CS 5805: Machine Learning I

FINAL TERM PROJECT

Submitted by: Shutonu Mitra

Table of Contents

Table of Figures

Table of Tables

Abstract

The Final Term Project (FTP) presents a comprehensive exploration of machine learning methodologies, unfolding in four phases to analyze and model a Smoking and Drinking Dataset with Body Signal. The project encompasses feature engineering, regression analysis, classification, clustering, and association rule mining. Notable phases include feature selection using Random Forest Analysis, PCA, SVD, and VIF, regression analysis emphasizing stepwise regression and model significance, and a thorough classification analysis involving various classifiers with a focus on misclassification patterns. The project also explores clustering with K-means and association rule mining using the Apriori algorithm. Key findings include the identification of influential features, the development of a robust regression model, and the recommendation of Bagging with Random Forest as the best-performing classifier. Future work suggestions encompass advanced feature engineering, model tuning, data augmentation, and exploration of deep learning techniques for improved classification performance. The project's systematic approach provides a holistic understanding of the dataset, yielding valuable insights and recommendations for future research.

Introduction:

Embarking on a comprehensive exploration of machine learning methodologies, the Final Term Project (FTP) unfolds in multiple phases to meticulously analyze and model a selected dataset. The overarching objectives encompass feature engineering, regression analysis, classification, and an independent study on clustering and association rule mining. This introduction provides an outline of the procedures and report structure envisioned to accomplish these FTP objectives in four phases:

I. Phase I: Feature Engineering & EDA

1. Implementing missing data imputation, resolve data duplications, and perform aggregation or down-sampling as necessary.
2. Utilizing Random Forest Analysis, PCA, SVD, VIF for feature selection, and provide observations on the chosen method.
3. Applying discretization, binarization, and variable transformation techniques like normalization, standardization, and differencing.
4. Generating heatmap displays for the sample covariance matrix and Pearson correlation coefficients matrix, including observations.
5. Addressing imbalanced data and present the methodology employed for achieving balance.

II. Phase II: Regression Analysis

1. Performing prediction on a continuous numerical feature and visualize the train, test, and predicted variables in a plot.
2. Presenting the final model and develop a table with key metrics such as R-squared, adjusted R-square, AIC, BIC, and MSE.
3. Conducting T-test and F-test analyses, confidence interval analysis, stepwise regression, and adjusted R-square assessment.

III. Phase III: Classification Analysis

1. Appling decision tree, logistic regression, KNN, SVM (linear, polynomial, radial base), naive Bayes, random forest, and neural network classifiers.
2. Conducting grid search to determine optimal parameters for each classifier.
3. Displaying confusion matrix, precision, sensitivity/recall, specificity, F-score, ROC and AUC curve, and Stratified K-fold cross-validation results.

IV. Phase IV: Clustering and Association

1. Appling K-mean or K-mean++, DBSCAN, and Apriori algorithms to the dataset.
2. Providing insights and observations derived from clustering and association rule mining analyses.

The comprehensive exploration across these phases aims to unravel the intricacies of the selected dataset, employing a robust set of machine learning techniques. Each phase contributes uniquely to the understanding, modeling, and evaluation of the dataset, culminating in a holistic analysis.

Dataset Description:

The Smoking and Drinking Dataset with Body Signal is sourced from the National Health Insurance Service in Korea, emphasizing privacy and data security. It is designed for predictive analysis, focusing on body signal data to identify smoking and drinking behaviors among individuals. The Dataset Purpose are given below:

1. **Analysis of Body Signals:** Comprehensive physiological and health-related attributes enable insights into the relationship between body signals and smoking/drinking behaviors.
2. **Classification of Smokers and Drinkers:** The dataset facilitates the development of predictive models to classify individuals as smokers or drinkers based on their body signal data.

The dataset comprises various variables, including age, height, weight, blood pressure, cholesterol levels, and more. Essential for understanding physiological characteristics, it also features information on smoking status, drinking habits, and diverse body signal measurements. The detailed description of the dataset is given below:

Table 1: Dataset Feature Description

Column/Feature	Description
Sex	male, female
Age	round up to 5 years
Height	round up to 5 cm (cm)
Weight	(kg)
Sight_left	eyesight (left)
Sight_right	eyesight (right)
Hear_left	hearing left, 1 (normal), 2 (abnormal)
Hear_right	hearing right, 1 (normal), 2 (abnormal)
SBP	Systolic blood pressure (mmHg)
DBP	Diastolic blood pressure (mmHg)
BLDS	Blood Sugar Level or Fasting Blood Glucose (mg/dL)
Tot_chole	total cholesterol (mg/dL)
HDL_chole	High-Density Lipoprotein cholesterol (mg/dL)
LDL_chole	Low-Density Lipoprotein cholesterol (mg/dL)
Triglyceride	triglyceride (mg/dL)
Hemoglobin	hemoglobin (g/dL)
Urine_protein	protein in urine, 1 (-), 2 (+/-), 3 (+1), 4 (+2), 5 (+3), 6 (+4)

Column/Feature	Description
Serum_creatinine	serum (blood) creatinine (mg/dL)
SGOT_AST	Serum Glutamate Oxaloacetate Transaminase Aspartate Transaminase (IU/L)
SGOT_ALT	Serum Glutamate Oxaloacetate Transaminase Alanine Transaminase (IU/L)
Gamma_GTP	γ-Glutamyl Transpeptidase (IU/L)
SMK_stat_type_cd	Smoking state, 1 (never), 2 (used to smoke but quit), 3 (still smoke)
DRK_YN	Drinker or Not

Selected Dependent and Independent Variables*:*

1. **Dependent Variables:**

    - **Regression:** tot_chole (Total Cholesterol)
    - **Classification:** SMK_stat_type_cd (Smoking State)

2. **Independent Variables:**

    - Sex, Age, Height, Weight
    - Sight_left, Sight_right
    - Hear_left, Hear_right
    - SBP (Systolic Blood Pressure), DBP (Diastolic Blood Pressure)
    - BLDS (Blood Sugar Level), Hemoglobin
    - Urine_protein, Serum_creatinine
    - SGOT_AST, SGOT_ALT, Gamma_GTP
    - DRK_YN (Drinker or Not)

This dataset is crucial for health analytics, aiding in understanding the impact of body signals on smoking and drinking behaviors. The predictive models developed can contribute to targeted interventions and personalized health strategies. Additionally, it provides valuable insights for public health policies and interventions.

Phase 1:

- Target Feature: **SMK_stat_type_cd**

Attributes:

- Attributes: All the columns in the dataset, including **sex, age, height, weight, waistline, sight_left, sight_right, SBP, DBP, BLDS, tot_chole, HDL_chole, LDL_chole, triglyceride, hemoglobin, urine_protein, serum_creatinine, SGOT_AST, SGOT_ALT, gamma_GTP,**.

Data Preprocessing:

Data Cleaning:

Checking for missing values is a critical step in the data preprocessing phase of machine learning tasks. It ensures data quality, model performance, and the ethical and regulatory compliance of machine learning applications. Proper handling of missing values contributes to the robustness and reliability of machine learning models. The dataset used in this project has no missing values. The presence of missing values in every column have been checked and no missing instances are found

Duplicate Data:

Duplicate data can introduce several problems in machine learning tasks, and it's crucial to address them for reliable model training and accurate predictions. Duplicate data can lead to overfitting, where the model learns the training data too well, capturing noise and specific patterns that are unique to the duplicated instances. As a result, the model may not generalize well to unseen data, leading to poor performance on new and real-world examples. The dataset has been checked for the presence of duplicated rows and no duplicate instances are found.

Anomaly Detection/Outlier Analysis:

Anomaly detection is crucial for identifying rare and potentially harmful events or outliers within datasets. By plotting the histogram plots of the columns, we can find significant anomaly with several features. The goal of the below matrix of histograms (Figure 1) is to provide a visual representation of the data distribution for each independent variable. On first sight, it is pretty clear that the given dataset is not clean and contains outliers. Major outliers can be detected in the columns waistline, HDL_chole, LDL_chole, triglyceride, and most of the other columns.

In Python, various methods are available to detect and manage outliers in a dataset. These methods play a crucial role in identifying and addressing data points that deviate significantly from the majority of the data. Outliers can distort statistical analyses and machine learning models, making their detection and handling essential in data preprocessing. The method used

The Z-Score method is a statistical tool used to standardize and quantify the deviation of a data point from the mean of a dataset. It is calculated by subtracting the mean of the dataset from the individual data point and then dividing this difference by the standard deviation. This process results in a dimensionless score known as the Z-Score, which indicates how many standard deviations a data point is away from the mean. A Z-Score of 0 means the data point is exactly at the mean, while positive and negative Z-Scores represent data points above and below the mean, respectively. Typically, Z-Scores are used for outlier detection, with values significantly higher (greater than a predefined threshold) or lower (less than the negative of the threshold) suggesting potential outliers in the dataset. It provides a standardized way to assess and compare data points' deviations across different datasets, making it a versatile tool in statistical analysis and anomaly detection. About 5604 entries were removed due to the removal of outliers.

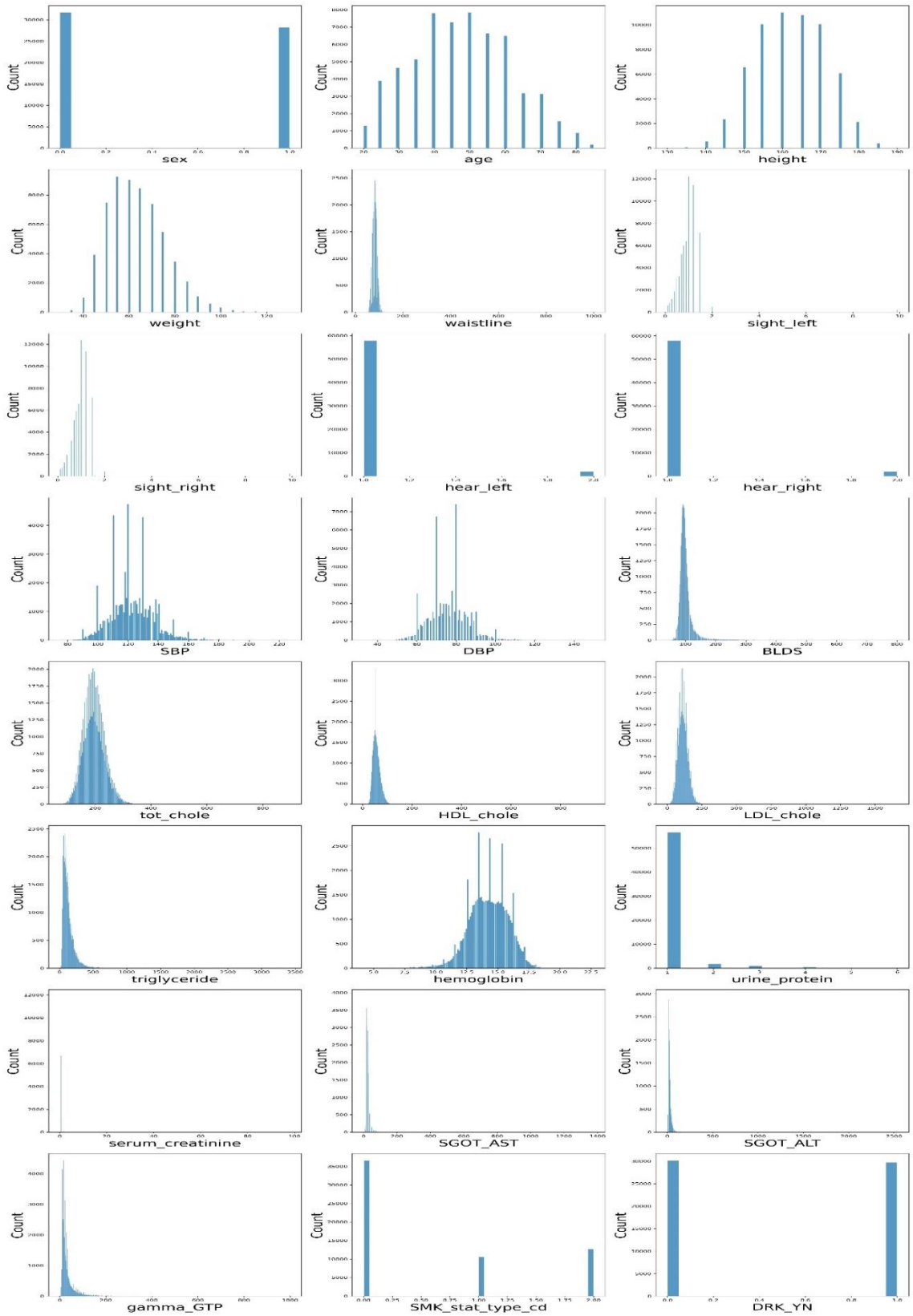


Figure 1: Histogram Plot of Numerical Variables

In Python, various methods are available to detect and manage outliers in a dataset. These methods play a crucial role in identifying and addressing data points that deviate significantly from the majority of the data. Outliers can distort statistical analyses and machine learning models, making their detection and handling essential in data preprocessing. The method used

The Z-Score method is a statistical tool used to standardize and quantify the deviation of a data point from the mean of a dataset. It is calculated by subtracting the mean of the dataset from the individual data point and then dividing this difference by the standard deviation. This process results in a dimensionless score known as the Z-Score, which indicates how many standard deviations a data point is away from the mean. A Z-Score of 0 means the data point is exactly at the mean, while positive and negative Z-Scores represent data points above and below the mean, respectively. Typically, Z-Scores are used for outlier detection, with values significantly higher (greater than a predefined threshold) or lower (less than the negative of the threshold) suggesting potential outliers in the dataset. It provides a standardized way to assess and compare data points' deviations across different datasets, making it a versatile tool in statistical analysis and anomaly detection. About 5604 entries were removed due to the removal of outliers.

Balanced or Imbalanced Data:

Balancing a dataset is crucial for developing fair, unbiased, and generalizable machine learning models that accurately represent the underlying patterns in the data, especially when dealing with imbalanced class distributions. For the target class SMK_stat_typ_cd, the dataset is highly imbalanced (Figure 2).

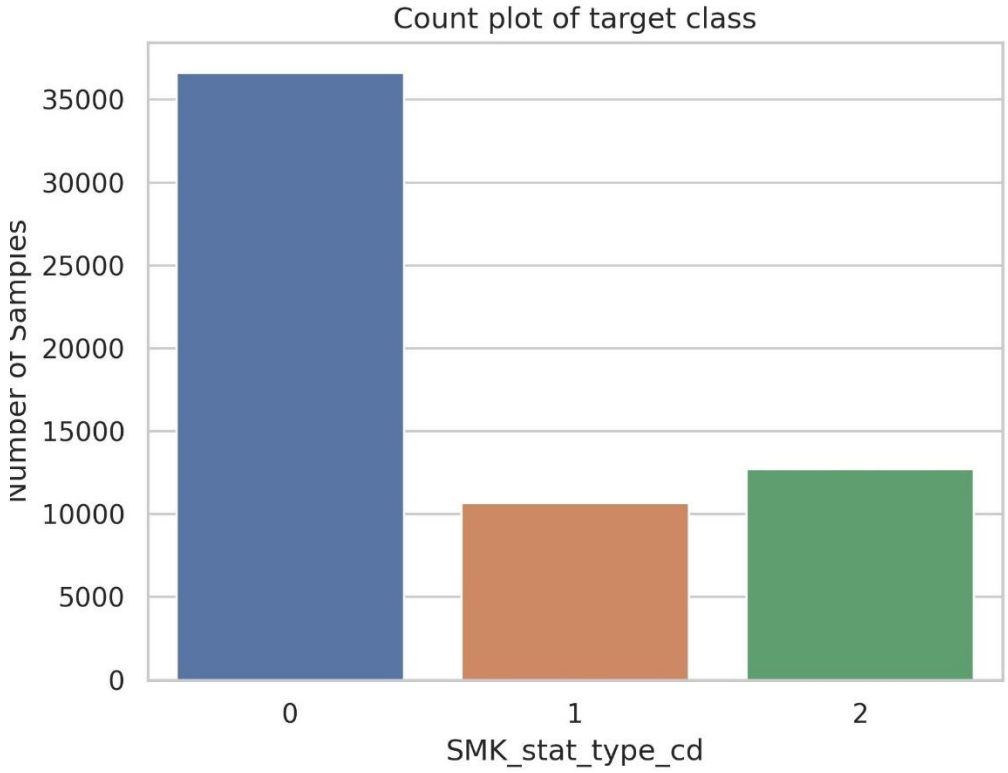


Figure 2: Count Plot of Target Class

Down Sampling:

Downsampling addresses imbalanced data by reducing the number of instances in the majority class to match the minority class. This helps balance class distribution, preventing the model from being biased towards the majority class and improving its ability to learn and generalize across all classes. In order to balance the dataset, it was downsampled, reducing the entries from 54,396 to 31,962. The count plot of the target column after downsampling is given below:

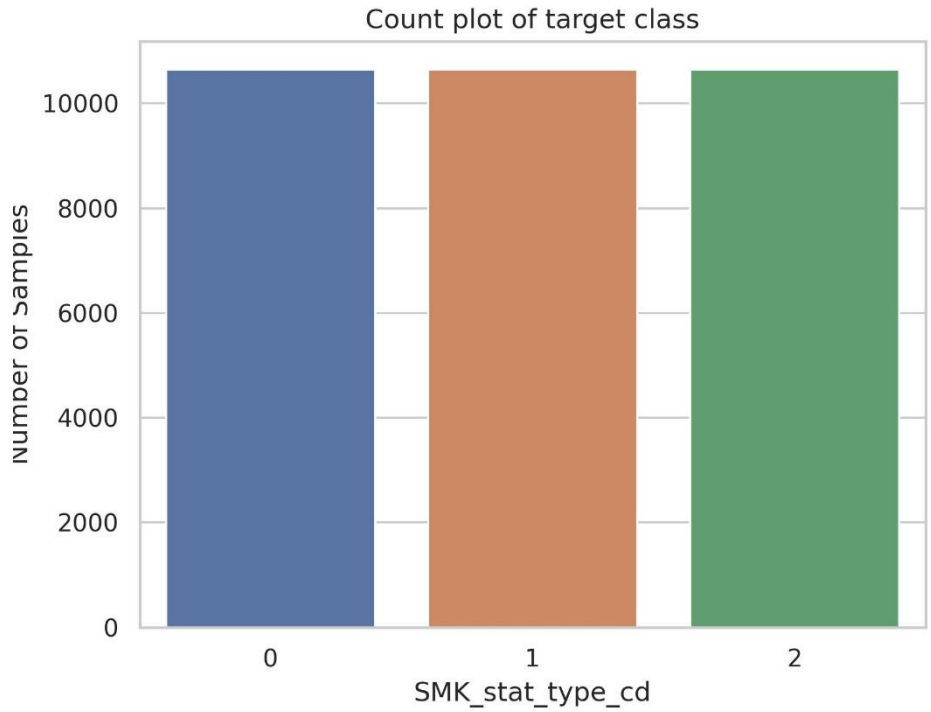


Figure 3: Count plot of target class (Balanced)

Discretization & Binarization/One hot Encoding:

Discretization involves converting continuous variables into discrete bins, aiding the handling of numeric data. Binarization transforms variables into binary values based on a specified threshold. On the other hand, one-hot encoding enhances machine learning tasks by converting categorical variables into binary columns, enabling algorithms to efficiently interpret and utilize categorical information in the dataset. In this project no discretization /Binarization was used. One-hot encoding was applied to four categorical variables : 'sex', 'hear_left','hear_right' and 'urine_protein', creating binary columns for each category. The dummy variable trap in One-Hot Encoding (OHE) occurs when one categorical variable's values can be predicted from the others. This redundancy can lead to multicollinearity issues in regression models. To avoid the trap, one binary column is dropped for each categorical feature during OHE. The dataset of after OHE is provided below:

```
============Step5:Discretization & Binarization:one hot encoding==================
Data after OHE:    age  height  weight  waistline  sight_left  sight_right  SBP  DBP  BLDS  tot_chole  HDL_chole  LDL_chole  triglyceride  hemoglobin  serum_creatinine
0   70     165      65          91           0            0  119   67   141        170         32        102           179          14                 0        22
1   55     170      90          97           1            1  124   79   104        187         43         95           244          16                 1        21
2   40     160      60          72           9            1  136   82   151        180         55         86           195          15                 1        24
3   55     160      70          82           1            0  136   87   127        192         41        107           222          15                 0        19
4   55     170      70          76           1            1  120   80   106        168         53        100            73          14                 0        27
```

```
  SGOT_AST  SGOT_ALT  gamma_GTP  SMK_stat_type_cd  DRK_YN  sex_1  hear_left_2.0  hear_right_2.0  urine_protein_2.0  urine_protein_3.0  urine_protein_4.0  urine_protein_5.0
  27         32                2                 0       0      0              0               0                  0                  0                  0
  34         35                2                 1       0      0              0               0                  0                  0                  0
  19         31                2                 1       0      0              0               0                  0                  0                  0
  22         24                0                 1       1      0              0               0                  0                  0                  0
  19         30                1                 0       0      0              0               0                  0                  0                  0

 urine_protein_6.0
 0
 0
 0
 0
 0
```

Variable Transformation:

Normalization, standardization, and differencing were applied to the 20 numerical variables.

1. **Normalization:** Scales numerical features to a standard range (usually [0, 1]) to ensure uniformity, preventing the dominance of certain variables due to their scale differences.
2. **Standardization:** Centers numerical features around a mean of 0 and scales them based on standard deviation. It transforms data to have zero mean and unit variance, making it suitable for algorithms sensitive to feature scales.
3. **Differencing:** Involves calculating the differences between consecutive data points. It's often used in time series analysis to make the data stationary and remove trends or seasonality, aiding in model stability. The dataset after differencing is given below:

The dataset after standardization, normalization, differencing are given below:

```
Data after Normalization:
        age    height  weight  waistline  sight_left  sight_right       SBP       DBP      BLDS  tot_chole  HDL_chole  LDL_chole  triglyceride  hemoglobin
0  0.769231  0.583333    0.35   0.040169    0.000000     0.000000  0.270270  0.216981  0.129288   0.182857   0.029095   0.144080      0.049839    0.555556
1  0.538462  0.666667    0.60   0.046512    0.111111     0.111111  0.304054  0.330189  0.080475   0.215238   0.040948   0.134894      0.068895    0.666667
2  0.307692  0.500000    0.30   0.020085    1.000000     0.111111  0.385135  0.358491  0.142480   0.201905   0.053879   0.121255      0.054529    0.611111
3  0.538462  0.500000    0.40   0.030655    0.111111     0.000000  0.385135  0.405660  0.110818   0.224762   0.038793   0.151213      0.062445    0.611111
4  0.538462  0.666667    0.40   0.024313    0.111111     0.111111  0.277027  0.339623  0.083113   0.179048   0.051724   0.141227      0.018763    0.555556
Data after Standardization:
        age    height    weight  waistline  sight_left  sight_right       SBP       DBP      BLDS  tot_chole  HDL_chole  LDL_chole  triglyceride  hemoglobi
0  1.657990  0.026960 -0.081383   0.773286   -0.802826    -0.849031 -0.311301 -0.996605  1.463702  -0.669701  -1.491517  -0.309656      0.312356   -0.121785
1  0.562398  0.593868  1.906657   1.335470    0.552834     0.596796  0.039264  0.215134  0.067571  -0.227500  -0.788366  -0.511160      0.899253    1.181140
2 -0.533193 -0.539948 -0.478991  -1.006963   11.398111     0.596796  0.880619  0.518069  1.841035  -0.409583  -0.021292  -0.770237      0.456823    0.529678
3  0.562398 -0.539948  0.316225  -0.069990    0.552834    -0.849031  0.880619  1.022960  0.935436  -0.097441  -0.916211  -0.165725      0.700611    0.529678
4  0.562398  0.593868  0.316225  -0.632174    0.552834     0.596796 -0.241188  0.316112  0.143037  -0.721725  -0.149137  -0.367229     -0.644737   -0.121785
Data after Differenciation:
        age    height    weight  waistline  sight_left  sight_right       SBP       DBP      BLDS  tot_chole  HDL_chole  LDL_chole  triglyceride  hemoglobi
0  0.000000  0.000000  0.000000   0.000000    0.000000     0.000000  0.000000  0.000000  0.000000   0.000000   0.000000   0.000000      0.000000   0.000000
1 -1.095592  0.566908  1.988040   0.562184    1.355660     1.445827  0.350565  1.211739 -1.396131   0.442201   0.703151  -0.201504      0.586897    1.302925
2 -1.095592 -1.133816 -2.385648  -2.342433   10.845278     0.000000  0.841355  0.302935  1.773464  -0.182083   0.767074  -0.259076     -0.442430   -0.651463
3  1.095592  0.000000  0.795216   0.936973  -10.845278    -1.445827  0.000000  0.504891 -0.905598   0.312142  -0.894920   0.604512      0.243788   0.000000
4  0.000000  1.133816  0.000000  -0.562184    0.000000     1.445827 -1.121807 -0.706848 -0.792399  -0.624284   0.767074  -0.201504     -1.345348   -0.651463
```

```
serum_creatinine  SGOT_AST  SGOT_ALT  gamma_GTP  S
        0.000000  0.015262  0.010260   0.029116
        0.025641  0.014535  0.013023   0.032129
        0.025641  0.016715  0.007103   0.028112
        0.000000  0.013081  0.008287   0.021084
        0.000000  0.018895  0.007103   0.027108

 serum_creatinine  SGOT_AST  SGOT_ALT  gamma_GTP
        -0.665435 -0.260158 -0.026110  -0.203603
         1.090851 -0.313625  0.210802  -0.150177
         1.090851 -0.153223 -0.296867  -0.221412
        -0.665435 -0.420559 -0.195333  -0.346072
        -0.665435  0.007178 -0.296867  -0.239221

 serum_creatinine  SGOT_AST  SGOT_ALT  gamma_GTP
         0.000000  0.000000  0.000000   0.000000
         1.756286 -0.053467  0.236912   0.053426
         0.000000  0.160401 -0.507669  -0.071234
        -1.756286 -0.267336  0.101534  -0.124660
         0.000000  0.427737 -0.101534   0.106852
```

Of the three methods, standardization is preferred in this project as it eliminates the influence of different scales, preventing features with larger magnitudes from dominating the model. This ensures algorithms like SVMs and k-means, sensitive to scale, perform consistently. It aids convergence speed in iterative optimization algorithms and enhances interpretability, as coefficients represent changes in the dependent variable per one-standard-deviation change in the standardized feature.

Covariance Analysis:

Covariance measures the degree to which two variables change together, indicating the direction of their linear relationship. A positive covariance suggests a direct relationship, while a negative covariance indicates an inverse relationship. Heatmap for the Sample Covariance Matrix is displayed below in Figure 4.

Observing the covariance matrix we can point out some significant relation between the variables.:

- **Positive Covariance:**
  - Features with positive covariance values indicate a positive linear relationship. For example, 'weight' and 'waistline' exhibit a strong positive covariance of 0.704, suggesting that as weight increases, the waistline also tends to increase.
- **Negative Covariance:**
  - Features with negative covariance values indicate a negative linear relationship. For instance, 'height' and 'age' have a negative covariance of -0.375, suggesting a tendency for shorter individuals to be older.
- **Low Covariance:**
  - Covariance values close to zero suggest a weak linear relationship. For example, 'sight_left' and 'sight_right' have low covariance, indicating weak correlation.
- **Strong Relationships:**
  - Some pairs exhibit strong relationships, such as 'DBP' (Diastolic Blood Pressure) and 'SBP' (Systolic Blood Pressure) with a covariance of 0.737, reflecting a positive correlation.
- **Multicollinearity:**
  - Features related to cholesterol levels ('tot_chole', 'HDL_chole', 'LDL_chole') show high positive covariances, suggesting multicollinearity. This could be a consideration for regression models.

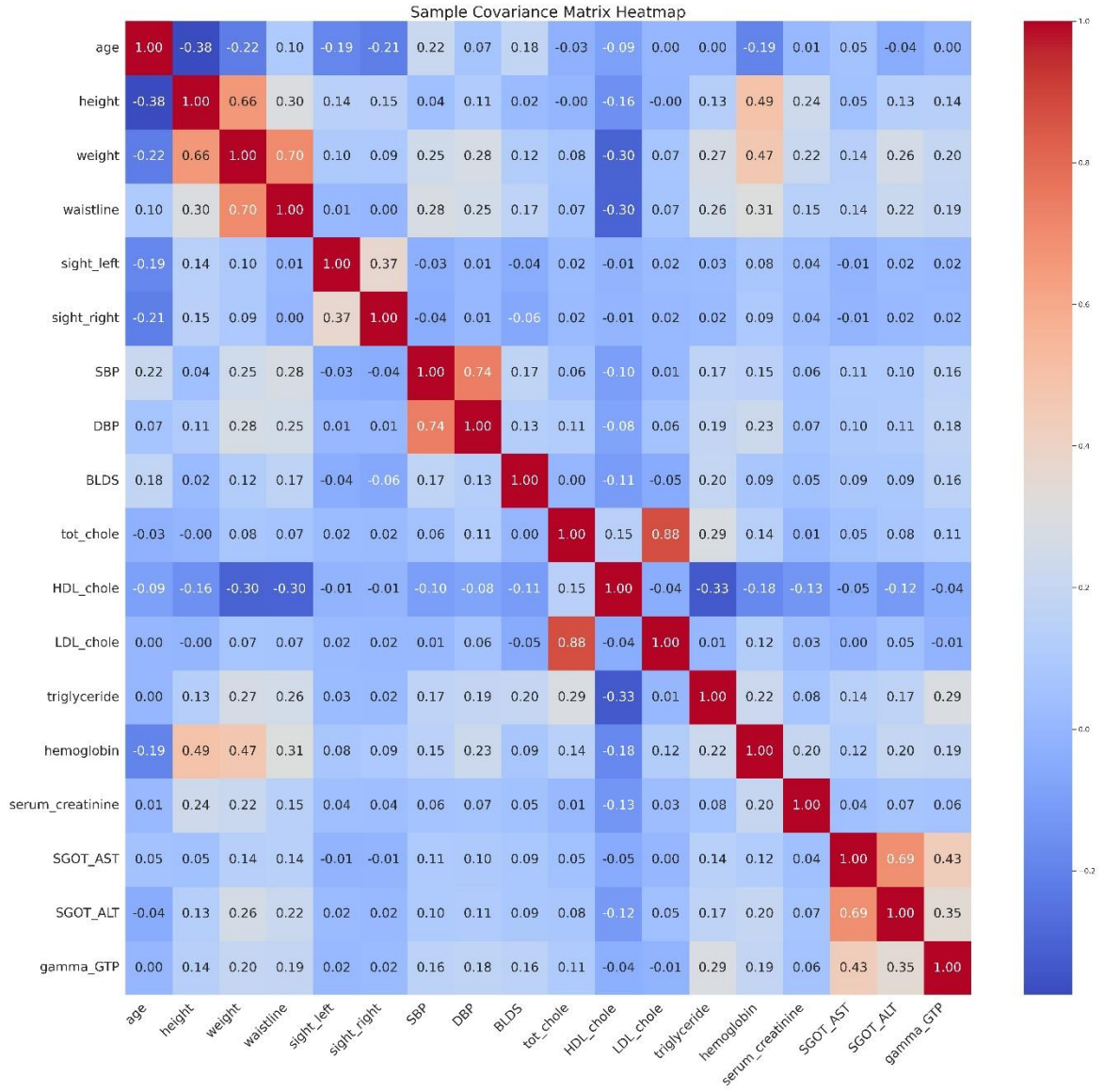


Figure 4: Covariance Matrix

Correlation Analysis:

Correlation quantifies the strength and direction of the linear relationship between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation. It provides insights into how changes in one variable correspond to changes in another, facilitating the analysis of dependencies between variables in a dataset. Heatmap for the Sample Correlation Matrix is displayed in Figure 5.

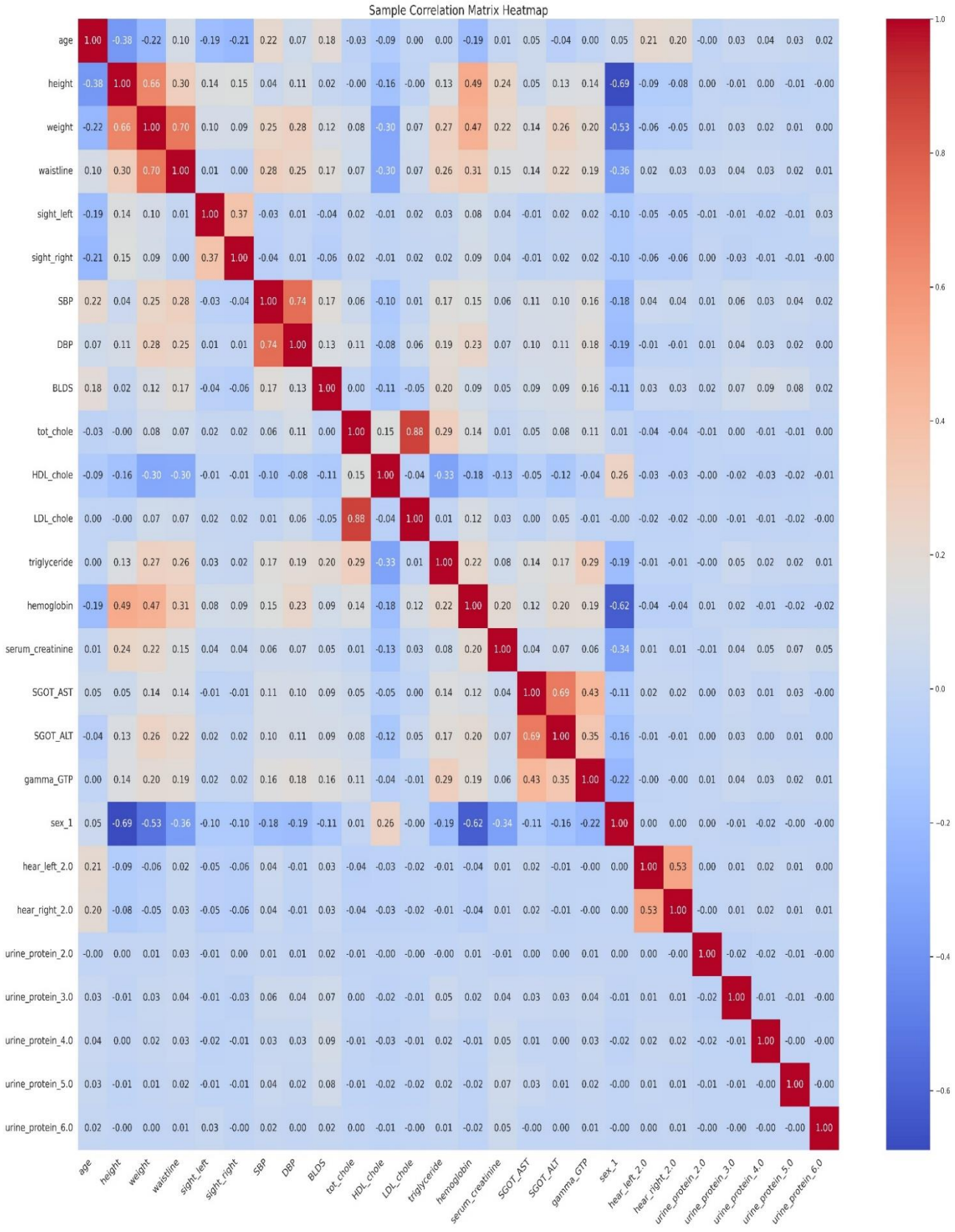


Figure 5: Correlation Matrix

Analyzing the correlation matrix, here are some insights:

1. **Age-related insights:**
    - Age is negatively correlated with height (-0.38) and positively correlated with weight (0.22), waistline (0.10), and SBP (0.22).
    - There is a moderate positive correlation between age and cholesterol levels (tot_chole, HDL_chole, LDL_chole, and triglycerides).
2. **Body measurements:**
    - Height and weight have a strong positive correlation (0.66).
    - Weight and waistline also have a strong positive correlation (0.70).
    - Height is negatively correlated with both SBP (-0.04) and DBP (-0.11).
3. **Blood pressure:**
    - SBP and DBP have a strong positive correlation (0.74).
    - Both SBP and DBP have a positive correlation with weight, waistline, and age.
4. **Cholesterol levels:**
    - Cholesterol levels (tot_chole, HDL_chole, LDL_chole, triglycerides) are positively correlated with each other.
    - HDL_chole (good cholesterol) has a negative correlation with weight and waistline.
5. **Vision:**
    - There is a positive correlation between sight_left and sight_right (0.37).
6. **Gender-related insights:**
    - There is a strong negative correlation between height and the binary variable 'sex_1' (-0.69), indicating that females (coded as 1) tend to be shorter.
    - 'sex_1' has a negative correlation with weight and waistline, indicating that females tend to have lower weights and smaller waistlines.
7. **Liver-related insights:**
    - There is a positive correlation between SGOT_AST and SGOT_ALT, indicating a correlation in these blood enzyme levels.
8. **Urine protein levels:**
    - There are weak positive correlations between urine protein levels (urine_protein_2.0 to urine_protein_6.0), suggesting a potential association between these variables.
9. **Other insights:**
    - Hemoglobin has a strong positive correlation with height (0.49) and a negative correlation with 'sex_1', indicating potential gender differences in hemoglobin levels.
    - Serum creatinine has positive correlations with various body measurements.

Dimensionality Reduction/Feature Selection:

Random Forest Analysis:

Random Forest Feature Importance is a technique used to assess the significance of different features (variables or columns) in a dataset when using a Random Forest algorithm for predictive modeling. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of the individual trees for regression problems or a majority vote for classification problems. Features were ranked using Random Forest Feature Importance, and features

with importance below the threshold (0.01) were eliminated. The random forest feature importance vs feature bar plot is given in Figure 6.

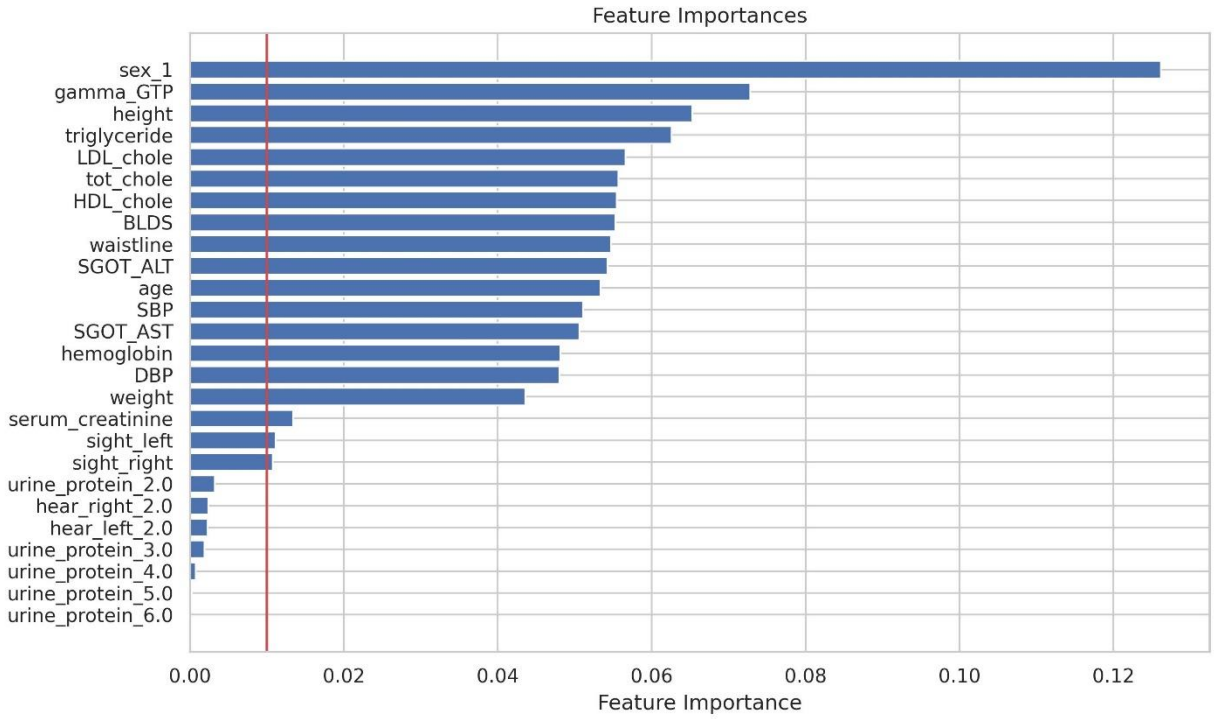


Figure 6: Random Forest Feature Importance

The threshold=0.01 suggests to remove: ['hear_left_2.0', 'hear_right_2.0', 'urine_protein_2.0', 'urine_protein_3.0', 'urine_protein_4.0', 'urine_protein_5.0', 'urine_protein_6.0'] and selected features are: ['age', 'height', 'weight', 'waistline', 'sight_left', 'sight_right', 'SBP', 'DBP', 'BLDS', 'tot_chole', 'HDL_chole', 'LDL_chole', 'triglyceride', 'hemoglobin', 'serum_creatinine', 'SGOT_AST', 'SGOT_ALT', 'gamma_GTP', 'sex_1']

Principal Component Analysis (PCA):

Principal Component Analysis (PCA) is a dimensionality reduction technique commonly used in machine learning and data analysis. The main goal of PCA is to transform the original features of a dataset into a new set of uncorrelated features, known as principal components, which capture the maximum variance in the data. This reduction in dimensionality can help simplify the dataset, improve computational efficiency, and retain the most significant information. In this project, PCA was applied to identify the number of features needed to explain more than 90% of the dependent variance. The PCA explained variances vs the number features is given below in Figure 7.

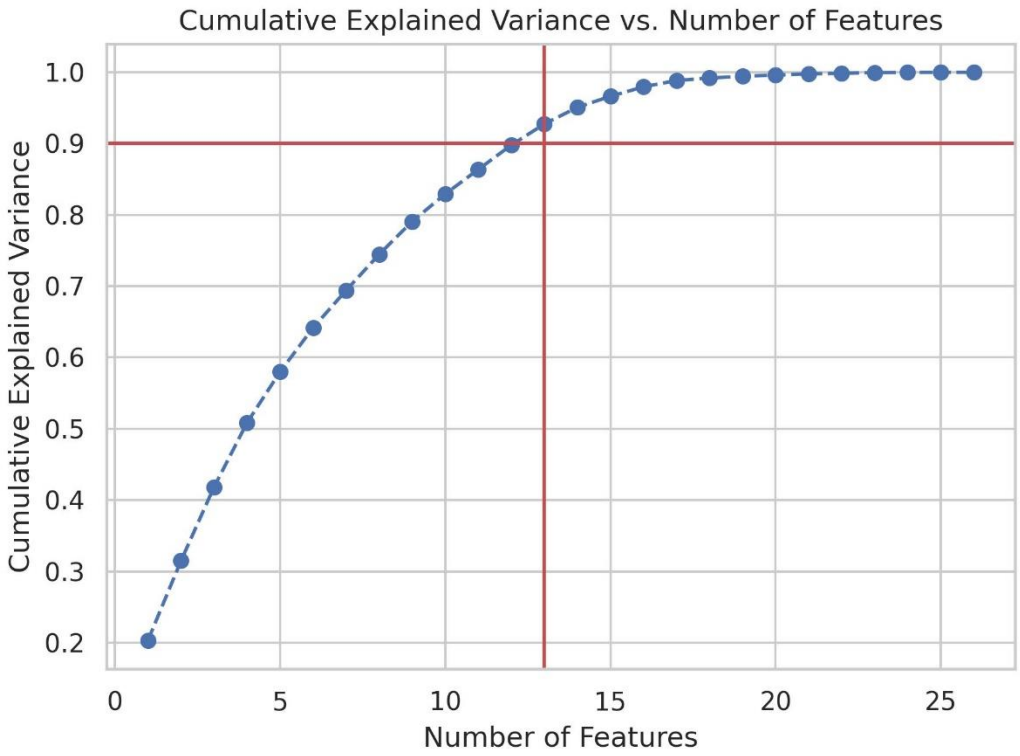


Figure 7: Variance vs number of features in PCA

The threshold=90% suggests that only 13 features are need to explain 90% dependent variances among the features. This analysis of PCA can only say the number of features needed not which features are needed.

Singular Value Decomposition Analysis:

Singular Value Decomposition (SVD) is a fundamental matrix decomposition method used in linear algebra and numerical analysis. It has applications in various fields, including signal processing, image compression, data analysis, and machine learning. SVD decomposes a matrix into three other matrices, revealing the underlying structure of the original matrix. SVD is often used for dimensionality reduction by keeping only the most significant singular values and corresponding singular vectors.

At first, Singular Value Decomposition was performed, and the number of components for explaining variance was determined using Logistic Regression. The average accuracy and standard deviation of the accuracy after 5 fold cross validation with components from 1 to 24 is given below:

```
------Method 3: SVD; Finding best n_components with Logistic Regression model-------
>1 0.502 (0.007)
>2 0.541 (0.006)
>3 0.545 (0.007)
>4 0.549 (0.007)
>5 0.550 (0.007)
>6 0.553 (0.006)
>7 0.574 (0.008)
>8 0.582 (0.008)
>9 0.589 (0.009)
>10 0.591 (0.009)
>11 0.599 (0.012)
>12 0.607 (0.009)
>13 0.611 (0.009)
>14 0.619 (0.009)
>15 0.619 (0.008)
>16 0.620 (0.007)
>17 0.628 (0.009)
>18 0.639 (0.008)
>19 0.639 (0.008)
>20 0.639 (0.008)
>21 0.639 (0.007)
>22 0.639 (0.007)
>23 0.639 (0.007)
>24 0.639 (0.008)
```

We can see after n_components=18 the accuracy does not improve. This can be better visualized by the box-plot given in Figure 8.

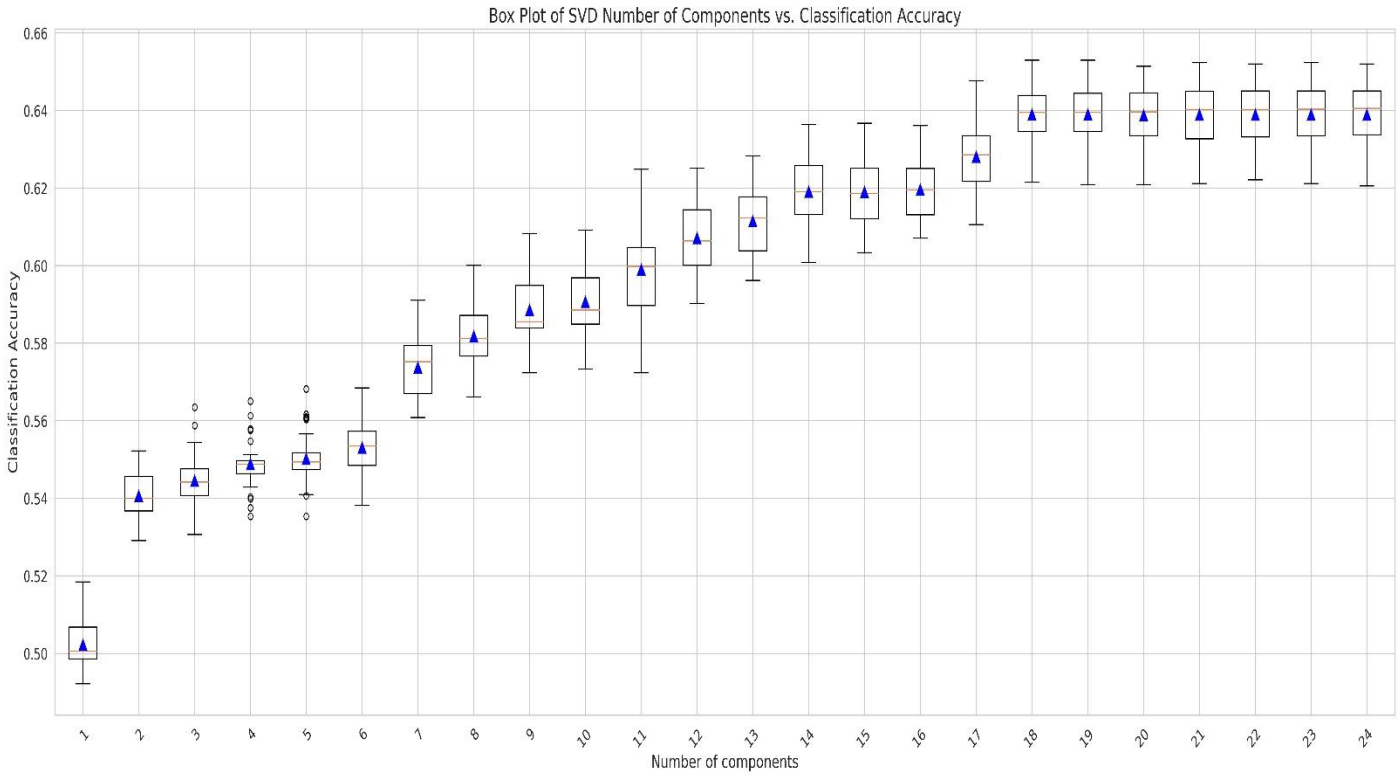


Figure 8: Box plot of SVD Number of Components vs Classification Accuracy

The explained variances vs the singular values plot after choosing the n_components=18 is given below:

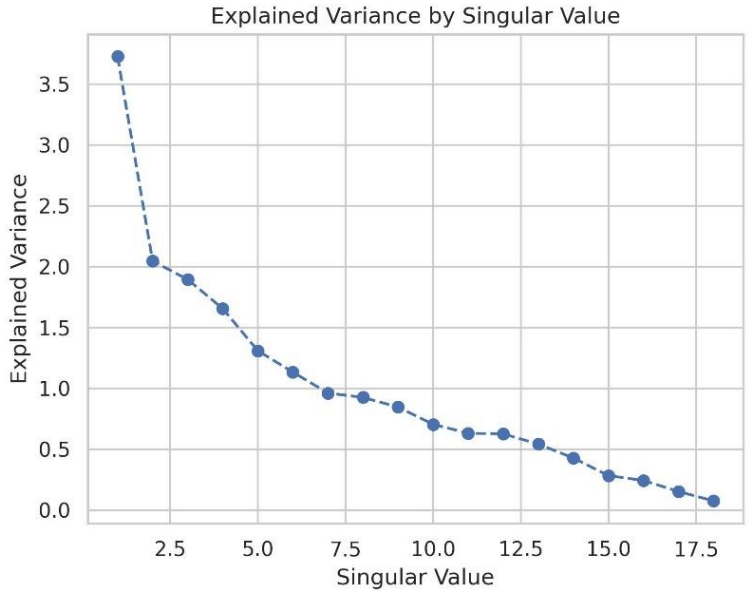


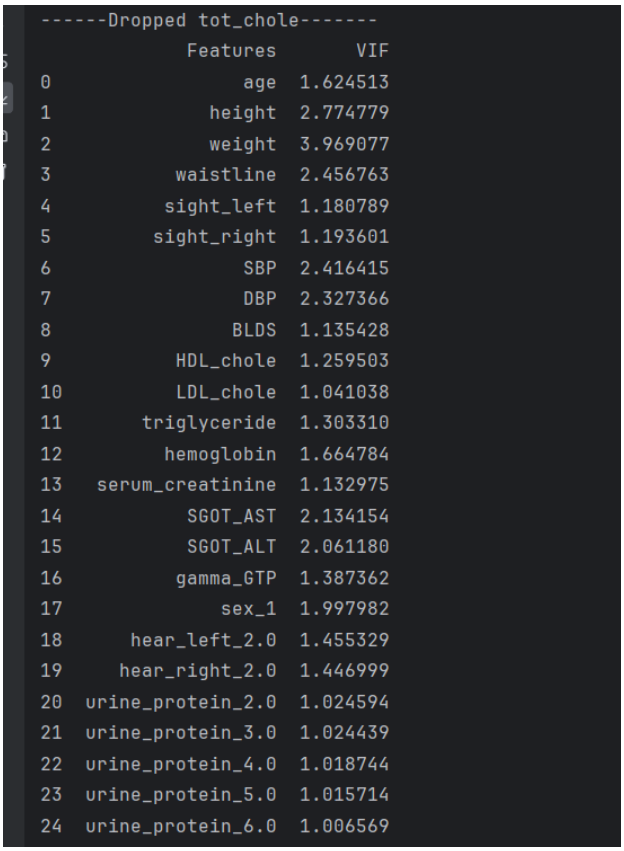
Figure 9: Explained Variance by Singular Values

Variance Inflation Factor (VIF):

The Variance Inflation Factor (VIF) is a measure used to quantify the severity of multicollinearity in a regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can lead to issues in interpreting the results and can affect the stability and reliability of the estimated coefficients. VIF assesses how much the variance of an estimated regression coefficient increases if the predictors are correlated. Features were dropped based on the VIF threshold (5.00). The performed dimensionality reduction step by step is given below. At each step, the feature with highest VIF has been removed.

Initial VIF

	Features	VIF
0	age	1.624536
1	height	2.775352
2	weight	3.969093
3	waistline	2.456782
4	sight_left	1.180798
5	sight_right	1.193610
6	SBP	2.416795
7	DBP	2.331383
8	BLDS	1.135487
9	tot_chole	16.056313
10	HDL_chole	2.801652
11	LDL_chole	13.564762
12	triglyceride	3.629448
13	hemoglobin	1.667036
14	serum_creatinine	1.133028
15	SGOT_AST	2.134408
16	SGOT_ALT	2.061569
17	gamma_GTP	1.392481
18	sex_1	2.000831
19	hear_left_2.0	1.455650
20	hear_right_2.0	1.447035
21	urine_protein_2.0	1.024826
22	urine_protein_3.0	1.024515
23	urine_protein_4.0	1.018857
24	urine_protein_5.0	1.015770
25	urine_protein_6.0	1.006570



```
------Dropped tot_chole-------
             Features       VIF
0                 age  1.624513
1              height  2.774779
2              weight  3.969077
3           waistline  2.456763
4          sight_left  1.180789
5         sight_right  1.193601
6                 SBP  2.416415
7                 DBP  2.327366
8                BLDS  1.135428
9           HDL_chole  1.259503
10          LDL_chole  1.041038
11      triglyceride  1.303310
12         hemoglobin  1.664784
13   serum_creatinine  1.132975
14           SGOT_AST  2.134154
15           SGOT_ALT  2.061180
16          gamma_GTP  1.387362
17              sex_1  1.997982
18     hear_left_2.0  1.455329
19     hear_right_2.0  1.446999
20  urine_protein_2.0  1.024594
21  urine_protein_3.0  1.024439
22  urine_protein_4.0  1.018744
23  urine_protein_5.0  1.015714
24  urine_protein_6.0  1.006569
```

The last snapshot shows no VIF greater than 5.00, which says that all multicollinearity problem has been addressed. This can be further made sure by checking if any feature remains in the dataset having correlation coefficient greater than 0.05. The result says the dataset is free from any multicollinearity issues.

Observations and Choosing dimensionality reduction technique:

- Dimensionality reduction techniques like PCA and SVD were applied to know the number of features needed for explaining certain variances. Since theses methods cannot tell which features to remove they have not been chosen for final dimensionality reduction.
- The Random forest feature importance is highly intuitive and tells specifically which features are important for supervised machine learning problems. So this method has been picked for the final dimensionality reduction of the dataset.
- The dimensionality reduction technique VIF can address all the multicollinearity issues and thus this technique is selected for the later phases of the project.

Random Forest Feature Importance and VIF have been chosen as the primary dimensionality reduction techniques for the project. Random Forest provides specific insights into feature importance, making it suitable for supervised learning. VIF effectively handles multicollinearity

issues, ensuring stability in regression analysis. The combination of these methods ensures a comprehensive approach to dimensionality reduction, considering both feature importance and multicollinearity concerns.

Phase 2:

For this phase prediction on a selected continuous numerical feature “tot_chole” has been done using a multiple linear regression model. The following operations has been included in the analysis:

Stepwise regression and adjusted R-square analysis:

Stepwise regression is a variable selection technique that systematically adds or removes predictors to achieve the most statistically significant and parsimonious model. It iteratively evaluates variables based on their impact, and the process continues until no further improvement is observed. Adjusted R-square analysis assesses the goodness of fit of a regression model, considering the trade-off between model complexity and explanatory power. It adjusts the R-square value for the number of predictors, providing a more accurate representation of the model's predictive capability. In this method the features are dropped one by one until the p-value becomes zero and there is significant drop in adjusted R-square value.

The step-by-step stepwise regression is given below:

```
============Step5: Step wise regression =============
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     4151.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:38   Log-Likelihood:                -23475.
No. Observations:               54396   AIC:                         4.700e+04
Df Residuals:                   54371   BIC:                         4.722e+04
Df Model:                          24
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0619      0.003    -23.125      0.000      -0.067      -0.057
age            0.0155      0.003      5.120      0.000       0.010       0.021
height        -0.0248      0.003     -7.371      0.000      -0.031      -0.018
weight         0.0124      0.003      4.107      0.000       0.006       0.018
sight_left     0.0144      0.008      1.821      0.069      -0.001       0.030
sight_right   -0.0034      0.008     -0.424      0.672      -0.019       0.012
SBP            0.0113      0.003      4.117      0.000       0.006       0.017
BLDS          -0.0125      0.005     -2.736      0.006      -0.021      -0.004
HDL_chole      0.2287      0.003     79.528      0.000       0.223       0.234
LDL_chole      0.7858      0.003    298.577      0.000       0.781       0.791
triglyceride   0.3304      0.005     71.805      0.000       0.321       0.339
hemoglobin     0.0271      0.003      9.323      0.000       0.021       0.033
serum_creatinine -0.0651   0.014     -4.732      0.000      -0.092      -0.038
SGOT_AST      -0.0135      0.007     -1.858      0.063      -0.028       0.001
SGOT_ALT       0.0097      0.007      1.444      0.149      -0.003       0.023
gamma_GTP      0.0523      0.006      8.619      0.000       0.040       0.064
hear_left_2.0 -0.0232      0.009     -2.525      0.012      -0.041      -0.005
urine_protein_2.0 -0.0110  0.009     -1.202      0.229      -0.029       0.007
urine_protein_3.0  0.0005  0.014      0.040      0.968      -0.026       0.027
urine_protein_4.0 -0.0009  0.022     -0.041      0.968      -0.044       0.042
urine_protein_5.0  0.0893  0.045      1.985      0.047       0.001       0.177
urine_protein_6.0  0.1031  0.104      0.993      0.320      -0.100       0.306
DRK_YN_1       0.0163      0.004      4.510      0.000       0.009       0.023
SMK_stat_type_cd_2.0 -0.0169 0.005    -3.563      0.000      -0.026      -0.008
SMK_stat_type_cd_3.0  0.0021 0.005     0.454      0.650      -0.007       0.011
==============================================================================
Omnibus:                     3130.746   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7778.950
Skew:                          -0.344   Prob(JB):                         0.00
Kurtosis:                       4.720   Cond. No.                         77.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped urine_protein_3.0============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     4331.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:38   Log-Likelihood:                -23475.
No. Observations:               54396   AIC:                         4.700e+04
Df Residuals:                   54372   BIC:                         4.721e+04
Df Model:                          23
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0619      0.003    -23.183      0.000      -0.067      -0.057
age            0.0155      0.003      5.121      0.000       0.010       0.021
height        -0.0248      0.003     -7.371      0.000      -0.031      -0.018
weight         0.0124      0.003      4.107      0.000       0.006       0.018
sight_left     0.0144      0.008      1.821      0.069      -0.001       0.030
sight_right   -0.0034      0.008     -0.424      0.672      -0.019       0.012
SBP            0.0113      0.003      4.118      0.000       0.006       0.017
BLDS          -0.0125      0.005     -2.737      0.006      -0.021      -0.004
HDL_chole      0.2287      0.003     79.529      0.000       0.223       0.234
LDL_chole      0.7858      0.003    298.585      0.000       0.781       0.791
triglyceride   0.3304      0.005     71.808      0.000       0.321       0.339
hemoglobin     0.0271      0.003      9.324      0.000       0.021       0.033
serum_creatinine -0.0651   0.014     -4.733      0.000      -0.092      -0.038
SGOT_AST      -0.0135      0.007     -1.858      0.063      -0.028       0.001
SGOT_ALT       0.0097      0.007      1.444      0.149      -0.003       0.023
gamma_GTP      0.0523      0.006      8.619      0.000       0.040       0.064
hear_left_2.0 -0.0232      0.009     -2.524      0.012      -0.041      -0.005
urine_protein_2.0 -0.0110  0.009     -1.204      0.229      -0.029       0.007
urine_protein_4.0 -0.0009  0.022     -0.041      0.967      -0.044       0.042
urine_protein_5.0  0.0893  0.045      1.985      0.047       0.001       0.177
urine_protein_6.0  0.1030  0.104      0.993      0.321      -0.100       0.306
DRK_YN_1       0.0163      0.004      4.510      0.000       0.009       0.023
SMK_stat_type_cd_2.0 -0.0169 0.005    -3.563      0.000      -0.026      -0.008
SMK_stat_type_cd_3.0  0.0021 0.005     0.454      0.650      -0.007       0.011
==============================================================================
Omnibus:                     3130.751   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7778.890
Skew:                          -0.344   Prob(JB):                         0.00
Kurtosis:                       4.720   Cond. No.                         77.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped urine_protein_4.0============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     4528.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23475.
No. Observations:               54396   AIC:                          4.700e+04
Df Residuals:                   54373   BIC:                          4.720e+04
Df Model:                          22
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0619      0.003    -23.199      0.000      -0.067      -0.057
age            0.0155      0.003      5.121      0.000       0.010       0.021
height        -0.0248      0.003     -7.371      0.000      -0.031      -0.018
weight         0.0124      0.003      4.107      0.000       0.006       0.018
sight_left     0.0144      0.008      1.820      0.069      -0.001       0.030
sight_right   -0.0034      0.008     -0.424      0.672      -0.019       0.012
SBP            0.0113      0.003      4.118      0.000       0.006       0.017
BLDS          -0.0125      0.005     -2.740      0.006      -0.021      -0.004
HDL_chole      0.2287      0.003     79.532      0.000       0.223       0.234
LDL_chole      0.7858      0.003    298.589      0.000       0.781       0.791
triglyceride   0.3304      0.005     71.810      0.000       0.321       0.339
hemoglobin     0.0271      0.003      9.325      0.000       0.021       0.033
serum_creatinine   -0.0651      0.014     -4.751      0.000      -0.092      -0.038
SGOT_AST           -0.0135      0.007     -1.859      0.063      -0.028       0.001
SGOT_ALT            0.0097      0.007      1.445      0.149      -0.003       0.023
gamma_GTP           0.0523      0.006      8.619      0.000       0.040       0.064
hear_left_2.0      -0.0233      0.009     -2.525      0.012      -0.041      -0.005
urine_protein_2.0  -0.0110      0.009     -1.203      0.229      -0.029       0.007
urine_protein_5.0   0.0893      0.045      1.985      0.047       0.001       0.177
urine_protein_6.0   0.1031      0.104      0.994      0.320      -0.100       0.306
DRK_YN_1            0.0163      0.004      4.510      0.000       0.009       0.023
SMK_stat_type_cd_2.0   -0.0169      0.005     -3.563      0.000      -0.026      -0.008
SMK_stat_type_cd_3.0    0.0021      0.005      0.454      0.650      -0.007       0.011
==============================================================================
Omnibus:                     3130.783   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7778.811
Skew:                          -0.344   Prob(JB):                         0.00
Kurtosis:                       4.720   Cond. No.                         77.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped sight_right============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     4744.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23475.
No. Observations:               54396   AIC:                         4.699e+04
Df Residuals:                   54374   BIC:                         4.719e+04
Df Model:                          21
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0619      0.003    -23.227      0.000      -0.067      -0.057
age            0.0156      0.003      5.153      0.000       0.010       0.022
height        -0.0248      0.003     -7.381      0.000      -0.031      -0.018
weight         0.0124      0.003      4.102      0.000       0.006       0.018
sight_left     0.0131      0.007      1.801      0.072      -0.001       0.027
SBP            0.0113      0.003      4.118      0.000       0.006       0.017
BLDS          -0.0125      0.005     -2.738      0.006      -0.021      -0.004
HDL_chole      0.2287      0.003     79.532      0.000       0.223       0.234
LDL_chole      0.7858      0.003    298.604      0.000       0.781       0.791
triglyceride   0.3304      0.005     71.809      0.000       0.321       0.339
hemoglobin     0.0271      0.003      9.321      0.000       0.021       0.033
serum_creatinine -0.0651   0.014     -4.748      0.000      -0.092      -0.038
SGOT_AST      -0.0135      0.007     -1.858      0.063      -0.028       0.001
SGOT_ALT       0.0097      0.007      1.444      0.149      -0.003       0.023
gamma_GTP      0.0523      0.006      8.620      0.000       0.040       0.064
hear_left_2.0 -0.0232      0.009     -2.518      0.012      -0.041      -0.005
urine_protein_2.0 -0.0110  0.009     -1.202      0.229      -0.029       0.007
urine_protein_5.0  0.0893  0.045      1.985      0.047       0.001       0.177
urine_protein_6.0  0.1032  0.104      0.995      0.320      -0.100       0.307
DRK_YN_1       0.0162      0.004      4.507      0.000       0.009       0.023
SMK_stat_type_cd_2.0 -0.0169 0.005    -3.569      0.000      -0.026      -0.008
SMK_stat_type_cd_3.0  0.0021 0.005     0.454      0.650      -0.007       0.011
==============================================================================
Omnibus:                     3131.018   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7779.133
Skew:                          -0.344   Prob(JB):                         0.00
Kurtosis:                       4.720   Cond. No.                         77.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped SMK_stat_type_cd_3.0============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     4981.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23475.
No. Observations:               54396   AIC:                         4.699e+04
Df Residuals:                   54375   BIC:                         4.718e+04
Df Model:                          20
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0615      0.003    -24.316      0.000      -0.066      -0.057
age            0.0156      0.003      5.171      0.000       0.010       0.022
height        -0.0246      0.003     -7.401      0.000      -0.031      -0.018
weight         0.0124      0.003      4.134      0.000       0.007       0.018
sight_left     0.0131      0.007      1.805      0.071      -0.001       0.027
SBP            0.0113      0.003      4.123      0.000       0.006       0.017
BLDS          -0.0125      0.005     -2.735      0.006      -0.021      -0.004
HDL_chole      0.2286      0.003     79.824      0.000       0.223       0.234
LDL_chole      0.7857      0.003    298.654      0.000       0.781       0.791
triglyceride   0.3305      0.005     71.943      0.000       0.322       0.340
hemoglobin     0.0274      0.003      9.592      0.000       0.022       0.033
serum_creatinine -0.0650   0.014     -4.742      0.000      -0.092      -0.038
SGOT_AST      -0.0136      0.007     -1.865      0.062      -0.028       0.001
SGOT_ALT       0.0097      0.007      1.443      0.149      -0.003       0.023
gamma_GTP      0.0525      0.006      8.676      0.000       0.041       0.064
hear_left_2.0 -0.0232      0.009     -2.516      0.012      -0.041      -0.005
urine_protein_2.0 -0.0111  0.009     -1.204      0.229      -0.029       0.007
urine_protein_5.0  0.0893  0.045      1.986      0.047       0.001       0.177
urine_protein_6.0  0.1033  0.104      0.996      0.319      -0.100       0.307
DRK_YN_1       0.0166      0.004      4.740      0.000       0.010       0.023
SMK_stat_type_cd_2.0 -0.0177 0.004    -4.023      0.000      -0.026      -0.009
==============================================================================
Omnibus:                     3130.627   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7779.577
Skew:                          -0.344   Prob(JB):                         0.00
Kurtosis:                       4.720   Cond. No.                         76.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped urine_protein_6.0============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     5243.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23475.
No. Observations:               54396   AIC:                         4.699e+04
Df Residuals:                   54376   BIC:                         4.717e+04
Df Model:                          19
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0615      0.003    -24.310      0.000      -0.066      -0.057
age            0.0156      0.003      5.167      0.000       0.010       0.022
height        -0.0246      0.003     -7.393      0.000      -0.031      -0.018
weight         0.0124      0.003      4.137      0.000       0.007       0.018
sight_left     0.0131      0.007      1.809      0.070      -0.001       0.027
SBP            0.0113      0.003      4.132      0.000       0.006       0.017
BLDS          -0.0124      0.005     -2.724      0.006      -0.021      -0.003
HDL_chole      0.2286      0.003     79.822      0.000       0.223       0.234
LDL_chole      0.7857      0.003    298.653      0.000       0.781       0.791
triglyceride   0.3305      0.005     71.954      0.000       0.322       0.340
hemoglobin     0.0273      0.003      9.577      0.000       0.022       0.033
serum_creatinine -0.0640   0.014     -4.679      0.000      -0.091      -0.037
SGOT_AST      -0.0136      0.007     -1.867      0.062      -0.028       0.001
SGOT_ALT       0.0096      0.007      1.440      0.150      -0.003       0.023
gamma_GTP      0.0525      0.006      8.677      0.000       0.041       0.064
hear_left_2.0 -0.0231      0.009     -2.508      0.012      -0.041      -0.005
urine_protein_2.0 -0.0111  0.009     -1.208      0.227      -0.029       0.007
urine_protein_5.0  0.0891  0.045      1.981      0.048       0.001       0.177
DRK_YN_1       0.0166      0.004      4.740      0.000       0.010       0.023
SMK_stat_type_cd_2.0 -0.0178 0.004    -4.032      0.000      -0.026      -0.009
==============================================================================
Omnibus:                     3130.131   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7776.998
Skew:                          -0.344   Prob(JB):                         0.00
Kurtosis:                       4.720   Cond. No.                         32.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped urine_protein_2.0============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     5535.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23476.
No. Observations:               54396   AIC:                         4.699e+04
Df Residuals:                   54377   BIC:                         4.716e+04
Df Model:                          18
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0618      0.003    -24.592      0.000      -0.067      -0.057
age            0.0156      0.003      5.175      0.000       0.010       0.022
height        -0.0246      0.003     -7.399      0.000      -0.031      -0.018
weight         0.0125      0.003      4.144      0.000       0.007       0.018
sight_left     0.0131      0.007      1.808      0.071      -0.001       0.027
SBP            0.0113      0.003      4.124      0.000       0.006       0.017
BLDS          -0.0126      0.005     -2.749      0.006      -0.022      -0.004
HDL_chole      0.2286      0.003     79.824      0.000       0.223       0.234
LDL_chole      0.7857      0.003    298.652      0.000       0.781       0.791
triglyceride   0.3306      0.005     71.960      0.000       0.322       0.340
hemoglobin     0.0273      0.003      9.564      0.000       0.022       0.033
serum_creatinine -0.0642   0.014     -4.694      0.000      -0.091      -0.037
SGOT_AST      -0.0136      0.007     -1.872      0.061      -0.028       0.001
SGOT_ALT       0.0097      0.007      1.443      0.149      -0.003       0.023
gamma_GTP      0.0525      0.006      8.677      0.000       0.041       0.064
hear_left_2.0 -0.0231      0.009     -2.514      0.012      -0.041      -0.005
urine_protein_5.0  0.0895  0.045      1.990      0.047       0.001       0.178
DRK_YN_1       0.0166      0.004      4.738      0.000       0.010       0.023
SMK_stat_type_cd_2.0 -0.0178 0.004    -4.034      0.000      -0.026      -0.009
==============================================================================
Omnibus:                     3129.819   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7776.360
Skew:                          -0.344   Prob(JB):                         0.00
Kurtosis:                       4.720   Cond. No.                         32.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped SGOT_ALT============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     5860.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23477.
No. Observations:               54396   AIC:                         4.699e+04
Df Residuals:                   54378   BIC:                         4.715e+04
Df Model:                          17
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0614      0.002    -24.583      0.000      -0.066      -0.057
age            0.0154      0.003      5.104      0.000       0.009       0.021
height        -0.0247      0.003     -7.435      0.000      -0.031      -0.018
weight         0.0130      0.003      4.341      0.000       0.007       0.019
sight_left     0.0131      0.007      1.813      0.070      -0.001       0.027
SBP            0.0113      0.003      4.134      0.000       0.006       0.017
BLDS          -0.0125      0.005     -2.737      0.006      -0.021      -0.004
HDL_chole      0.2284      0.003     79.823      0.000       0.223       0.234
LDL_chole      0.7858      0.003    298.811      0.000       0.781       0.791
triglyceride   0.3309      0.005     72.105      0.000       0.322       0.340
hemoglobin     0.0275      0.003      9.664      0.000       0.022       0.033
serum_creatinine -0.0643   0.014     -4.702      0.000      -0.091      -0.037
SGOT_AST      -0.0090      0.007     -1.377      0.168      -0.022       0.004
gamma_GTP      0.0534      0.006      8.871      0.000       0.042       0.065
hear_left_2.0 -0.0232      0.009     -2.517      0.012      -0.041      -0.005
urine_protein_5.0  0.0891  0.045      1.981      0.048       0.001       0.177
DRK_YN_1       0.0164      0.004      4.685      0.000       0.010       0.023
SMK_stat_type_cd_2.0 -0.0177 0.004    -4.022      0.000      -0.026      -0.009
==============================================================================
Omnibus:                     3132.184   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7783.628
Skew:                          -0.344   Prob(JB):                         0.00
Kurtosis:                       4.721   Cond. No.                         32.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped SGOT_AST============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     6226.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23478.
No. Observations:               54396   AIC:                         4.699e+04
Df Residuals:                   54379   BIC:                         4.714e+04
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0617      0.002    -24.837      0.000      -0.067      -0.057
age            0.0153      0.003      5.075      0.000       0.009       0.021
height        -0.0246      0.003     -7.406      0.000      -0.031      -0.018
weight         0.0128      0.003      4.290      0.000       0.007       0.019
sight_left     0.0132      0.007      1.826      0.068      -0.001       0.027
SBP            0.0113      0.003      4.115      0.000       0.006       0.017
BLDS          -0.0127      0.005     -2.788      0.005      -0.022      -0.004
HDL_chole      0.2285      0.003     79.834      0.000       0.223       0.234
LDL_chole      0.7859      0.003    298.857      0.000       0.781       0.791
triglyceride   0.3308      0.005     72.093      0.000       0.322       0.340
hemoglobin     0.0274      0.003      9.634      0.000       0.022       0.033
serum_creatinine -0.0641   0.014     -4.692      0.000      -0.091      -0.037
gamma_GTP      0.0521      0.006      8.764      0.000       0.040       0.064
hear_left_2.0 -0.0232      0.009     -2.517      0.012      -0.041      -0.005
urine_protein_5.0  0.0884  0.045      1.965      0.049       0.000       0.176
DRK_YN_1       0.0164      0.004      4.694      0.000       0.010       0.023
SMK_stat_type_cd_2.0 -0.0178 0.004    -4.042      0.000      -0.026      -0.009
==============================================================================
Omnibus:                     3133.171   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7783.037
Skew:                          -0.344   Prob(JB):                         0.00
Kurtosis:                       4.720   Cond. No.                         32.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
==============================Dropped sight_left==============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     6640.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23480.
No. Observations:               54396   AIC:                         4.699e+04
Df Residuals:                   54380   BIC:                         4.713e+04
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0622      0.002    -25.161      0.000      -0.067      -0.057
age            0.0148      0.003      4.922      0.000       0.009       0.021
height        -0.0244      0.003     -7.345      0.000      -0.031      -0.018
weight         0.0129      0.003      4.334      0.000       0.007       0.019
SBP            0.0112      0.003      4.102      0.000       0.006       0.017
BLDS          -0.0128      0.005     -2.814      0.005      -0.022      -0.004
HDL_chole      0.2285      0.003     79.830      0.000       0.223       0.234
LDL_chole      0.7860      0.003    298.898      0.000       0.781       0.791
triglyceride   0.3307      0.005     72.080      0.000       0.322       0.340
hemoglobin     0.0275      0.003      9.664      0.000       0.022       0.033
serum_creatinine -0.0643   0.014     -4.704      0.000      -0.091      -0.038
gamma_GTP      0.0521      0.006      8.777      0.000       0.040       0.064
hear_left_2.0 -0.0236      0.009     -2.569      0.010      -0.042      -0.006
urine_protein_5.0  0.0885  0.045      1.967      0.049       0.000       0.177
DRK_YN_1       0.0165      0.004      4.712      0.000       0.010       0.023
SMK_stat_type_cd_2.0 -0.0177 0.004    -4.009      0.000      -0.026      -0.009
==============================================================================
Omnibus:                     3135.514   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7790.613
Skew:                          -0.345   Prob(JB):                         0.00
Kurtosis:                       4.721   Cond. No.                         32.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped urine_protein_5.0============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     7114.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23482.
No. Observations:               54396   AIC:                         4.699e+04
Df Residuals:                   54381   BIC:                         4.713e+04
Df Model:                          14
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                     -0.0621      0.002    -25.123      0.000      -0.067      -0.057
age                        0.0148      0.003      4.928      0.000       0.009       0.021
height                    -0.0244      0.003     -7.359      0.000      -0.031      -0.018
weight                     0.0130      0.003      4.340      0.000       0.007       0.019
SBP                        0.0113      0.003      4.129      0.000       0.006       0.017
BLDS                      -0.0127      0.005     -2.773      0.006      -0.022      -0.004
HDL_chole                  0.2284      0.003     79.823      0.000       0.223       0.234
LDL_chole                  0.7860      0.003    298.891      0.000       0.781       0.791
triglyceride               0.3308      0.005     72.097      0.000       0.322       0.340
hemoglobin                 0.0275      0.003      9.664      0.000       0.022       0.033
serum_creatinine          -0.0630      0.014     -4.616      0.000      -0.090      -0.036
gamma_GTP                  0.0522      0.006      8.779      0.000       0.041       0.064
hear_left_2.0             -0.0238      0.009     -2.581      0.010      -0.042      -0.006
DRK_YN_1                   0.0165      0.004      4.702      0.000       0.010       0.023
SMK_stat_type_cd_2.0      -0.0177      0.004     -4.016      0.000      -0.026      -0.009
==============================================================================
Omnibus:                     3135.736   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7797.880
Skew:                          -0.344   Prob(JB):                         0.00
Kurtosis:                       4.722   Cond. No.                         9.97
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped hear_left_2.0============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     7660.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23485.
No. Observations:               54396   AIC:                         4.700e+04
Df Residuals:                   54382   BIC:                         4.712e+04
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                  -0.0629      0.002    -25.637      0.000      -0.068      -0.058
age                     0.0135      0.003      4.557      0.000       0.008       0.019
height                 -0.0243      0.003     -7.308      0.000      -0.031      -0.018
weight                  0.0130      0.003      4.356      0.000       0.007       0.019
SBP                     0.0112      0.003      4.092      0.000       0.006       0.017
BLDS                   -0.0128      0.005     -2.812      0.005      -0.022      -0.004
HDL_chole               0.2285      0.003     79.848      0.000       0.223       0.234
LDL_chole               0.7861      0.003    298.969      0.000       0.781       0.791
triglyceride            0.3309      0.005     72.111      0.000       0.322       0.340
hemoglobin              0.0275      0.003      9.647      0.000       0.022       0.033
serum_creatinine       -0.0638      0.014     -4.671      0.000      -0.091      -0.037
gamma_GTP               0.0522      0.006      8.792      0.000       0.041       0.064
DRK_YN_1                0.0165      0.004      4.718      0.000       0.010       0.023
SMK_stat_type_cd_2.0   -0.0178      0.004     -4.030      0.000      -0.026      -0.009
==============================================================================
Omnibus:                     3138.407   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7804.982
Skew:                          -0.345   Prob(JB):                         0.00
Kurtosis:                       4.723   Cond. No.                         9.96
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
============================Dropped BLDS============================
                            OLS Regression Results
==============================================================================
Dep. Variable:              tot_chole   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.647
Method:                 Least Squares   F-statistic:                     8297.
Date:                Thu, 07 Dec 2023   Prob (F-statistic):               0.00
Time:                        01:44:39   Log-Likelihood:                -23489.
No. Observations:               54396   AIC:                         4.700e+04
Df Residuals:                   54383   BIC:                         4.712e+04
Df Model:                          12
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                      -0.0633      0.002    -25.866      0.000      -0.068      -0.059
age                         0.0127      0.003      4.305      0.000       0.007       0.018
height                     -0.0240      0.003     -7.240      0.000      -0.031      -0.018
weight                      0.0126      0.003      4.227      0.000       0.007       0.018
SBP                         0.0108      0.003      3.949      0.000       0.005       0.016
HDL_chole                   0.2288      0.003     79.965      0.000       0.223       0.234
LDL_chole                   0.7865      0.003    299.559      0.000       0.781       0.792
triglyceride                0.3305      0.005     72.052      0.000       0.321       0.339
hemoglobin                  0.0273      0.003      9.598      0.000       0.022       0.033
serum_creatinine           -0.0648      0.014     -4.747      0.000      -0.092      -0.038
gamma_GTP                   0.0516      0.006      8.691      0.000       0.040       0.063
DRK_YN_1                    0.0165      0.004      4.708      0.000       0.010       0.023
SMK_stat_type_cd_2.0       -0.0181      0.004     -4.098      0.000      -0.027      -0.009
```

```
==============================================================================
Omnibus:                     3145.712   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7816.249
Skew:                          -0.346   Prob(JB):                         0.00
Kurtosis:                       4.723   Cond. No.                         9.95
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The final selected features are: ['const', 'age', 'height', 'weight', 'SBP', 'HDL_chole', 'LDL_chole', 'triglyceride', 'hemoglobin', 'serum_creatinine', 'gamma_GTP', 'DRK_YN_1', 'SMK_stat_type_cd_2.0']

The eliminated features along with their p-values and adjusted r-square are shown below:

Results Table:

	Removed Feature	AIC	BIC	Adjusted R-squared	P-value
0	urine_protein_3.0	46999.014585	47221.615732	0.646763	0.967998
1	urine_protein_4.0	46997.016195	47210.713297	0.646770	0.967081
2	sight_right	46995.017899	47199.810955	0.646776	0.671888
3	SMK_stat_type_cd_3.0	46993.197378	47189.086388	0.646782	0.649556
4	urine_protein_6.0	46991.403923	47178.388887	0.646787	0.319471
5	urine_protein_2.0	46990.395426	47168.476344	0.646787	0.227065
6	SGOT_ALT	46989.855126	47159.031998	0.646784	0.149062
7	SGOT_AST	46989.937698	47150.210524	0.646777	0.168456
8	sight_left	46989.834972	47141.203752	0.646771	0.067856
9	urine_protein_5.0	46991.170184	47133.634918	0.646756	0.049131
10	hear_left_2.0	46993.042225	47126.602914	0.646737	0.009864
11	BLDS	46997.703384	47122.360027	0.646700	0.004923

Final regression model and prediction of dependent variable:

The final regression model has been developed on the final selected feature. The OLS summery of the final regression model is given below:

(43516, 13)

OLS Regression Results

Dep. Variable:	tot_chole	R-squared:	0.647
Model:	OLS	Adj. R-squared:	0.647
Method:	Least Squares	F-statistic:	6641.
Date:	Thu, 07 Dec 2023	Prob (F-statistic):	0.00
Time:	02:00:10	Log-Likelihood:	-18814.
No. Observations:	43516	AIC:	3.765e+04
Df Residuals:	43503	BIC:	3.777e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0630	0.003	-23.002	0.000	-0.068	-0.058
age	0.0135	0.003	4.109	0.000	0.007	0.020
height	-0.0235	0.004	-6.323	0.000	-0.031	-0.016
weight	0.0104	0.003	3.111	0.002	0.004	0.017
SBP	0.0120	0.003	3.934	0.000	0.006	0.018
HDL_chole	0.2266	0.003	70.929	0.000	0.220	0.233
LDL_chole	0.7884	0.003	268.537	0.000	0.783	0.794
triglyceride	0.3261	0.005	63.360	0.000	0.316	0.336
hemoglobin	0.0278	0.003	8.716	0.000	0.022	0.034
serum_creatinine	-0.0797	0.015	-5.217	0.000	-0.110	-0.050
gamma_GTP	0.0502	0.007	7.504	0.000	0.037	0.063
DRK_YN_1	0.0173	0.004	4.402	0.000	0.010	0.025
SMK_stat_type_cd_2.0	-0.0179	0.005	-3.635	0.000	-0.028	-0.008

Omnibus:	2571.222	Durbin-Watson:	2.019
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6450.986
Skew:	-0.351	Prob(JB):	0.00
Kurtosis:	4.751	Cond. No.	9.97

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The develop table showing the R-squared, adjusted R-square, AIC, BIC and MSE is given below:

Final results:

	AIC	BIC	R-squared	Adjusted R-squared	MSE
0	37654.867465	37767.718956	0.646886	0.646789	202.973845

The plot of train-test and predicted variable is given below:

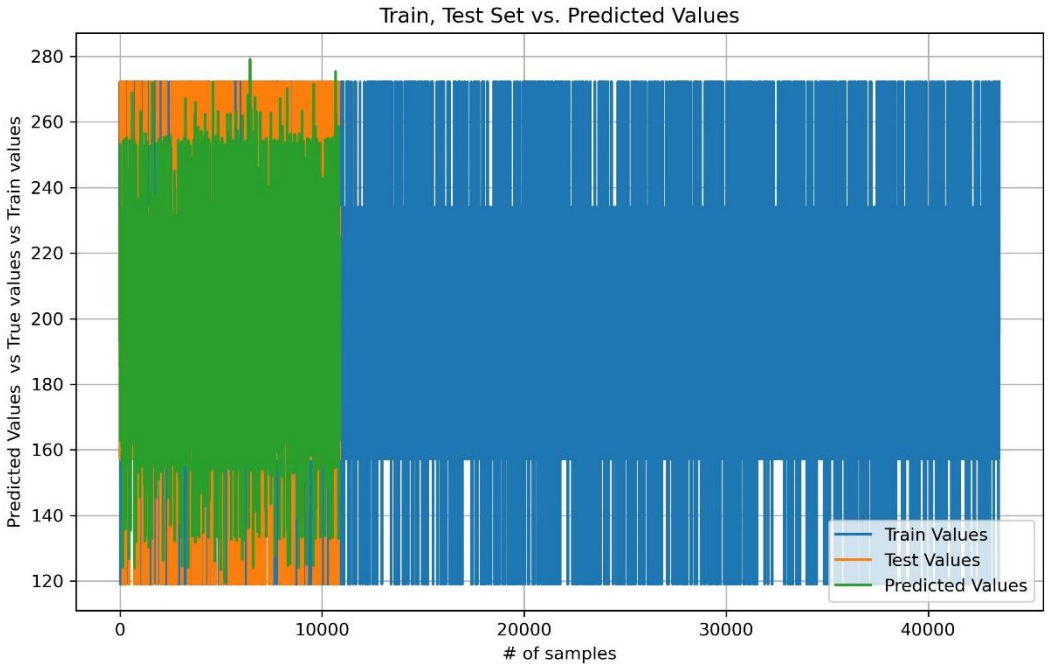


Figure 10: Train vs Test vs Predicted values

Confidence interval analysis:

The confidence interval analysis of the predicted values is shown below:

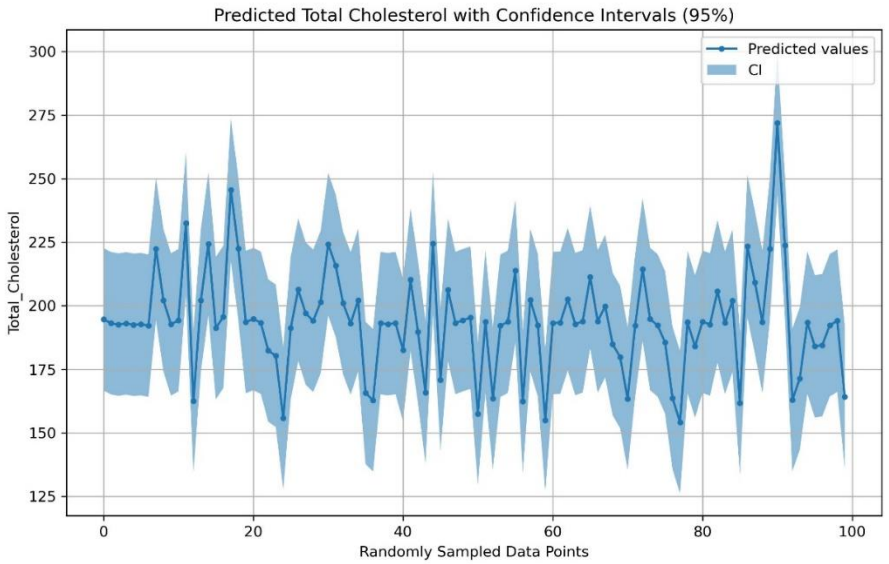


Figure 11: 95% Confidence Intervals of the Predicted values

T-test analysis:

The T-test asses the statistical significance of each coefficient based on the p-value. If the p-value is less than the model's chosen significance level (commonly 0.05), we consider the coefficient statistically significant, suggesting that the variable has a meaningful impact on the dependent variable. The t-value gives us an idea of the strength and direction of that impact.

The t-test analysis shows the following result:

```
T-Test Analysis:
                     Coefficient     t-value        p-value
const                  -0.062963  -23.002107  2.203442e-116
age                     0.013537    4.108779   3.984841e-05
height                 -0.023529   -6.323199   2.586844e-10
weight                  0.010403    3.111164   1.864720e-03
SBP                     0.012015    3.934096   8.364009e-05
HDL_chole               0.226604   70.929445   0.000000e+00
LDL_chole               0.788398  268.536596   0.000000e+00
triglyceride            0.326083   63.359652   0.000000e+00
hemoglobin              0.027823    8.715939   2.983303e-18
serum_creatinine       -0.079721   -5.216990   1.826860e-07
gamma_GTP               0.050241    7.504434   6.286888e-14
DRK_YN_1                0.017263    4.401958   1.075333e-05
SMK_stat_type_cd_2.0   -0.017935   -3.635191   2.780944e-04
```

The T-test analysis in the output is showing the results of t-tests for each coefficient in the final linear regression model. Let's break down the key components:

- **Coefficient**: This is the estimated value for the coefficient of each variable in the linear regression model. It represents the change in the dependent variable for a one-unit change in the corresponding independent variable, assuming all other variables are held constant.
- **t-value**: The t-value is a measure of how many standard deviations a coefficient estimate is from zero. It is calculated as the coefficient divided by its standard error. The larger the absolute value of the t-value, the more evidence we have against the null hypothesis that the true value of the coefficient is zero.
- **p-value**: The p-value is the probability of observing a t-statistic as extreme as the one computed from your data, assuming that the null hypothesis is true (i.e., the true coefficient is zero). If the p-value is below a certain significance level (commonly 0.05), we reject the null hypothesis.

Now, let's interpret the results for a few variables:

1. **HDL_chole**: The coefficient is 0.2266, and the t-value is very high (70.93). The p-value is essentially zero (scientific notation is used). This suggests a strong indication that HDL_chole has a significant and positive impact on the dependent variable.
2. **LDL_chole**: Similar to HDL_chole, the coefficient is 0.7884, the t-value is extremely high (268.54), and the p-value is zero. This indicates a highly significant and positive impact of LDL_chole on the dependent variable.
3. **serum_creatinine**: The coefficient is -0.0797, the t-value is -5.22, and the p-value is small (1.83e-07). This suggests a significant negative impact of serum_creatinine on the dependent variable.

4. **SMK_stat_type_cd_2.0**: The coefficient is -0.0179, the t-value is -3.64, and the p-value is 0.000278. This indicates that the variable is statistically significant, and its effect is negative.

   Lastly, we can say that, since none of the p-values are greater than 0.05, we can finally reject the null hypothesis, meaning all the coefficients chosen by the model are statistically significant.

F-test analysis:

The F-test analysis used to test the overall significance of the regression model. It measures the ratio of the variance explained by the model to the variance not explained. In simple terms, it assesses whether the linear regression model as a whole is statistically significant. The output of the F-test is given below:

```
F-Test Analysis:
<F test: F=6133.716622712181, p=0.0, df_denom=4.35e+04, df_num=11>
```

Let's break down the key components:

- **p-value (p)**: The p-value associated with the F-statistic is used to test the null hypothesis that all the coefficients in the model are equal to zero (i.e., none of the independent variables have a significant effect on the dependent variable). A low p-value (typically less than 0.05) suggests that we can reject this null hypothesis.
- **Degrees of freedom (df_denom and df_num)**:
  - df_denom: Degrees of freedom for the denominator, which is related to the error term in your model. It represents the number of observations minus the number of parameters estimated.
  - df_num: Degrees of freedom for the numerator, which is related to the model itself. It represents the number of parameters estimated.

Now, let's interpret the results:

- **F-statistic**: The value of the F-statistic is 6133.72. This is a large value, indicating that the variance explained by the model is much greater than the unexplained variance.
- **p-value**: The p-value is very close to zero (scientific notation is used). This suggests strong evidence against the null hypothesis that all coefficients are zero, indicating that at least one of the independent variables has a significant effect on the dependent variable.

In summary, the low p-value associated with the F-statistic indicates that your overall regression model is statistically significant. In other words, there is evidence that at least one of the independent variables in the model has a significant effect on the dependent variable.

Conclusions after phase 2:

- **Statistical Significance (T-test):** All coefficients in the model were found to be statistically significant (p-values < 0.05).

- **Selected Features:** The stepwise regression selected features for the final regression model are: ['const', 'age', 'height', 'weight', 'SBP', 'HDL_chole', 'LDL_chole', 'triglyceride', 'hemoglobin', 'serum_creatinine', 'gamma_GTP', 'DRK_YN_1', 'SMK_stat_type_cd_2.0'] .
- **Predictive Power:** The model demonstrated a good predictive capability, as indicated by R-squared 64.68% and adjusted R-squared 64.67% with an MSE of 202.97.
- **Feature Contributions:** Features like HDL_chole, LDL_chole, serum_creatinine, and SMK_stat_type_cd_2.0 significantly influenced the dependent variable.
- **Overall Model Significance:** The F-test confirmed the overall significance of the regression model, reinforcing its reliability for predictive analysis.
- **Rejecting Null Hypothesis:** Given all p-values were below 0.05, the null hypothesis was rejected, affirming the statistical significance of all chosen coefficients and the model.

In summary, the detailed analysis using stepwise regression, adjusted R-square, t-tests, and F-tests provided a comprehensive understanding of the selected dataset's predictive model for the continuous numerical feature "tot_chole." The results emphasize the reliability, significance, and interpretability of the developed regression model.

Phase 3

In this phase, various machine learning classifiers are applied to the selected dataset and the best technique has been picked to recommend a classifier with the highest performance. The purpose of this phase is to improve the performance of classification and make sure that the classifier is not overfitted or underfitted. Grid search for hyper parameter search has been performed for each classifier. The following classifiers are explored for this project:

- **Decision Tree:** Decision trees are versatile models with parameters to be optimized. Pre-pruning involves stopping the tree's growth early, while post-pruning prunes branches afterward. Grid search helps find optimal hyperparameters like 'criterion,' 'splitter,' 'max_depth,' 'min_samples_split,' 'max_features,' and 'ccp_alpha.' Optimizing the Cost Complexity Pruning (CCP) alpha parameter is crucial for controlling the trade-off between complexity and accuracy.

  Without pruning the Decision Trees clearly overfits and displays a training accuracy of 100% and testing accuracy of 52%:

```
Train accuracy :1.00
Test accuracy :  0.52
```

In order to improve the intuitiveness of the tree, the feature importance of the features has been found out and some features have been eliminated on threshold=0.05:

```
Feature Importances:
             Feature  Importance
18             sex_1    0.230581
12      triglyceride    0.070841
17         gamma_GTP    0.060305
10         HDL_chole    0.059973
8               BLDS    0.055175
11         LDL_chole    0.054585
16          SGOT_ALT    0.054494
9          tot_chole    0.053226
7                DBP    0.052312
15          SGOT_AST    0.051499
6                SBP    0.051010
3          waistline    0.050354
0                age    0.047561
2             weight    0.029581
13        hemoglobin    0.026250
1             height    0.025580
5        sight_right    0.009625
14  serum_creatinine    0.008564
4         sight_left    0.008484
Eliminated Features:
['age', 'height', 'weight', 'sight_left', 'sight_right', 'hemoglobin', 'serum_creatinine']

Final Selected Features:
['waistline', 'SBP', 'DBP', 'BLDS', 'tot_chole', 'HDL_chole', 'LDL_chole', 'triglyceride', 'SGOT_AST', 'SGOT_ALT', 'gamma_GTP', 'sex_1']
```

After pre pruning with the best hyper-parameters the tree becomes optimized and does not overfit. The train and test accuracy after pre-pruning is given below:

```
Train accuracy on best model :0.62
Test accuracy on best model:  0.60
```

The best hyper-parameters are : {'criterion': 'gini', 'max_depth': 7, 'max_features': None, 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'best'}. The Pre pruned Decision Tree is shown below:

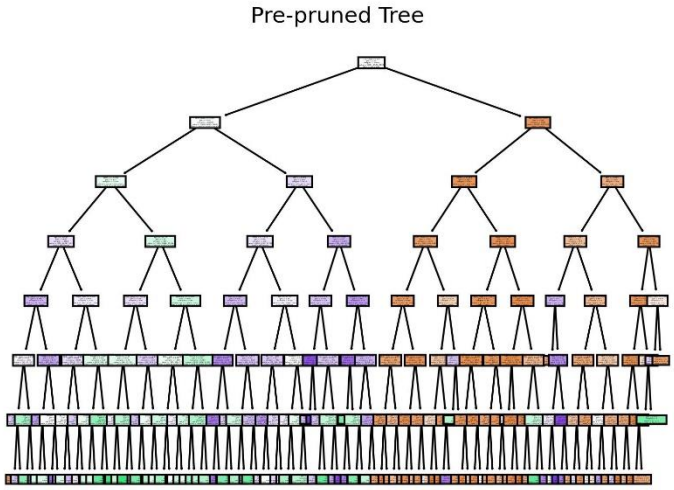


Figure 12: Pre-Pruned Decision Tree

The post pruning optimizes the cost complexity function by selecting the optimized alpha by grid-search. The plot for accuracy vs alpha is given below:

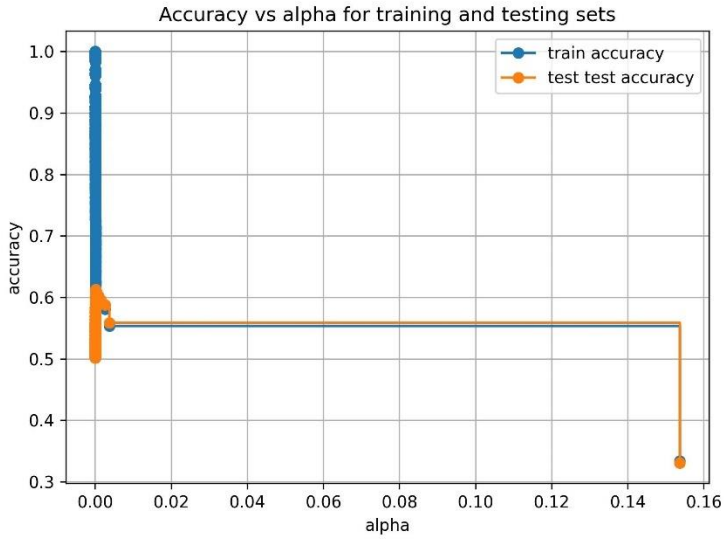


Figure 13: Alpha vs Accuracy for train and test sets

Optimized alpha is shown below:

```
Optimal Alpha: 0.00017839864883602806
```

The post pruned tree train and test accuracy is given below:

```
Train accuracy on pruned model :0.62
Test accuracy on pruned model:  0.61
```

We can see test accuracy has improved by post-pruning.

Finally, the post pruned tree:

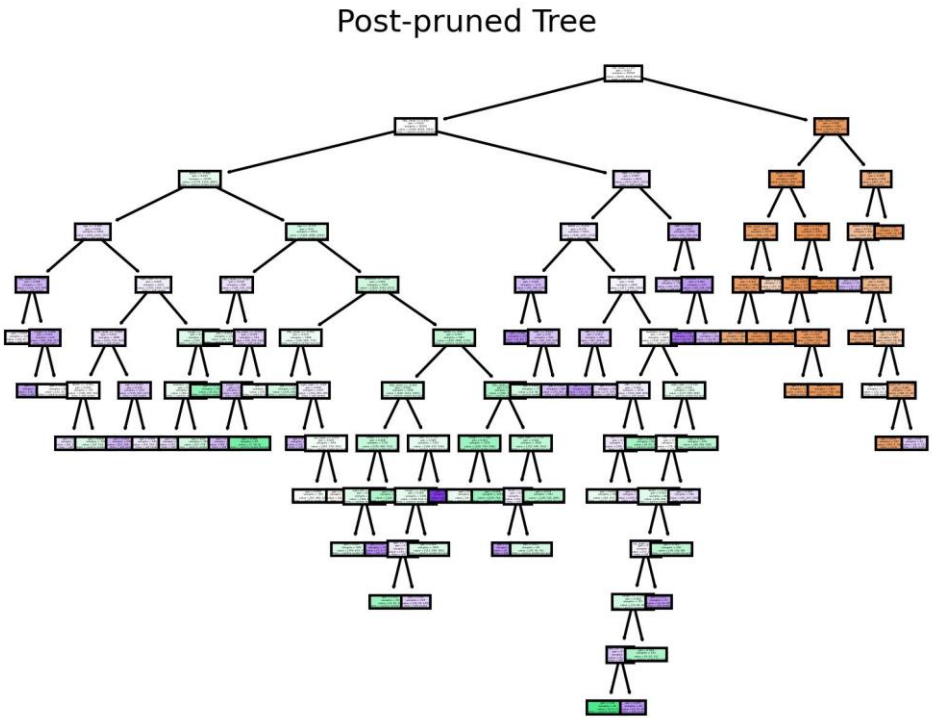


Figure 14: Post Pruned Decision Tree

- **Logistic Regression:** Logistic regression is a linear model for binary classification. It involves optimizing coefficients through techniques like gradient descent. Regularization parameters, such as L1 or L2 penalties, can be tuned for optimal performance. Logistic regression is suitable for probability estimation and is sensitive to feature scaling. For finding the best logistic regression technique, the grid search has been performed to find the best parameters:

```
Best Parameters for Logistic Regression: {'C': 1, 'penalty': 'l1', 'solver': 'liblinear'}
```

- **K-Nearest Neighbors (KNN):** KNN is a non-parametric algorithm relying on the choice of 'K.' The optimum 'K' can be determined using the elbow method, which plots accuracy against different 'K' values. It is essential to consider the balance between overfitting (low 'K') and underfitting (high 'K').

For finding best K, wcss- elbow method has been used. The plot of wcss vs k-values are given below:

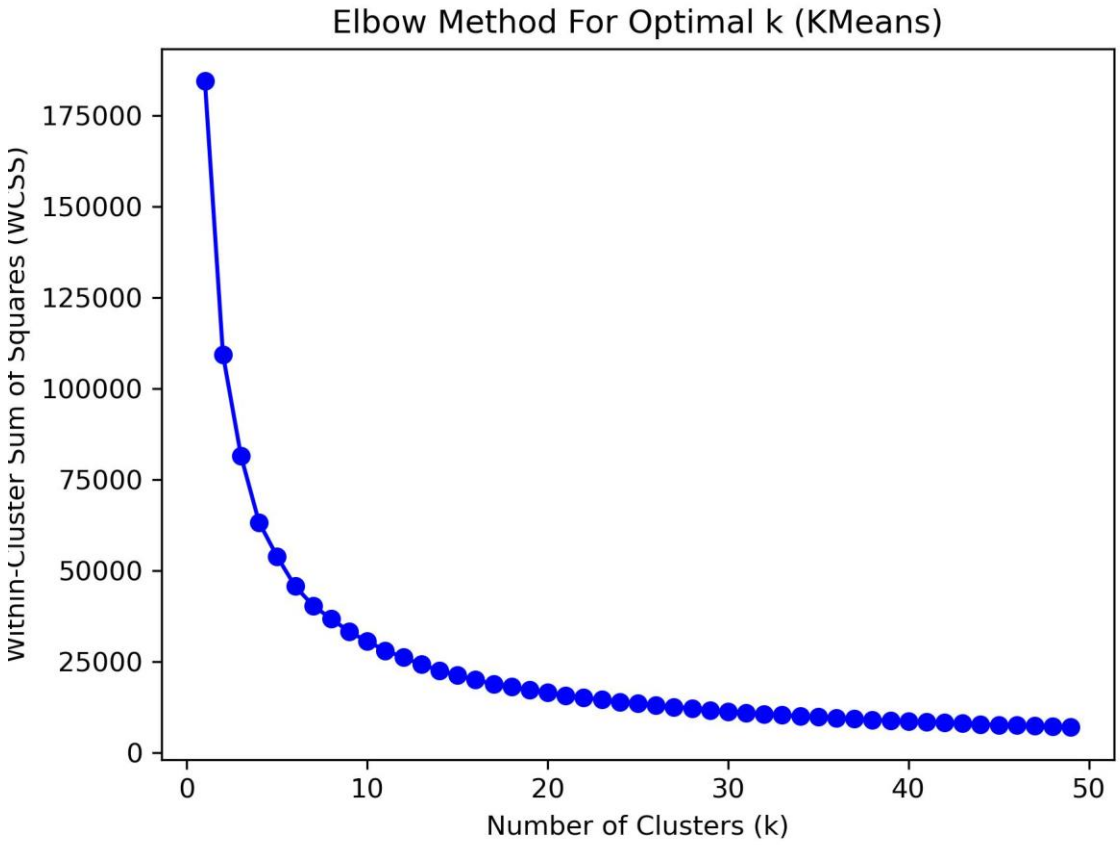


Figure 15: Elbow Method for Optimal K

But finding best k from the plot is somewhat difficult as the elbow is not sharp enough. But it can be seen that the k lies somewhat between 10-25 range. So to find the best k, the k-values in range of 10-25 has been taken to test and train KNN and best-k has been found by he measure of accuracy scores:

```
k = 10: Accuracy = 0.5673
k = 11: Accuracy = 0.5686
k = 12: Accuracy = 0.5703
k = 13: Accuracy = 0.5745
k = 14: Accuracy = 0.5772
k = 15: Accuracy = 0.5752
k = 16: Accuracy = 0.5795
k = 17: Accuracy = 0.5809
k = 18: Accuracy = 0.5811
k = 19: Accuracy = 0.5838
k = 20: Accuracy = 0.5838
k = 21: Accuracy = 0.5844
k = 22: Accuracy = 0.5867
k = 23: Accuracy = 0.5889
k = 24: Accuracy = 0.5910
k = 25: Accuracy = 0.5889
Best k value: 24 with Accuracy = 0.5910
```

- **Support Vector Machine (SVM):** SVM supports different kernels, including linear, polynomial, and radial basis function (RBF). Parameter tuning involves optimizing the kernel choice and adjusting parameters like 'C' for regularization and 'gamma' for the RBF kernel. SVM is effective for both linear and non-linear classification tasks. The result of the grid search of SVM is given below:

```
Best Params of SVC: {'C': 1, 'kernel': 'linear'}
```

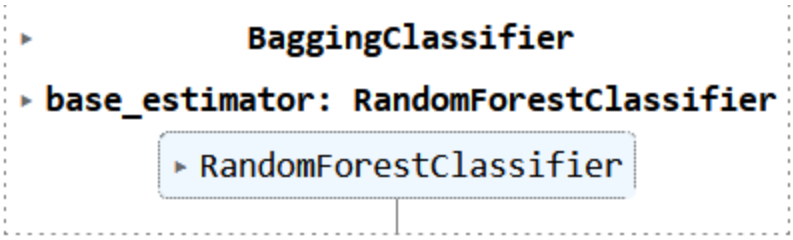
- **Naïve Bayes:** Naïve Bayes is a probabilistic model based on Bayes' theorem. It assumes independence between features, making it computationally efficient. Hyperparameter tuning is minimal, and it works well for text classification and simple classification tasks. For this project, the best parameters of naïve bayes are:

```
Best Hyperparameters for Naive bayes: {'var_smoothing': 1e-09}
```

- **Random Forest:** Random Forest is an ensemble model that supports bagging, stacking, and boosting. It consists of multiple decision trees and can be fine-tuned by adjusting hyperparameters like the number of trees, depth of trees, and feature selection criteria. Random Forest provides robust performance and is less prone to overfitting. The best parameters of the Random Forest classifier are:

```
Best Params of RF: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 200}
```

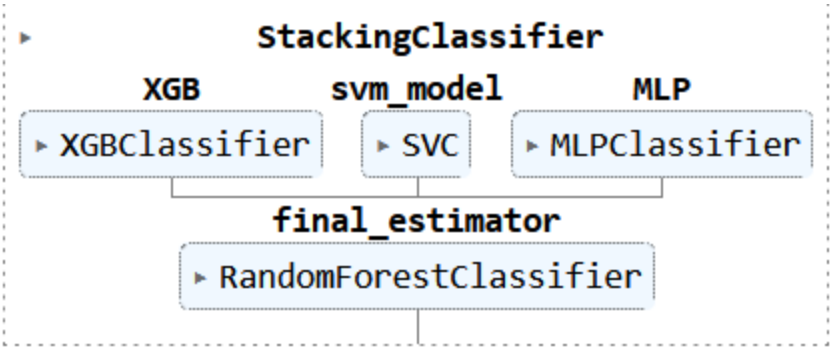
For this project, the bagging, boosting and the stacking methods are explored. Bagging has been done with the Random Forest.



The boosting has been done with Extra Gradient Boosting algorithm. The best parameters for the boosting technique has been found by performing grid search:

```
Best Params of XGB: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}
```

In case of stacking, the base classifiers are XGB, MLP and Support Vector Classifier with best parameters from the grid search. The final estimator is the Random Forest Classifier.



- **Neural Network (Multi-layered Perceptron):** Neural Networks, particularly Multi-layered Perceptrons (MLPs), involve multiple layers of interconnected nodes. Training involves optimizing weights using techniques like backpropagation. Hyperparameters include the number of layers, nodes per layer, activation functions, and learning rate. Proper initialization and regularization techniques help prevent overfitting. The best parameters of the MLP Classifier are:

```
Best Parameters for MLPC: {'activation': 'logistic', 'alpha': 0.001, 'hidden_layer_sizes': (50, 50), 'max_iter': 300}
```

Results of the Grid Search:

For each model a grid search has been performed to find the best parameters. The results of the grid search are depicted in the table below:

Table 2: Results of the Grid Search

Algorithm	Best Parameters
Decision Trees Pre Pruning	Best Parameters of DT after Pre Pruning: {'criterion': 'gini', 'max_depth': 7, 'max_features': None, 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'best'}
Decision Trees Post Pruning	Optimal Alpha: 0.00017839864883602806
Logistic Regression	Best Parameters for Logistic Regression: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
K Nearest Neighbours	Best k value: 24
Support Vector Machine	Best Params of SVC: {'C': 1, 'kernel': 'linear'}
Naïve Bayes	Best Hyperparameters for Naive bayes: {'var_smoothing': 1e-09}
Random Forest	Best Params of RF: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 200}
Extra Gradient Boosting	Best Params of XGB: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}
Multiple Perceptron Classifier	Best Parameters for MLPC: {'activation': 'logistic', 'alpha': 0.001, 'hidden_layer_sizes': (50, 50), 'max_iter': 300}

Evaluating the models:

For evaluation of the models the following metrics have been used. The evaluation has been performed by Stratified k-fold cross validation. Stratified K-Fold Cross-Validation ensures that each fold maintains the same class distribution as the original dataset. It enhances the robustness of model evaluation, especially when dealing with imbalanced datasets, by providing more reliable performance metrics across different subsets of data.

Confusion Matrix: The confusion matrix provides a detailed breakdown of model performance, showing true positives, true negatives, false positives, and false negatives. It's a valuable tool for understanding the distribution of predictions.

Precision: Precision is the ratio of true positives to the sum of true positives and false positives. It indicates the accuracy of positive predictions, emphasizing the relevance of the identified positive instances.

Sensitivity or Recall: Sensitivity, or recall, measures the ratio of true positives to the sum of true positives and false negatives. It quantifies the model's ability to capture all positive instances, minimizing false negatives.

Specificity: Specificity is the ratio of true negatives to the sum of true negatives and false positives. It evaluates the model's capacity to correctly identify negative instances, reducing false positives.

F-score: The F-score, or F1-score, is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance, particularly when there's an imbalance between positive and negative classes.

ROC and AUC Curve: The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate. The Area Under the Curve (AUC) summarizes the ROC curve's performance, with higher values indicating better discrimination ability.

The precision, Recall, ROC-AUC, Specificity and F1 scores are given in the table below:

Table 3: Performance Evaluation Table

Model	Precision	Recall	Specificity	F-score	Roc Auc
Logistic Regression	0.6264	0.6098	{0: 0.93, 1: 0.70, 2: 0.78}	0.6149	{0: 0.85, 1: 0.72, 2: 0.72},
K-Nearest Neighbors	0.6083	0.5909	{0: 0.92, 1: 0.70, 2: 0.75}	0.5965	{0: 0.83, 1: 0.70, 2: 0.70}
Naive Bayes	0.5264	0.5294	{0: 0.73, 1: 0.62, 2: 0.93}	0.4843	{0: 0.82, 1: 0.70, 2: 0.67}
Random Forest	0.6392	0.6247	{0: 0.93, 1: 0.74, 2: 0.76}	0.6300	{0: 0.86, 1: 0.74, 2: 0.74}
DT Pre Pruned	0.6208	0.6036	{0: 0.93, 1: 0.74, 2: 0.72}	0.6099	{0: 0.85, 1: 0.72, 2: 0.72}
DT Post Pruned	0.6291	0.6128	{0: 0.93, 1: 0.73, 2: 0.75}	0.6187	{0: 0.85, 1: 0.72, 2: 0.72}
Bagging	0.6603	0.6463	{0: 0.93, 1: 0.79, 2: 0.74}	0.6508	{0: 0.87, 1: 0.78, 2: 0.77}
Stacking	0.6472	0.6350	{0: 0.92, 1: 0.76, 2: 0.76}	0.6395	{0: 0.86, 1: 0.76, 2: 0.76}
Boosting	0.6415	0.6272	{0: 0.93, 1: 0.74, 2: 0.76}	0.6323	{0: 0.86, 1: 0.75, 2: 0.74}
SVC	0.6280	0.6108	{0: 0.93, 1: 0.68, 2: 0.80}	0.6144	{0: 0.82, 1: 0.72, 2: 0.72}
MLP	0.6384	0.6222	{0: 0.93, 1: 0.70, 2: 0.79}	0.6264	{0: 0.86, 1: 0.74, 2: 0.74}

Analysis of Performance Evaluation Table:

Based on the provided performance metrics, let's assess the models and declare the best-performing one:

1. **Precision:**
    - **Highest Precision:** Bagging achieves the highest precision of 0.6603, indicating its ability to correctly classify positive instances.
2. **Recall:**
    - **Highest Recall:** Bagging also demonstrates the highest recall of 0.6463, showcasing its effectiveness in capturing positive instances.

3. **Specificity:**
    - **Highest Specificity:** Random Forest attains the highest specificity values for all classes, indicating its proficiency in correctly identifying negative instances.
4. **F-score:**
    - **Highest F-score:** Bagging leads with the highest F-score of 0.6508, considering both precision and recall.
5. **Roc Auc:**
    - **Highest Roc Auc:** Bagging and Stacking both achieve the highest Roc Auc values across classes, indicating their strong overall performance.

The Confusion matrices after stratified k-fold cross validation are given below:

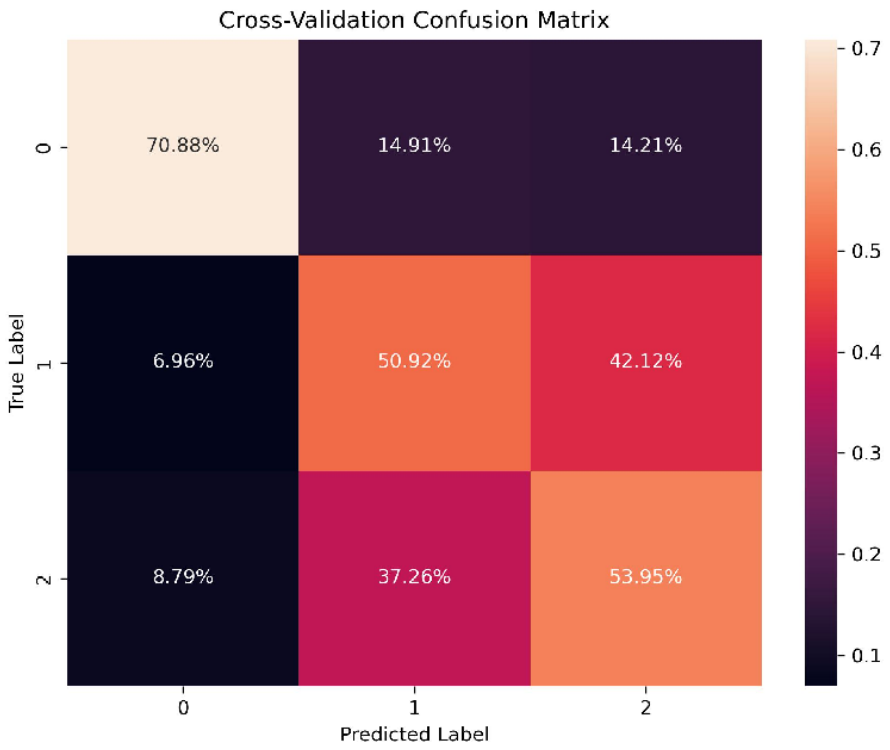


Figure 16: Confusion Matrix of Pre-Pruned DT

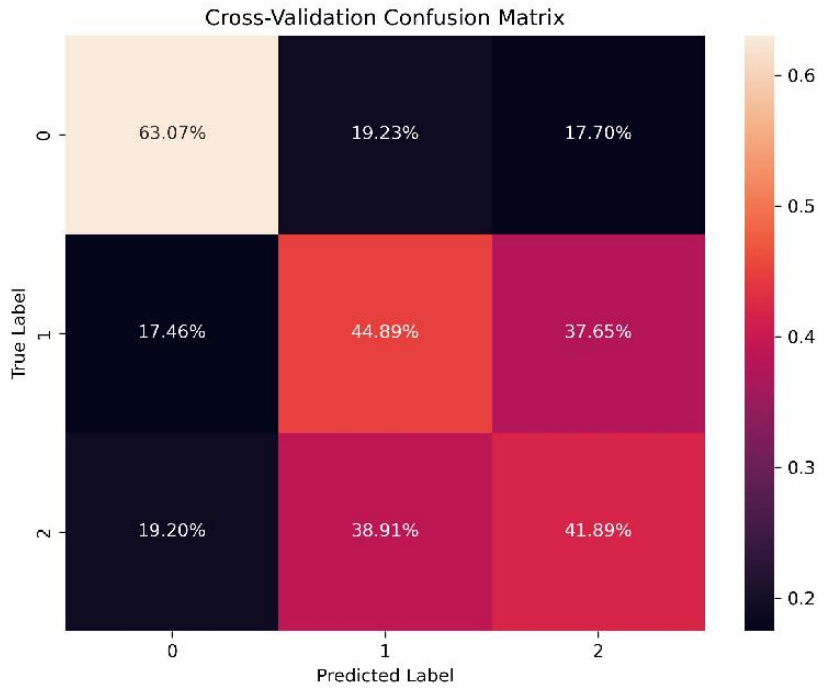


Figure 17: Confusion Matrix of Post Pruned DT

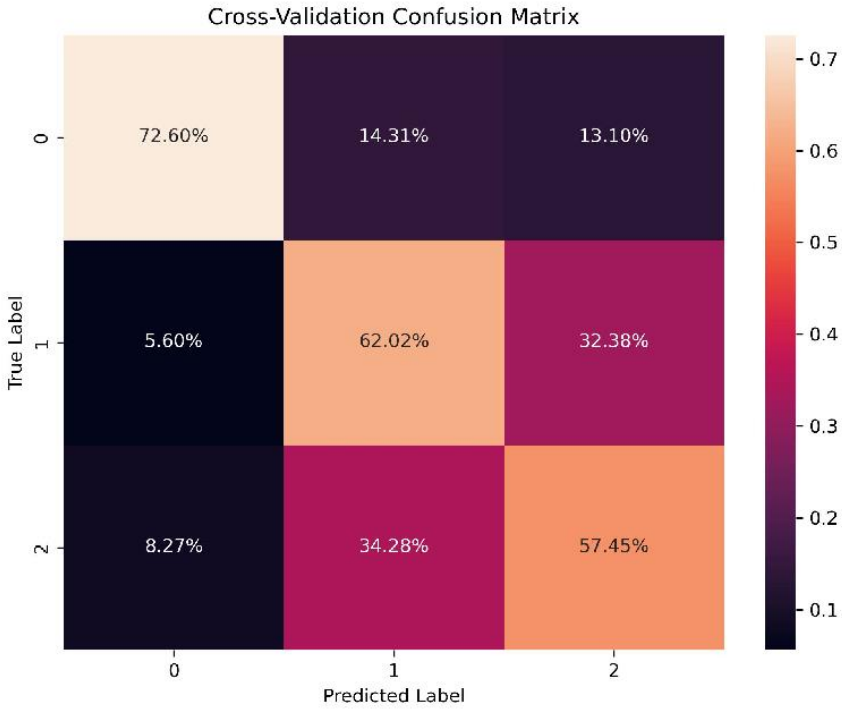


Figure 18: Confusion Matrix of Logistic Regression

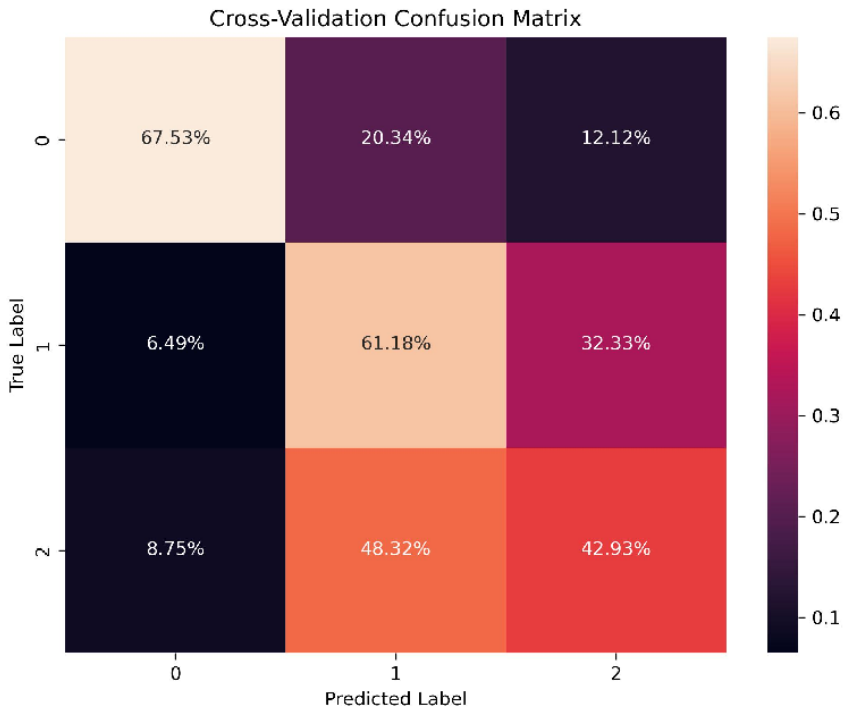


Figure 19: Confusion Matrix of KNN

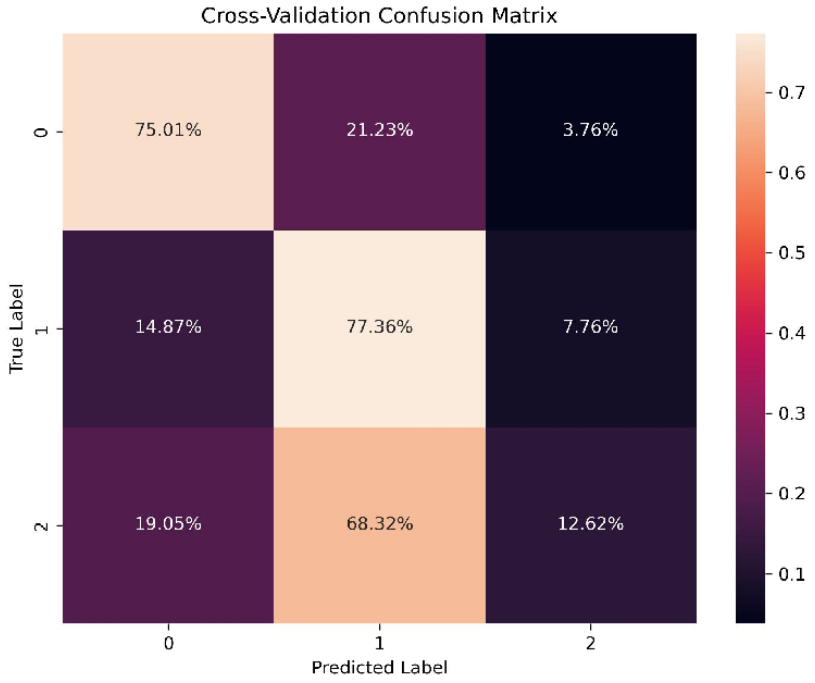


Figure 20: Confusion matrix of Naïve Bayes

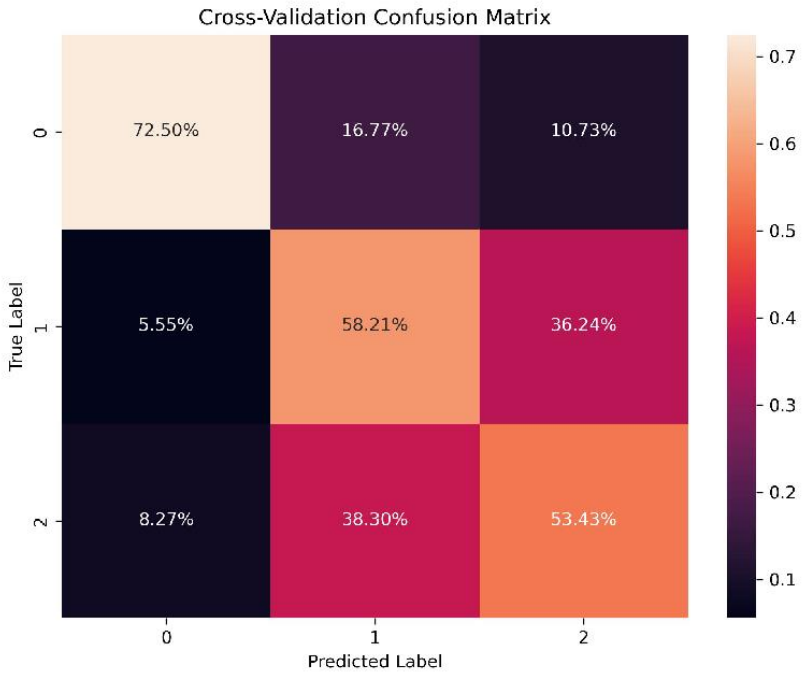


Figure 21: Confusion Matrix of Random Forest

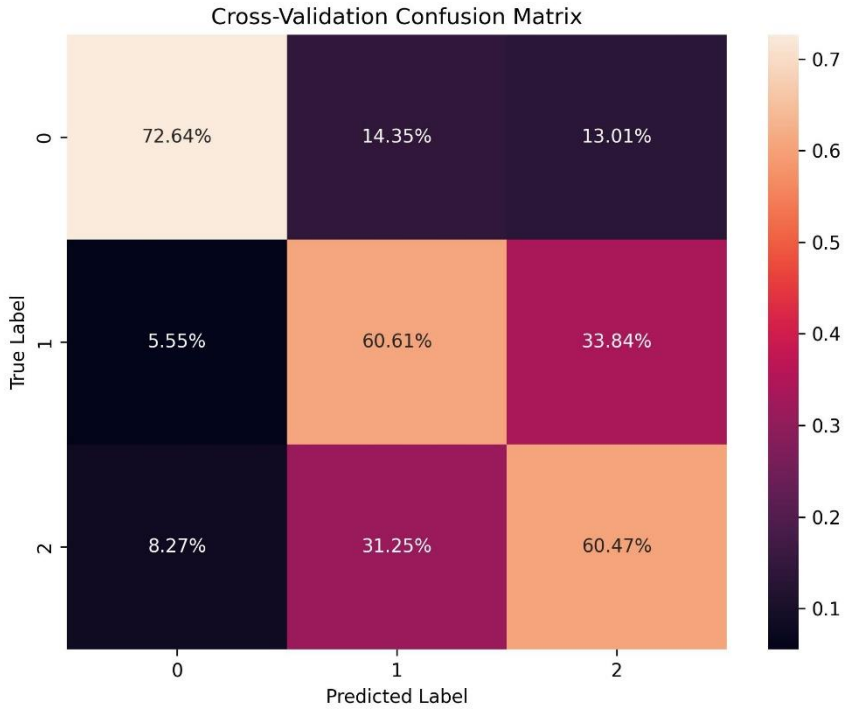


Figure 22: Confusion Matrix of Bagging Classifier

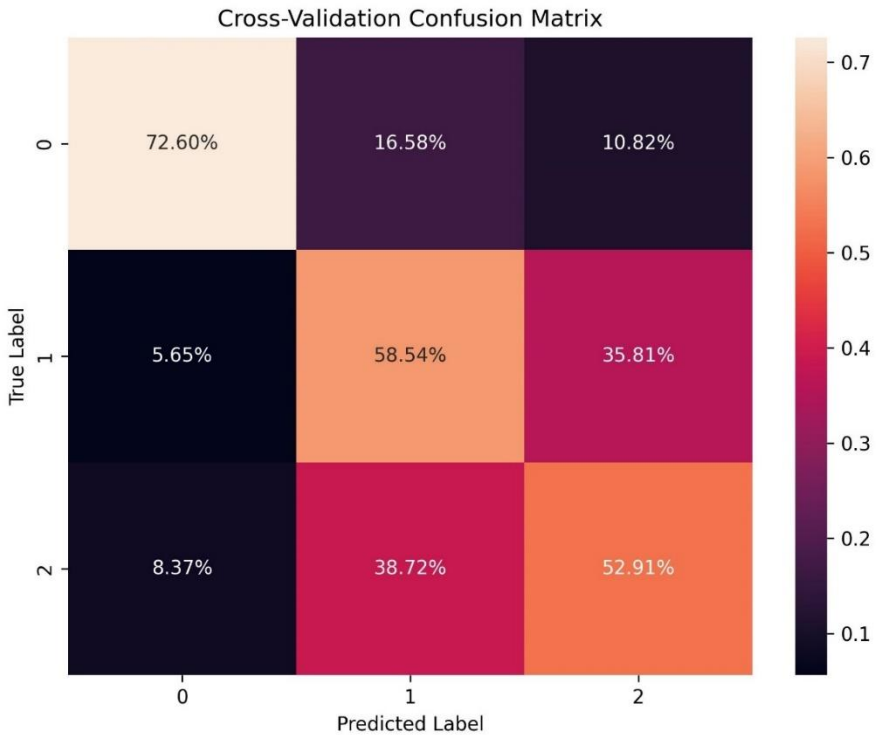


Figure 23: Confusion Matrix of Boosting Classifier

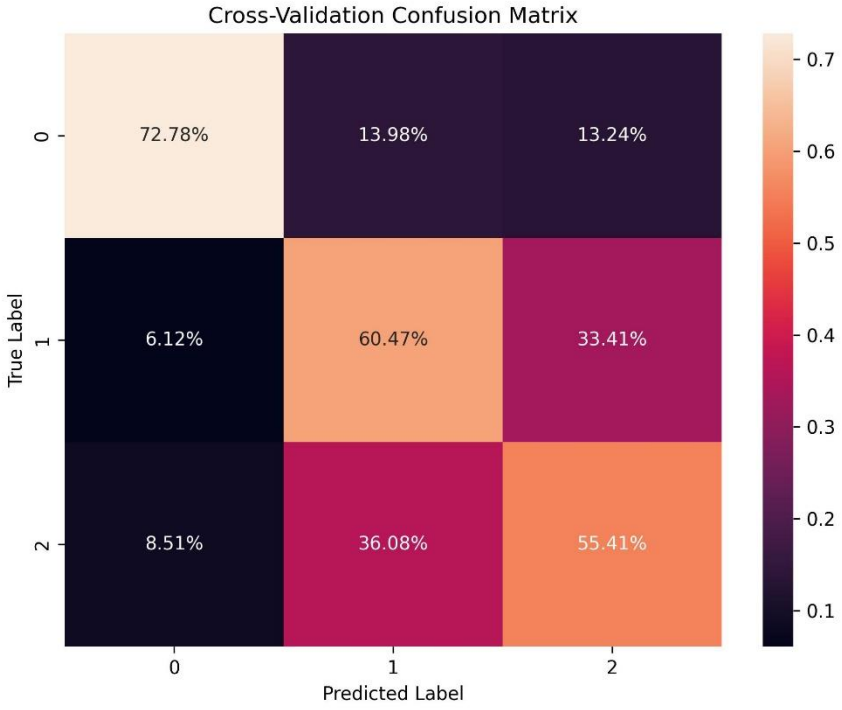


Figure 24: Confusion Matrix of Stacking Classifier

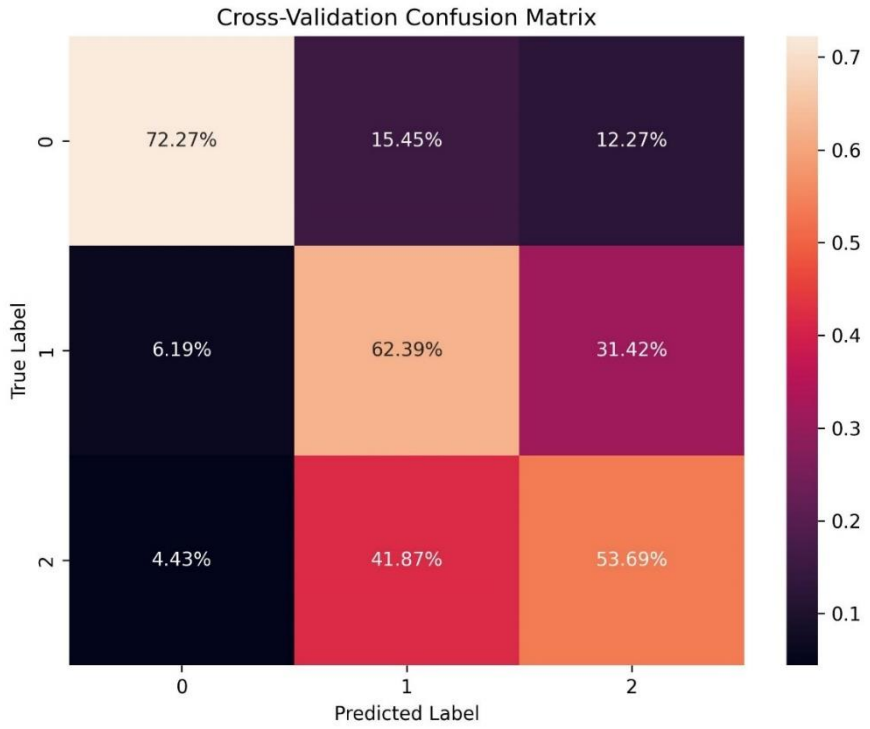


Figure 25: Confusion Matrix of the Stacking Classifier

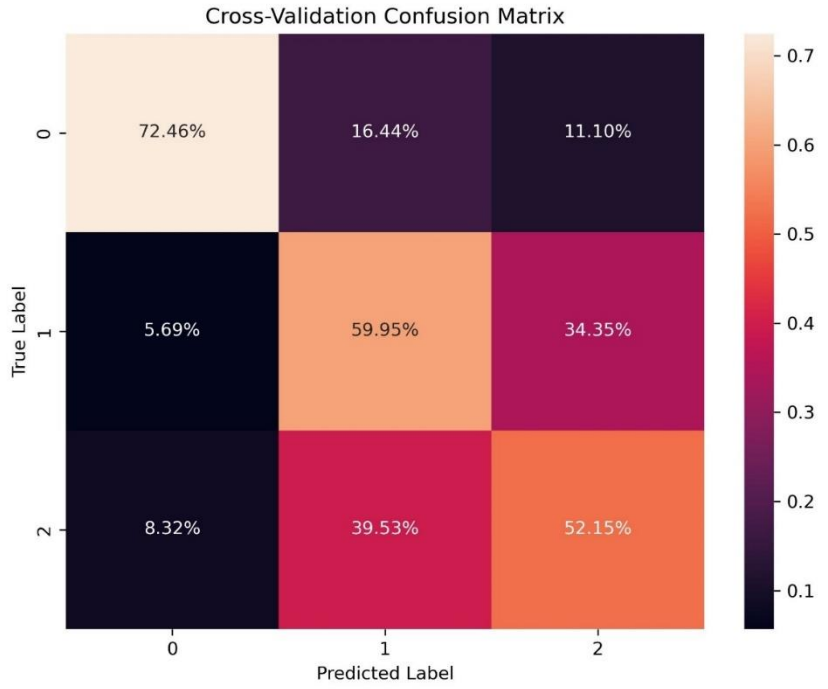


Figure 26: Confusion Matrix of the MLP Classifier

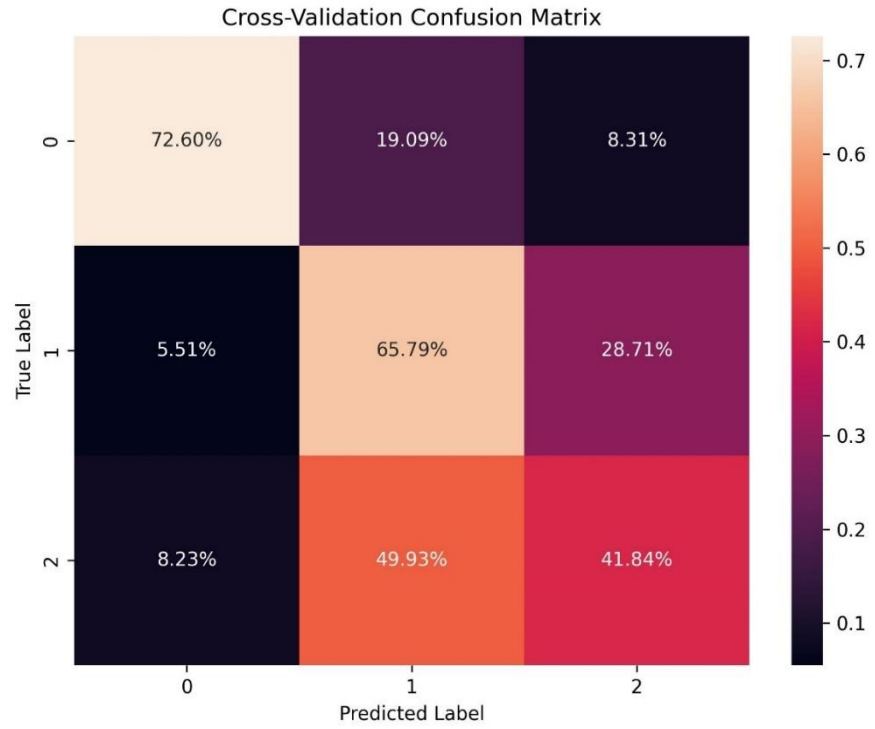


Figure 27: Confusion Matrix of SVC

The ROC-AUC plots after stratified k-fold cross validation are given below:

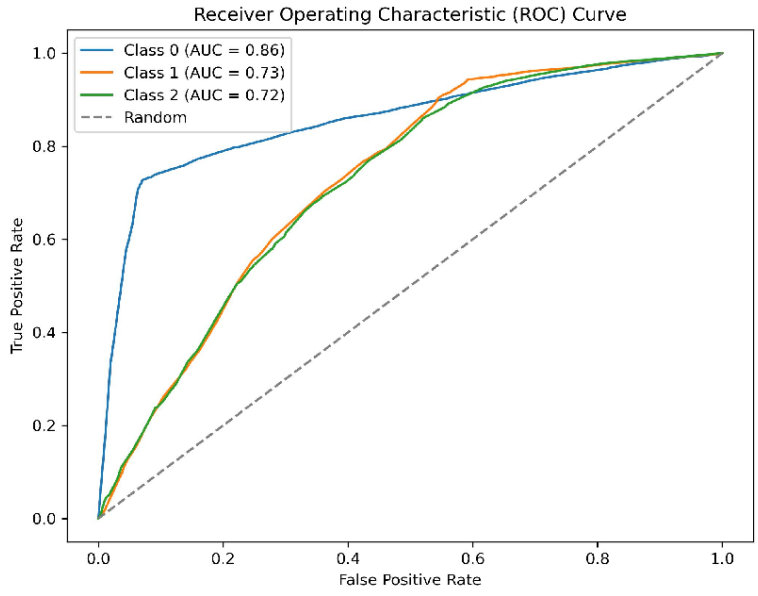


Figure 28: ROC of Post Pruned DT

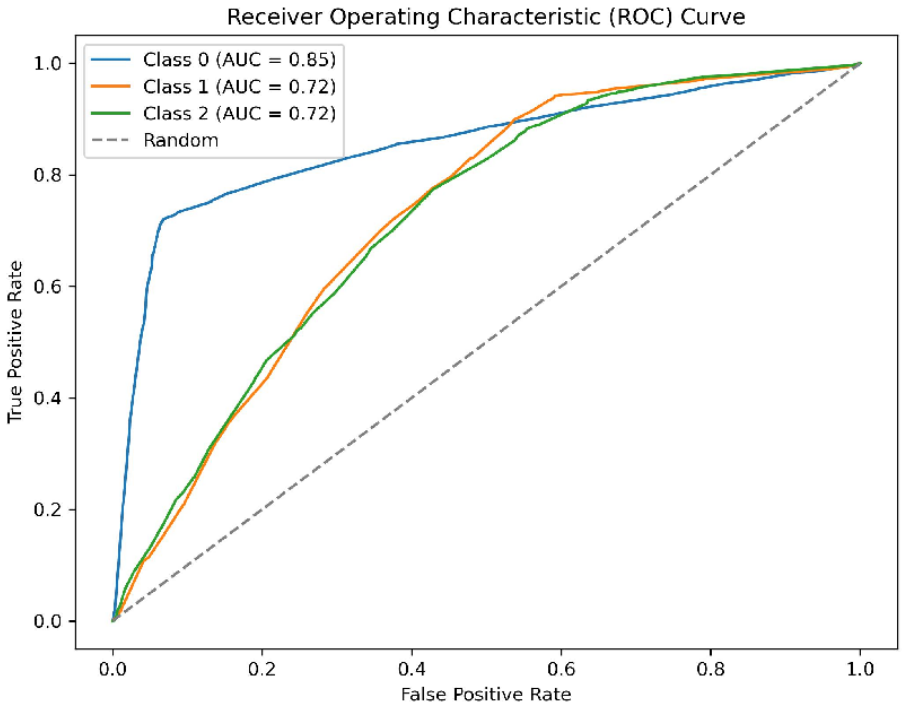


Figure 29: ROC of Pre Pruned DT

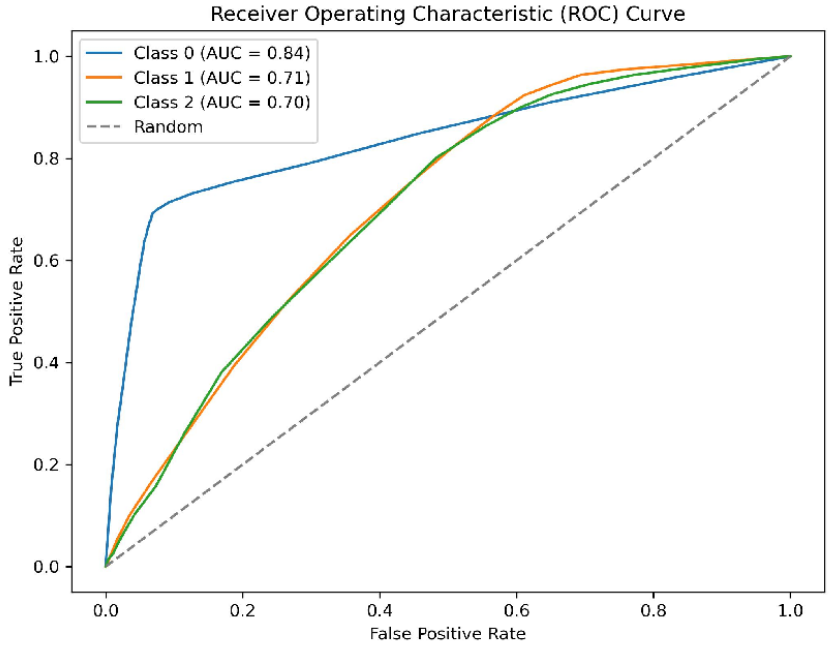


Figure 30: ROC of KNN

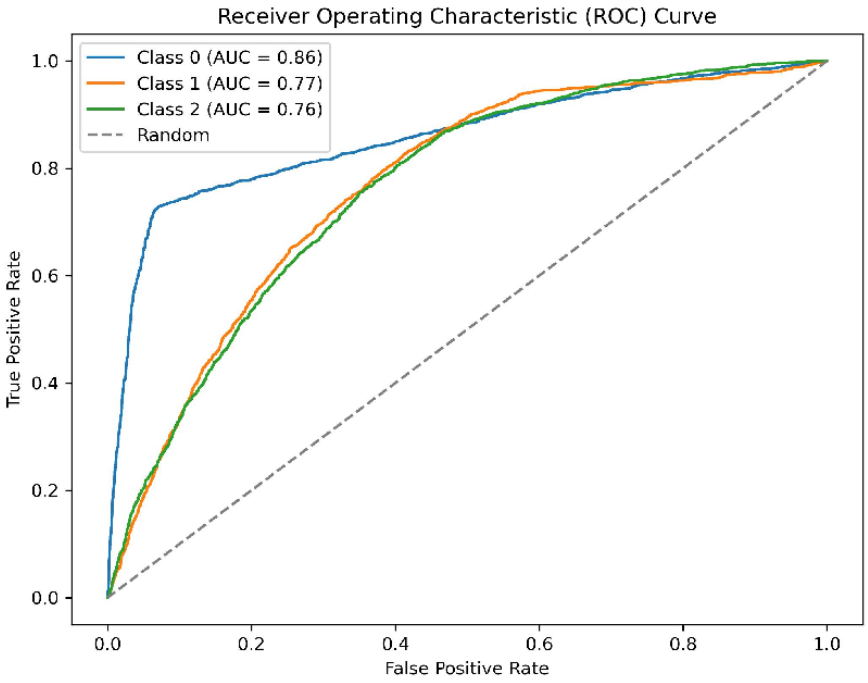


Figure 31: ROC of Logistic Regression

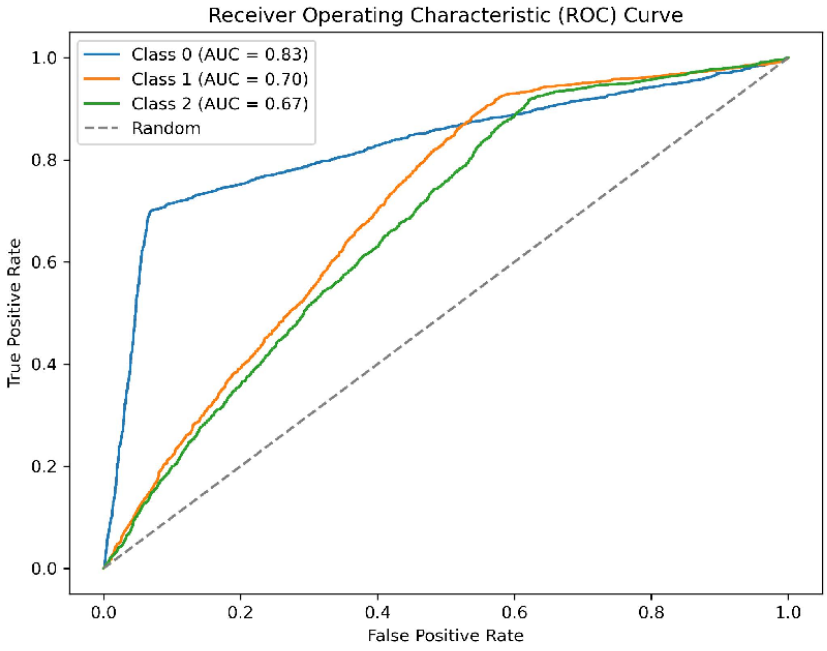


Figure 32: ROC of Naïve Bayes

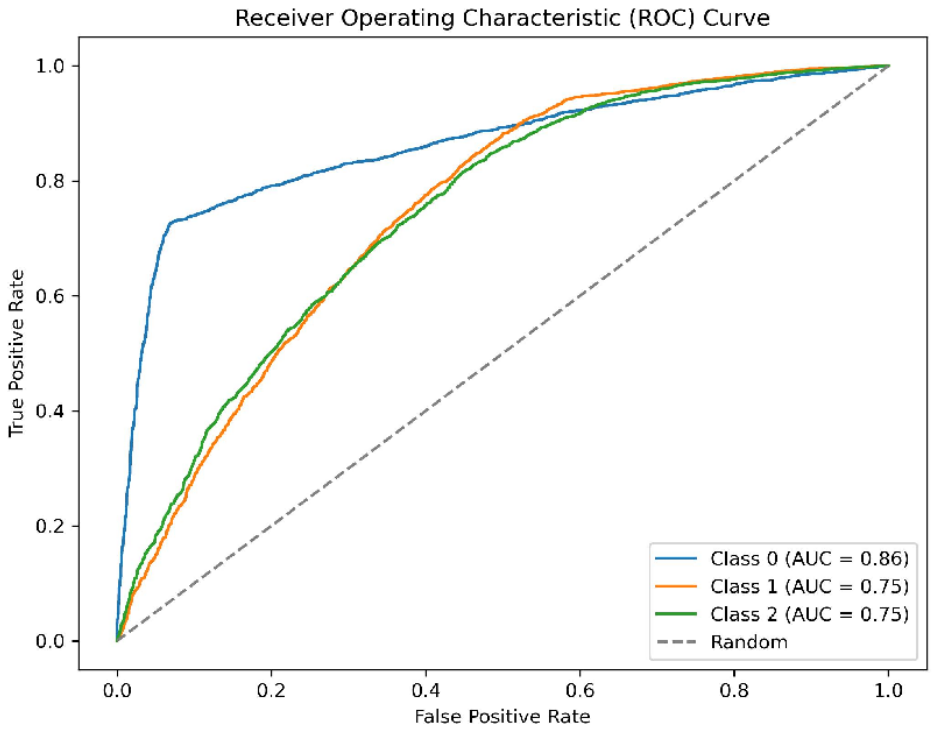


Figure 33: ROC of Random Forest

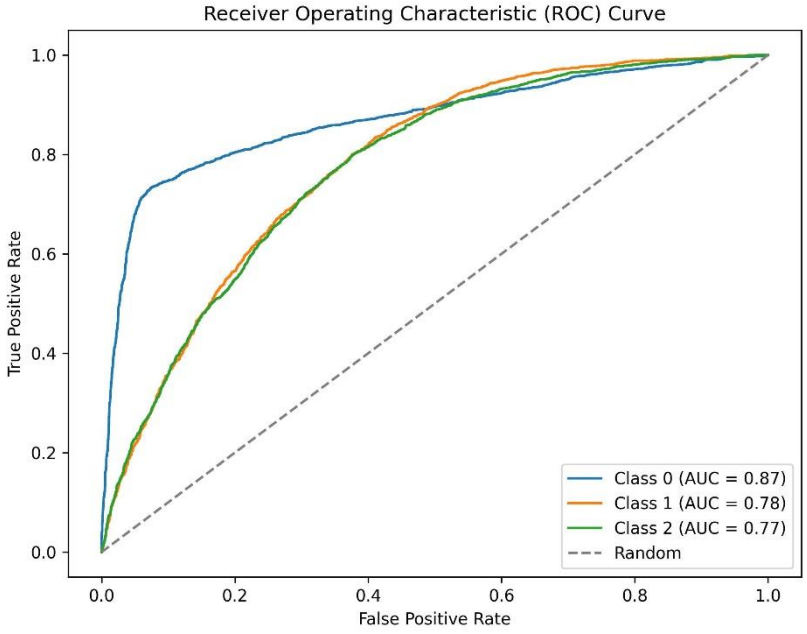


Figure 34: ROC of Bagging Classifier

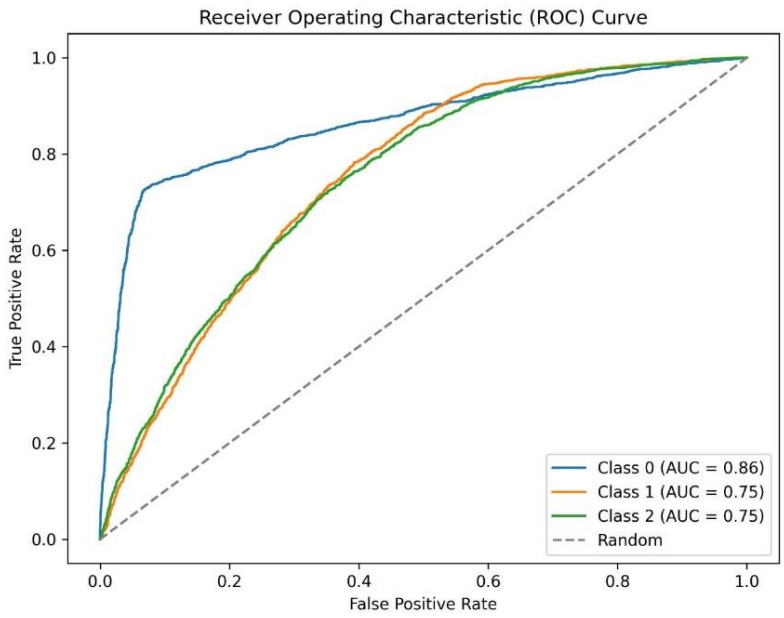


Figure 35: ROC of Boosting Classifier

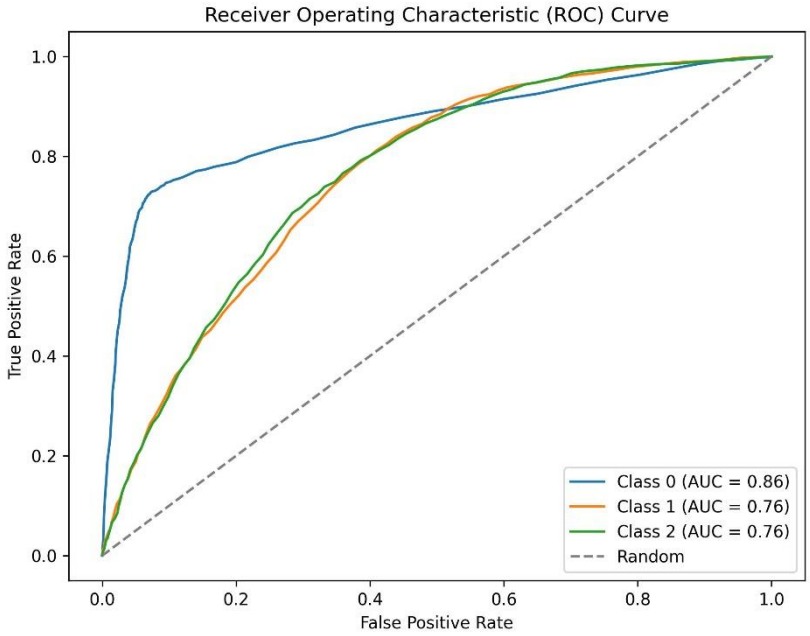


Figure 36: ROC of Stacking Classifier

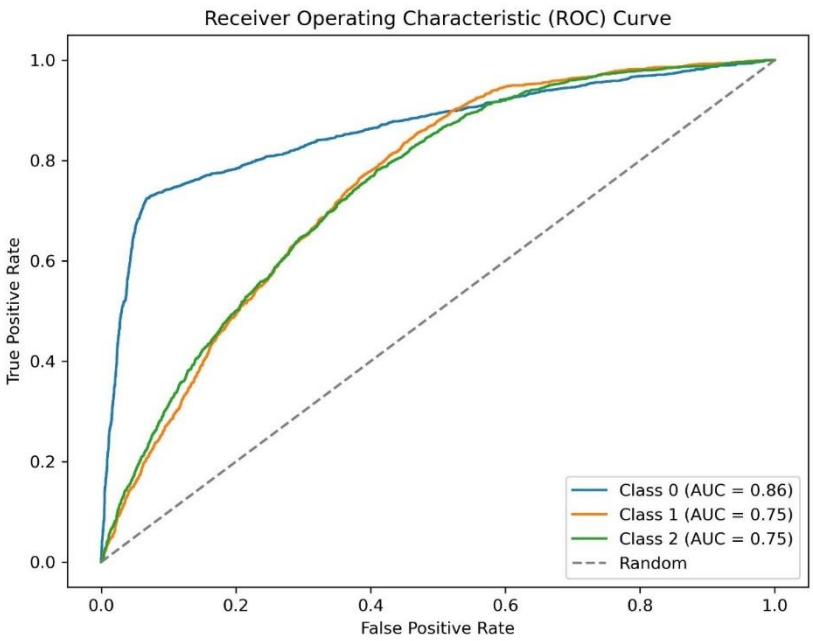


Figure 37: ROC of MLP Classifier

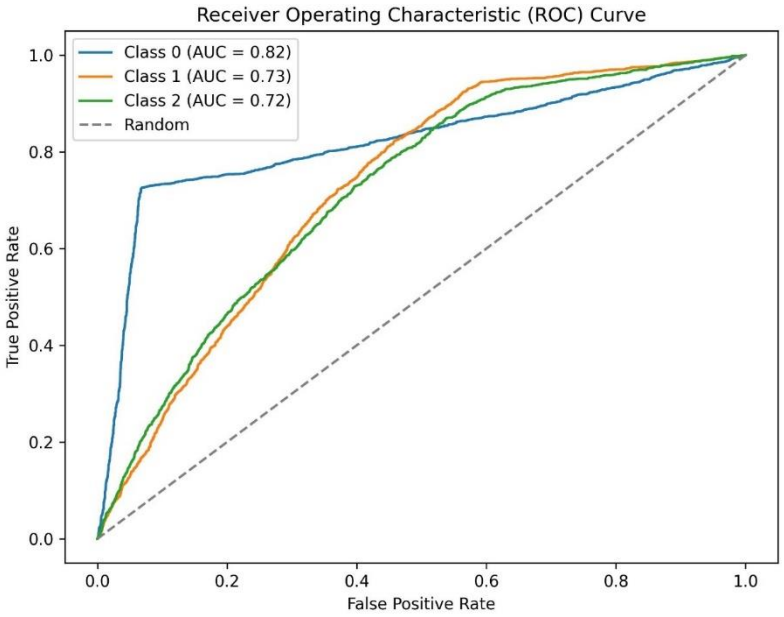


Figure 38: ROC of SVC

Observations:

Analyzing the Confusion Matrices and the ROC-AUC plots we can see most of the confusion arises in class label 1 and 2 which is actually the encoded labels for classes “used to smoke but quit” and “still

smokes". The misclassification rate of class label 0 which is the encoded label for the class "never smoked" is lower than the other two.

Conclusion:

- Bagging emerges as the best-performing model, showcasing top-tier performance in precision, recall, F-score, and Roc Auc.
- Random Forest also stands out with high specificity values, emphasizing its ability to correctly identify negative instances.
- Stacking, Boosting, and MLP also demonstrate competitive performance across multiple metrics.

Final Declaration:

The best-performing model is Bagging with RF, as it consistently achieves high values across precision, recall, F-score, and Roc AUC, indicating their overall effectiveness in classification tasks.

Phase 4

K-mean :

K-means is a popular unsupervised machine learning algorithm used for clustering data points into distinct groups, called clusters, based on similarity. It operates by iteratively assigning data points to clusters and updating the cluster centroids to minimize the sum of squared distances between data points and their assigned centroids. The algorithm converges when the centroids no longer change significantly. K-means requires specifying the number of clusters (k) beforehand and is sensitive to the initial placement of centroids. It is widely used for tasks such as customer segmentation, image compression, and document categorization.

Within-cluster variation plot for k-selection:

Within-cluster variation plot, often represented as an elbow plot, helps in choosing an appropriate number of clusters by observing the rate of decrease in within-cluster variation as the number of clusters increases. The "elbow point" signifies the optimal k where adding more clusters does not significantly reduce the within-cluster variation, indicating a good balance between model complexity and performance. The within-cluster variation plot is given below:

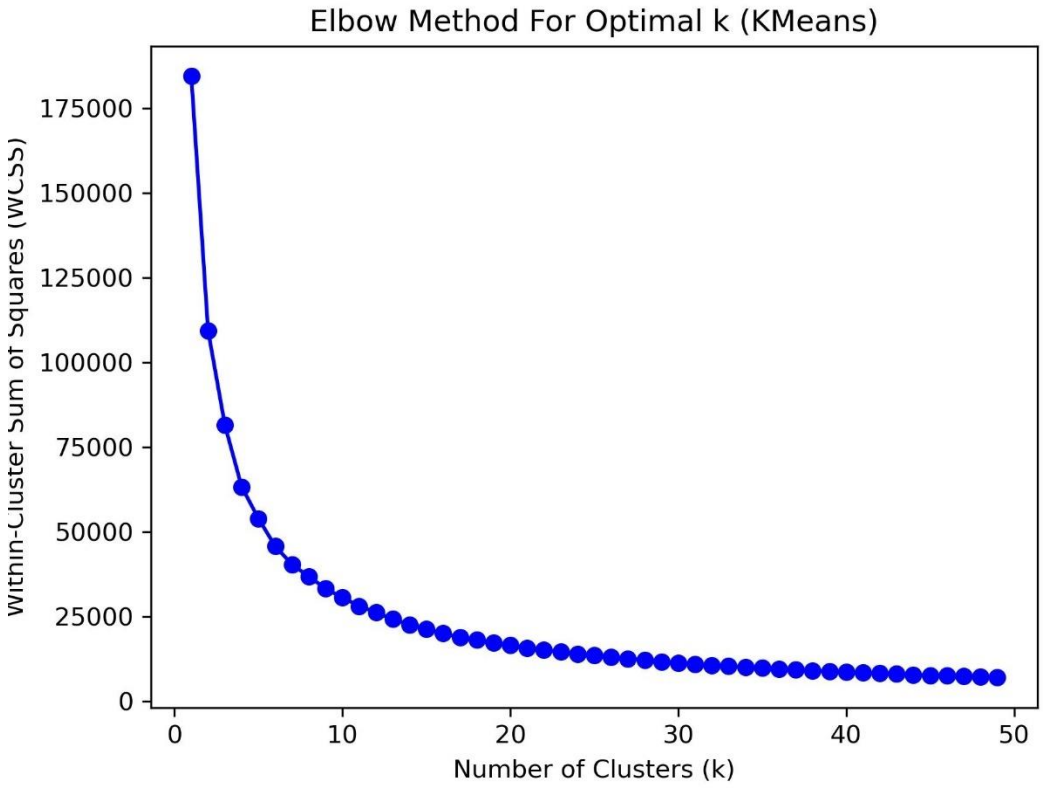


Figure 39: Elbow method for optimal K

The elbow point would be somewhere between 0-10 k-value range. To determine it specifically the Silhouette analysis for the k selection has been done.

Silhouette analysis for the k selection:

Silhouette analysis is a method for evaluating the optimal number of clusters (k) in a dataset by measuring how well-defined and separated the clusters are. It assigns a silhouette score to each data point, indicating its similarity to its own cluster compared to other clusters. A higher average silhouette score suggests a better-defined clustering structure. Silhouette analysis plot is shown below:

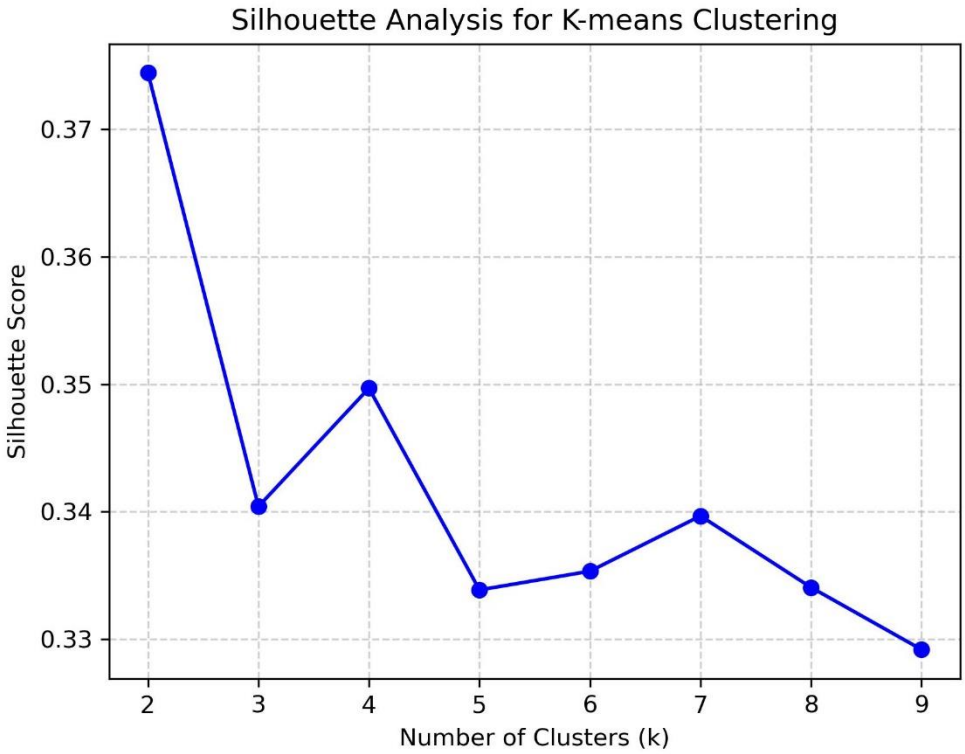


Figure 40 : Silhouette Analysis for K-means Clustering

From the plot we can see the silhouette score for k=2 is the highest. So the chosen k is 2.

The K-means cluster for k=2 is given below:

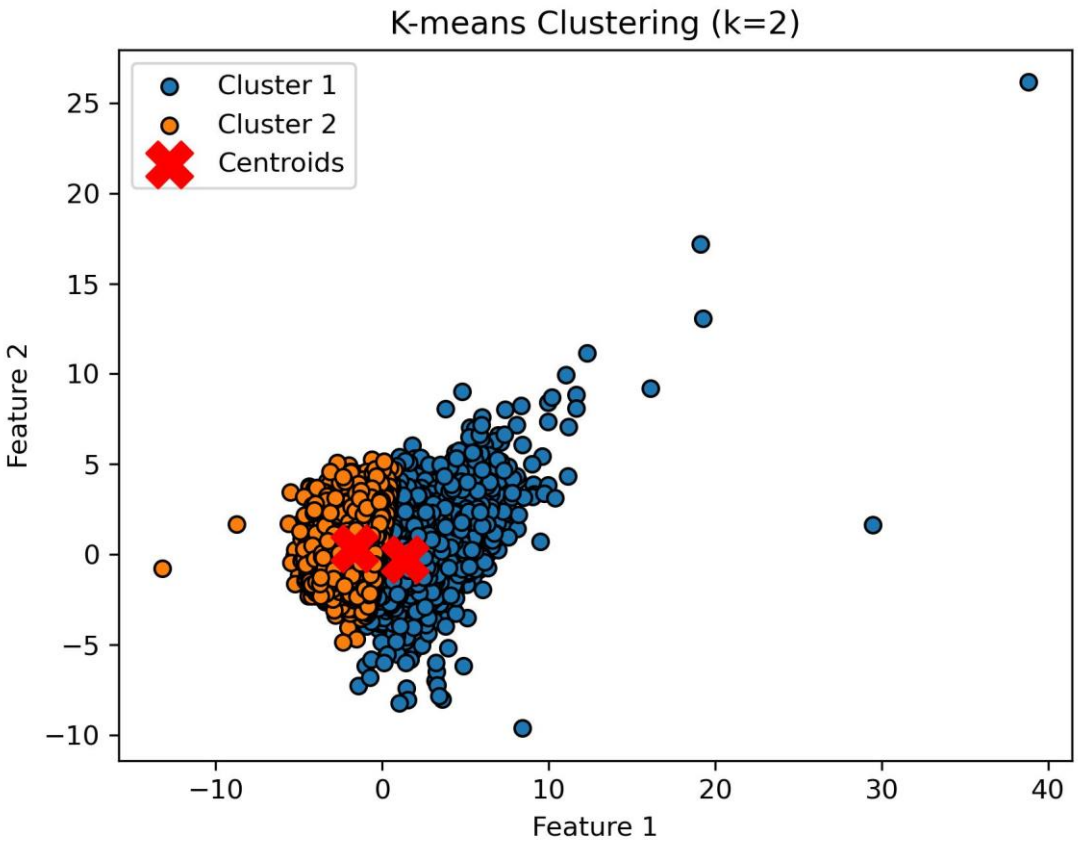


Figure 41: KMeans Clustering K=2

Apriori algorithm:

Apriori is a popular algorithm used for association rule mining in data mining and market basket analysis. It identifies frequent itemsets in a transactional dataset, where an itemset is a collection of items purchased together. The algorithm employs a bottom-up approach, iteratively discovering frequent itemsets and generating association rules based on their support and confidence. Support measures the frequency of occurrence, while confidence quantifies the reliability of the association. Apriori is efficient due to the "apriori property," which states that any subset of a frequent itemset must also be frequent. The algorithm helps uncover valuable insights into patterns and relationships within large datasets, particularly in retail and e-commerce analyses.

The algorithm provides with 2 frequent item sets :

```
[5 rows x 28 columns]
Processing 2 combinations | Sampling itemset size 2

DataFrame with Frequent Itemsets:
    support itemsets
0  0.000133      (_)
1  0.000100      (e)
2  0.000100   (_, e)
```

The association rules of the itemsets are given below:

```
Association Rules:
  antecedents consequents  antecedent support  consequent support  support  confidence    lift  leverage  conviction  zhangs_metric
1        (e)         (_)            0.000100            0.000133   0.0001        1.00  7500.0    0.0001         inf       0.999967
0        (_)         (e)            0.000133            0.000100   0.0001        0.75  7500.0    0.0001      3.9996       1.000000

Process finished with exit code 0
```

Let's break down the information:

DataFrame with Frequent Itemsets:

1. **Support:** This column shows the support for each itemset, which represents the proportion of transactions in the dataset that contain the itemset.
    - **Itemset (_):** Appears in 0.0133% of transactions.
    - **Itemset (e):** Appears in 0.01% of transactions.
    - **Itemset (_, e):** Appears in 0.01% of transactions.

Association Rules:

These rules describe relationships between antecedents and consequents, providing insights into how frequently certain items (antecedents) are associated with others (consequents).

1. **Antecedents and Consequents:**
    - Rule 1: If (e) is present, then (_) is also present.
    - Rule 2: If (_) is present, there is a 75% chance that (e) is also present.
2. **Support and Confidence:**
    - **Antecedent Support and Consequent Support:** The support of (e) is 0.01%, and (_) is 0.0133%.
    - **Support:** The combined support of both (e) and (_) is 0.01%.
    - **Confidence:** Rule 1 has 100% confidence, indicating that whenever (e) is present, (*) is also present. Rule 2 has 75% confidence, suggesting that (*) is present when (e) is present.
3. **Lift:**
    - **Lift:** Lift measures how much more likely the consequent is given the antecedent than without it. A lift of 7500 indicates a strong positive association in both rules.

4. **Leverage:**
    - **Leverage:** It measures the difference in the joint occurrence of antecedent and consequent compared to what would be expected if they were statistically independent.
5. **Conviction:**
    - **Conviction:** It provides information on how much the consequent relies on the antecedent. A higher conviction value suggests a stronger relationship.
6. **Zhang's Metric:**
    - **Zhang's Metric:** This metric is another measure of the interestingness of the association rule. A higher value indicates a more interesting rule.

Conclusions:

1. **K-mean Algorithm:**
    - **Elbow Point:** The within-cluster variation plot suggests an elbow point between 0-10 for k-value, indicating optimal k.
    - **Silhouette Analysis:** Silhouette scores peak at k=2, further validating the selection.
    - **K-means Clustering:** Applied K-means clustering for k=2, providing distinct clusters as illustrated.
2. **Apriori Algorithm:**
    - **Frequent Itemsets:** The algorithm identifies two frequent itemsets, denoted by (_), (e), and (_e).
    - **Association Rules:** Extracted rules indicating strong associations between items (antecedents) and their consequents. The Apriori algorithm successfully identifies frequent itemsets and association rules, providing valuable insights into item associations in the dataset.

These analyses contribute to a comprehensive understanding of clustering patterns and association rules within the dataset, aiding in meaningful decision-making and pattern recognition.

Recommendations

1. **Learning from the Project:**
    - Dimensionality reduction techniques like PCA and SVD were explored, but the Random Forest's feature importance proved to be more intuitive and specifically identified influential features for supervised learning.
    - The VIF technique effectively addressed multicollinearity concerns, providing a reliable method for dimensionality reduction in later project phases.
    - The application of stepwise regression, adjusted R-square, t-tests, and F-tests offered a comprehensive understanding of the dataset's predictive model, ensuring reliability, significance, and interpretability.
    - Rejecting the null hypothesis (p-values < 0.05) confirmed the statistical significance of all chosen coefficients, reinforcing the credibility of the developed regression model.
    - A systematic and thorough approach to classification analysis and clustering/association mining provides comprehensive insights into dataset characteristics.

- The combination of multiple classifiers and clustering algorithms ensures a well-rounded understanding of the data's structure and relationships.
- Recommendations for further improvements and model enhancements can be derived from a deep analysis of both supervised and unsupervised learning techniques.

2. **Best Performing Classifiers:**
   - Bagging with Random Forest consistently demonstrated top-tier performance across various metrics, making them strong candidates for the best models.
   - Random Forest also stands out with high specificity values, emphasizing its ability to correctly identify negative instances.
   - Stacking, Boosting, and MLP also demonstrate competitive performance across multiple metrics.

1. **Improving Classification Performance:**

   For addressing the misclassification of class label 1 (“Used to smoke”) and class label 2 (“Still smokes”) the following measures can be taken as future work:

   - **Feature Engineering:** Explore additional features or refine existing features that may provide more discriminatory power, especially for Class Labels 1 and 2.
   - **Model Tuning:** Fine-tune hyperparameters of the classifiers, particularly focusing on parameters that influence the decision boundaries for the challenging classes.
   - **Data Augmentation:** Consider augmenting the dataset, especially with more instances of Class Labels 1 and 2, to improve the model's ability to generalize for these classes.
   - Exploring advanced techniques, such as deep learning architectures, might provide further improvements in model performance and generalization.

2. **Features Associated with the Target Variable:**
   - Identified features like HDL_chole, LDL_chole, serum_creatinine, and SMK_stat_type_cd_2.0 demonstrated significant influence on the dependent variable "tot_chole."
   - Continuous monitoring and understanding of these features can contribute to better management and prediction of the target variable.
3. **Number of Clusters in Feature Space:**
   - For the K-means clustering algorithm, the elbow point analysis and silhouette scores suggested an optimal number of clusters (k=2).
   - Understanding and utilizing the identified clusters can aid in segmentation and targeted strategies.
4. **Association Rule Mining:**
   - The Apriori algorithm successfully identified frequent itemsets and association rules, providing valuable insights into item associations in the dataset.
   - Further exploration of association rules could reveal hidden patterns and relationships for informed decision-making.

In conclusion, the project's comprehensive analyses have uncovered key insights into both regression and classification aspects, leading to informed recommendations for future work and strategies based on the dataset's characteristics and patterns. The utilization of robust techniques and thorough

statistical analysis ensures the reliability and relevance of the project's findings.

Appendix:

All supporting code can be found with submitted files named as:

- phase1.py
- Phase2.py
- Phase3.py
- Phase4.py

References:

[1] James G, Witten D, Hastie T, Tibshirani R, Taylor J. An introduction to statistical learning: With applications in python. (No Title). 2023.