Insurance Applicant Risk Analysis

Problem Statement

Insurance firms evaluate applicants through a process known as underwriting. This involves assessing the risk associated with insuring an individual and determining the terms of coverage, including premiums and policy limits.

Of the various components of this analysis, a vital one is the health assessment, where applicants are asked about their medical history, lifestyle choices (e.g., smoking, drinking), and current health status. This helps flag any potential health risks.

Since a drinking and/or smoking habit could lead to a higher risk score, and consequently a higher premium or application rejection, applicants might hide this information, leading to potential losses for the insurance firms. We seek to be able to accurately classify a person as a smoker or drinker based on common health metrics used by insurance agencies.

Dataset

This dataset is collected from South Korea’s National Health Insurance Service, via Kaggle. It consolidates information from 1 million patients, with two columns indicating if each patient is an active drinker and/or smoker.

There are 20+ variables for each patient. Apart from the standard sex/age/height/weight, there are blood pressure readings, cholesterol and triglyceride numbers, and enzyme readings that indicate liver and kidney function.

The goal is to leverage these variables to predict if a given patient could be a drinker and/or smoker, and to identify markers among them that could be strong indicators of a drinking/smoking habit.

Exploratory Data Analysis

To begin with, we used a randomly generated subset of the original dataset, which had nearly a million records, as we felt that 100k was a good enough number for our purpose while also reducing runtimes for our code. We had some ideas for feature engineering, but we decided to do our preliminary EDA based only on the original columns.

Preliminary EDA

There's a near equal split of males and females. Most smokers in the dataset are males (with nearly 70% of males having smoked at some point in their lives, as compared to ~5% of the females). A person who smokes (or has at some point) is almost thrice as likely to be a drinker as opposed to not. On the contrary, someone who has never smoked is unlikely to drink. Most of the people who have smoked at least once are in their 40s or 50s, and almost half the people in their 30s are smokers.

Most drinkers are in their 40s. And taller, and heavier. Males have a larger percentage of drinkers in their ranks. The percentage of drinkers tends to get higher (>50%) with higher total cholesterol levels (180+). Age has an inverse relation, with the percentage of drinkers decreasing in higher age groups (50+). Higher the gamma-GTP (which indicates liver function), higher is the percentage of drinkers in that category. Hemoglobin levels are high for both drinkers & smokers.

Post Feature Engineering

We added some features to the dataset with a view to potentially achieving greater prediction accuracy, including Liver Enzyme Ratio, Liver Damage Score, BMI, and HDL:LDL ratio (please refer to the Appendix for an exhaustive list of variables).

While we did get marginally higher out of sample testing accuracies in these post-feature engineering scenarios, not all of these new features were high up in the feature importance lists. The exceptions were the 2 liver function features mentioned above, which makes sense as alcohol damages the liver.

That aside, the findings from the preliminary dataset mentioned above were further validated when the same checks were done on the feature-engineered dataset.

Solution & Insights

To help us find the optimal solution for predicting a patient's drinking and/or smoking habit, we considered 6 different algorithms for classification - logistic regression, KNN, Naive Bayes, decision tree, random forest, bagging.

Each model was run for 5 scenarios - to determine drinking status and smoking status separately, once each on the original and the engineered datasets, and once on the latter to determine drinking and/or smoking.

Baseline

If we assume everyone to be a drinker, about 50% of them actually turn out to be one. If we assume everyone isn't a smoker, about 60% of them turn out to be one. And if we assume everyone either drinks or smokes, about 60% actually fall in that category.

Model Performance

See the Appendix for the performance matrix of our models. When predicting if the person is a Drinker, the Random Forest Model pre-feature engineering works the best. This is likely to be overfitting due to the increased number of variables used in the prediction. When predicting Smoker statuses, the best model is the Decision Tree while including our engineered features. The best model to say if a person is a Smoker and/or a Drinker would be the Random Forest model. Each result is roughly 20% better than the baseline. As one would expect, because of the different ways in which each of the above algorithms work, we found slightly varying results. The models were trained using a 70%-30% training test split.

1. Logistic Regression

The sex(Male) category was an important feature consistently across all scenarios. In the pre-feature engineered models, hemoglobin, weight and serum creatinine were other important features. However, in the post-feature engineered models, gamma GTP, HDL cholesterol & Liver Enzyme Ratio became more prominent.

2. KNN

In the KNN model with 10-fold cross-validation, the key features were Sex (Male), Hemoglobin, and Triglyceride levels. After feature engineering, Sex (Male), Hemoglobin, and Liver Damage Score emerged as the most significant features.

3. Naive Bayes

After applying feature engineering, the Naive Bayes model showed a marked improvement in accuracy. Individual probabilities for the classification were calculated using key variables such as age, sex, cholesterol levels, and liver damage scores, which were identified as the most important features.

4. Decision Trees

Similar to previous models, the decision trees identified age, sex, both cholesterol types, and gamma-GTP as key predictors. After feature engineering, tree depth increased as new variables like cholesterol ratios, hemoglobin, BLDS, SBP, and organ damage indicators were included. The trees were evaluated using 10-fold cross-validation, with depths ranging from 1-15. Initially, tree depths were four to five levels, increasing to five to six after adding the new features.

5. Random Forest

The Random Forest models identified similar important features, including age, sex, hemoglobin, gamma-GTP, and cholesterol types. Age was crucial for drinkers but not smokers, while sex was more important for smokers. The liver_damage_score was the most important feature across all models. Initially, four or five variables were used, but after feature engineering, this increased to five or six due to using the square root of the total variables. Interestingly, prediction accuracy decreased slightly (0.1%) after feature engineering, likely due to overfitting, as it performed better in-sample but worse out-of-sample.

6. Bagging

Like all the models above, the Bagging model had similar important features: age, sex, gamma_GTP, etc. It is a variant of the Random Forest Model that takes into account all of the categories when predicting, not just a subset. The model performs slightly worse than the Random Forest, likely due to overfitting by accounting for all variables in its classification.

Insights

Age is quite important for the drinking predictions models but not so much for the smoking predictions. This could be a cultural trend as people tend to drink when they are younger universally while smokers are more constant throughout their lives. Liver_damage_score is important for some models like decision trees, random forest and KNN while another variation of this liver_enzyme_ratio is important for logistic regression. These are different variations of the presence of enzymes signaling organ damage. This shows some models use the information differently. The smoker predictions are better than those for drinking across the board - this might be due to the nearly even split of drinkers and non-drinkers, which makes it more difficult to predict.

Appendix:

Variable Explanation:

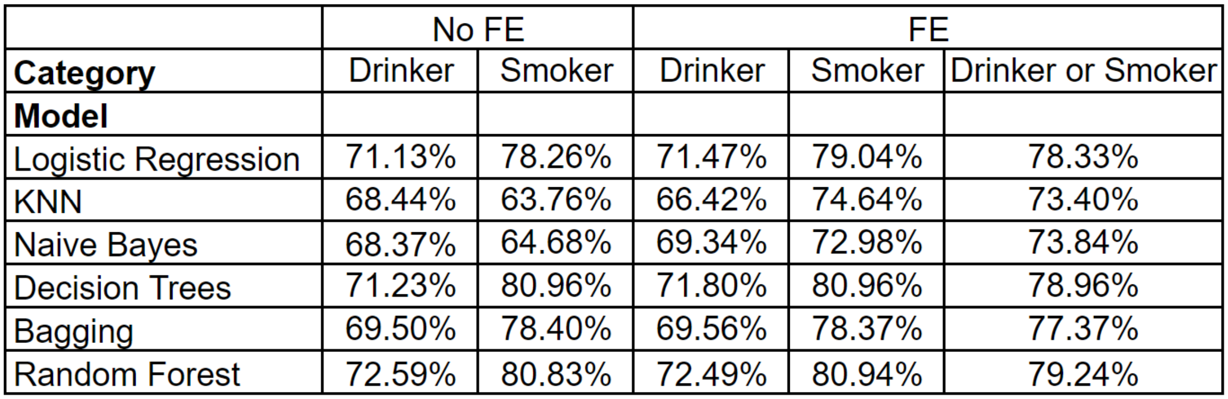
NAME - TYPE - UNITS/MEANING

Sex - Categorical - Male or Female
age - Numeric - (years)
height - Numeric - (cm)
weight - Numeric - (kg)
sight_left - Boolean - indicating normal or abnormal sight in left eye
sight_right - Boolean - indicating normal or abnormal sight in right eye
hear_left - Boolean - indicating normal or abnormal hearing in left ear
hear_right - Boolean - indicating normal or abnormal hearing in right ear
SBP - Numeric - Systolic blood pressure (mmHg)
DBP - Numeric - Diastolic Blood Pressure (mmHg)
BLDS - Numeric - Fasting blood glucose (mg/dL)
tot_chole - Numeric - Total Cholesterol (mg/dL)
HDL_chole - Numeric - HDL Cholesterol (mg/dL)
LDL_chole - Numeric - LDL cholesterol (mg/dL)
triglyceride - Numeric - triglyceride (mg/dL)
hemoglobin - Numeric - hemoglobin (g/dL)
urine_protein - Numeric - types of protein found in urine
serum_creatinine - Numeric - blood creatinine (mg/dL)
SGOT_AST - Numeric - Glutamate-oxaloacetate transaminase Aspartate transaminase (IU/L)
SGOT_ALT - Numeric - Glutamate-oxaloacetate transaminase Alanine transaminase (IU/L)
gamma_GTP - Numeric - y-glutamyl transpeptidase (IU/L)
SMK_state_type_cd - Categorical - Never smoked, used to smoke or smokes
DRK_YM - Categorical - Drinker or non-drinker

Feature Engineered Variables:

Hearing_State - Boolean - Classifies hearing as faulty if present in either ear
HDL_LDL_Ratio - Numeric - HDL Cholesterol per LDL Cholesterol
BMI - Numeric - (Weight/Height^2)
Total_HDL_Ratio - Numeric - Total Cholesterol per HDL Cholesterol
Liver_Enzyme_Ratio - Numeric - SGOT_AST per SGOT_ALT (Both are enzymes indicating the presence of organ damage)
Liver_damage_score - Numeric - Gamma_GTP + SGOT_ALT (Both are enzymes indicating the presence of liver damage)
Smoking_Status - Boolean - Smoker or Not
Drinking_Status - Boolean - Drinker or Not
DorS - Boolean - Drinker and/or Smoker or Not

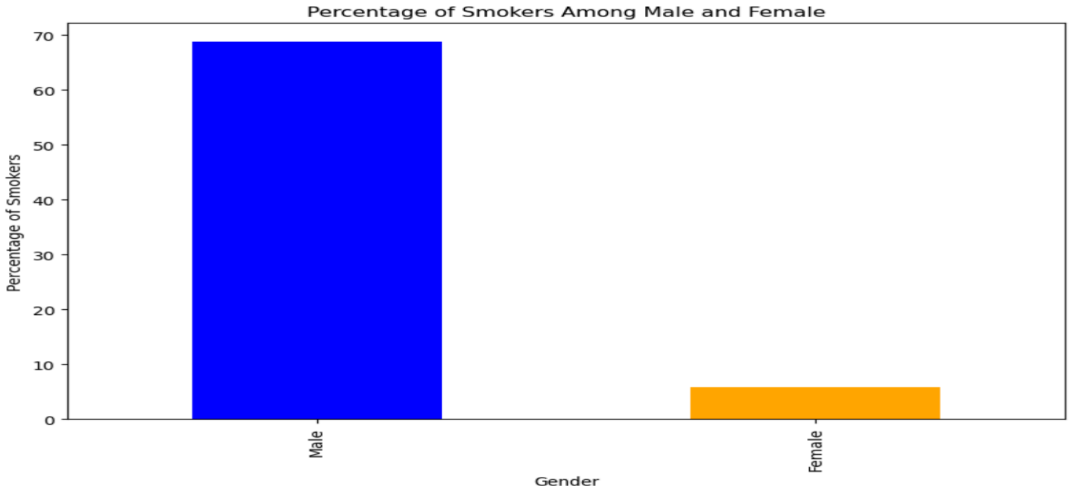
Prediction Accuracy Matrix:

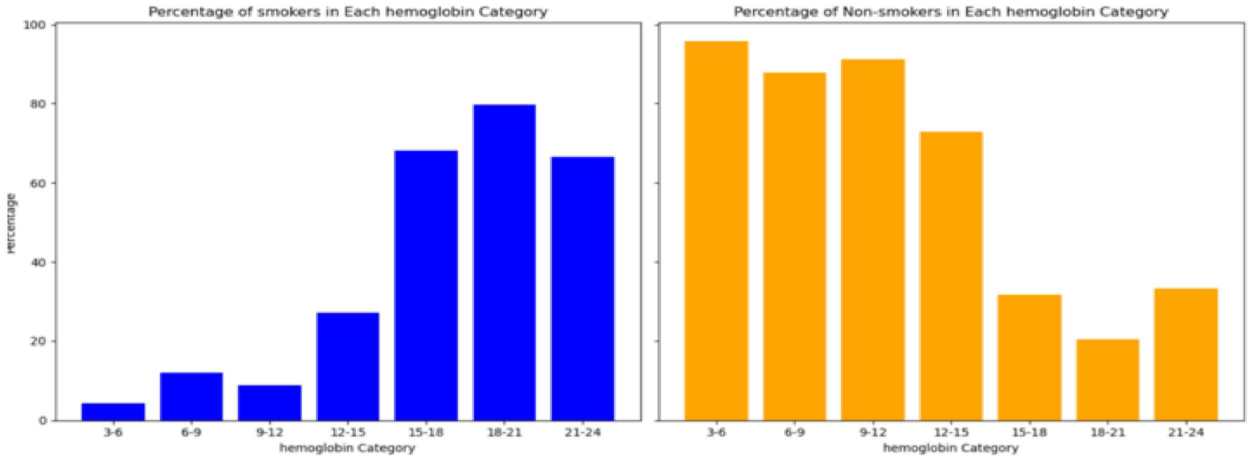


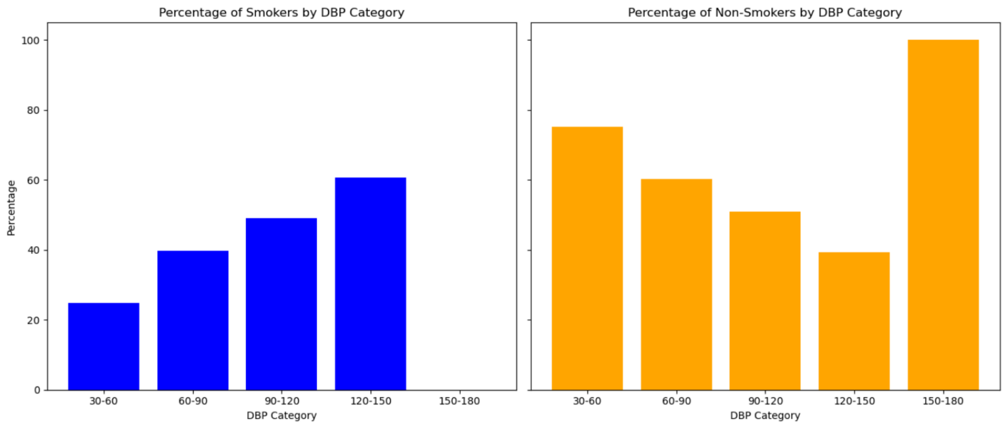
	No FE		FE		
Category	Drinker	Smoker	Drinker	Smoker	Drinker or Smoker
Model					
Logistic Regression	71.13%	78.26%	71.47%	79.04%	78.33%
KNN	68.44%	63.76%	66.42%	74.64%	73.40%
Naive Bayes	68.37%	64.68%	69.34%	72.98%	73.84%
Decision Trees	71.23%	80.96%	71.80%	80.96%	78.96%
Bagging	69.50%	78.40%	69.56%	78.37%	77.37%
Random Forest	72.59%	80.83%	72.49%	80.94%	79.24%

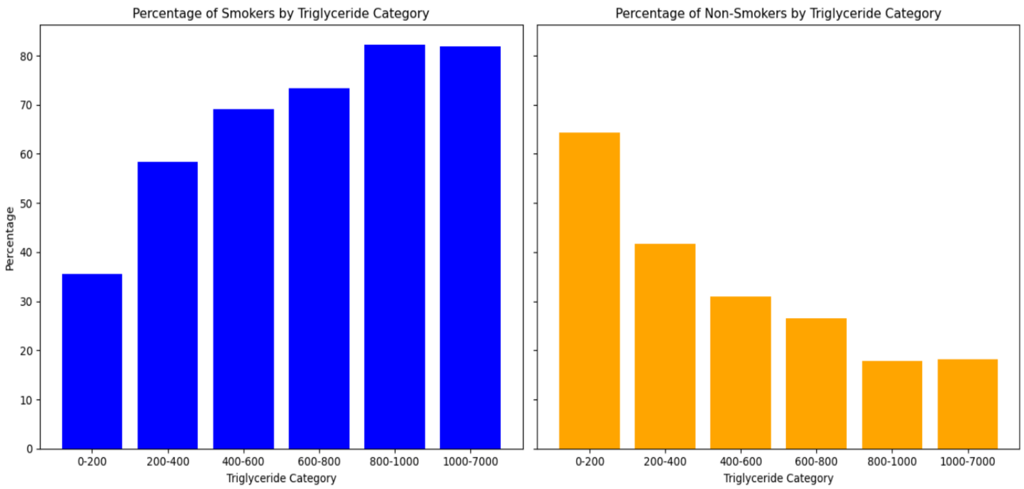
Preliminary Data Insights and Visualizations:

Smoking:

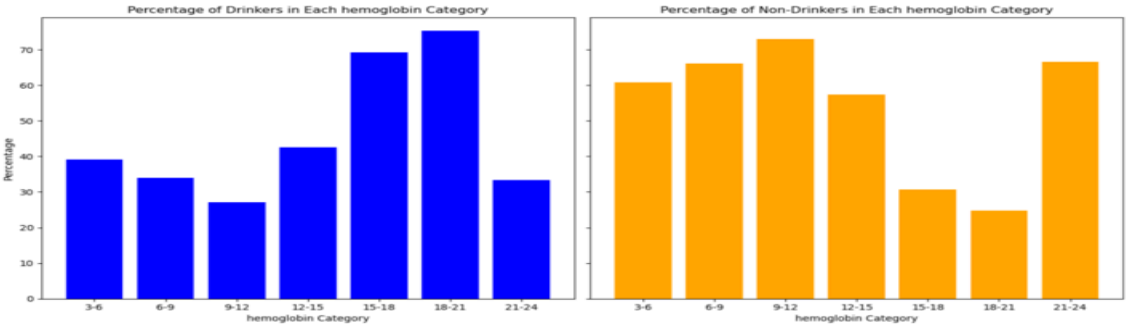


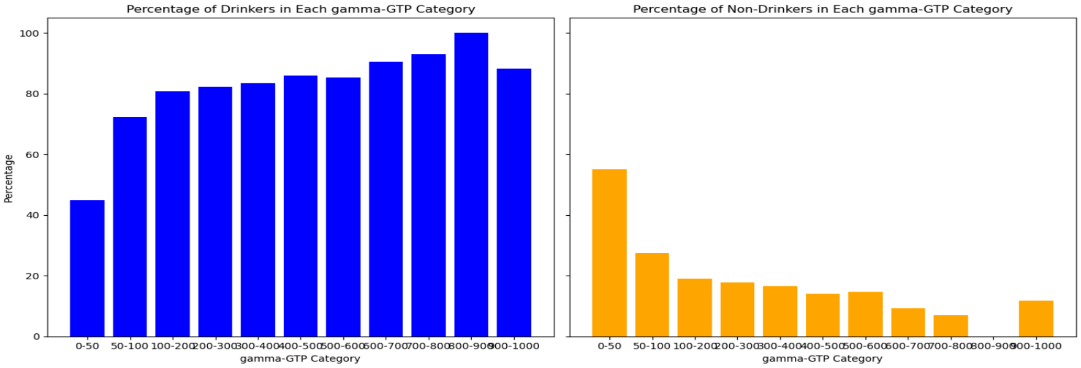


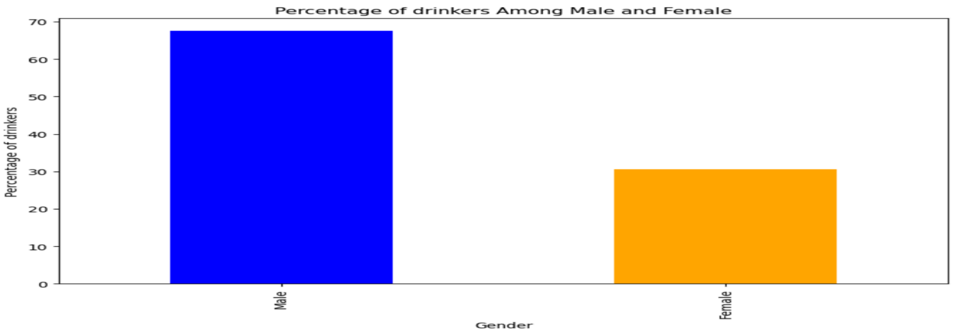



Percentage of Smokers by Triglyceride Category
Percentage of Non-Smokers by Triglyceride Category
Percentage
0
10
20
30
40
50
60
70
80
0-200
200-400
400-600
600-800
800-1000
1000-7000
Triglyceride Category
Triglyceride Category

Drinking:

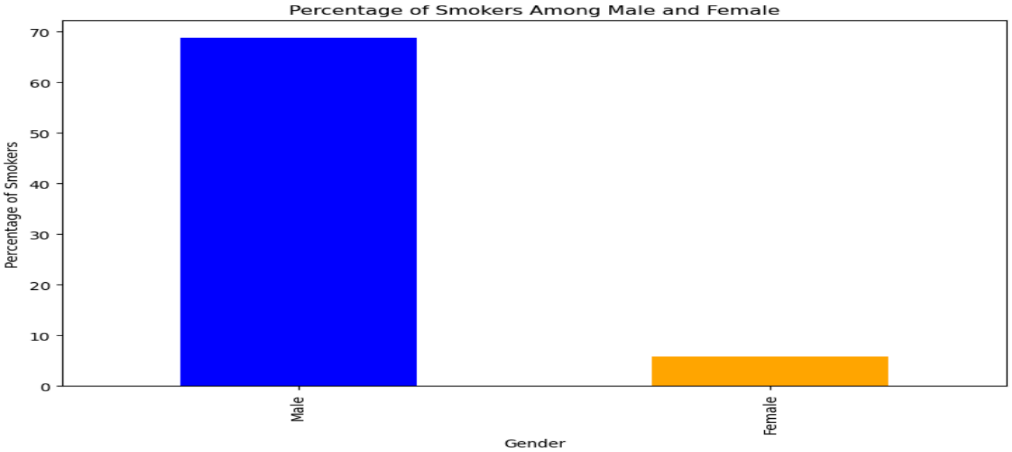

Percentage of Drinkers in Each hemoglobin Category
Percentage of Non-Drinkers in Each hemoglobin Category
Percentage
0
10
20
30
40
50
60
70
3-6
6-9
9-12
12-15
15-18
18-21
21-24
hemoglobin Category
hemoglobin Category

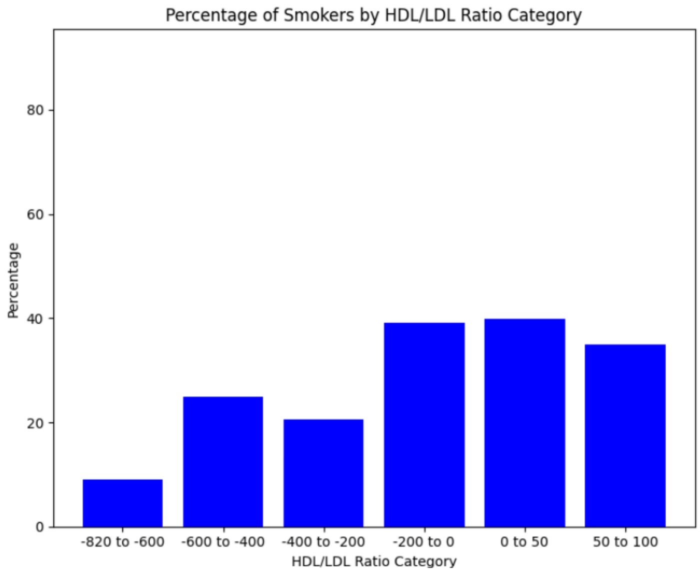

Percentage of Drinkers in Each gamma-GTP Category
Percentage of Non-Drinkers in Each gamma-GTP Category
Percentage
0
20
40
60
80
100
0-50
50-100
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000
gamma-GTP Category
gamma-GTP Category

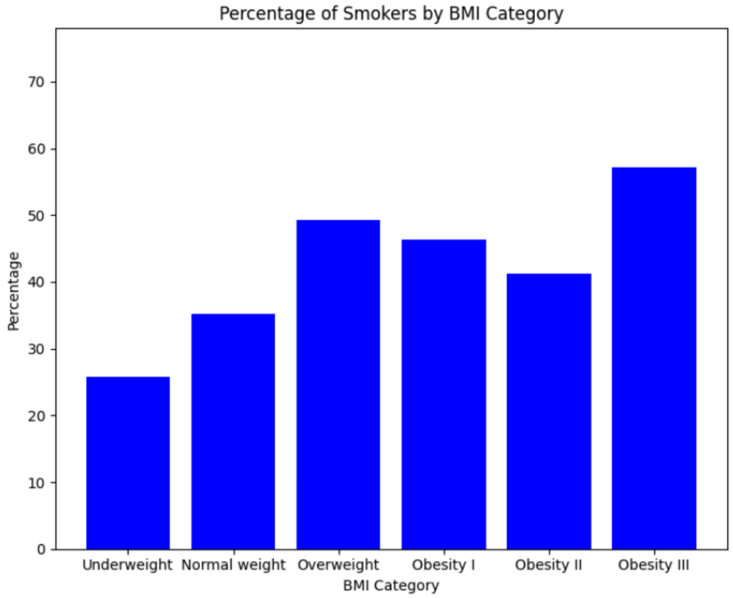


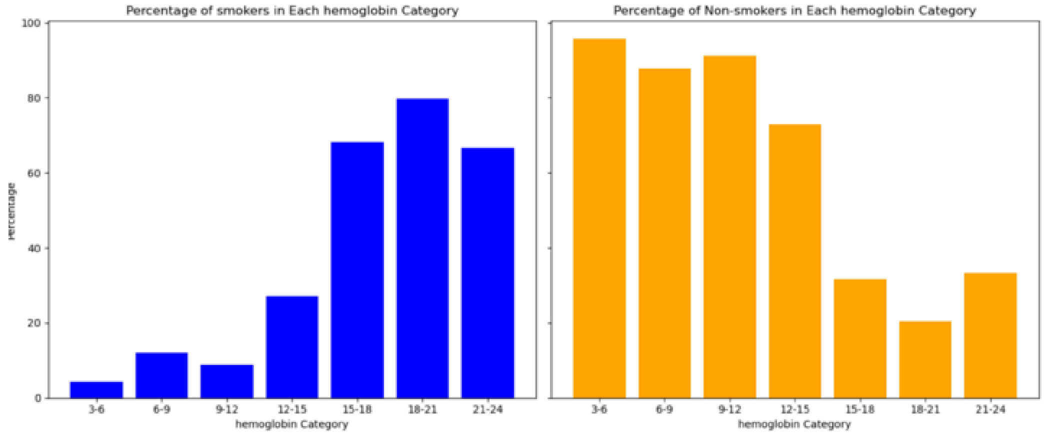
Post EDA Analysis and Insights:

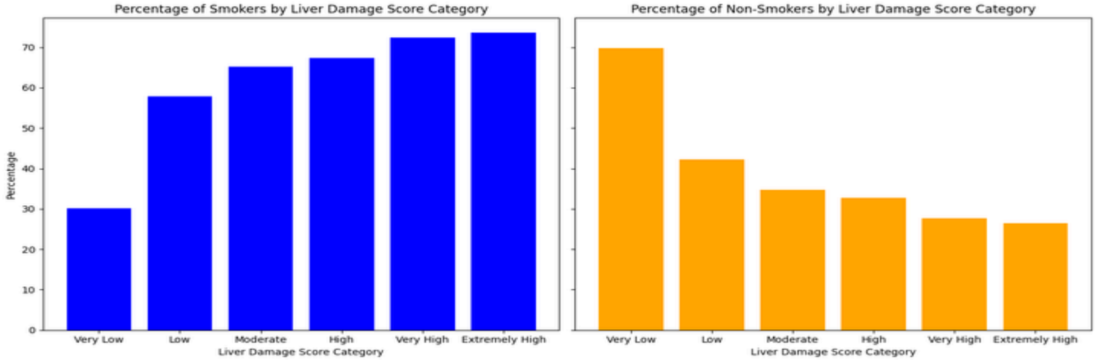
Smoking:



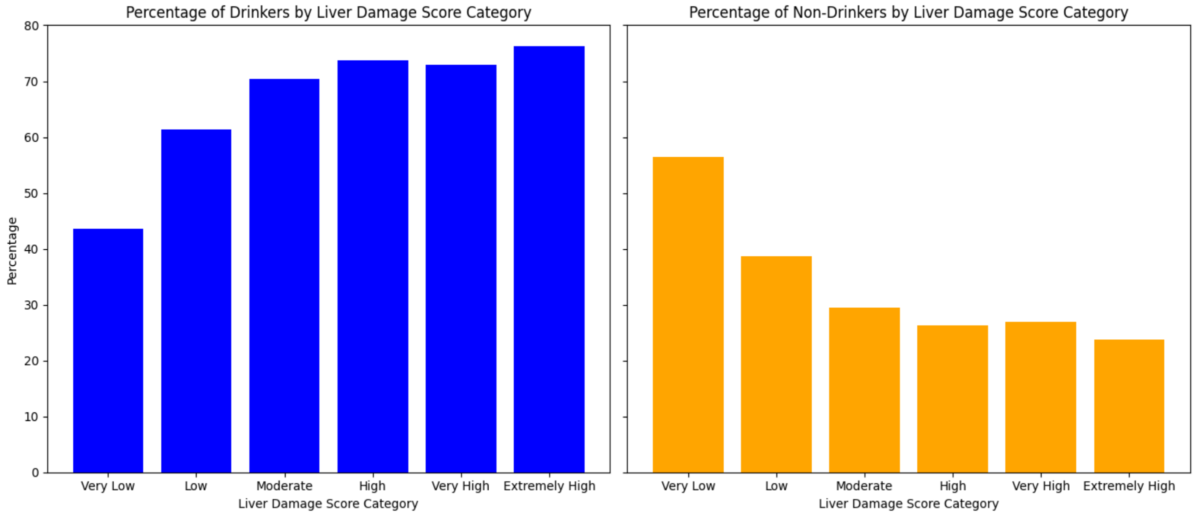

Percentage of Smokers by HDL/LDL Ratio Category
Percentage
0
20
40
60
80
-820 to -600
-600 to -400
-400 to -200
-200 to 0
0 to 50
50 to 100
HDL/LDL Ratio Category

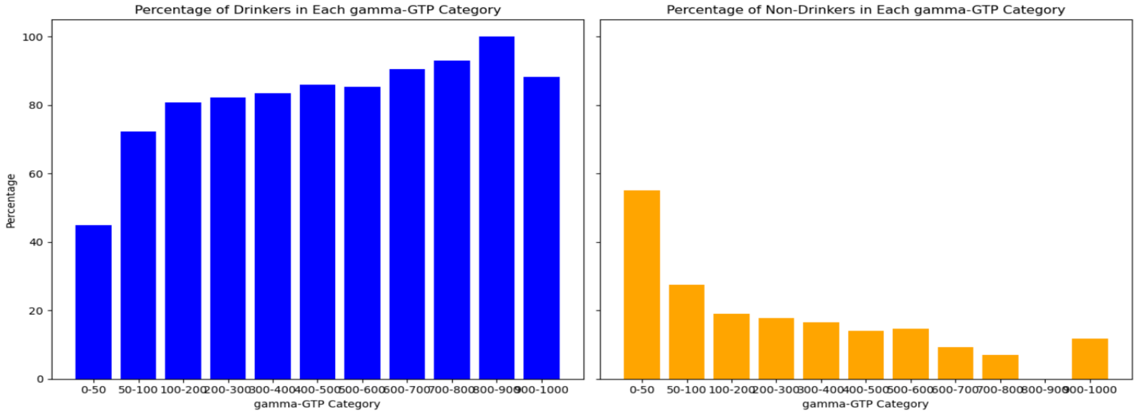

Percentage of Smokers by BMI Category
Percentage
0
10
20
30
40
50
60
70
Underweight
Normal weight
Overweight
Obesity I
Obesity II
Obesity III
BMI Category

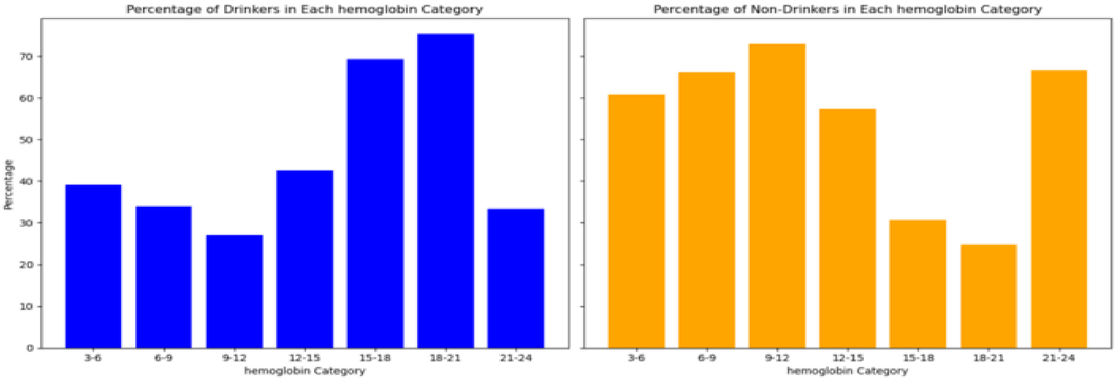

Percentage of smokers in Each hemoglobin Category
Percentage
0
20
40
60
80
100
3-6
6-9
9-12
12-15
15-18
18-21
21-24
hemoglobin Category
Percentage of Non-smokers in Each hemoglobin Category
3-6
6-9
9-12
12-15
15-18
18-21
21-24
hemoglobin Category


Percentage of Smokers by Liver Damage Score Category
Percentage
0
10
20
30
40
50
60
70
Very Low
Low
Moderate
High
Very High
Extremely High
Liver Damage Score Category
Percentage of Non-Smokers by Liver Damage Score Category
Very Low
Low
Moderate
High
Very High
Extremely High
Liver Damage Score Category

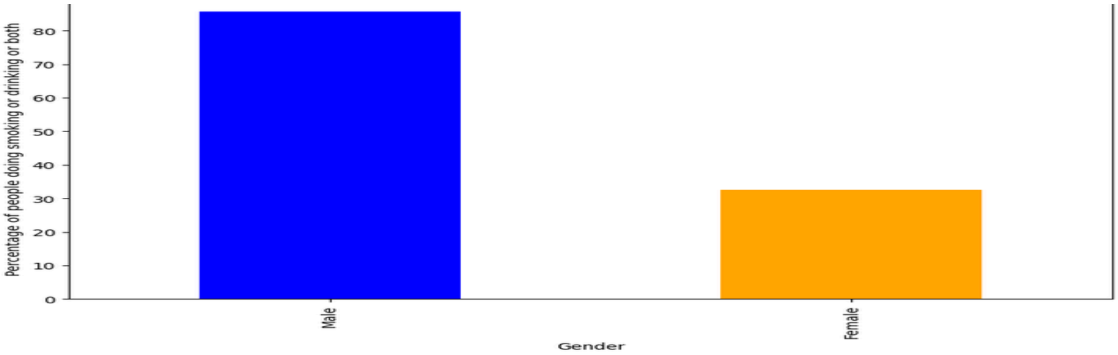
Drinking:

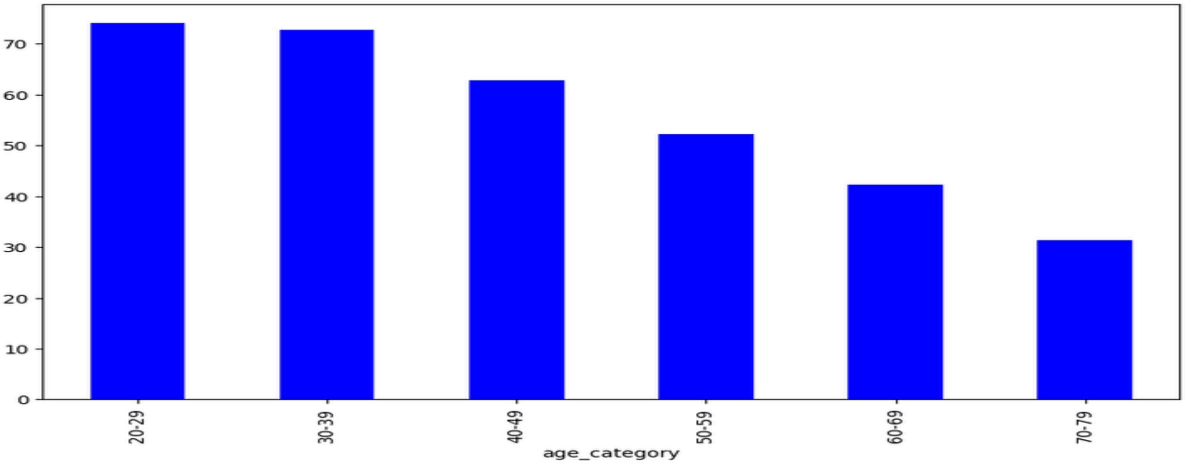

Percentage of Drinkers by Liver Damage Score Category
Percentage
0
10
20
30
40
50
60
70
80
Very Low
Low
Moderate
High
Very High
Extremely High
Liver Damage Score Category
Percentage of Non-Drinkers by Liver Damage Score Category
Very Low
Low
Moderate
High
Very High
Extremely High
Liver Damage Score Category

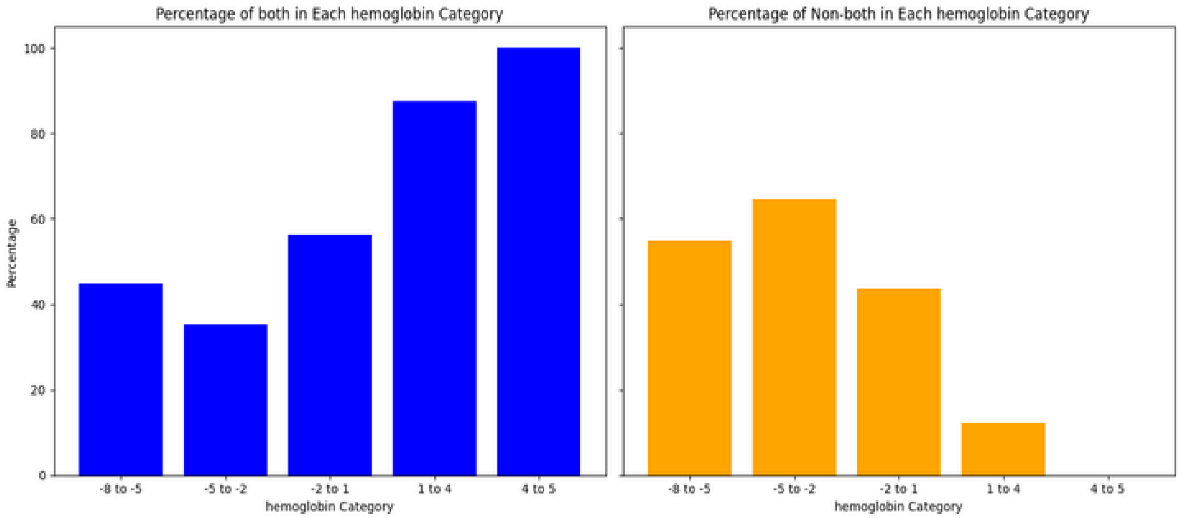

Percentage of Drinkers in Each gamma-GTP Category
Percentage
0
20
40
60
80
100
0-50
50-100
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000
gamma-GTP Category
Percentage of Non-Drinkers in Each gamma-GTP Category
0-50
50-100
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000
gamma-GTP Category

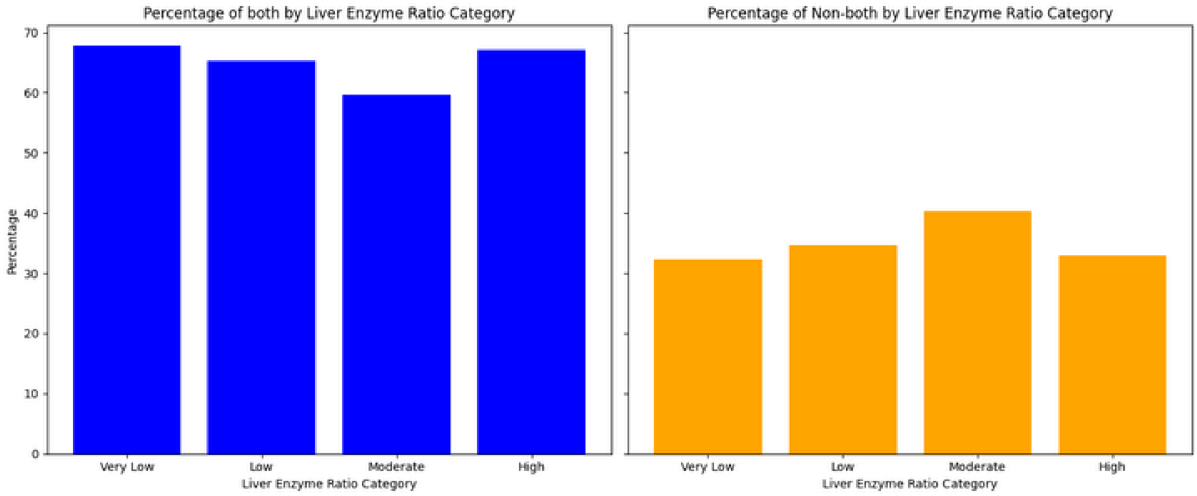


Both smoking and drinking:

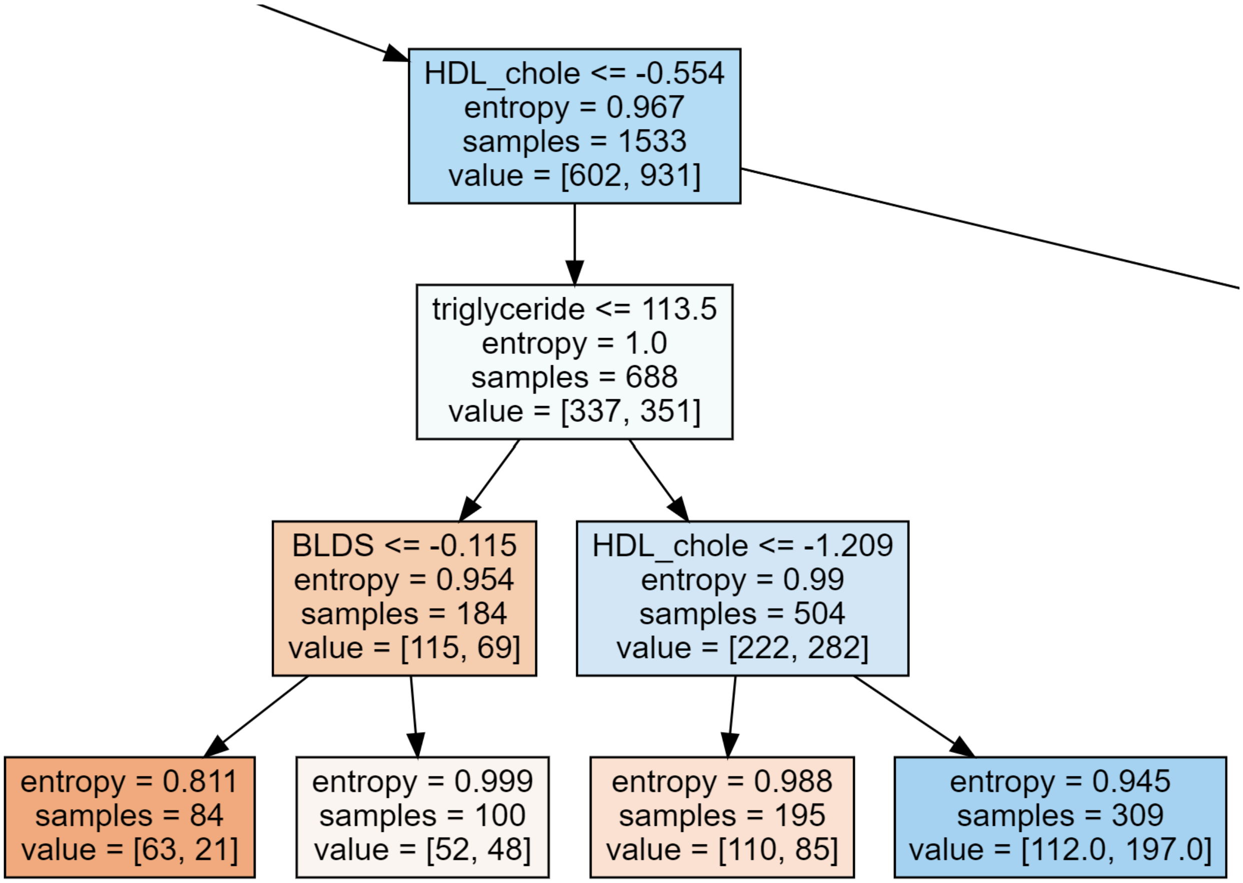




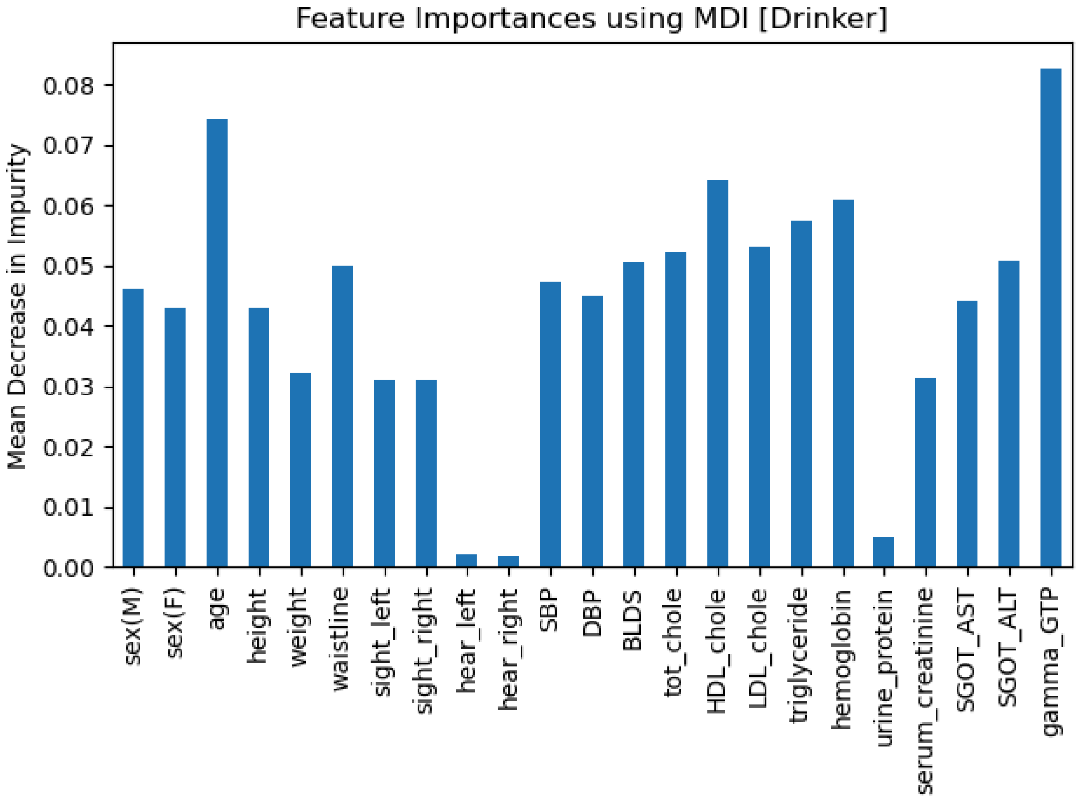


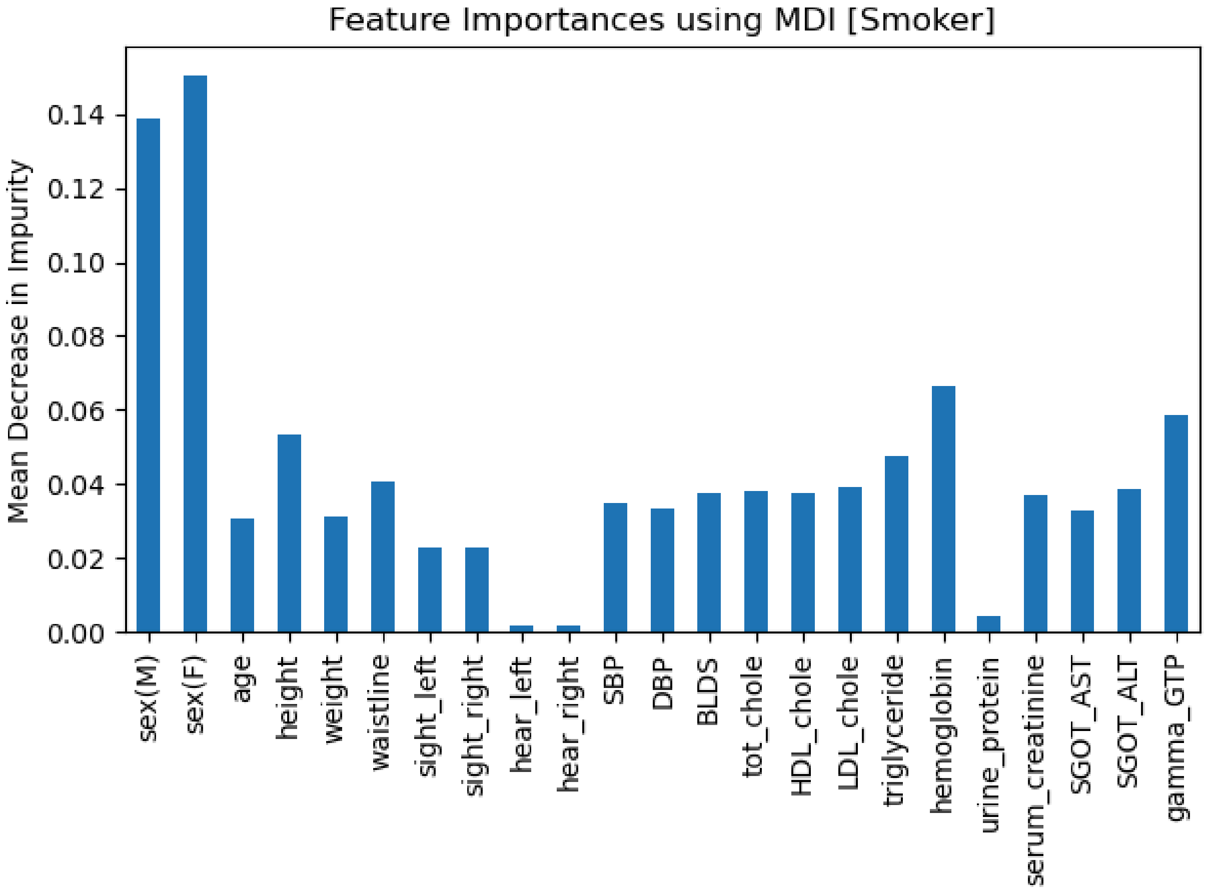


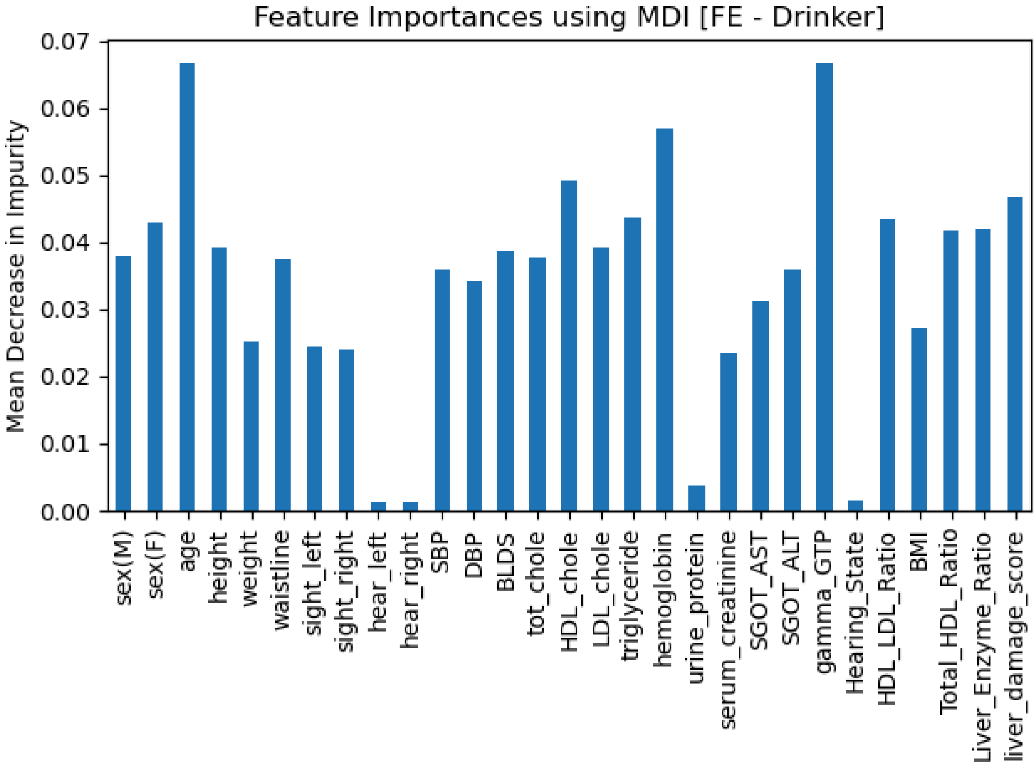
Decision Tree Partial Example (Whole Trees are too big to show)

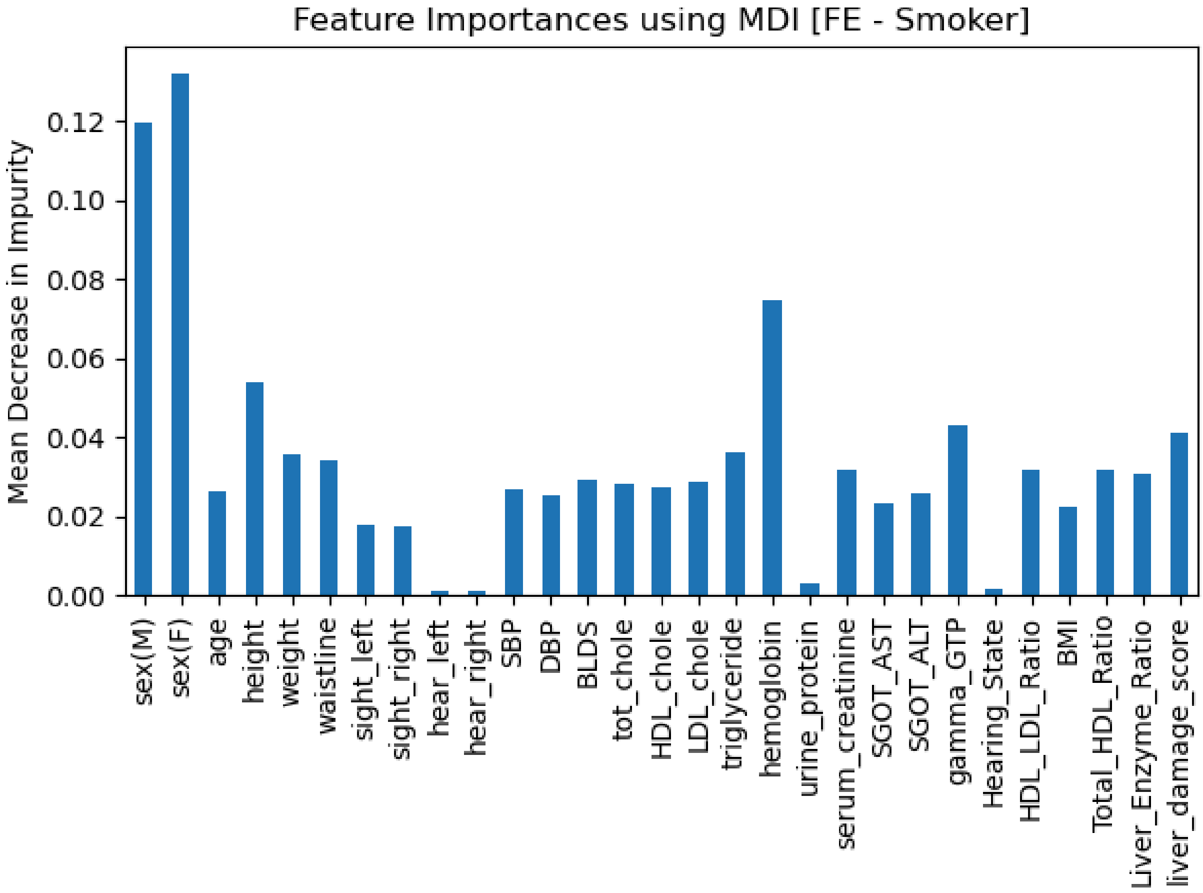


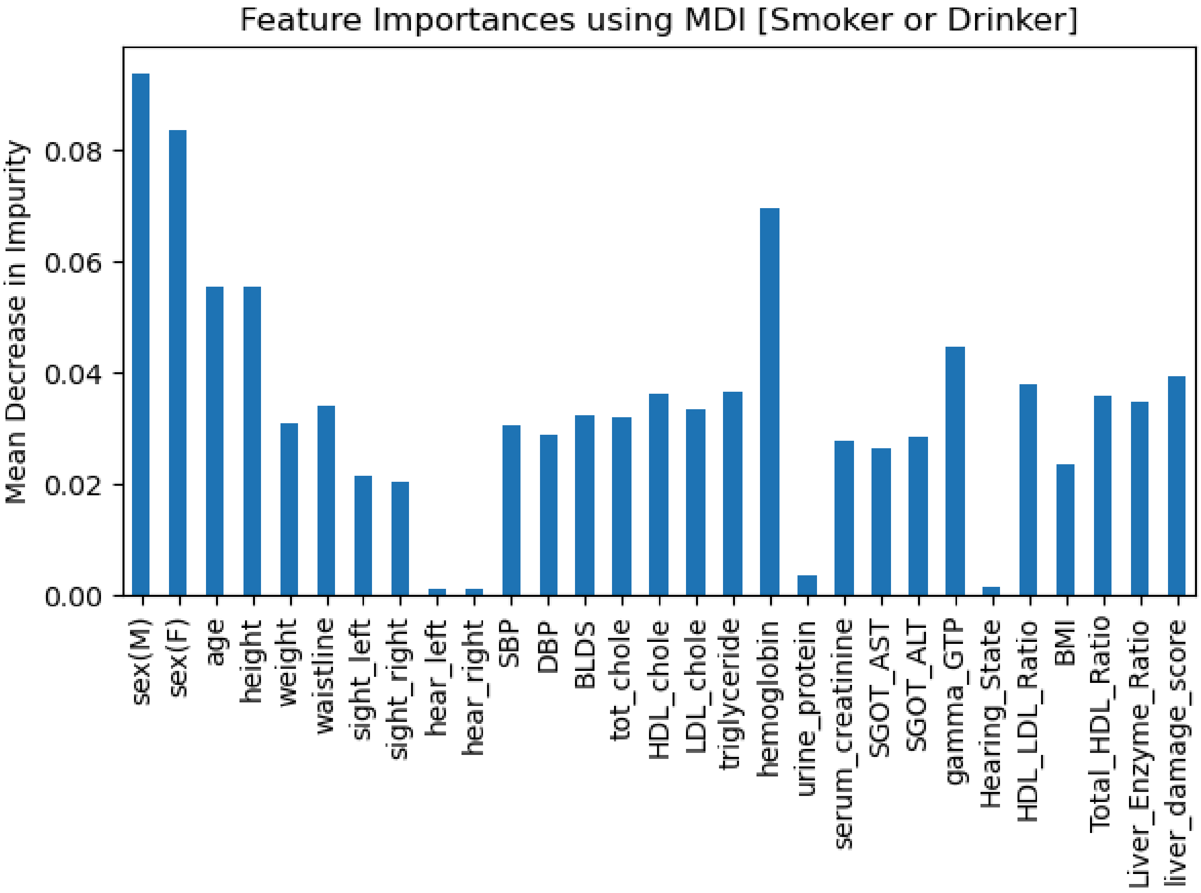
Feature Importance For Random Forest


Feature Importances using MDI [Drinker]
Mean Decrease in Impurity
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
sex(M)
sex(F)
age
height
weight
waistline
sight_left
sight_right
hear_left
hear_right
SBP
DBP
BLDS
tot_chole
HDL_chole
LDL_chole
triglyceride
hemoglobin
urine_protein
serum_creatinine
SGOT_AST
SGOT_ALT
gamma_GTP


Feature Importances using MDI [Smoker]
Mean Decrease in Impurity
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
sex(M)
sex(F)
age
height
weight
waistline
sight_left
sight_right
hear_left
hear_right
SBP
DBP
BLDS
tot_chole
HDL_chole
LDL_chole
triglyceride
hemoglobin
urine_protein
serum_creatinine
SGOT_AST
SGOT_ALT
gamma_GTP


Feature Importances using MDI [FE - Drinker]
Mean Decrease in Impurity
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
sex(M)
sex(F)
age
height
weight
waistline
sight_left
sight_right
hear_left
hear_right
SBP
DBP
BLDS
tot_chole
HDL_chole
LDL_chole
triglyceride
hemoglobin
urine_protein
serum_creatinine
SGOT_AST
SGOT_ALT
gamma_GTP
Hearing_State
HDL_LDL_Ratio
BMI
Total_HDL_Ratio
Liver_Enzyme_Ratio
liver_damage_score


Feature Importances using MDI [FE - Smoker]
Mean Decrease in Impurity
0.00
0.02
0.04
0.06
0.08
0.10
0.12
sex(M)
sex(F)
age
height
weight
waistline
sight_left
sight_right
hear_left
hear_right
SBP
DBP
BLDS
tot_chole
HDL_chole
LDL_chole
triglyceride
hemoglobin
urine_protein
serum_creatinine
SGOT_AST
SGOT_ALT
gamma_GTP
Hearing_State
HDL_LDL_Ratio
BMI
Total_HDL_Ratio
Liver_Enzyme_Ratio
liver_damage_score



Dataset Source

https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset/data