Therefore, there is a need for effective interventions to help individuals quit smoking and prevent the initiation of smoking in young people.

Problem Definition:

Despite the availability of evidence-based treatment for smoking cessation, success rates remain low, with only a small percentage of participants achieving abstinence. It is important to educate them about the serious medical issues brought on by smoking in order to increase the likelihood of success. The model analyses the inputs and determines if the subject smokes or not. However, using each of these variables separately can provide results that are hard for patients and medical professionals to understand and use. To understand the likelihood that each smoker would successfully quit, a prediction model employing machine learning techniques may be more efficient.

Data Source:

Development of a smoking cessation support system data set has been taken from smoker status prediction dataset using Kaggle datasets. The data description is from the below mentioned source.
(**https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction**)

Data Description:

The dataset consists of 38,985 rows and 23 columns. It contains 5 categorical variables and 18 numerical variables. The columns are comprised with data like age : 5-years gap, height(cm), weight(kg), waist(cm) : Waist circumference length, eyesight(left), eyesight(right),

hearing(left), hearing(right), systolic : Blood pressure, relaxation : Blood pressure, fasting blood sugar, Cholesterol : total, triglyceride, HDL : cholesterol type, LDL : cholesterol type, haemoglobin, Urine protein, serum creatinine, AST : glutamic oxaloacetic transaminase type, ALT : glutamic oxaloacetic transaminase type, Gtp : γ-GTP, dental caries, smoking.

Data Mining Tasks:

A. Data Understanding

The original dataset contains 38983 instances, with 22 parameters and one target variable named "Smoking". The dataset comprises 4 categorical variables and 18 numeric variables. Four of the five categorical variables are binary, meaning they have values of either 0 or 1. The fifth variable has a range of values from 1 to 6. The target variable is also binary, where the value of 1 indicates smoking and 0 indicates non-smoking. There are 14318 instances of smoking and 24666 instances of non-smoking in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38984 entries, 0 to 38983
Data columns (total 23 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   age                  38984 non-null  int64
 1   height(cm)           38984 non-null  int64
 2   weight(kg)           38984 non-null  int64
 3   waist(cm)            38984 non-null  float64
 4   eyesight(left)       38984 non-null  float64
 5   eyesight(right)      38984 non-null  float64
 6   hearing(left)        38984 non-null  int64
 7   hearing(right)       38984 non-null  int64
 8   systolic             38984 non-null  int64
 9   relaxation           38984 non-null  int64
 10  fasting blood sugar  38984 non-null  int64
 11  Cholesterol          38984 non-null  int64
 12  triglyceride         38984 non-null  int64
 13  HDL                  38984 non-null  int64
 14  LDL                  38984 non-null  int64
 15  hemoglobin           38984 non-null  float64
 16  Urine protein        38984 non-null  int64
 17  serum creatinine     38984 non-null  float64
 18  AST                  38984 non-null  int64
 19  ALT                  38984 non-null  int64
 20  Gtp                  38984 non-null  int64
 21  dental caries        38984 non-null  int64
 22  smoking              38984 non-null  int64
```

Figure 1..Data Information

B. Data Pre-processing

Since there are no missing values, null values, or inconsistencies in the dataset, and the data is already in a standard format, further tasks can be carried out using the same dataset. The dataset contains 38983 rows and 23 columns.

Data Exploration:

A. Categorical Variables

The Seaborn library's 'countplot' function is used to create graphical representations of categorical variables, showing the frequency of each category. The 'hearing(left)' and 'hearing(right)' variables have two categories, where '1' denotes a person who can hear and '2' indicates a person who cannot. For the 'smoking' and 'dental caries' variables, '0' indicates non-existence and '1' indicates existence. The 'Urine protein' variable has six categories, ranging from '1' to '6', which describe different levels of protein found in the urine.

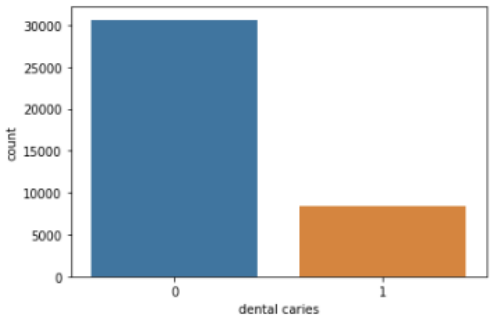

Figure 2..Dental Caries vs Count.

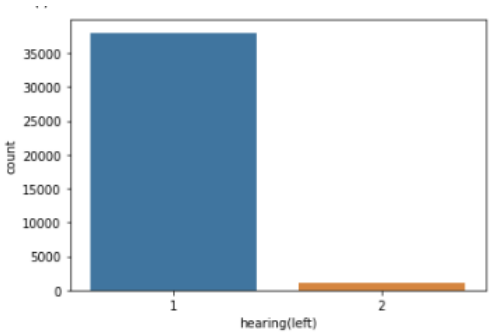

Figure 3..Hearing(left) vs Count

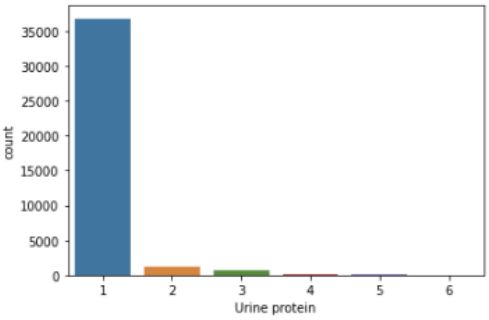

Figure4..Urine Protein vs Count.

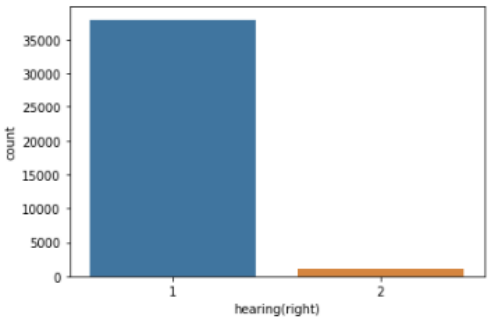

Figure 5..Hearing(right) vs Count

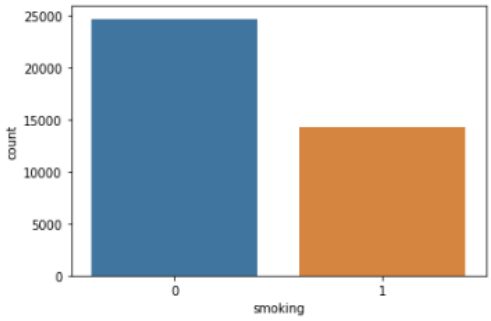

Figure 6..Smoking vs Count

Chi-Square test for correlation analysis of categorical variables

The Chi-Square test involves the formulation of a null hypothesis and an alternative hypothesis related to the correlation between variables.

1. Null hypothesis H0: The variable has no correlation
2. Alternate hypothesis H1: The variable has a correlation

If the resulting p-value from the test is greater than 0.05, we fail to reject Null hypothesis, and it is concluded that the variables are not correlated. However, if the p-value is less than 0.05, the null hypothesis is rejected, and it is concluded that the variables are correlated.

The chi-Square test is performed between target categorical variable ‘Smoking’ and other categorical variables.

Result and conclusion

The p-value for all the categorical variables is less than 0.05, it indicates that there is a statistically significant correlation between the categorical variables and the target variable 'smoking'. Therefore, it can be concluded that the categorical variables are correlated with the 'smoking' variable.

```
Chi-Square test result for Smoking and Dental caries
p value is 4.8096792160578O4e-100
Corelated (reject H0)
Chi-Square test result for Smoking and Urine protein
p value is 0.049511278397280714
Corelated (reject H0)
Chi-Square test result for Smoking and hearing(right)
p value is 0.00020202395358092 26
Corelated (reject H0)
Chi-Square test result for Smoking and hearing(left)
p value is 1.5214418715754427e-05
Corelated (reject H0)
```

Figure 7..Chi-Square test results

B. Numeric Variables

Numeric variables are represented using histograms, with the majority of variables showing right-skewed distributions. However, the 'haemoglobin' and 'height' variables are slightly left-skewed. Additionally, the 'age', 'height', and 'weight' variables exhibit multimodal distributions.

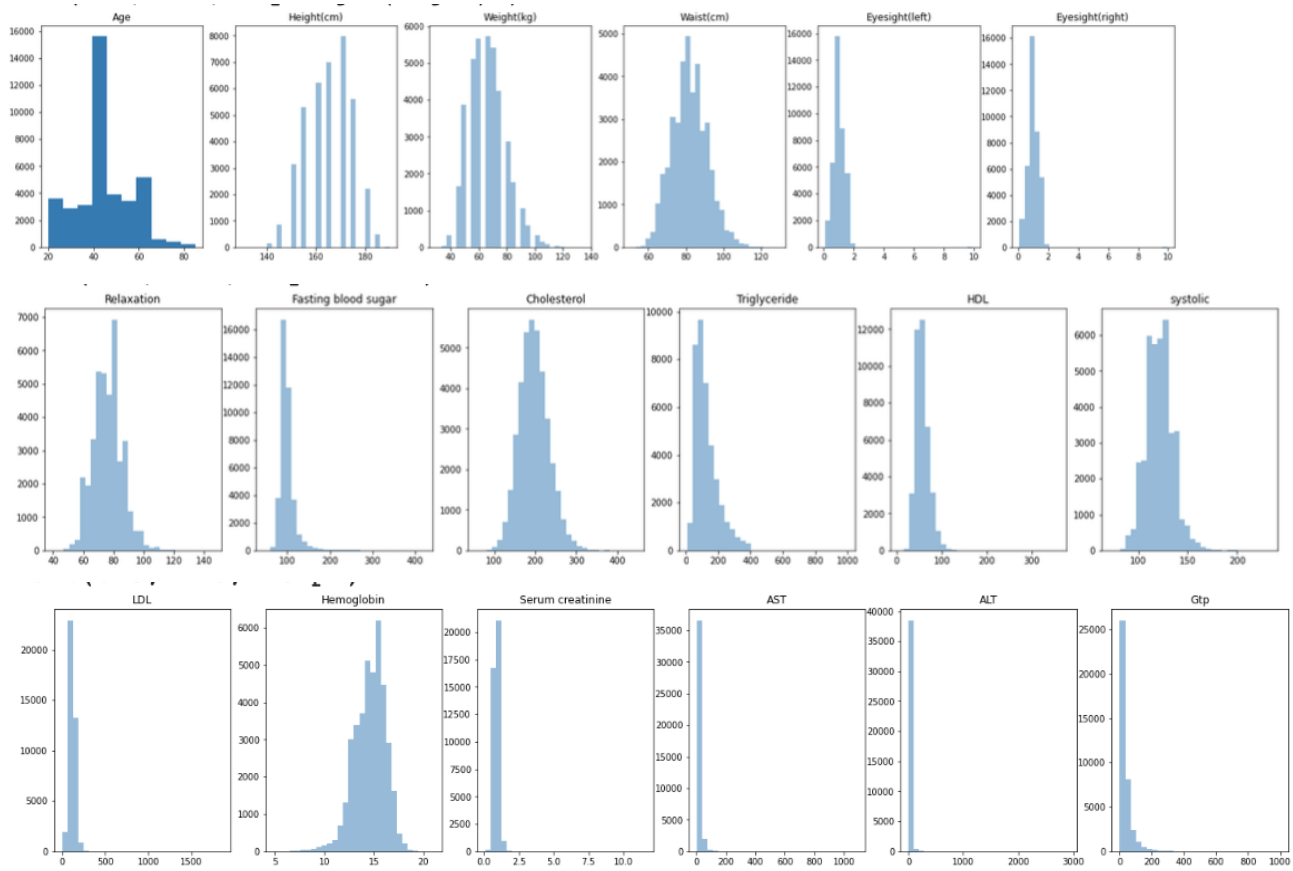

Figure 8..Numeric variables Plots before removing outliers

Based on the visualization presented above, it appears that there are a few data points that lie far outside the typical range. It may be beneficial to exclude these outliers in order to create a more effective model with improved performance. Upon removal of the outliers from the dataset, the data size decreased by 1171 observations. These are significantly less when compared to the original dataset. It appears that the original dataset contained duplicate rows, and after removing these 5362 duplicates, the new dataset has 32451 unique rows. The below graphs shows the visualization after removal of outliers.

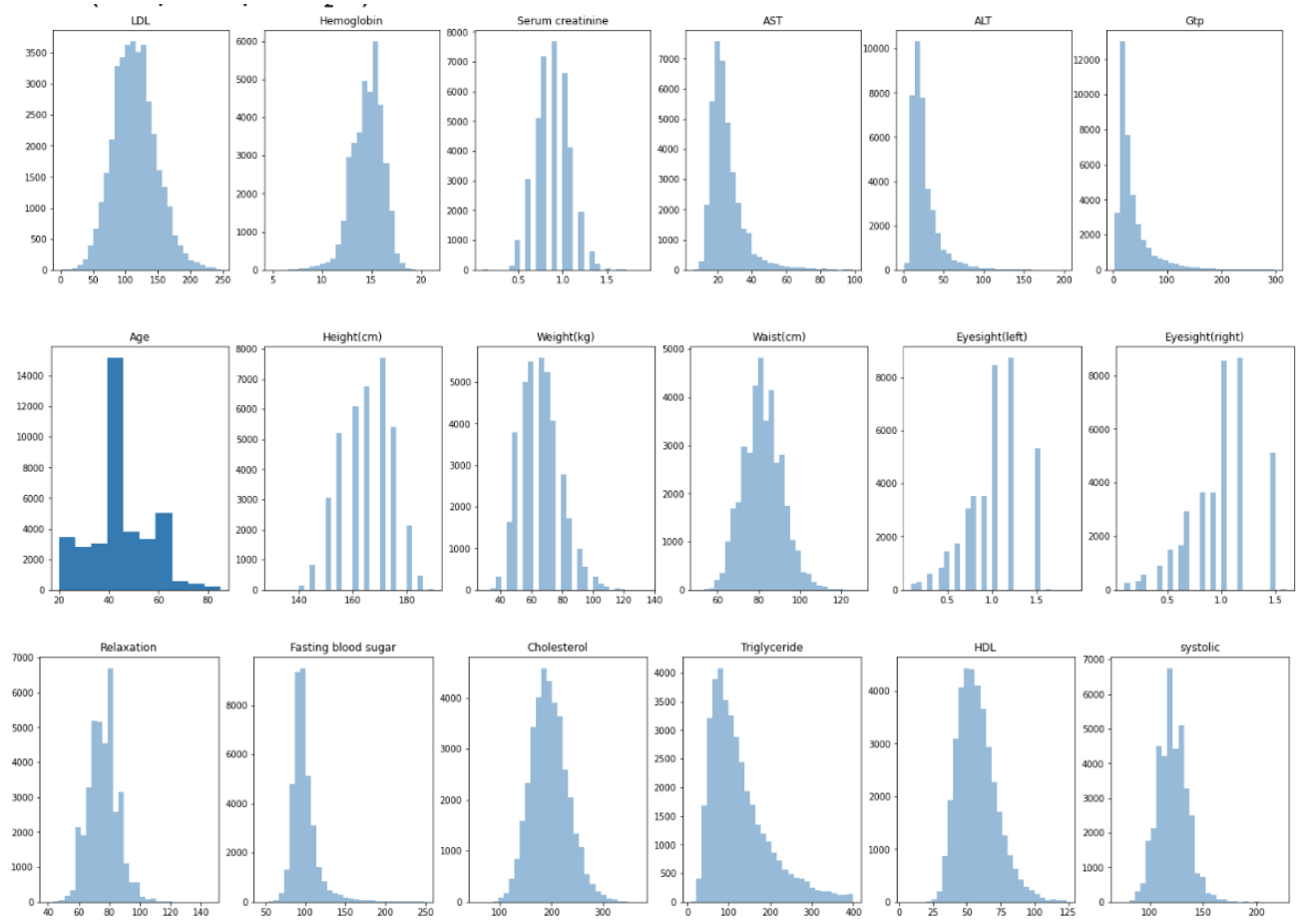

Figure 9..Numeric variables plots after removing outliers

Correlation analysis of numeric variables using Pearson's correlation

The heatmap illustrates the correlation between numerical variables. It indicates that there are certain variables that have a strong correlation with each other like Waist and Weight with correlation coefficient of 0.82, Relaxation and Systolic with correlation coefficient of 0.76, LDL and Cholesterol with correlation coefficient of 0.9. There are variables which are negatively correlated like Triglyceride and HDL with correlation coefficient of -0.43, Height

and Age with correlation coefficient of -0.48. These highly correlated variables can provide redundant information and may not be necessary to include in the analysis. Therefore, it could be advantageous to remove some of these variables in order to simplify the model and potentially improve its accuracy.

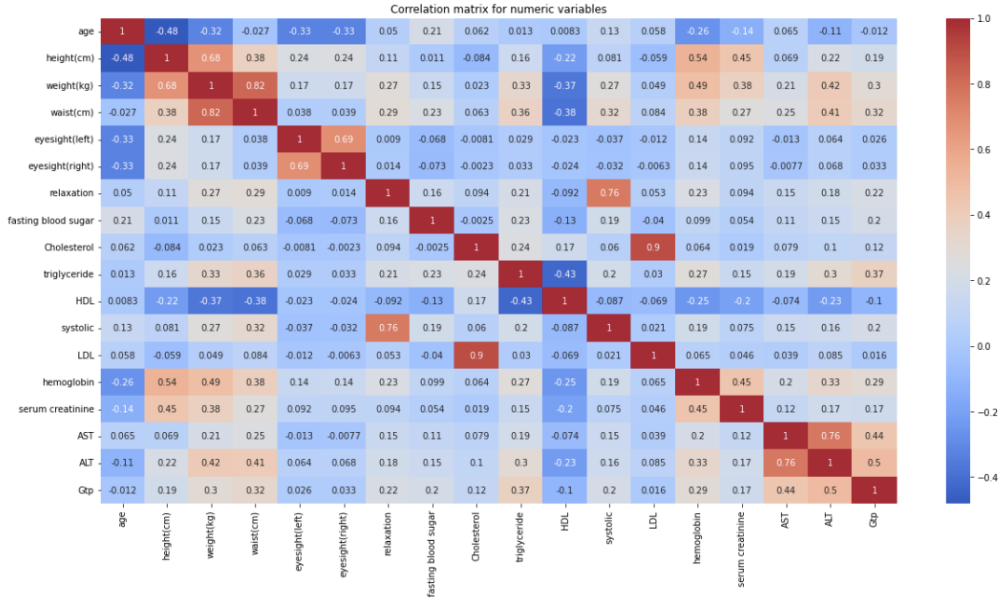

Figure 10..Heatmap showing correlation between numerical variables

Dimension Reduction:

The four variables 'waist(cm)', 'systolic', 'Cholesterol', 'AST' from the above data set are removed. The heatmap shows the correlation between variables after removing the variables.

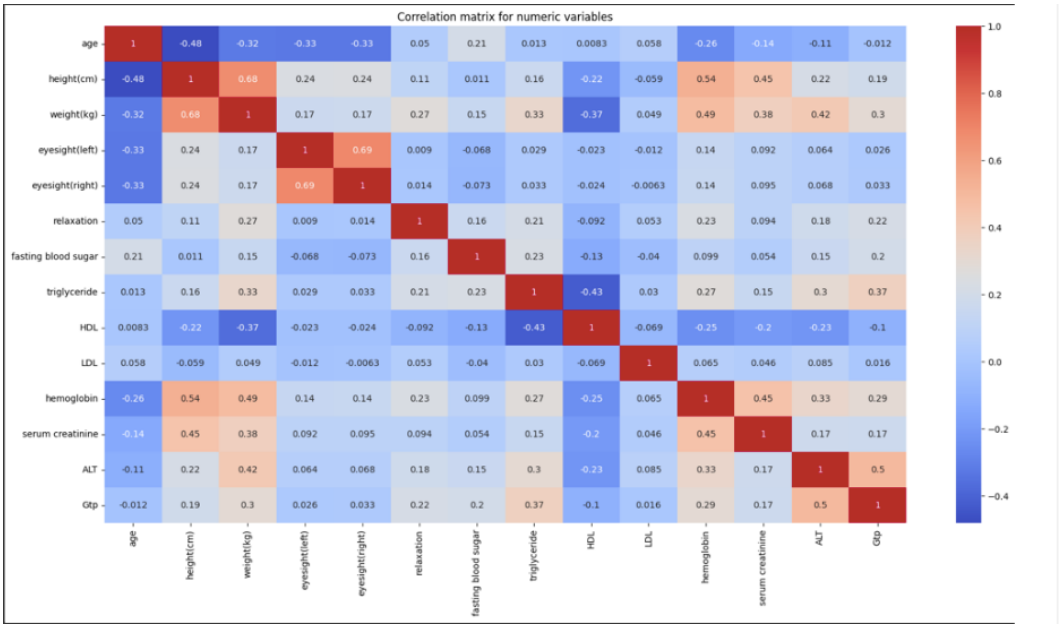

Figure 11..Heatmap showing correlation between numerical variables

Point Biserial test for correlation analysis of numeric and categorical target variable

To evaluate the correlation between a binary target variable and a continuous variable in a dataset, the Point Biserial Correlation test formulates a Null Hypothesis and an Alternative Hypothesis:

1. Null Hypothesis (H0): There is no correlation between the continuous variable and the target variable.
2. Alternative Hypothesis (H1): There is a correlation between the continuous variable and the target variable.

If the resulting p-value from the test is greater than 0.05, we fail to reject Null hypothesis, and it is concluded that the variables are not correlated. However, if the p-value is less than 0.05, the null hypothesis is rejected, and it is concluded that the variables are correlated.

The Point Biserial test is performed between target categorical variable ‘Smoking’ and other numeric variables.

Result and conclusion

The p-value for all the categorical variables is less than 0.05, it indicates that there is a statistically significant correlation between the numerical variables and the target variable 'smoking'. Therefore, it can be concluded that the numeric variables are correlated with the 'smoking' variable.

```
Point-biserial correlation result for Smoking and age
Point-biserial correlation coefficient: -0.16932003540717697
P-value: 2.851830580380366e-207
Correlated (reject H0)
Point-biserial correlation result for Smoking and height(cm)
Point-biserial correlation coefficient: 0.3957231677997295
P-value: 0.0
Correlated (reject H0)
Point-biserial correlation result for Smoking and weight(kg)
Point-biserial correlation coefficient: 0.3030128002908918
P-value: 0.0
Correlated (reject H0)
Point-biserial correlation result for Smoking and eyesight(left)
Point-biserial correlation coefficient: 0.09535593715018403
P-value: 2.009153736703375e-66
Correlated (reject H0)
Point-biserial correlation result for Smoking and eyesight(right)
Point-biserial correlation coefficient: 0.10408779508383485
P-value: 7.423698239432036e-79
Correlated (reject H0)
Point-biserial correlation result for Smoking and relaxation
Point-biserial correlation coefficient: 0.10271846702778656
P-value: 7.821518695686961e-77
Correlated (reject H0)
Point-biserial correlation result for Smoking and fasting blood sugar
Point-biserial correlation coefficient: 0.09499050920611098
P-value: 6.298966635265718e-66
Correlated (reject H0)
Point-biserial correlation result for Smoking and triglyceride
Point-biserial correlation coefficient: 0.2489267522197282
P-value: 0.0
Correlated (reject H0)
Point-biserial correlation result for Smoking and HDL
Point-biserial correlation coefficient: -0.18321969058712761
P-value: 6.083515653313439e-243
Correlated (reject H0)
Point-biserial correlation result for Smoking and LDL
Point-biserial correlation coefficient: -0.05684539705552884
P-value: 1.2067548862549454e-24
Correlated (reject H0)
Point-biserial correlation result for Smoking and hemoglobin
Point-biserial correlation coefficient: 0.4008163424929112
P-value: 0.0
Correlated (reject H0)
Point-biserial correlation result for Smoking and serum creatinine
Point-biserial correlation coefficient: 0.25058092348469024
P-value: 0.0
Correlated (reject H0)
Point-biserial correlation result for Smoking and ALT
Point-biserial correlation coefficient: 0.15861909007699465
P-value: 7.840452173195513e-182
Correlated (reject H0)
Point-biserial correlation result for Smoking and Gtp
Point-biserial correlation coefficient: 0.2867833165025298
P-value: 0.0
Correlated (reject H0)
```

Figure 12..Point-biserial correlation test results

Data Partitioning:

The target variable, "Smoking," was represented using the variable "y," and the predictor variables as a whole were represented using the variable "X." The dataset was divided and resulting train and test sets have a ratio of 80:20. According to this, 80% of the dataset was used to train the classification models, while the remaining 20% used to evaluate the performance of models. X_test comprised 8281 records and 18 variables while, X_train contained 33,121 records and 18 variables. 33,121 records and 1 variable were present in y_train, while the model-building process employed the y_test selection.

Data Mining Methods:

The training data mentioned above was used to create four data mining classification models, which are explained below.

A. K-NN Classifier

The main concept behind the K-NN classification algorithm is to categorize new records from test data by comparing them to comparable records in the training data. The majority decision rule is used to categorize the new record as a member of the majority class of the K-neighbors after determining the similar records using the K-nearest neighbors.

Implementation

The nearest K-NN classifier (k=3, k=5, k=7) was used. The weights metrics "Uniform" and "Distance" and the distance metrics "Euclidean" and "Manhattan" were employed in each iteration as extra arguments for finding the ideal value of K. The nearest K-NN classifier k=3, weight metrics ‘Distance’ and distance metrics ‘Euclidean’ was used. The accuracy of the model is 76.4%.

B. Decision Tree Classifier

The decision tree classifier utilizes a tree-like model to predict outcomes and is a data-driven, non-parametric approach. Prediction splits are produced by trees. Trees separate records into

subgroups and produce logical rules that are simple to understand by splitting predictors in such a way that homogeneity grows with each split.

Implementation

The decision tree model was built with min_samples_split=5. The accuracy of the model is 74%.

C. Logistic Regression

A specific model linking the predictor variables with the target variables is what is required for the parametric classification model known as logistic regression, which produces an outcome variable that is categorical in nature. Limit on the likelihood of being assigned to either of the classes. The logit outcome variable can be conceptualized as a linear function of the predictors.

Implementation

The logistic regression model was built by considering random_state = 250. The accuracy of the model is 71.2%.

D. Random Forest

The Random Forest is a collective approach that makes predictions using a collection of several decision trees as opposed to individual models. Each model must be superior to a random classifier and must be able to produce predictions that are independent of the others.

Implementation

The Random Forest model was built by considering n_estimators = 100. The accuracy of the model is 81%.

Performance Evaluation:

A. K-NN Classifier

The model gives the best accuracy value of 76.4% with best metrics of k=3, weight metric = 'Distance', distance metric = 'Euclidean'. The sensitivity value showing the ability to correctly classify true positives is 88%. The specificity value showing the ability to correctly classify

true negatives is 64.4%. The value of F1 indicating low false positives and false negatives is 78.8%. The ROC Curve is moderately closer to the top left corner, indicating the model is good.

```
Accuracy: 0.7647627098176548
F1 score: 0.7883528900478053
Sensitivity: 0.8883447600391773
Specificity: 0.6445079818918275
```

Figure 13..Performance Evaluation results of KNN Classifier

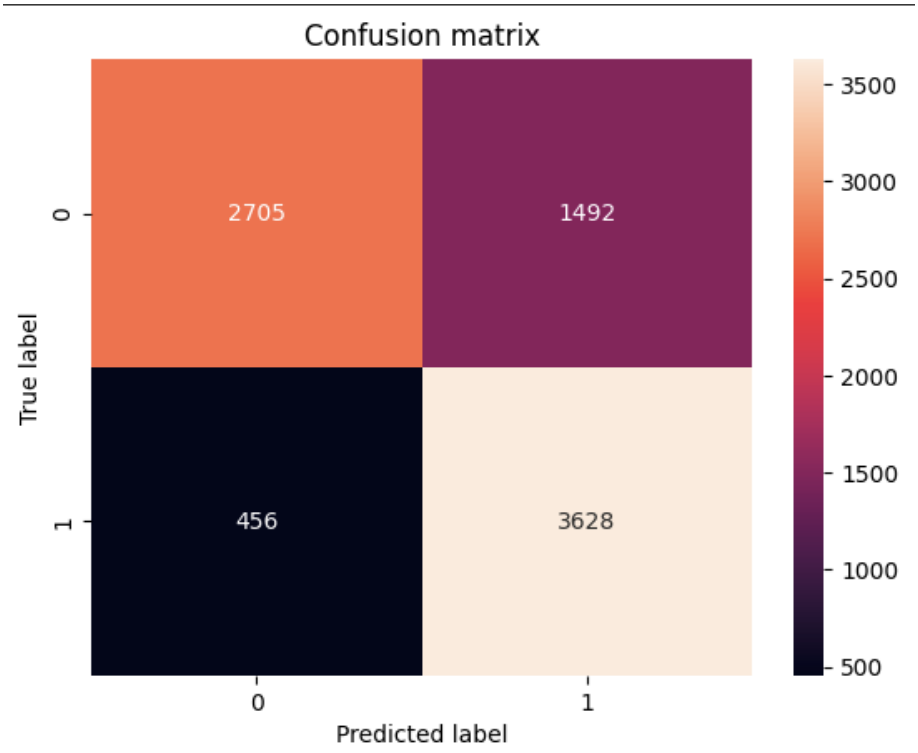

Figure 14..Confusion Matrix of KNN Classifier

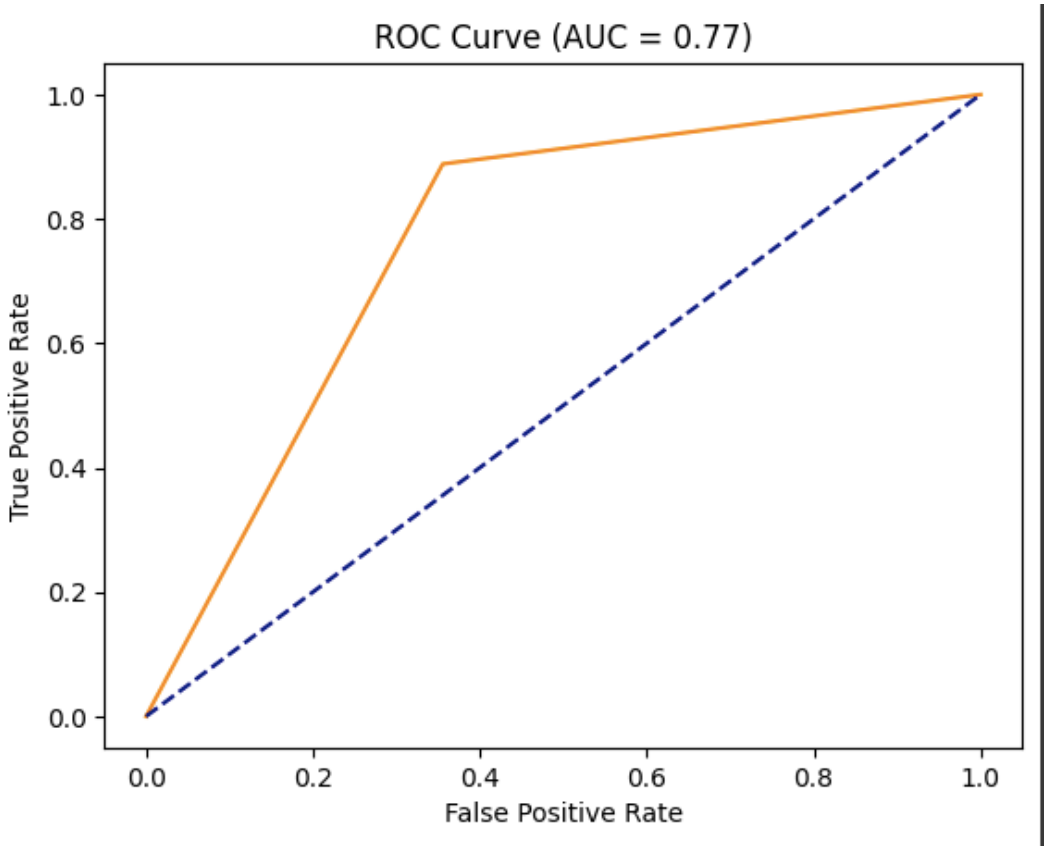

Figure 15..ROC Curve and AUC value of KNN Classifier

B. Decision Tree Classifier

The model gives the best accuracy value of 74% with min_samples_split=5. The sensitivity value showing the ability to correctly classify true positives is 74.1%. The specificity value showing the ability to correctly classify true negatives is 73.8%. The value of F1 indicating low false positives and false negatives is 73.8%. The ROC Curve is moderately closer to the top left corner, indicating the model is good.

```
Accuracy: 0.7402487622267842
F1 score: 0.7380343441724516
Sensitivity: 0.7419196865817825
Specificity: 0.7386228258279723
```

Figure 16..Performance Evaluation results of Decision Tree Classifier

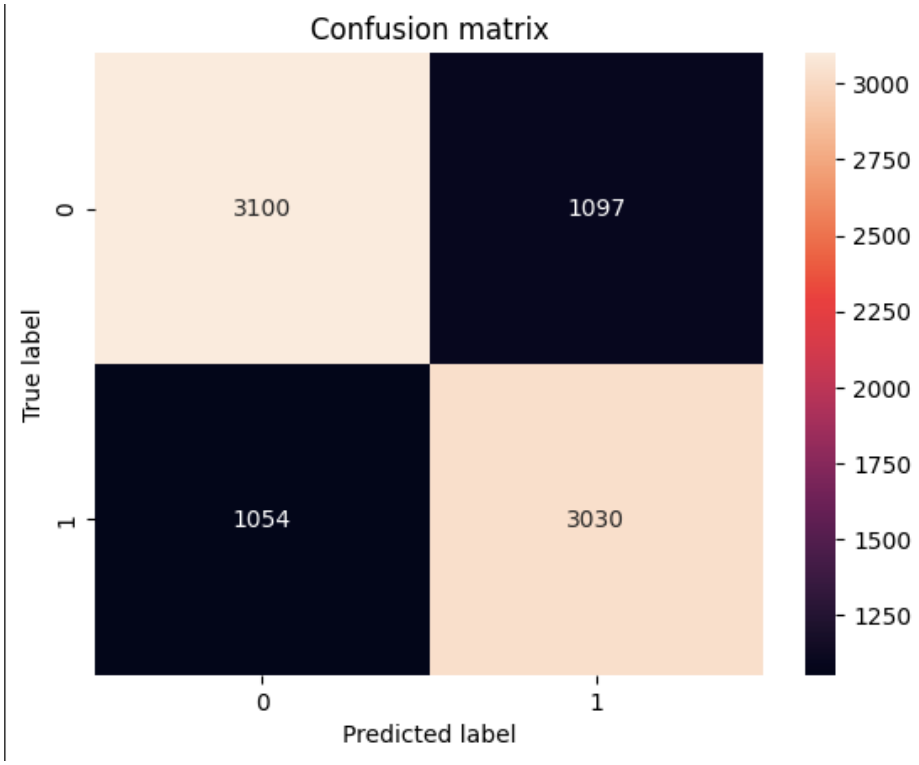

Figure 17..Confusion Matrix of Decision Tree Classifier

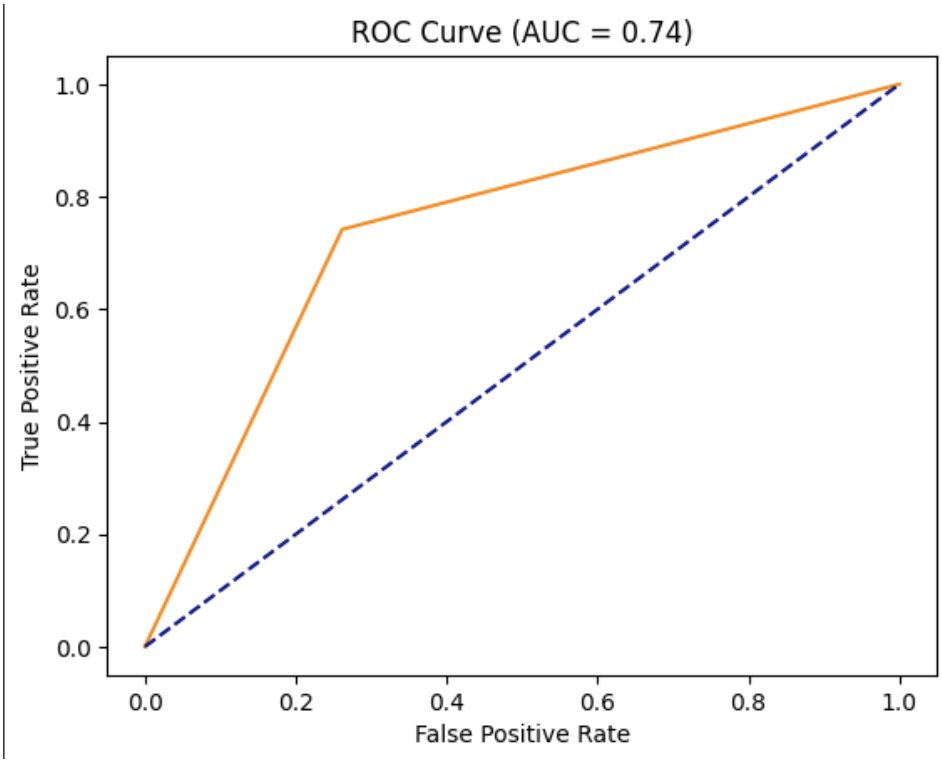

Figure 18..ROC Curve and AUC value of Decision Tree Classifier

C. Logistic Regression

The model gives the best accuracy value of 71.2% with random_state = 250. The sensitivity value showing the ability to correctly classify true positives is 73.7%. The specificity value showing the ability to correctly classify true negatives is 68.8%. The value of F1 indicating low false positives and false negatives is 71.7%. The ROC Curve is moderately closer to the top left corner, indicating the model is good.

```
Accuracy: 0.7128366139355l5
F1 score: 0.71703950499762O3
Sensitivity: 0.7377571008814887
Specificity: 0.6885870860138193
```

Figure 19..Performance Evaluation results of Logistic Regression

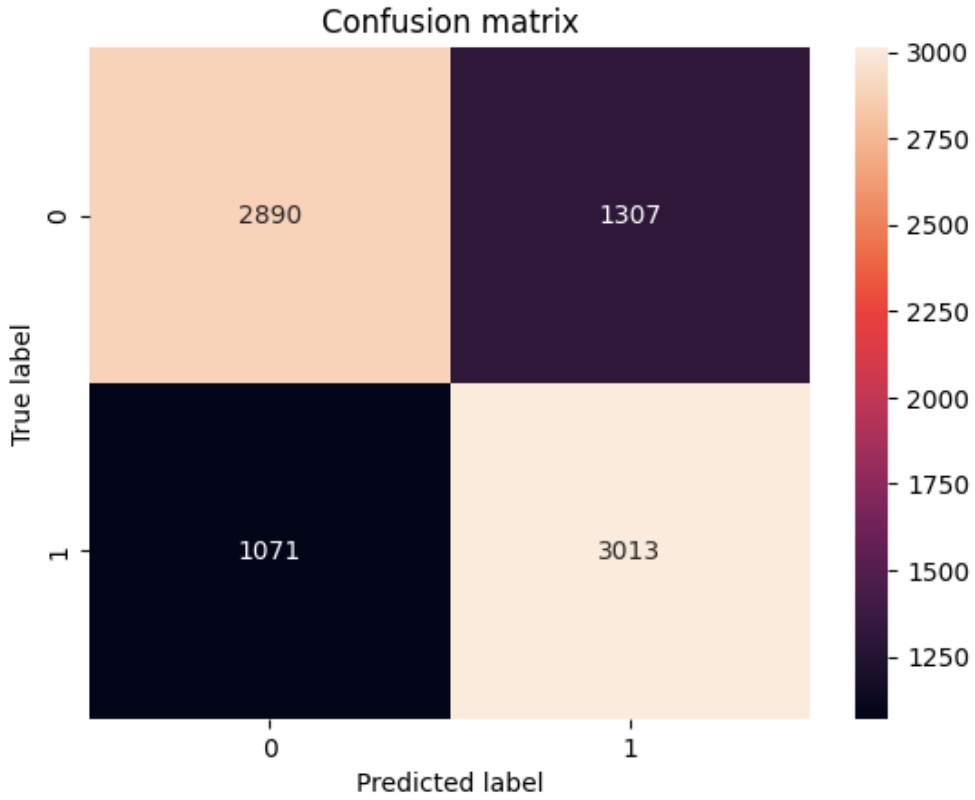

Figure 20..Confusion Matrix of Logistic Regression

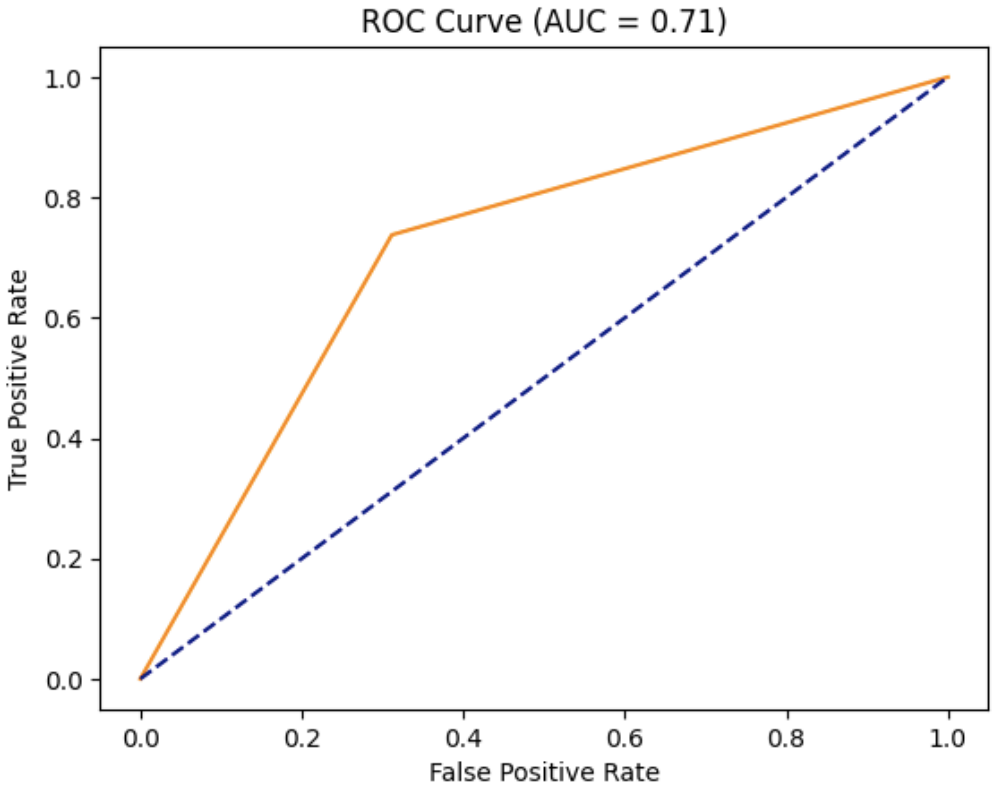

Figure 21..ROC Curve and AUC value of Logistic Regression

D. Random Forest

The model gives the best accuracy value of 81% with n_estimators = 100. The sensitivity value showing the ability to correctly classify true positives is 85.9%. The specificity value showing the ability to correctly classify true negatives is 76%. The value of F1 indicating low false positives and false negatives is 81.6%. The ROC Curve is moderately closer to the top left corner, indicating the model is good.

```
Accuracy: 0.8092017872237652
F1 score: 0.8163217856312485
Sensitivity: 0.8596963761018609
Specificity: 0.7600667143197523
```

Figure 22..Performance Evaluation results of Random Forest Classifier

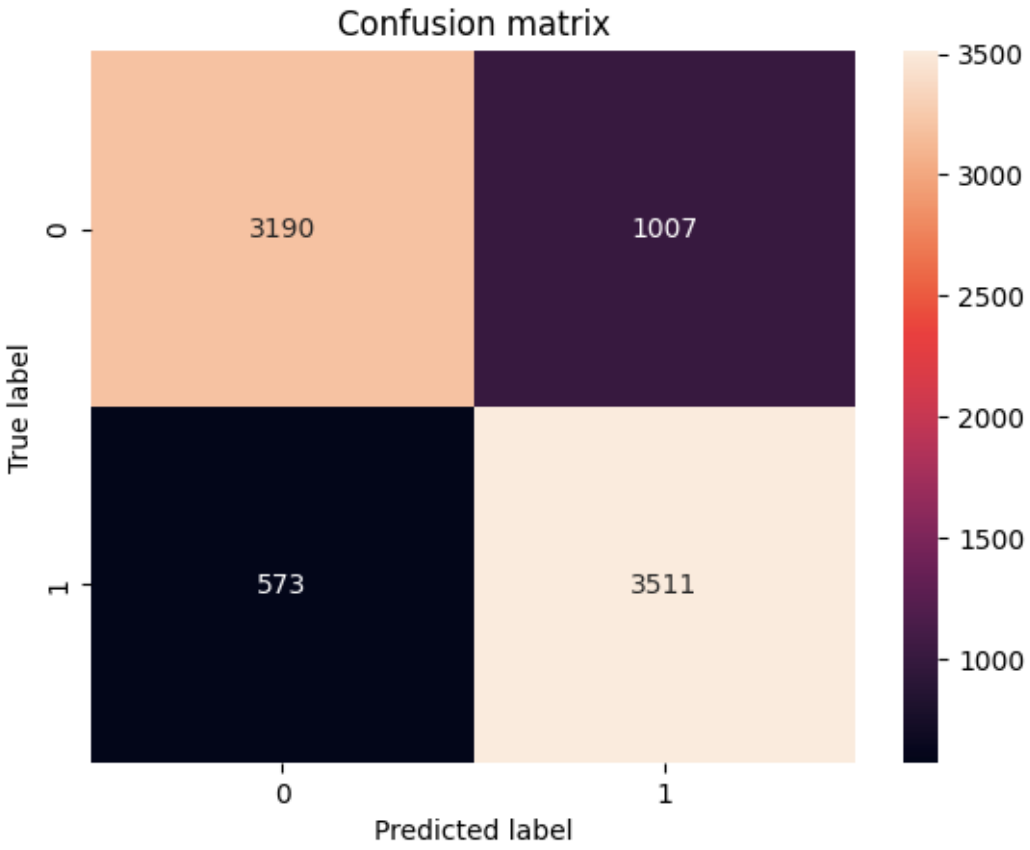

Figure 23..Confusion Matrix of Random Forest Classifier

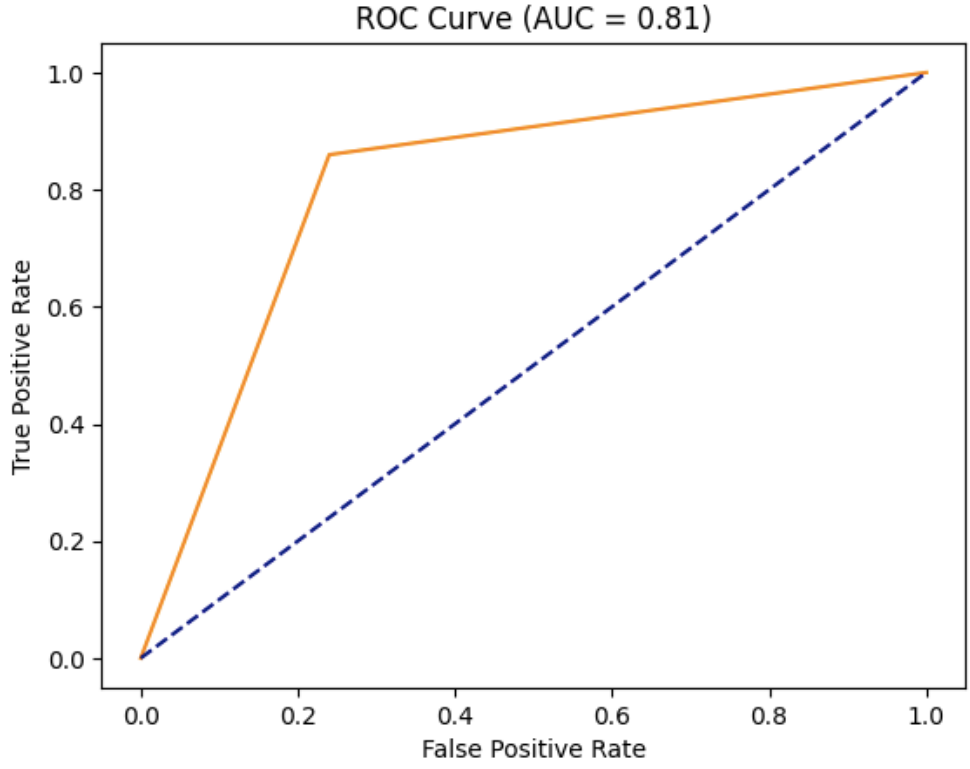

Figure 24..ROC Curve and AUC value of Random Forest Classifier

Project Results:

The Random forest model has high accuracy of 81% when compared to other models. The Sensitivity, Specificity and F1 values are also higher for this model with values 85.9%, 76%, 81.6 % respectively. The figure below shows the comparision of accuracy of each model.

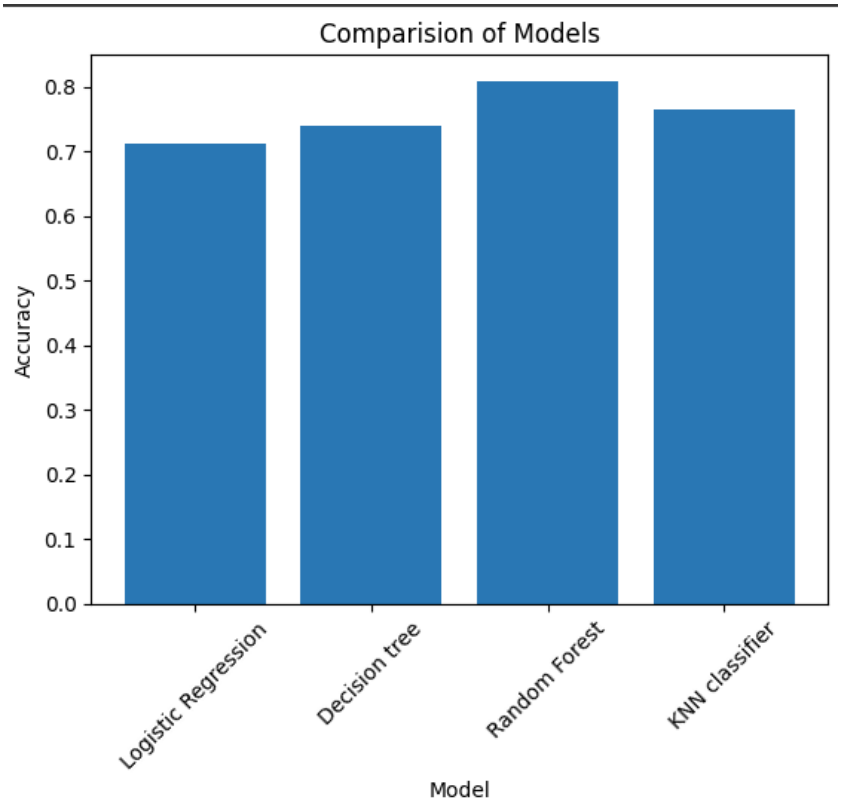

Figure 25..Comparision of accuracy of different models

Model	Accuracy	F1 Score	Sensitivity	Specificity
K-NN Classifier	76.4%	78.8%	88.8%	64.4%
Decision Tree Classifier	74.0%	73.8%	74.1%	73.8%
Logistic Regression	71.2%	71.7%	73.7%	68.8%
Random Forest	80.9%	81.6%	85.9%	76.0%

Figure 26..Comparision of different parameters of different models

Impact of Project Outcomes:

The project aimed to develop a model that can accurately identify whether an individual smokes or not. The purpose of this model is to provide medical assistance to people who have smoking habits. The results showed that the random forest model outperformed all other models with high accuracy, sensitivity, specificity, and F1 values.

The ability to accurately determine smoking status is important for medical professionals to develop effective treatment plans for individuals who want to quit smoking or manage their smoking habits. The model developed in this project can assist healthcare professionals in identifying smokers and provide them with targeted interventions to help them quit smoking.

The high accuracy, sensitivity, specificity, and F1 values of the random forest model indicate that it can accurately identify smokers and non-smokers with a high degree of confidence. The model's performance can be attributed to the ability of the random forest algorithm to handle complex relationships between features in the data set.

Overall, the success of this project in identifying smoking status has significant implications for public health interventions aimed at reducing the prevalence of smoking. The use of this model can provide targeted support to individuals who smoke and help them quit smoking, thus improving their overall health outcomes.

References:

Gaurav Dutt. (2021). Smoker Status Prediction [Data set]. Kaggle. https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction