Body signal of smoking

FINAL CASE STUDY : PREDICTIVE ANALYTICS AND DATA MINING

TABLE OF CONTENTS

01

Our team

OUR TEAM

Yash Kumar Jain

Le Thi Truc Linh

Qinghua Ye

Julian Oliveros Forero

02

Introduction

About the Data set

This dataset is a collection of basic health biological signal data.

The goal is to determine the presence or absence of smoking through bio-signals. It will give us an alert for the condition of our body.

Dataset

Shape

55,692 rows, 27 columns

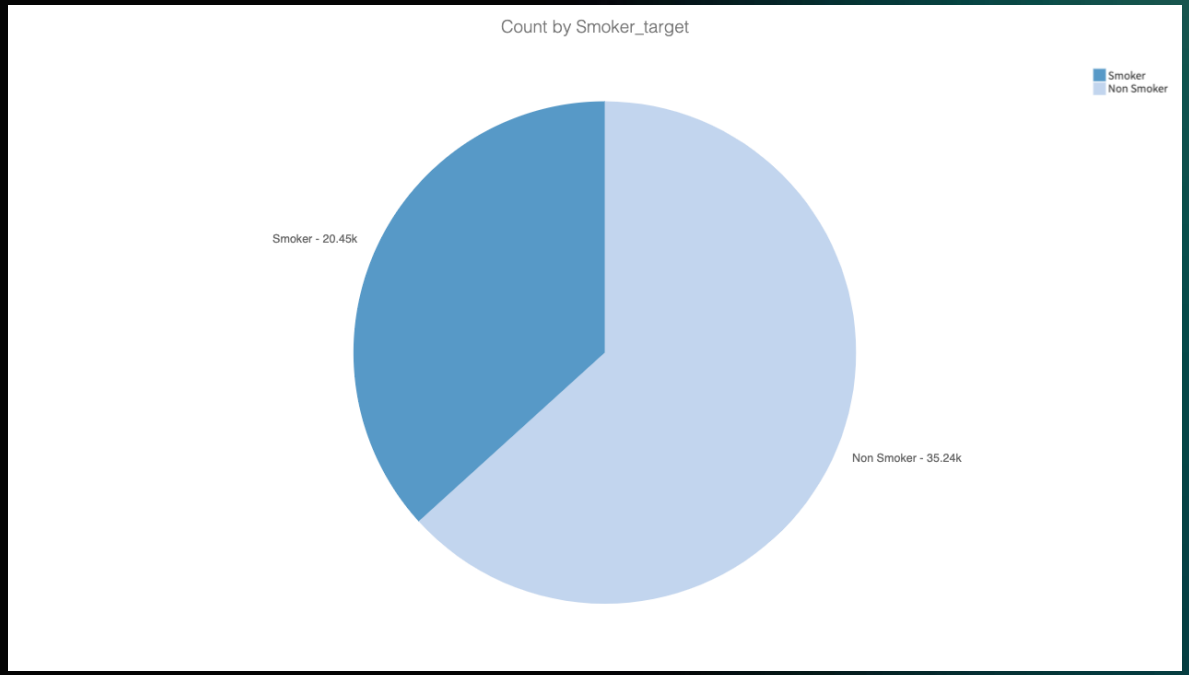
Target Variable

Smoking

Source of Dataset

Kaggle

Target Variable Analysis



From the pie chart, we can get the statistics that there are 37% of smokers and 63% of non-smokers in the dataset according to the calculation.

Features in the data set

- gender
- age : 5-years gap
- height(cm)
- weight(kg)
- waist(cm) : Waist circumference length
- eyesight(left)
- eyesight(right)
- systolic : Blood pressure
- relaxation : Blood pressure
- fasting blood sugar
- Cholesterol : total
- Cholesterol : total
- triglyceride
- hemoglobin
- Urine protein
- serum creatinine
- AST : glutamic oxaloacetic transaminase type
- ALT : glutamic oxaloacetic transaminase type
- Gtp : γ-GTP
- dental caries
- tartar : tartar status

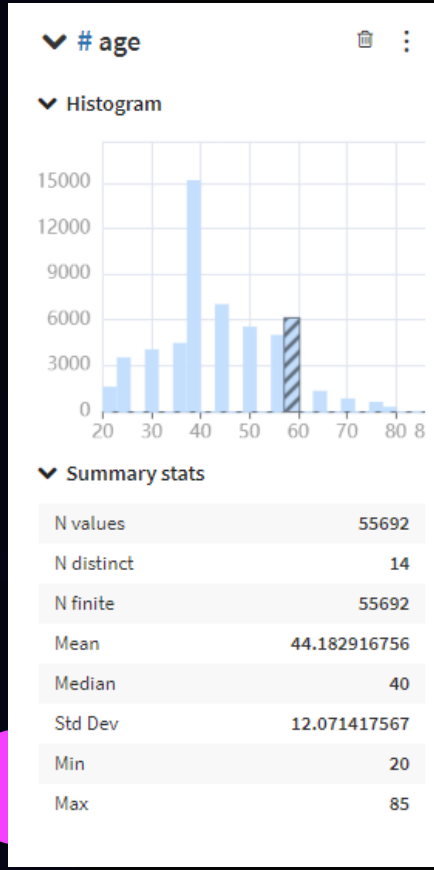
03

Exploratory analysis

Analysis Feature

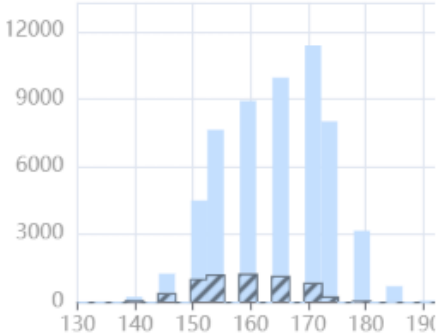
Let's look at some of the private data in the dataset.

We can observe that 64% of the data set is made up of men, whereas just 34% of it is made up of women. They may want to boost the male sample size because it's probable that the majority of males smoke frequently. Additionally, we can observe that our dataset's average age is 44 years old, with the oldest person being 85 years old and the youngest person being 20.



height(cm)

Histogram

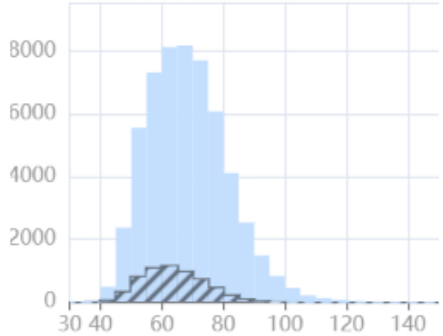


Summary stats

N values	55692
N distinct	13
N finite	55692
Mean	164.64932127
Median	165
Std Dev	9.1945969126
Min	130
Max	190

weight(kg)

Histogram



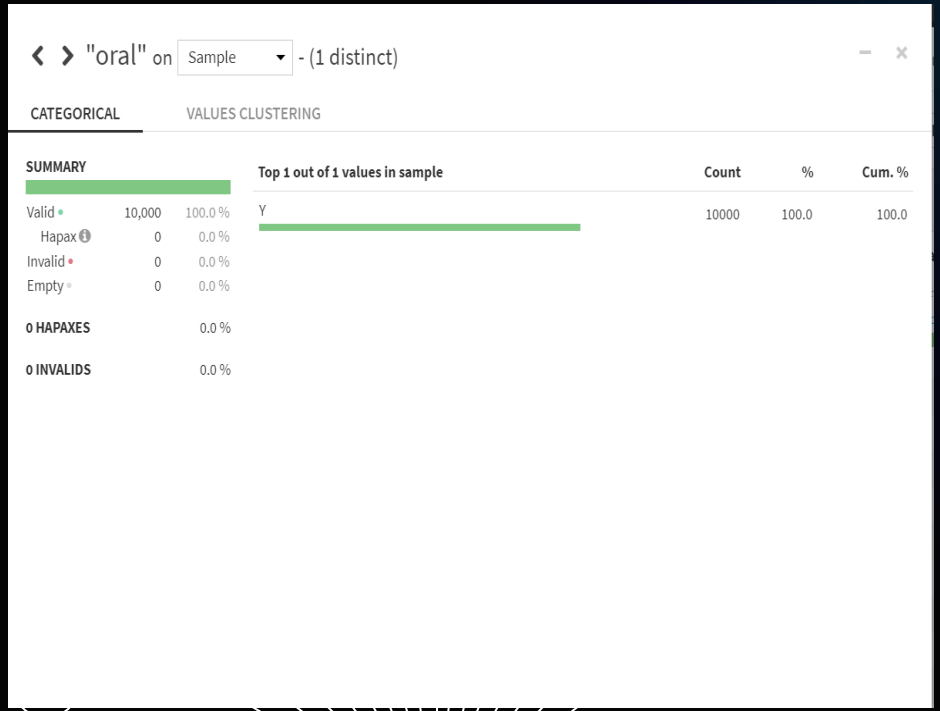
Summary stats

N values	55692
N distinct	22
N finite	55692
Mean	65.864935718
Median	65
Std Dev	12.820305713
Min	30
Max	135

Analysis Feature

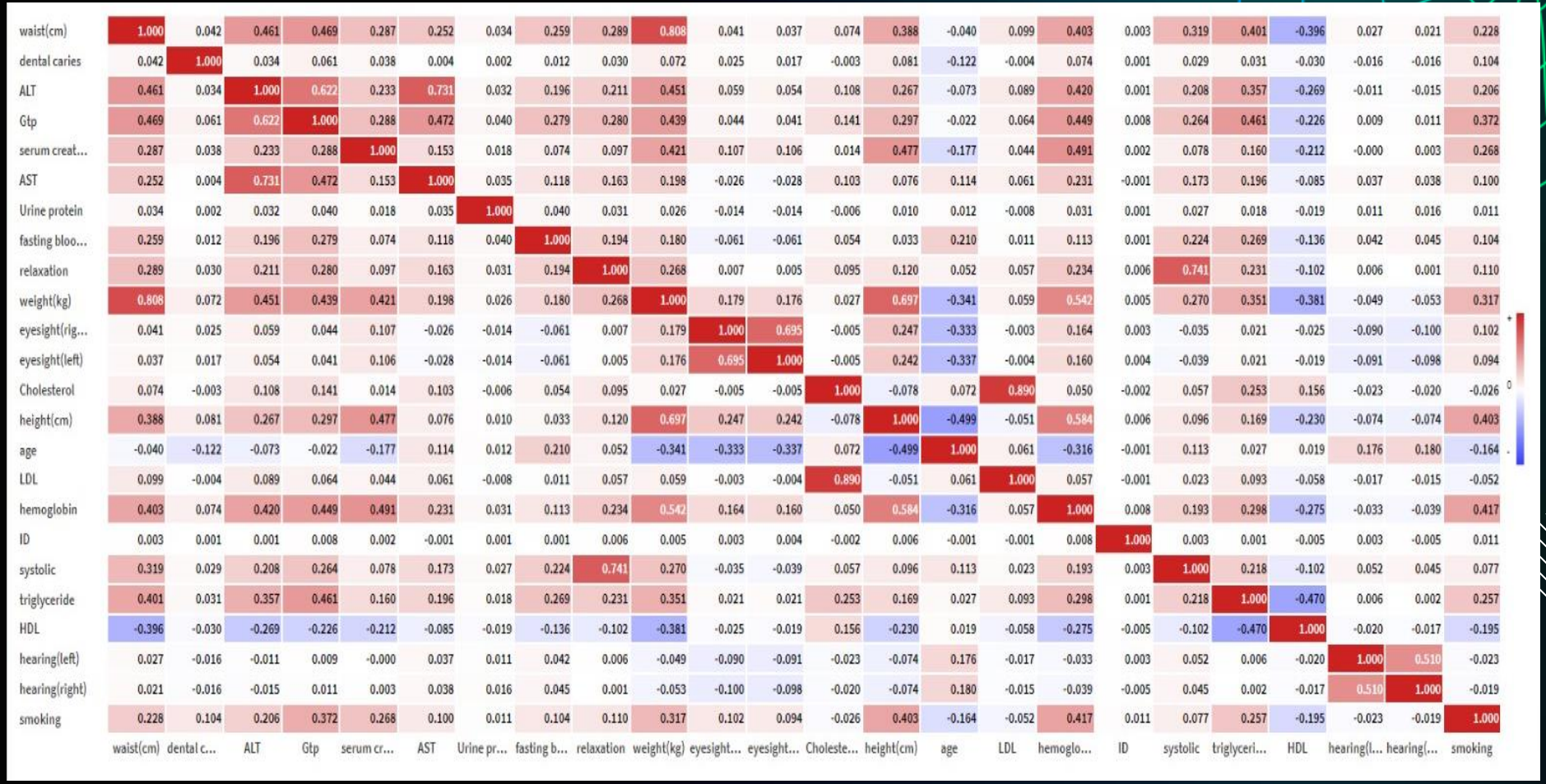
Based on the data from the chart below, we can see that the people surveyed in the data set have an average height mainly between 1m60 and 1m70, with an average height of 1m65. Most people's weight is between 60 and 80 kg, and the average weight is 65 kg.

Analysis Feature



After exploring all features of the dataset we see that The Oral column would not be useful for our model because it just has only one value in the sample. This value can't distribute more useful information to the algorithm. In other words, this feature hasn't predictive power. Hence, We decided to drop it out of the dataset.

Correlation matrix



Correlation Matrix Conclusion

Based on Correlation matrix we can see that:

Some features that are highly relevant to the target (Smoking) are respectively hemoglobin (0.417), height(0.403), Gtp(0.372), weight(0.317), serum creatinine(0.268), triglyceride(0.257), waist(0.228) and ALT(0.206). It means that this feature might have big predictive power to the target.

Click to add text

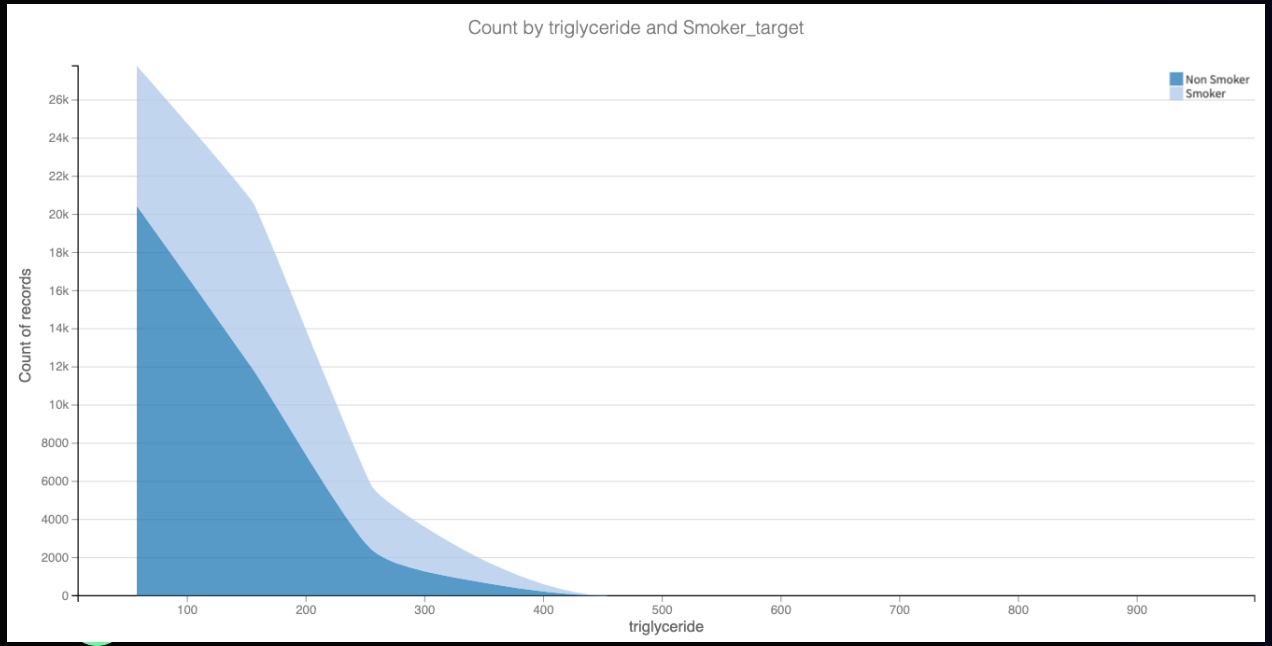
Some features that unrelated to the target (Smoking) are respectively Cholesterol (-0.026), Age(-0.164), LDL(-0.052), HDL(-0.195), hearing left (-0.023), hearing right (-0.019).

The correlation matrix works very well with the continuous variable, but the value type of our target is binary value. It's one of the categorical variables, so maybe the correlation matrix may not really accurately reflects the correlations of the predictors and target. So let's create some graphs between the predictors and the target in the detailed analysis step and see their right relationship.

04

Detailed analysis

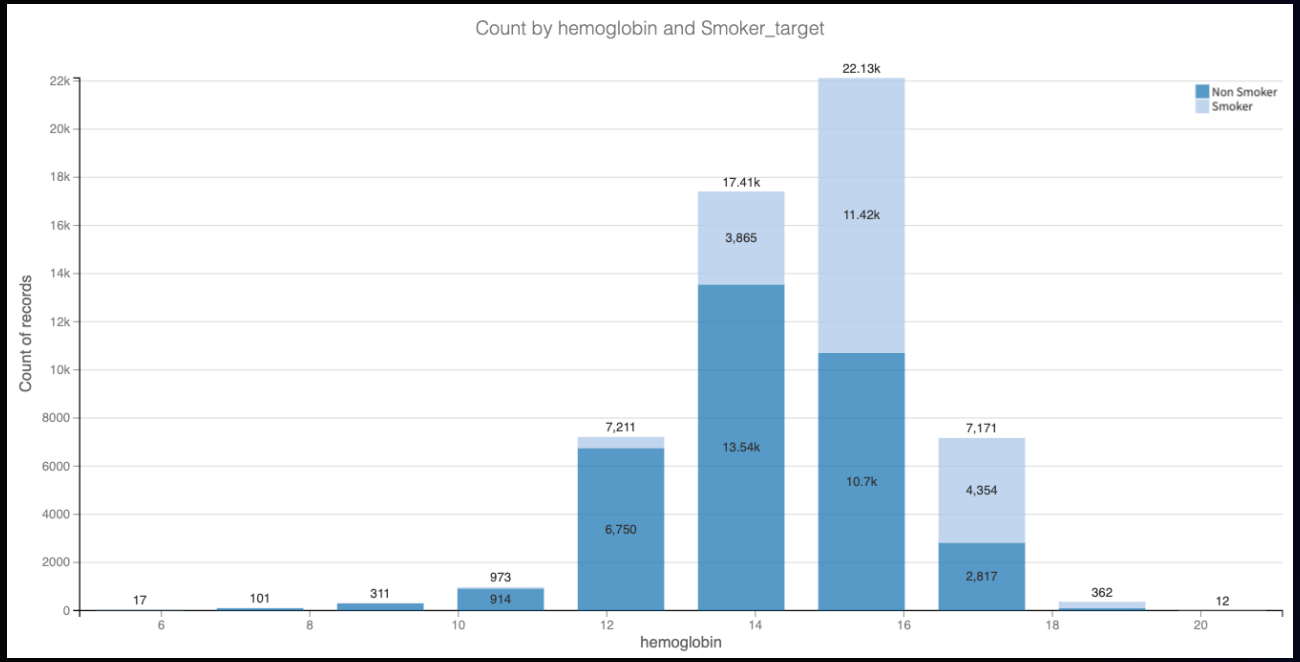
Bar Chart Gender Analysis



As per the graph majority of people who smoke has low triglyceride level than the people who does not smoke.

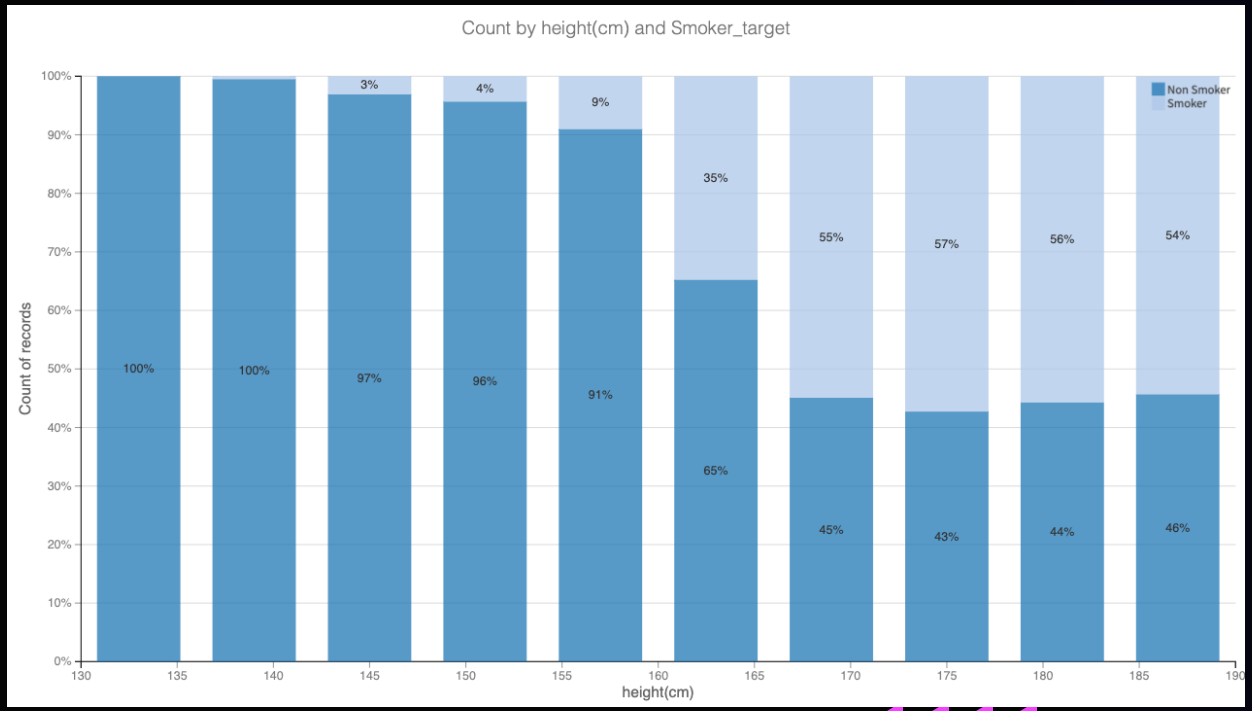
So , if we find low triglyceride level we can assume that he might be smoking.

Bar Chart Hemoglobin Analysis



As per the given chart majority of people who smoke their hemoglobin ranges between 15-16 compared to non - smoker hemoglobin ranges between 13-15. So we can assume that if person's hemoglobin level is between 15-16, he might be smoking.

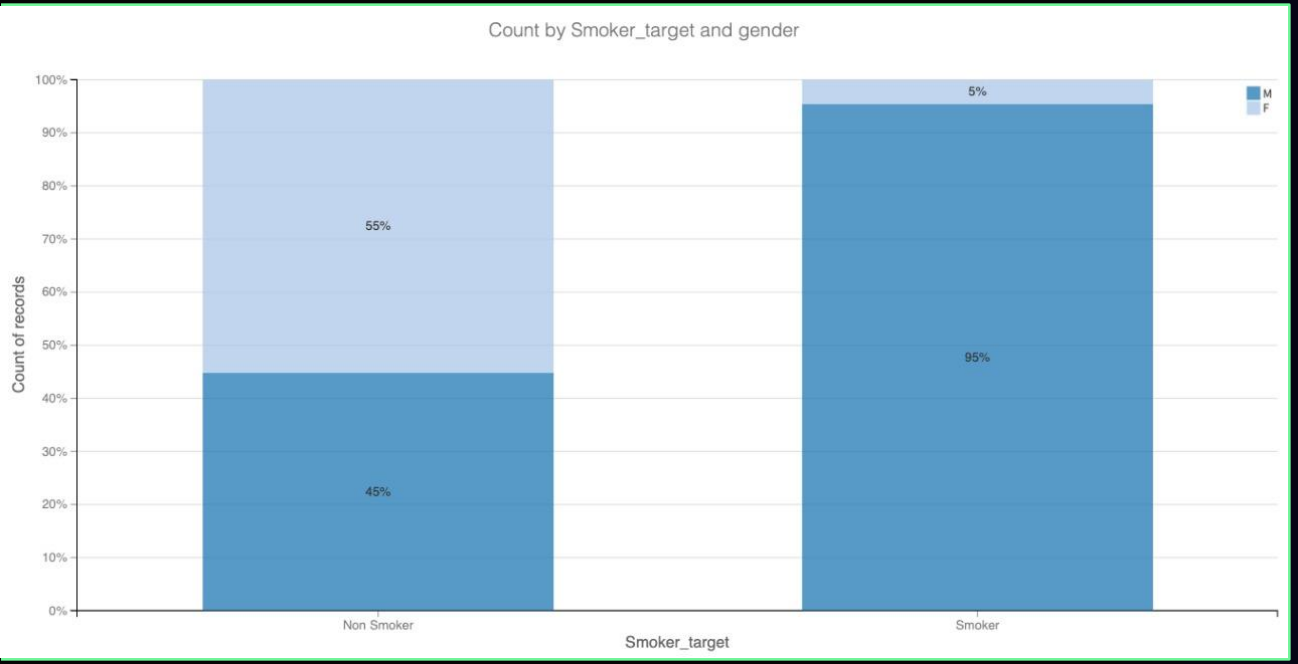
Bar Chart Gender Analysis



Majority of people who has height greater than 165 has high high chances that they smoke and majority of people having height lesser than 160 are non smokers.

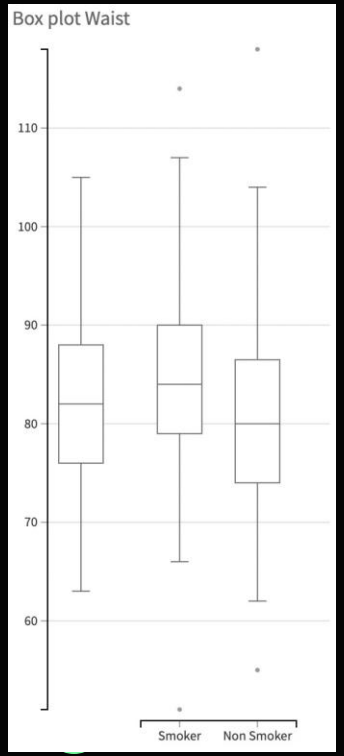
So we can assume than person whose height is greater than 165 might be a smoker.

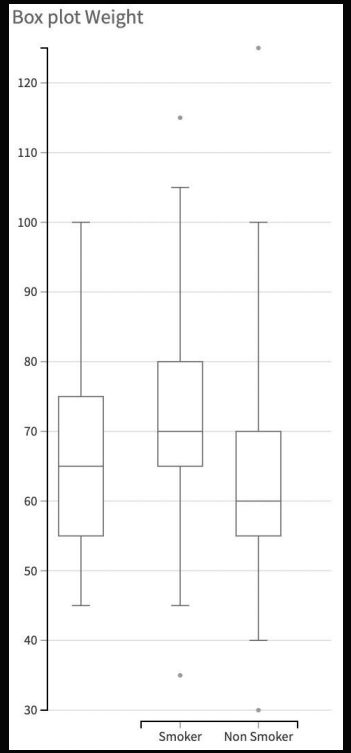
Bar Chart Gender Analysis

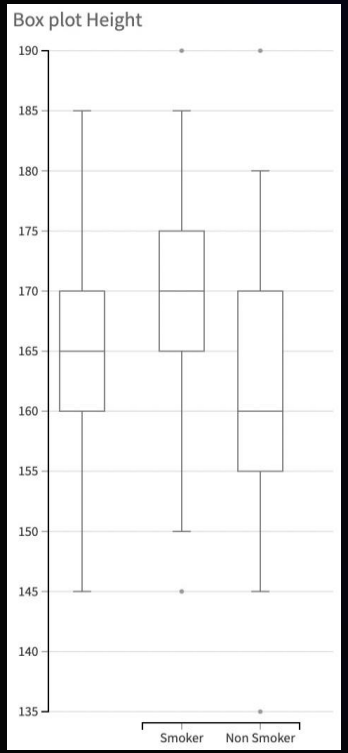


According to the graph, we can see the percentage of the male who doesn't smoke are almost same with that of the female. there are 90% more male smoker than female smoker. Hence, we can say gender has a big influence on people smoking or not.

BOX PLOT Waist, Weight, Height

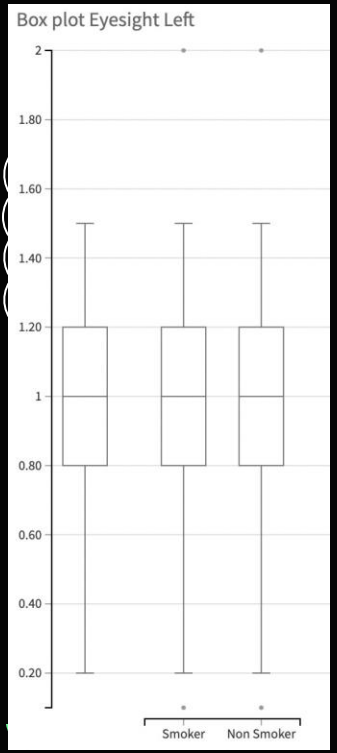


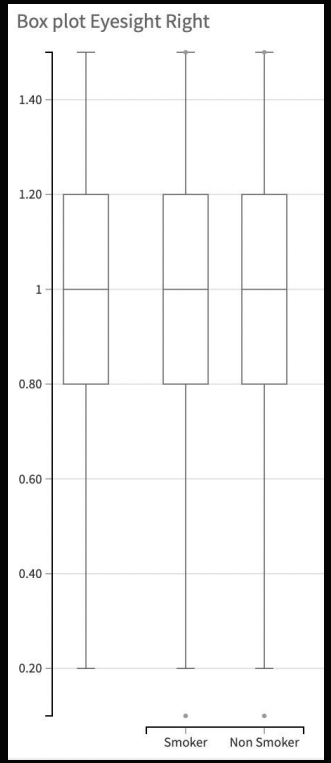


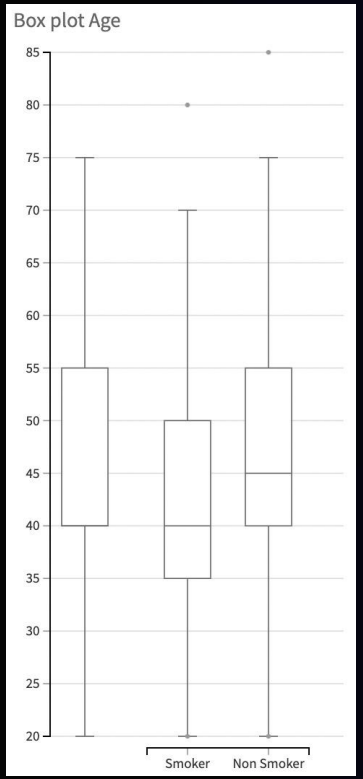


From these box plots, we can see that the variables have high importance in predicting if a person is a smoker since the difference between the values of Waist, Weight, and Height has different average values for smokers and non-smokers.

BOX PLOT Eyesight Left and Right, Age

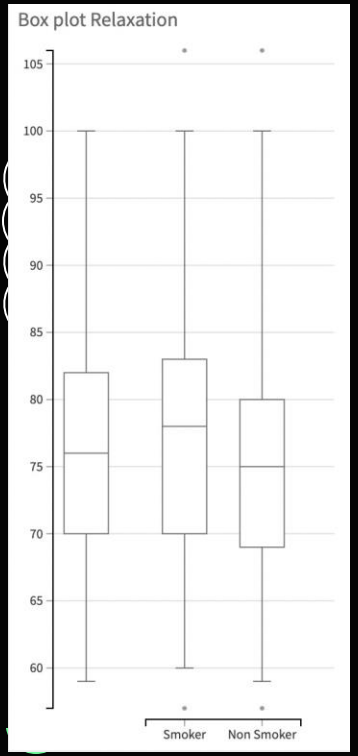


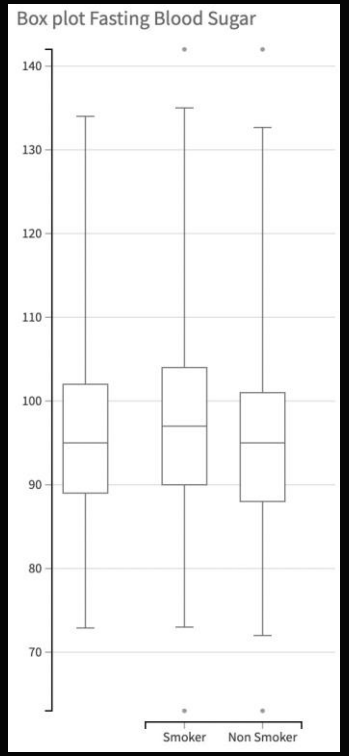


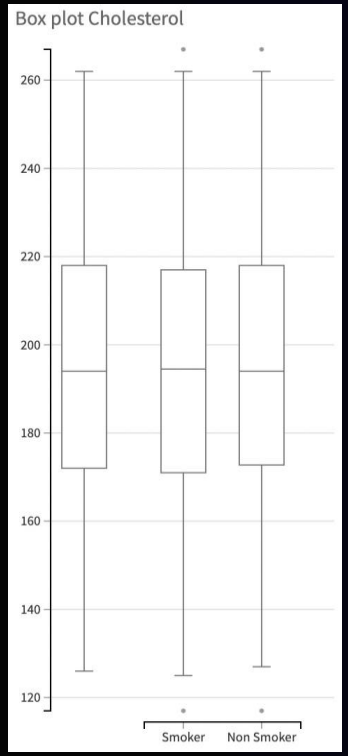


From these box plots, we can see that the variable age has small importance in predicting if a person is a smoker since the average age for smokers is 40 and for non-smokers is 45. On the other hand, for Eyesight left and right we see that these variables have non-relevance in knowing if a person is a smoker since the values for these variables are the same.

BOX PLOT Relaxation, Blood Sugar, Cholesterol

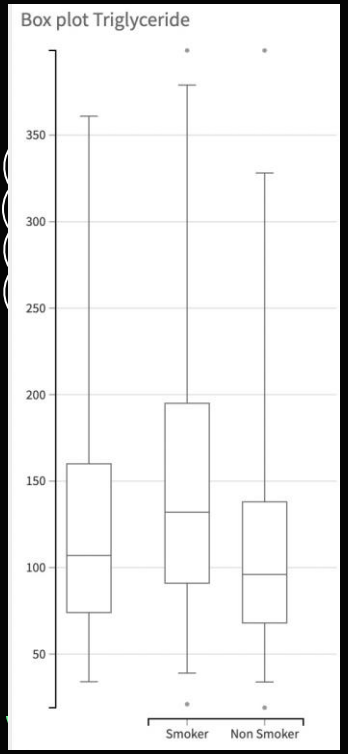


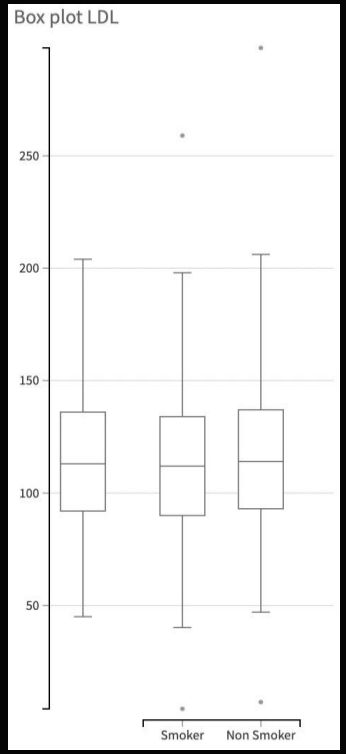


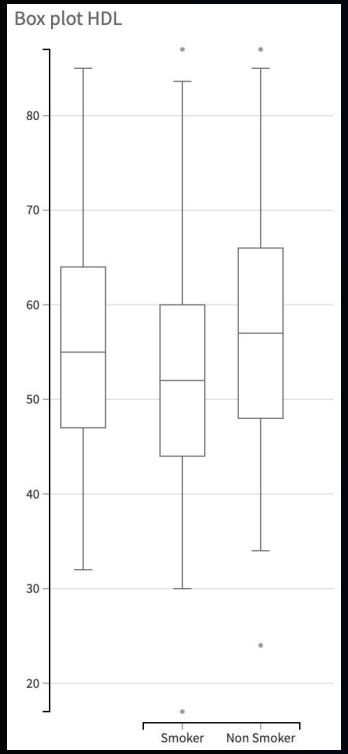


From these box plots, we can see that the variables relaxation and Fasting Blood Sugar have small importance in predicting if a person is a smoker since the average values have a very small difference between smokers and non-smokers. On the other hand, for Cholesterol we see that these variables have non-relevance in knowing if a person is a smoker since the values for these variables are the same for smokers and non-smokers.

BOX PLOT Triglyceride, LDL, HDL

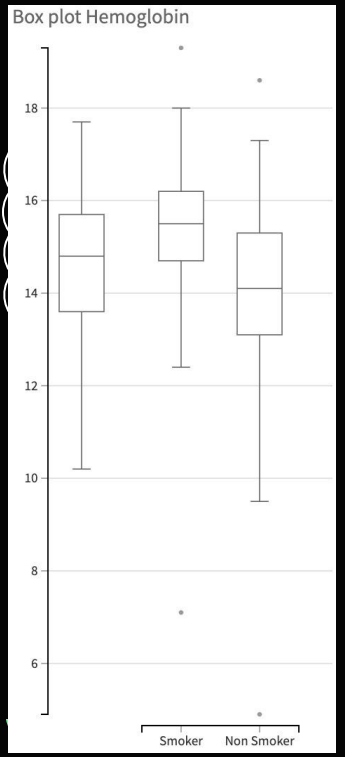


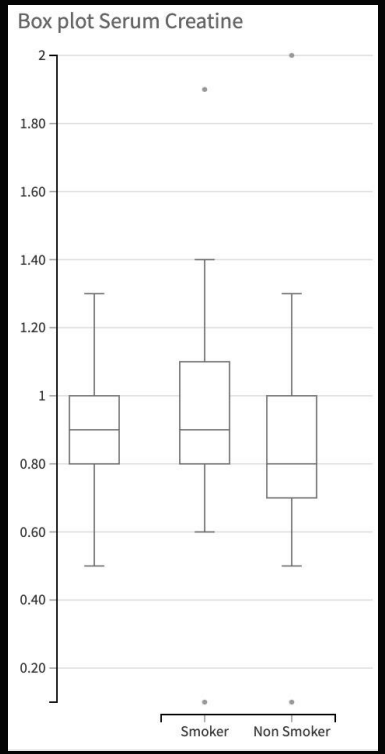


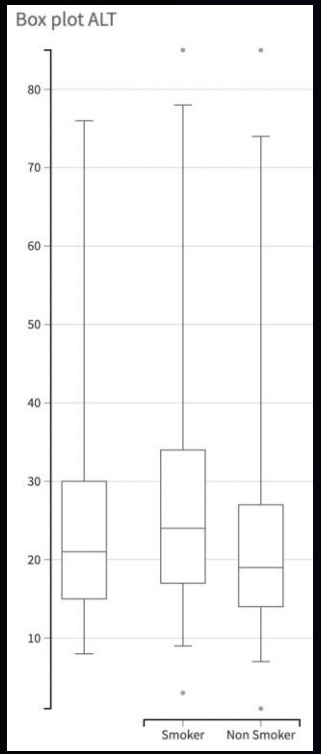


From these box plots, we can see that the variables Triglyceride and HDL have importance in predicting if a person is a smoker since the average values have a small difference between smokers and non-smokers. On the other hand, for LDL we see that these variables have non-relevance in knowing if a person is a smoker since the values for these variables are the same for smokers and non-smokers.

BOX PLOT Hemoglobin, Serum Creatine, ALT

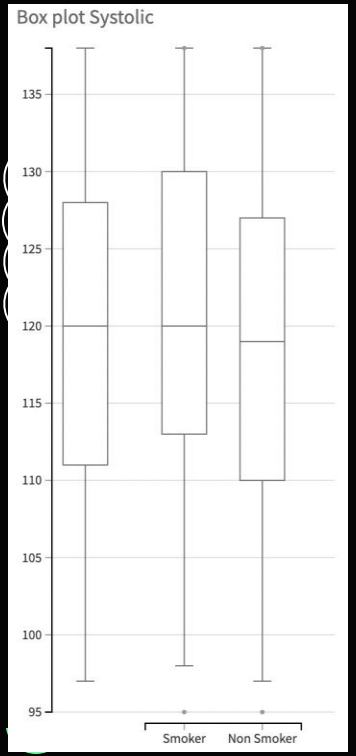


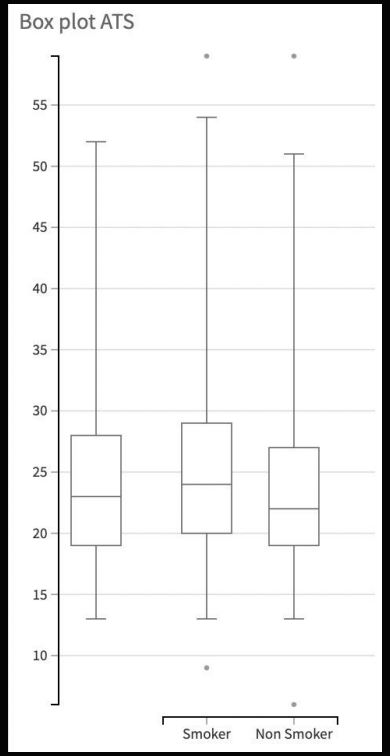




From these box plots, we can see that the variables Hemoglobin have importance in predicting if a person is a smoker since the average values have a big difference between smokers and non-smokers. On the other hand, for Serum Creatine and ALT we see that these variables have a small difference in knowing if a person is a smoker since the values for these variables only have a little difference for smokers and non-smokers.

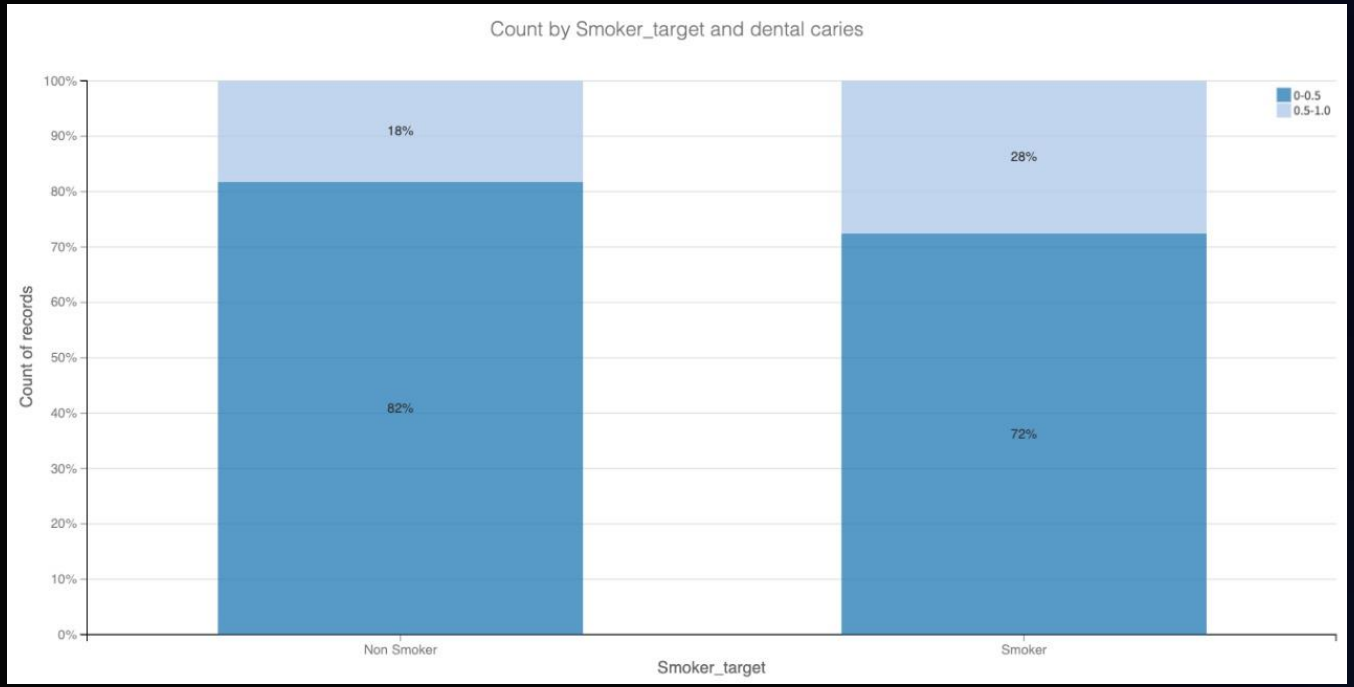
BOX PLOT Systolic, ATS





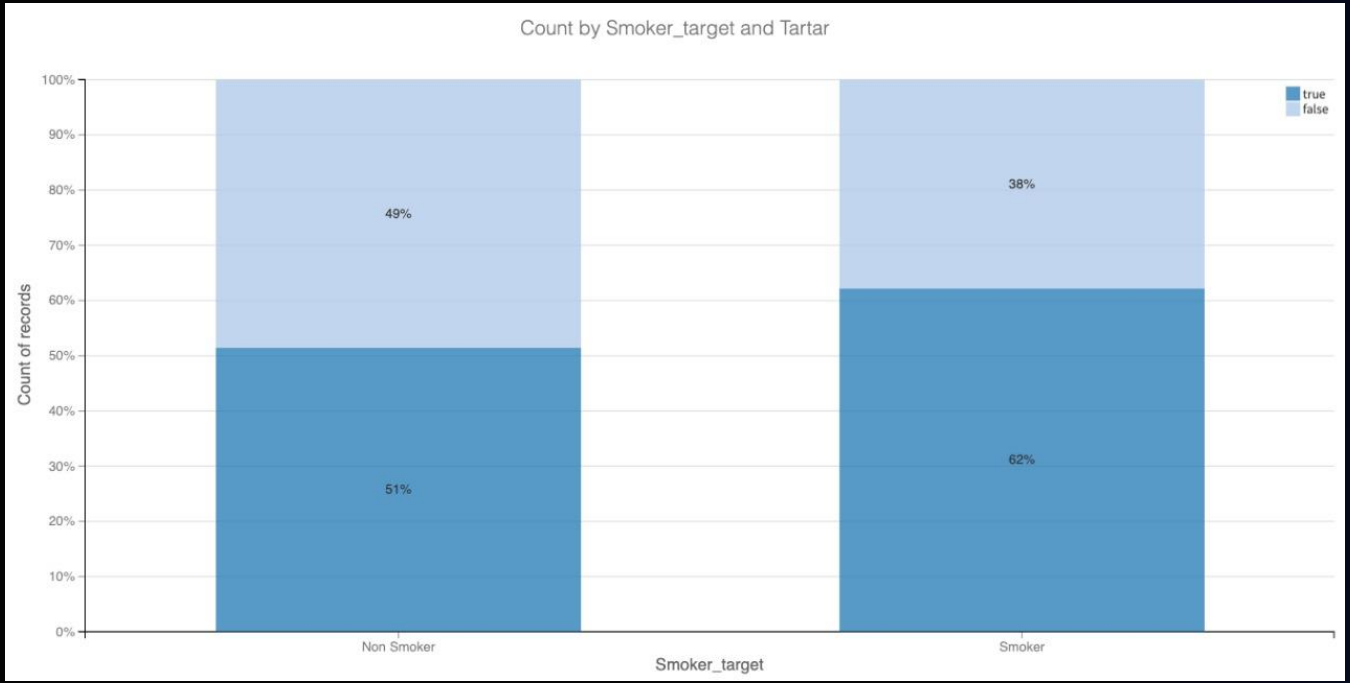
From the two box plots, we can see there is only a little difference for blood pressure and glutamic oxaloacetic transaminase type among people who smoke or not. So we can say smoking barely influence people's blood pressure and glutamic oxaloacetic transaminase(AST).

Bar Chart Dental Caries



According to the graph, we can see there is 18% of non-smokers instead of 28% of smokers who have more dental caries. So, we can say smoking decides slightly if having dental caries or not.

Bar Chart Tartar



According to the graph, we can see that 51% of non-smokers instead of 62% of smokers have tartar on the teeth. Hence, we can say smoking won't cause more tartar issues.

General results

Categorical

- We see that gender has high relevance in knowing if a person is a smoker.
- Dental caries and tartar have small importance in knowing if a person is a smoker.

Numerical

The following variables have a small relevance in knowing if a person is a smoker or not.

- Eyesight (left)
- Eyesight (Right)
- Hearing(left)
- Hearing (Right)
- Cholesterol
- LDL

05

Prediction

DESING

Algorithm

Since our target variable is binary, we decided to use random Forrest for our prediction model.

Sampling & Splitting

Since we have more people for non-smokers than non-smokers, we did a rebalance on the data set so that the target values can have a similar amount of values

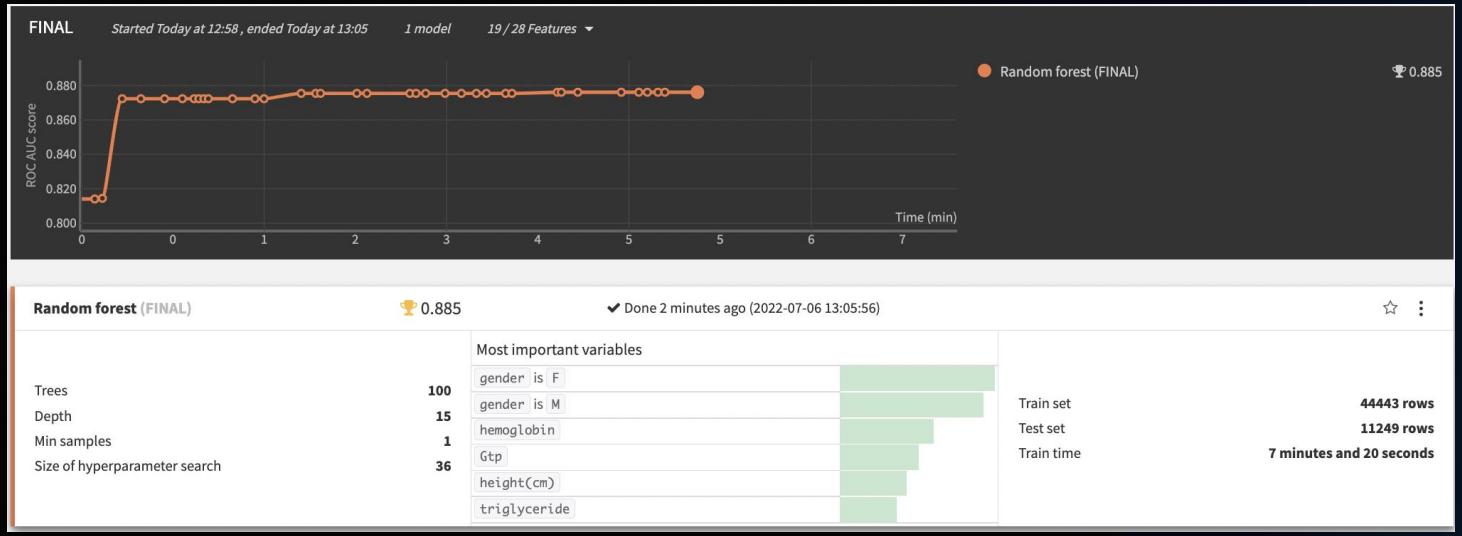
DESING

Features Handling

After the feature analysis, we decided to train the model with the following features.

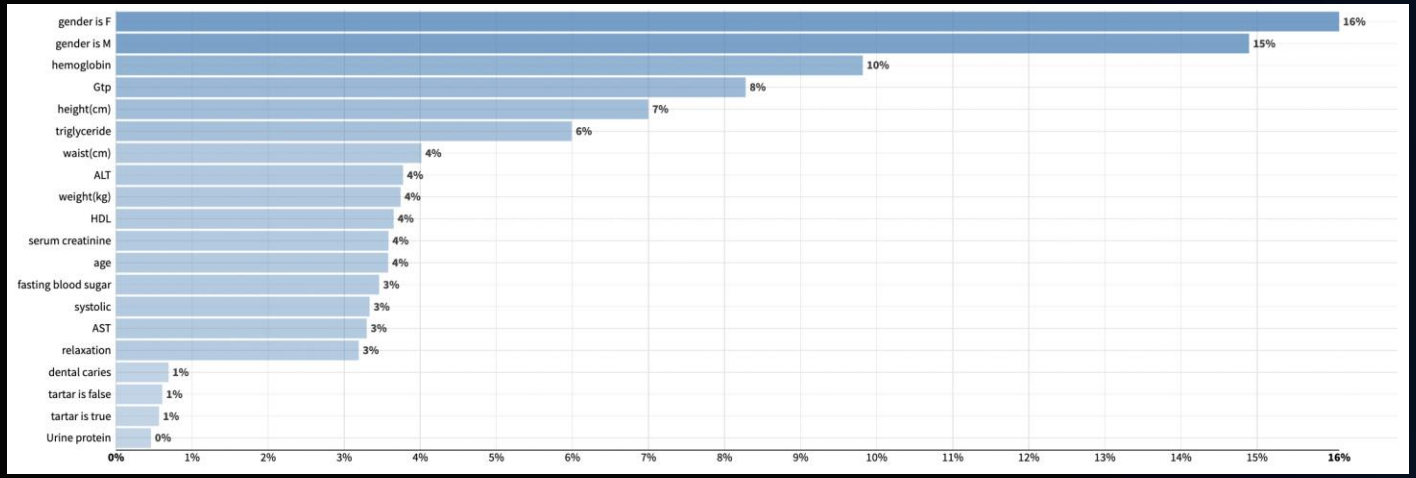
- gender
- age
- height(cm)
- weight(kg)
- waist(cm)
- systolic
- relaxation
- fasting blood sugar
- triglyceride
- hemoglobin
- Urine protein
- serum creatinine
- AST
- ALT
- Gtp
- dental caries
- tartar

Model performance



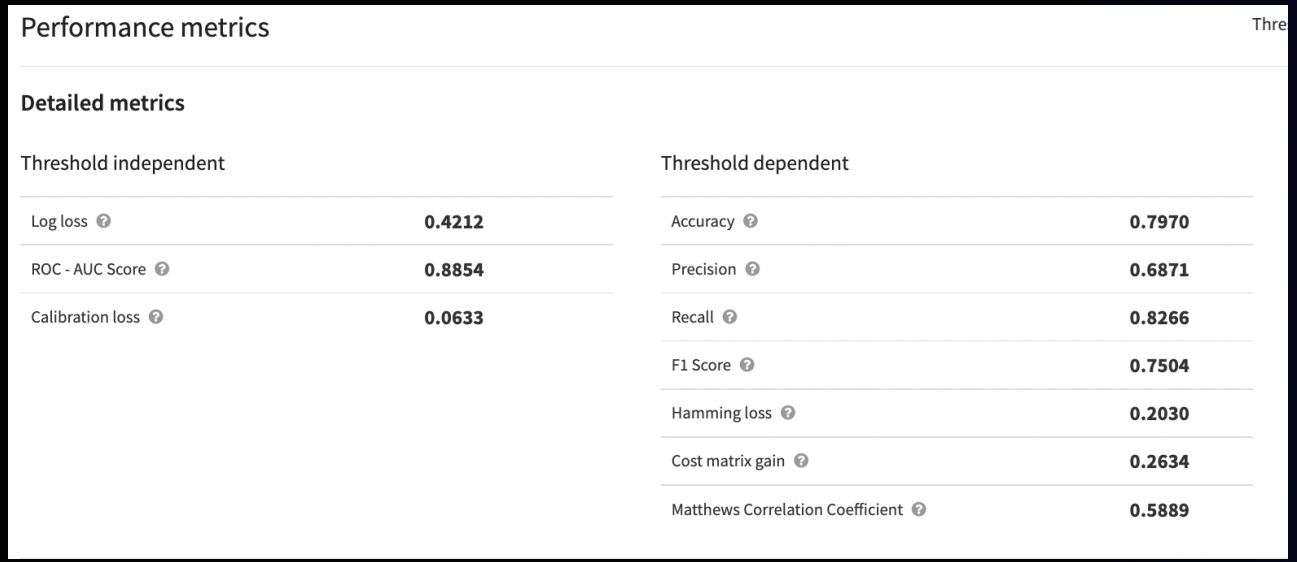
We have achieved 88.5% of ROC with most important features like gender, hemoglobin, height and triglyceride.

Variable Importance



The most important feature which has high impact on model training and its accuracy are gender, hemoglobin, GTP, height and Triglyceride.

Performance metrics



We have achieved overall accuracy of 79.7%. We majorly focused on recall to improve, because as per the use case if our model predicts that user is non-smoker but actual he was smoker it might cause a problem during medical treatments.

Algorithm Details

Algorithm details

Algorithm	**Random forest classification**	Split quality criterion	**Gini**
Number of trees	**100**	Use bootstrap	**Yes**
Max trees depth	**15**	Feature sampling strategy	**auto**
Min samples per leaf	**1**		
Min samples to split	**3**		

Training data

Rows (before preprocessing)	**44443**	Rows (after preprocessing)	**44443**
Columns (before preprocessing)	**28**	Columns (after preprocessing)	**24**
Matrix type	**dense**		
Estimated memory usage	**8.14 MB**		

We have used Random Forest Classifier for our model with 100 trees, max depth of 15 and at least 1 sample per leaf, split criterion is Gini and bootstrap enabled.

06

Conclusion

Conclusion on Analysis and prediction

- Our detailed analysis and model prediction are converging.
- We recommend that if Male smoker whose height is greater than 165 and hemoglobin level is 15-16 can be assumed as smoker.
- Later using non body signals and our model we can predict that if person is a smoker or not.
- We majorly concentrated on recall score because our model major use case can be in medical treatment and if my model predict user is non smoker but actually he is a smoker then it might cause problems in medical treatment which even might lead to death.

Thanks for your attention