Illinois Institute of Technology

Information Technology and Management

ITMD 522 Data Mining and Machine Learning

Presentation on
Health Insights: Integrating Data Mining and Machine Learning

First name	Last Name	IIT Email
Loka Adarsh	Dronamraju	idronamraju@hawk.iit.edu

Content

1. Abstract
2. Data Sets
3. Data Processing
4. Exploratory Data Analysis
5. Machine Learning
6. Predication
7. Association Rule
8. Clustering
9. Conclusion

Abstract

- When patients are diagnosed with respiratory issues or other health conditions, they often hesitate to disclose their smoking and drinking habits, including frequency. This reluctance may stem from various factors like fear of judgment or being perceived as negligent about personal health. Consequently, this lack of information leads to misinformation and complicates the task for medical professionals in providing appropriate treatment. Thus, there's a critical need for accurate knowledge of patients' smoking and drinking habits that doesn't solely rely on their self-reporting.

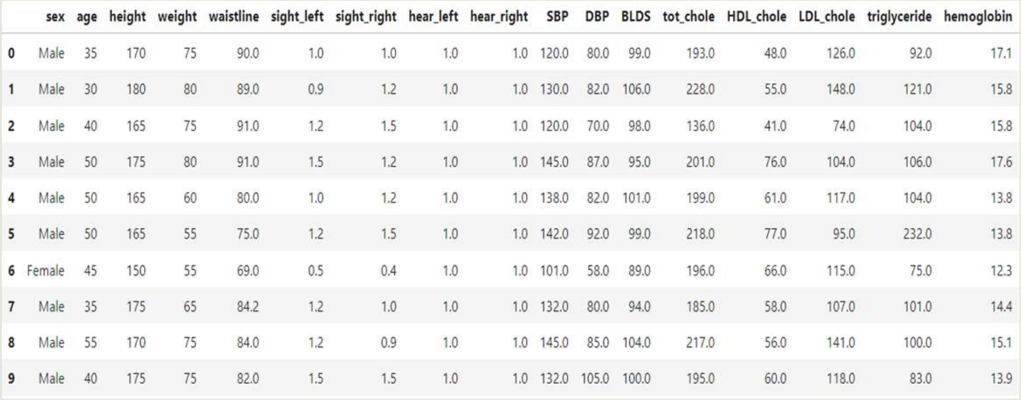
- Research indicates that this information can be inferred from various measurable body signals such as Blood Pressure, Cholesterol levels, Urine Proteins, and specific enzymes. Healthcare practitioners can utilize Machine Learning Models trained on datasets comprising diverse patient profiles to predict the habits of future patients.

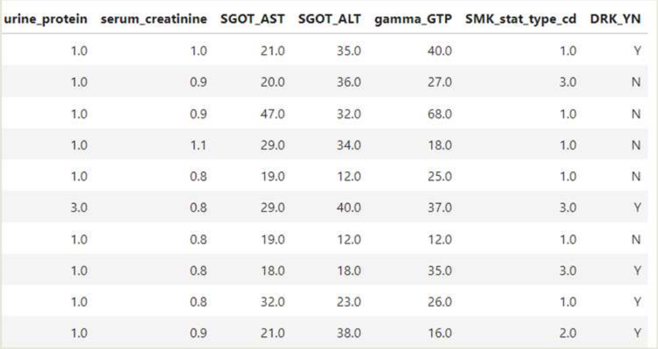
- This report presents an effort to address the issue, primarily focusing on smoking habits. However, the analytical approach and methodology discussed are equally applicable to assessing drinking habits.

Dataset

- The dataset utilized for the project is the Smoking and Drinking Dataset, comprising body signals, sourced from Kaggle. It was originally obtained from the National Health Insurance Service in Korea, with all personal and sensitive information omitted.

- This dataset consists of 991,320 rows and 24 columns





	sex	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	SMK_stat_type_cd	DRK_YN
0	Male	35	170	75	90.0	1.0	1.0	1.0	1.0	120.0	80.0	99.0	193.0	48.0	126.0	92.0	17.1	1.0	1.0	21.0	35.0	40.0	1.0	Y
1	Male	30	180	80	89.0	0.9	1.2	1.0	1.0	130.0	82.0	106.0	228.0	55.0	148.0	121.0	15.8	1.0	0.9	20.0	36.0	27.0	3.0	N
2	Male	40	165	75	91.0	1.2	1.5	1.0	1.0	120.0	70.0	98.0	136.0	41.0	74.0	104.0	15.8	1.0	0.9	47.0	32.0	68.0	1.0	N
3	Male	50	175	80	91.0	1.5	1.2	1.0	1.0	145.0	87.0	95.0	201.0	76.0	104.0	106.0	17.6	1.0	1.1	29.0	34.0	18.0	1.0	N
4	Male	50	165	60	80.0	1.0	1.2	1.0	1.0	138.0	82.0	101.0	199.0	61.0	117.0	104.0	13.8	1.0	0.8	19.0	12.0	25.0	1.0	N
5	Male	50	165	55	75.0	1.2	1.5	1.0	1.0	142.0	92.0	99.0	218.0	77.0	95.0	232.0	13.8	3.0	0.8	29.0	40.0	37.0	3.0	Y
6	Female	45	150	55	69.0	0.5	0.4	1.0	1.0	101.0	58.0	89.0	196.0	66.0	115.0	75.0	12.3	1.0	0.8	19.0	12.0	12.0	1.0	N
7	Male	35	175	65	84.2	1.2	1.0	1.0	1.0	132.0	80.0	94.0	185.0	58.0	107.0	101.0	14.4	1.0	0.8	18.0	18.0	35.0	3.0	Y
8	Male	55	170	75	84.0	1.2	0.9	1.0	1.0	145.0	85.0	104.0	217.0	56.0	141.0	100.0	15.1	1.0	0.8	32.0	23.0	26.0	1.0	Y
9	Male	40	175	75	82.0	1.5	1.5	1.0	1.0	132.0	105.0	100.0	195.0	60.0	118.0	83.0	13.9	1.0	0.9	21.0	38.0	16.0	2.0	Y

Dataset Columns

Column Name	Description
Sex	male, female
Age	years
Height	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	BLDS or FSG (fasting blood glucose) [mg/dL]
Tot_chole	total cholesterol [mg/dL]
HDL_chole	HDL cholesterol l[mg/dL]
LDL_chole	LDL 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)
Serum_creatinine	Serum (blood) creatinine [mg/dL]
SGOT_AST	SGOT (Glutamate-oxaloacetate transaminase) AST(Aspartate transaminase) [IU/L]
SGOT_ALT	ALT(Alanine transaminase) [IU/L]
Gamma_GTP	y-glutamyl transpeptidase [IU/L]
SMS_stat_type_cd	Smoking state, 1(never smoked), 2 (used to smoke now quit), 3(still smokes)
DRK_YN	Drinking or not drinking

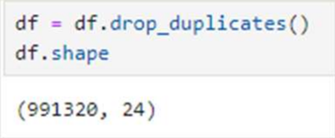
Data Processing

- Cleaning Data:

When data was observed we headed to data processing and cleaning.

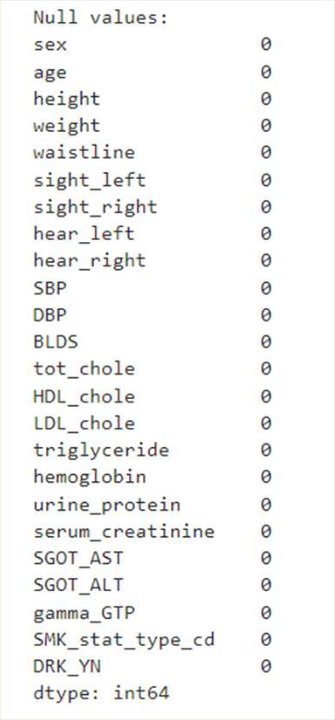
First duplicates were checked to see whether any duplicates are present in the dataset. The duplicates were found in the dataset and were removed. Now our new data has a shape 991,320 rows and 24 columns.

Then Null values were given a check, and no null values were found.



```
df = df.drop_duplicates()
df.shape
```

```
(991320, 24)
```



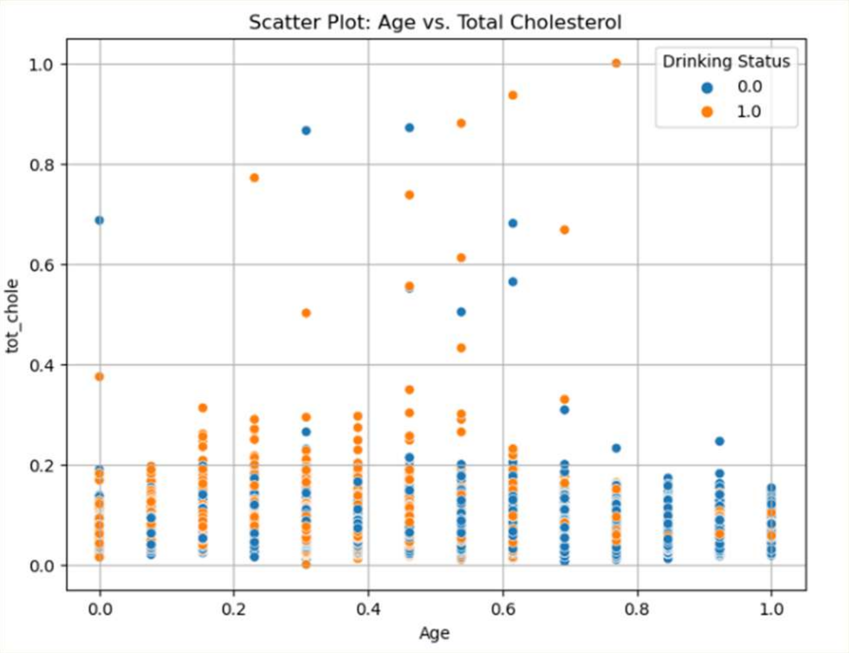
```
Null values:
sex                 0
age                 0
height              0
weight              0
waistline           0
sight_left          0
sight_right         0
hear_left           0
hear_right          0
SBP                 0
DBP                 0
BLDS                0
tot_chole           0
HDL_chole           0
LDL_chole           0
triglyceride        0
hemoglobin          0
urine_protein       0
serum_creatinine    0
SGOT_AST            0
SGOT_ALT            0
gamma_GTP           0
SMK_stat_type_cd    0
DRK_YN              0
dtype: int64
```

Exploratory Data Analysis

Observations revealed that the recorded values for 'height', 'weight', 'waistline', 'sight_left', 'sight_right', 'hear_left', 'hear_right', 'urine_protein', and 'serum_creatinine' were within the normal range for both smokers and non-smokers. As a result, these parameters were disregarded in the predictive analysis.

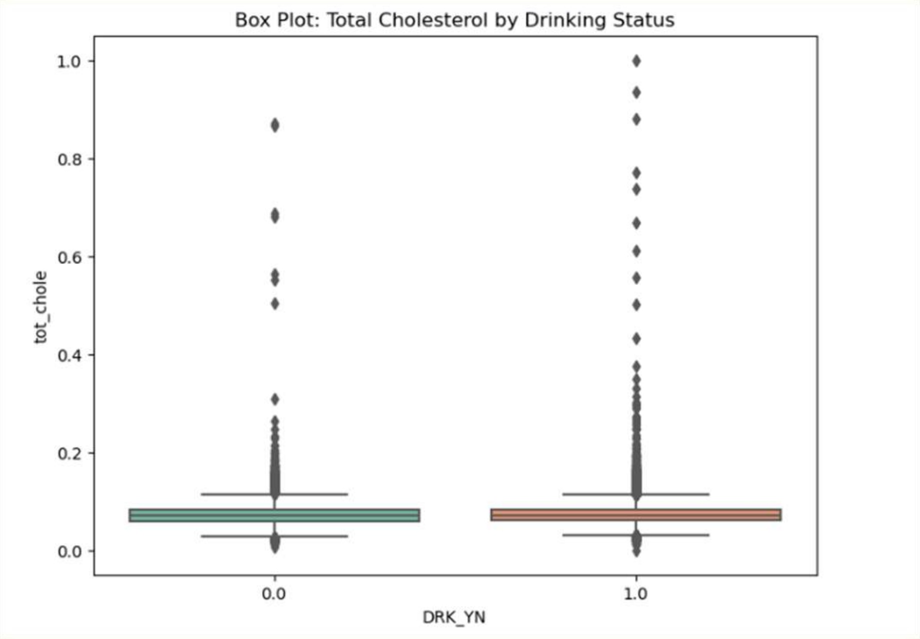
Scatter Plot

The data indicates a diverse distribution of total cholesterol levels among various age brackets. Although most observations fall within a lower range. There exist distinct anomalies displaying markedly elevated cholesterol concentrations, particularly within the 20 to 60 age category. The individuals belonging to the 'Y' (Yes) group are more likely to exhibit outliers with remarkably elevated cholesterol levels. This observation suggests a potential association between drinking habits and the variability of cholesterol levels.



Box Plot

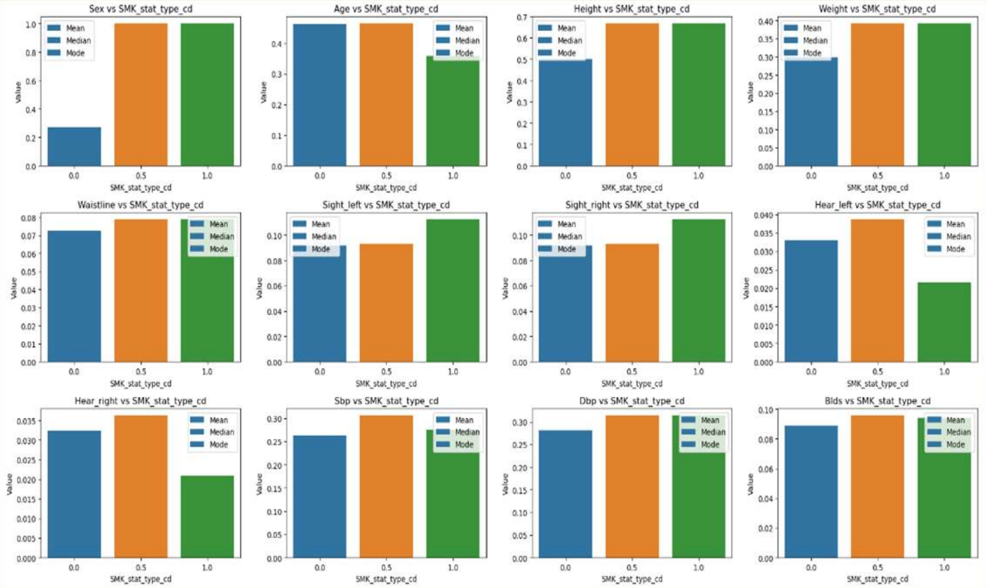
This plot indicates that individuals who have a drinking status of 'Y' (Yes) exhibit a broader spectrum of total cholesterol values, wherein certain data points deviate significantly with exceptionally high levels. Conversely, individuals with a drinking status of 'N' (No) typically demonstrate a more limited range with a reduced number of outliers



Categorical Variable Analysis

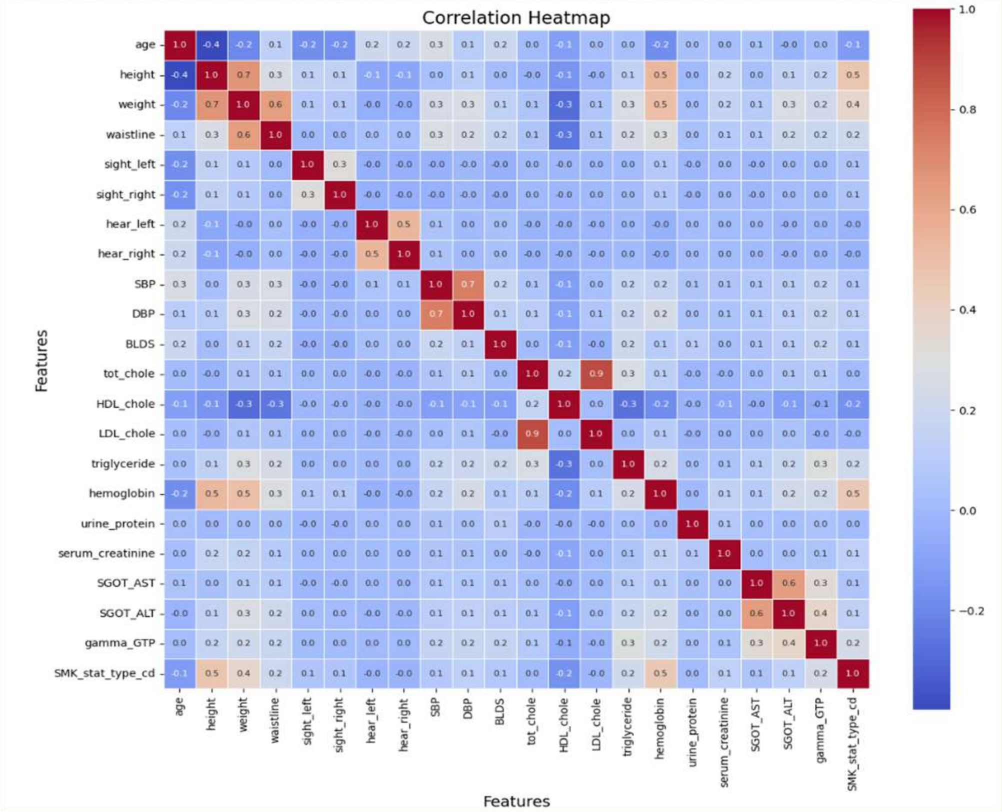
- In the following figure, we've graphed the mean, standard deviation, and median of the data points for each parameter. The horizontal axis categorizes individuals into 'Never Smoked,' 'Former Smoker,' and 'Current Smoker.' Interestingly, we observe a relatively consistent pattern across all smoker types, where the mean, standard deviation, and median values are quite similar.

1. Blue: Mean
2. Orange: Standard Deviation
3. Green: Median



Heatmap

There are strong positive correlations between SBP and DBP (0.74), indicating that as systolic blood pressure increases, diastolic blood pressure tends to increase too. Hemoglobin has a high correlation with SMK_stat_type_cd (0.45), suggesting that smoking status might be linked to hemoglobin levels. SGOT_AST and SGOT_ALT also share a strong correlation (0.64), indicating that liver enzyme levels tend to move together. HDL_chole has notable negative correlations with several variables. It has a -0.29 correlation with weight, suggesting that higher weight might be associated with lower HDL cholesterol.



Machine Learning

- **Logistic Regression**

Logistic regression is a linear classification algorithm commonly used for binary classification problems. It models the probability that a given input belongs to a certain class. Despite its simplicity, logistic regression can be very effective, especially when the relationship between the features and the target variable is roughly linear.

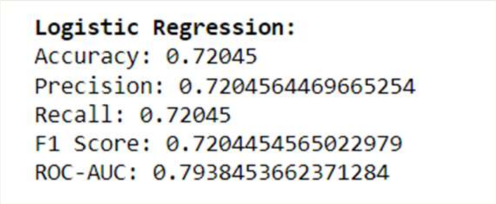
- **Gaussian Naive Bayes**

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with the "naive" assumption of independence between features. Gaussian Naive Bayes assumes that features follow a Gaussian (normal) distribution. It's simple, fast, and works well with high-dimensional data, although it may not capture complex relationships between features.

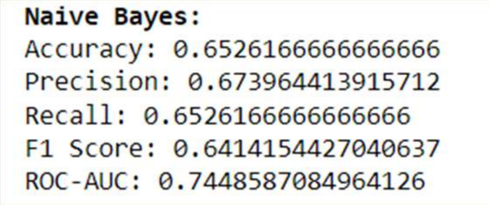
- **Decision Trees**

Decision trees are a versatile supervised learning algorithm used for both classification and regression tasks. They partition the feature space into regions based on the values of features, making decisions based on a series of if-else questions. Decision trees are easy to interpret and visualize, but they may suffer from overfitting if not properly regularized.

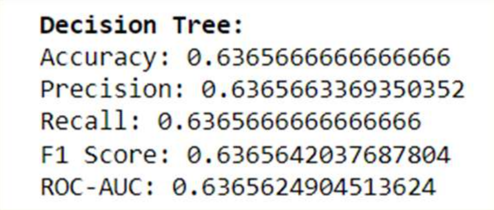
For DRK



```
Logistic Regression:
Accuracy: 0.72045
Precision: 0.7204564469665254
Recall: 0.72045
F1 Score: 0.7204454565022979
ROC-AUC: 0.7938453662371284
```

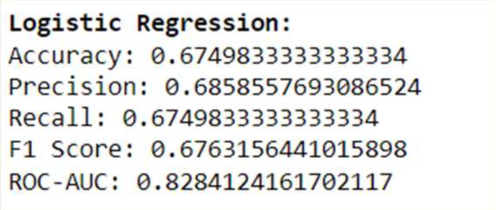


```
Naive Bayes:
Accuracy: 0.6526166666666666
Precision: 0.673964413915712
Recall: 0.6526166666666666
F1 Score: 0.6414154427040637
ROC-AUC: 0.7448587084964126
```

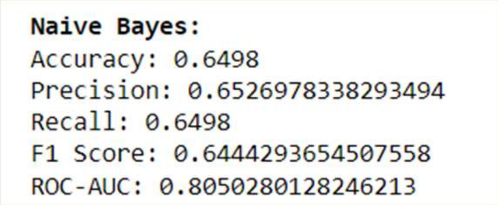


```
Decision Tree:
Accuracy: 0.6365666666666666
Precision: 0.6365663369350352
Recall: 0.6365666666666666
F1 Score: 0.6365642037687804
ROC-AUC: 0.6365624904513624
```

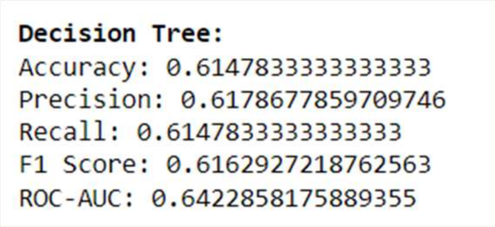
For SMK



```
Logistic Regression:
Accuracy: 0.6749833333333334
Precision: 0.6858557693086524
Recall: 0.6749833333333334
F1 Score: 0.6763156441015898
ROC-AUC: 0.828412416170211 7
```



```
Naive Bayes:
Accuracy: 0.6498
Precision: 0.6526978338293494
Recall: 0.6498
F1 Score: 0.6444293654507558
ROC-AUC: 0.805028012824621 3
```



```
Decision Tree:
Accuracy: 0.6147833333333333
Precision: 0.6178677859709746
Recall: 0.6147833333333333
F1 Score: 0.6162927218762563
ROC-AUC: 0.6422858175889355
```

- **Gradient Boosting**

Gradient boosting is an ensemble learning technique that combines multiple weak learners (typically decision trees) to create a strong predictive model. It builds trees sequentially, with each tree correcting the errors of the previous ones. Gradient boosting is known for its high predictive accuracy and robustness against overfitting.

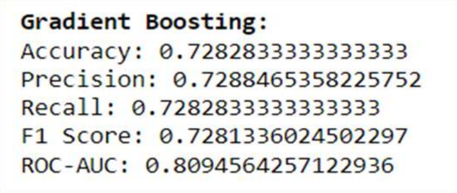
- **Random Forest**

Random forest is another ensemble learning method that builds multiple decision trees and combines their predictions through averaging (for regression) or voting (for classification). It improves upon decision trees by reducing overfitting and increasing predictive accuracy. Random forests are also capable of handling high-dimensional data and are less sensitive to noise and outliers.

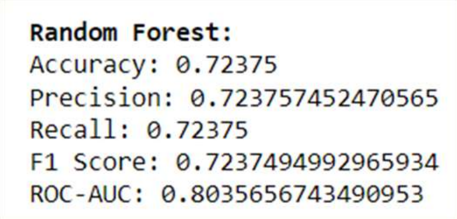
- **Artificial Neural Network**

Artificial Neural Networks (ANNs) are chosen for your dataset due to their effectiveness in handling complex relationships, especially with large datasets. ANNs excel in tasks like classification, regression, and pattern recognition. They consist of interconnected layers: input, hidden, and output. Activation functions, such as sigmoid, tanh, and ReLU, are applied to inputs. Training involves optimization algorithms like gradient descent to minimize loss. Post-training, performance is evaluated on a separate dataset to gauge generalization ability. Overall, ANNs offer a robust framework for exploring complex data structures and relationships

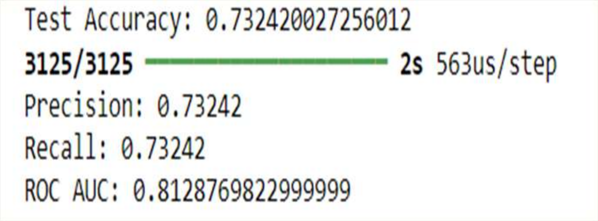
For DRK



```
Gradient Boosting:
Accuracy: 0.7282833333333333
Precision: 0.7288465358225752
Recall: 0.7282833333333333
F1 Score: 0.7281336024502297
ROC-AUC: 0.8094564257122936
```

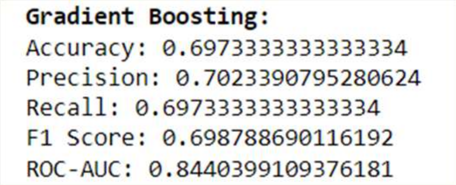


```
Random Forest:
Accuracy: 0.72375
Precision: 0.723757452470565
Recall: 0.72375
F1 Score: 0.7237494992965934
ROC-AUC: 0.8035656743490953
```

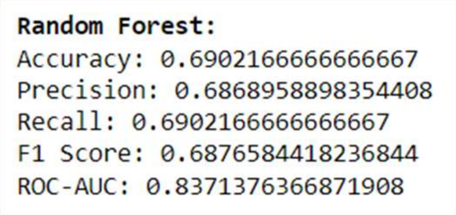


```
Test Accuracy: 0.732420027256012
3125/3125 ━━━━━━━━━━━━━━━━━━━━ 2s 563us/step
Precision: 0.73242
Recall: 0.73242
ROC AUC: 0.8128769822999999
```

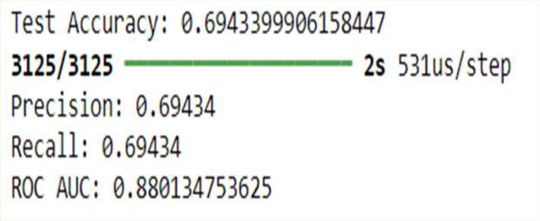
For SMK



```
Gradient Boosting:
Accuracy: 0.6973333333333334
Precision: 0.7023390795280624
Recall: 0.6973333333333334
F1 Score: 0.698788690116192
ROC-AUC: 0.8440399109376181
```



```
Random Forest:
Accuracy: 0.6902166666666667
Precision: 0.6868958898354408
Recall: 0.6902166666666667
F1 Score: 0.6876584418236844
ROC-AUC: 0.8371376366871908
```

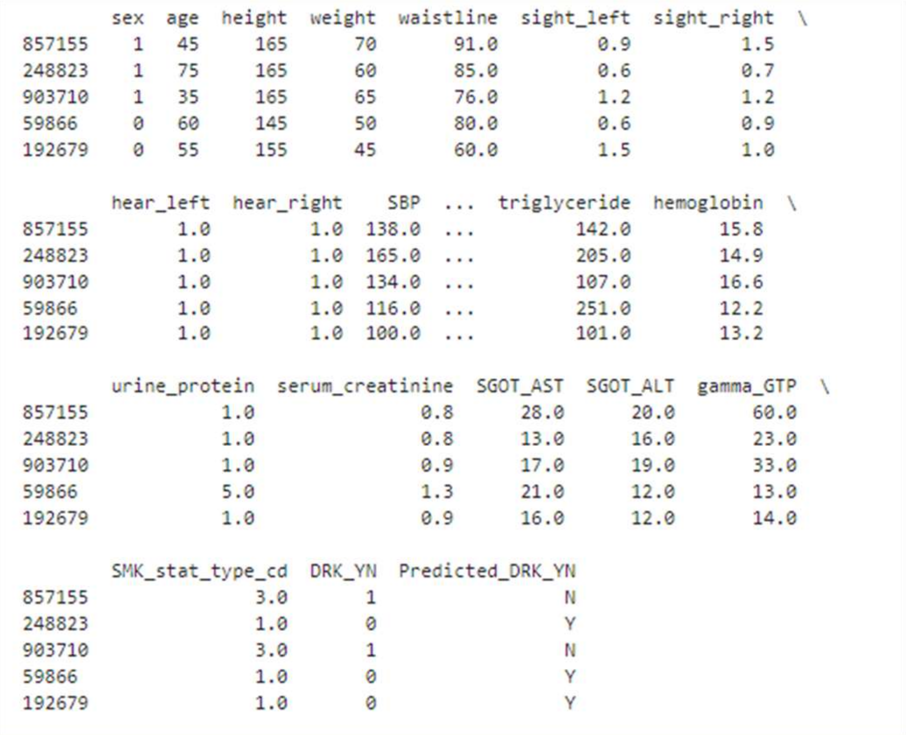


```
Test Accuracy: 0.6943399906158447
3125/3125 ━━━━━━━━━━━━━━━━━━━━ 2s 531us/step
Precision: 0.69434
Recall: 0.69434
ROC AUC: 0.880134753625
```

Prediction

Since we received highest accuracy and auc for Gradient Boosting, so we have chosen it for our prediction.

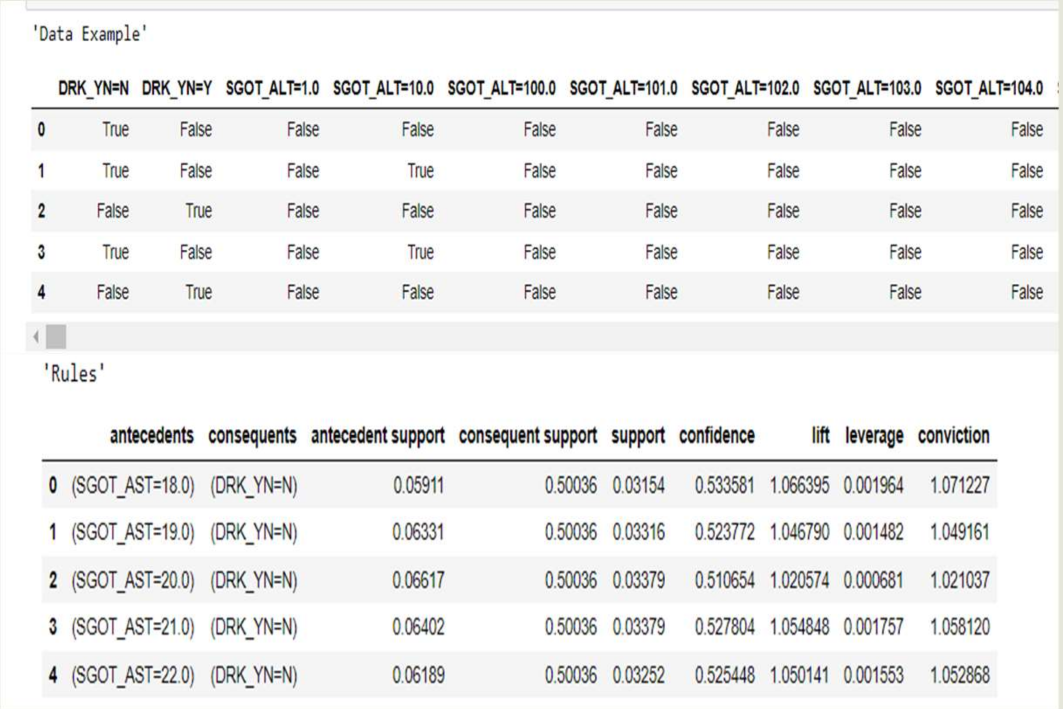
- Here 1 = N and 0 = Y



```
        sex  age  height  weight  waistline  sight_left  sight_right  \
857155    1   45     165      70       91.0         0.9          1.5
248823    1   75     165      60       85.0         0.6          0.7
903710    1   35     165      65       76.0         1.2          1.2
59866     0   60     145      50       80.0         0.6          0.9
192679    0   55     155      45       60.0         1.5          1.0

        hear_left  hear_right    SBP  ...  triglyceride  hemoglobin  \
857155        1.0         1.0  138.0  ...         142.0        15.8
248823        1.0         1.0  165.0  ...         205.0        14.9
903710        1.0         1.0  134.0  ...         107.0        16.6
59866         1.0         1.0  116.0  ...         251.0        12.2
192679        1.0         1.0  100.0  ...         101.0        13.2

        urine_protein  serum_creatinine  SGOT_AST  SGOT_ALT  gamma_GTP  \
857155            1.0               0.8      28.0      20.0       60.0
248823            1.0               0.8      13.0      16.0       23.0
903710            1.0               0.9      17.0      19.0       33.0
59866             5.0               1.3      21.0      12.0       13.0
192679            1.0               0.9      16.0      12.0       14.0

        SMK_stat_type_cd  DRK_YN  Predicted_DRK_YN
857155               3.0       1                 N
248823               1.0       0                 Y
903710               3.0       1                 N
59866                1.0       0                 Y
192679               1.0       0                 Y
```

Association Rule

- The antecedents demonstrate varying levels of SGOT_AST, with the support ranging from 0.030 to 0.034.
- The confidence values suggest a moderate likelihood of these antecedents leading to non-drinking status, indicating a certain relationship between SGOT_AST levels and abstinence from alcohol consumption.
- Lift values are approximately 1, signifying a relatively weak association between the antecedents and the consequents.
- These association rules imply only mild associations between SGOT_AST levels and non-drinking status. This might indicate a pattern where certain liver enzyme levels tend to be observed in individuals who do not drink



'Data Example'

	DRK_YN=N	DRK_YN=Y	SGOT_ALT=1.0	SGOT_ALT=10.0	SGOT_ALT=100.0	SGOT_ALT=101.0	SGOT_ALT=102.0	SGOT_ALT=103.0	SGOT_ALT=104.0
0	True	False	False	False	False	False	False	False	False
1	True	False	False	True	False	False	False	False	False
2	False	True	False	False	False	False	False	False	False
3	True	False	False	True	False	False	False	False	False
4	False	True	False	False	False	False	False	False	False

'Rules'

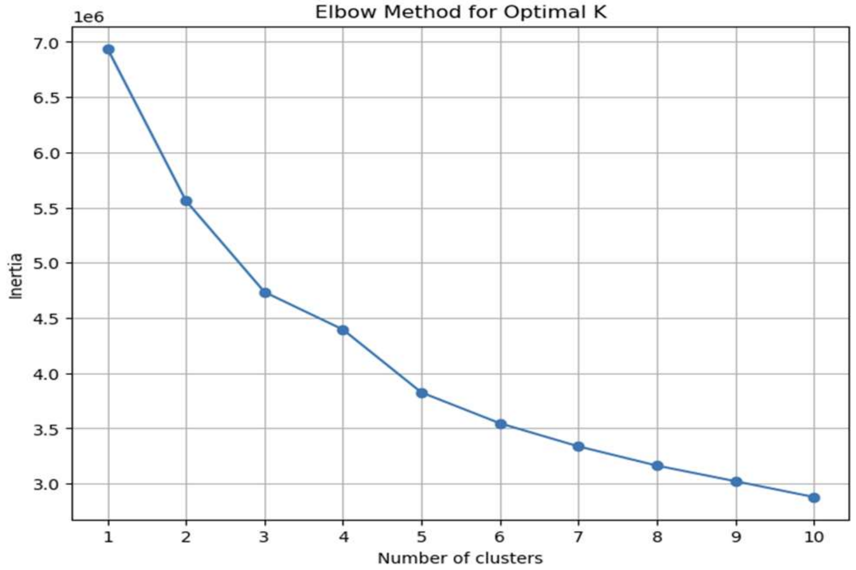
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(SGOT_AST=18.0)	(DRK_YN=N)	0.05911	0.50036	0.03154	0.533581	1.066395	0.001964	1.071227
1	(SGOT_AST=19.0)	(DRK_YN=N)	0.06331	0.50036	0.03316	0.523772	1.046790	0.001482	1.049161
2	(SGOT_AST=20.0)	(DRK_YN=N)	0.06617	0.50036	0.03379	0.510654	1.020574	0.000681	1.021037
3	(SGOT_AST=21.0)	(DRK_YN=N)	0.06402	0.50036	0.03379	0.527804	1.054848	0.001757	1.058120
4	(SGOT_AST=22.0)	(DRK_YN=N)	0.06189	0.50036	0.03252	0.525448	1.050141	0.001553	1.052868

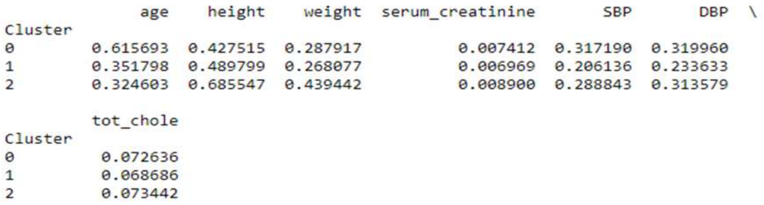
Clustering

- Cluster 0 comprises a subgroup that exhibits potential health issues related to the kidneys. Healthcare providers should prioritize interventions aimed at managing and improving kidney health within this cluster.

- Cluster 1 consists of a subgroup with generally healthier profiles. To maintain their healthy behaviors, targeted health promotion campaigns could prove effective in this cluster.

- Cluster 2 encompasses a subgroup of taller individuals who may have higher body mass and potential kidney-related health issues. In order to address these concerns, interventions prioritizing weight management and kidney health should be implemented within this cluster.





```
              age    height    weight  serum_creatinine       SBP       DBP  \
Cluster
0        0.615693  0.427515  0.287917          0.007412  0.317190  0.319960
1        0.351798  0.489799  0.268077          0.006969  0.206136  0.233633
2        0.324603  0.685547  0.439442          0.008900  0.288843  0.313579

         tot_chole
Cluster
0         0.072636
1         0.068686
2         0.073442
```

Conclusion

In conclusion, the project represents a significant effort to address the challenge of patients' reluctance to disclose their smoking and drinking habits in healthcare settings. By utilizing machine learning models trained on measurable body signals, the project aims to accurately infer these habits, thus providing healthcare practitioners with valuable information for better treatment decisions. Overall, the project has the potential to improve patient care and treatment outcomes related to respiratory issues and other health conditions associated with smoking and drinking habits

THANK YOU