Classification of Smokers and Drinkers

Project Report

Project By:

Ronit Kumar (48)

Paras Kumar(37)

Problem Statement

Smoking and drinking are detrimental to health due to their severe adverse effects. Smoking tobacco exposes the body to harmful chemicals that damage the lungs, heart, and blood vessels, leading to respiratory diseases, cardiovascular issues, and cancer.

Additionally, alcohol consumption in excess can harm the liver, and brain, and increase the risk of addiction, accident, and chronic conditions like liver disease and certain cancers.

Some people smoke or drink excessively to relieve stress or cope with problems. Studies have shown that social norms play a part in shaping behaviour. Often, people smoke or drink among friends who do so, to be socially accepted.
Also People hide this information from doctors, parents or life insurers because of fear of being looked down upon.

Our Model, on the basis of vitals of the body or on the basis of body signals, will classify whether a person is a smoker or drinker. The dataset is collected from the National Health Insurance Service in Korea. All personal information and sensitive data were excluded.

Research Paper 1

A study in alcohol: A comparison of data mining methods for identifying binge drinking risk factors in university students [1]

Introduction

A recent study by Patrick and Schulenberg (2013) showed that parents and peers, school and work, religiosity, exercise, and sports are among the predictors for binge drinking since 2011 for adolescents.

Aim

This study aims to discover which background factors are likely to contribute to students with hazardous drinking habits. The study aims to accomplish this by applying different classification techniques such as decision trees, random forest, and logistic regression on data collected from students studying at Linköping University.

The definition of hazardous consumption of alcohol

The Public Health Agency of Sweden (PHAS) (Andréasson et al., 2005) and on Karolinska Institutets website (Karlsson, 2017). They define the first guideline as the total weekly consumption which is 14 standardised drinks for men and 9 for women. The definition of binge drinking is 5 standardised drinks for men and 4 for women, respectively, at the same occasion.

Related work

1. The question regarding alcohol consumption was measured by two attributes wAlc and dAlc, weekly respectively daily alcohol consumption on a scale from 1 - very low to 5 - very high. The dataset was originally collected for EDM; the attributes consisted of student grades, demographic, social, family, and school related questions. Pagnotta and Amran (2016) defined a alcohol consumer by merging the attributes as in Equation 2.1

$$\sum Alc = (Walc \cdot 2 + Dalc \cdot 2)/7 \qquad (2.1)$$

   Pagnotta and Amran (2016) determines if a student is an alcohol consumer as when the value exceeds 3. The authors analysed the data using two different kinds of decision trees, namely C4.5 and random forest, and applied cross validation as test. The random forest technique provided the best results with 92 % accuracy and indicated that the most risky factors are social males with larger amounts of free time and less study time. A more recent study achieved an accuracy of 96.6 % with SVM and 98.5 % with the random forest algorithm.

2. A similar study from Spain carried out by Gervilla et al. (2010) studied the environmental, personal, and social variables in adolescents on the levels of alcohol consumption. This study defined drinking in measures of Standard Drink Units (SDU) meaning 10 grams of pure alcohol. The data consisted of drug use, demographics, academics scores, criminal records and other psycho-social and personality variables. In order to pinpoint alcohol consumption, the authors excluded all trial participants that were under any influence of earlier drug use. Gervilla et al. (2010) used several regression models to analyse and quantify the predictive value of social and environmental variables and QUEST (Quick, Unbiased, Efficient Statistical Tree) algorithm (Loh and Shih, 1997) to explore the organisation of the variables on the dependent variable SDU.

Data

Table 3.1: Table describing each attribute.

Attribute	Type	Value	Description of attribute
Gender	Categorical	Binary	Male or female.
Age	Categorical	Interval	18-20, 21-22, 23-25 or 26+.
Nation	Categorical	Binary	Yes or no.
Riskdrinker	Categorical	Binary	Yes or no
Faculty	Categorical	Multiple options	Tekfak, Filfak or medfak.
Semester	Scale	1-10	Where 10 equals 10 or more semesters.
FailExam	Categorical	Interval	0, 1-2, 3,4 or 5+ more courses.
SectionActive	Categorical	Binary	Yes or no.
StudieTim	Categorical	Interval	Hours of studying see question 8 Appendix A.
Trains	Categorical	Interval	Habits of training see Q.10 Appendix A
Badfood	Categorical	Interval	Appendix A, Q.11
Vegan	Categorical	Binary	Yes or no.
Sleep	Categorical	Interval	With: 5 or less, 6-8 or 9+ hours.
Relation	Categorical	Binary	Yes or no.
Computer	Categorical	Interval	Describing the computer use Appendix A Q.16
dAlc	Scale	1-5	Daily alcohol consumption
wAlc	Scale	1-5	Weekly alcohol consumption
FamRel	Scale	1-5	Family relation
Fhealth	Scale	1-5	Physical health
Mhealth	Scale	Interval	From 1-5 in mental health
Smoke	Categorical	Binary	Yes or no
Fsize	Categorical	Binary	Family size > 3 or less.
Ftog	Categorical	Binary	Yes or no. Asks if parents are together
pStudEvent	Categorical	Binary	Yes or no if student attends student events
SchemaUnd	Categorical	Binary	Yes, no or doesn't drink. Appendix A Q.26
DrickStudenEven	Categorical	Binary	Yes, no or doens't drink if student drink at student parties
SpareTimeFriend	Scale	1-5	In how much spare time student spend with friends
Fritid	Scale	1-5	In how much spare time student have

Predictive results

1. Decision Tree
   Decision trees were created based on a training set (70 %) and were later tested on the remaining 30 %. The decision tree was constructed by having the Riskdrinker variable as dependent and the remaining as predictors.
2. Random Forest
   The results suggest that 1000 trees perform best with an accuracy of 88 %. The table (4.1) shows that the out of bag (OOB) error rate is dropping with the increased number of trees. What is shown in table 4.1 is that the accuracy seems to increase by approximately 10 % from 1 to 1000 trees. Additionally, the precision increased with 13 % and recall increased with 22 %.

Table 4.1: The evaluation of the number of trees constructed in the Random Forest model.

Number of trees	Accuracy	Precision	Recall	F1	OOB error rate
1	79.50 %	54.05 %	45.45 %	49.38 %	23.53 %
10	87.00 %	56.76 %	67.74 %	61.76 %	21.63 %
100	87.00 %	62.16 %	65.71 %	63.88 %	13.74 %
500	87.50 %	67.57 %	65.79 %	66.66 %	13.61 %
1000	88.00 %	67.57 %	67.57 %	67.56 %	13.34 %

3. Logistic Regression

Logistic Regression models were also applied in order to identify the contributing factors to binge drinking.

Evaluation of all models

The random forest model with 1000 iterations resulted in the highest accuracy. However, the logistic regression model proved to be the most reliable in terms of precision and recall where it outperformed the random forest with nearly 20% .

Table 4.2: The evaluation of each model in terms on accuracy, precision, recall and F1 measure

Model	Accuracy	Precision	Recall	F1
Information Gain	87.00 %	59.46 %	66.67 %	62.86 %
Information Gain (Pruned)	87.50 %	70.27 %	65.00 %	67.54 %
Gini Index	85.00 %	56.76 %	60.00 %	58.34 %
Information Gain (K-fold 5)	87.50 %	70.27 %	65.00 %	67.54 %
C5.0	87.50 %	64.86 %	66.67 %	65.76 %
Random Forest (1000 trees)	88.00 %	67.57 %	67.57 %	67.56 %
Logistic Regression	85.50 %	93.87 %	88.95 %	91.34 %
Logistic Regression (K-fold 10)	86.50 %	92.64 %	90.96 %	91.80 %

Table 4.3: Table presenting the AUC values

Model	AUC
Decision trees (Information Gain)	0.83
Logistic regression	0.89
Random Forest	0.90

Research Paper 2

A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms [2]

Introduction

Lung cancer is the most commonly diagnosed cancer and the leading cause of cancer death globally. Lung cancer is mostly caused due to smoking. Such a disease needs to be predicted in advance to take the necessary remedial action against this disease. Despite all solutions attained by the previous research, they have all proven to be inadequate for the detection of the disease at an earlier stage.

Dataset

The Datasets used in this study are taken from the UCI Machine Learning Repository and data.world.

UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/lung+cancer
In this dataset: Number of Instances: 32
Number of Attributes: 57 (1 class attribute, 56 predictive)
Attribute Information: attribute 1 is the class label

data.world: https://data.world/cancerdatahp/lung-cancer
In this dataset: Number of Instances:1000
Number of Attributes:25(1 class attribute,24 predictive)
Attribute Information: attribute 25 is the class label

Evaluation of all Models

First, the given datasets are divided into training and test data by using k-fold cross validation technique.
Then using the classification algorithms such as SVM ,Logistic Regression, Naïve Bayes and Decision Tree, respective classification models are implemented using the given training data. The classification models are created using training data and the corresponding models are evaluated using test data to get the accuracy of the models. Finally, they compared the accuracy rates of each and every classification models that they implemented and arrived at a conclusion.

Table 1:Lung Cancer Dataset:UCI Machine Learning Repository

No of Folds	ML Algorithm	Accuracy(%)
7	Logistic Regression	96.9
7	Decision Tree	85.71

Table 2:Lung Cancer Dataset:data.world

ML Algorithm	Accuracy(%)
Logistic Regression	66.7
Decision Tree	90
Naïve Bayes	87.87
SVM	99.2

Table 1 shows that the performance of Logistic Regression exceeds the performance of Decision Tree, whereas Table 2 shows that the performance of SVM exceeds all other classification algorithms including Logistic Regression. So we can conclude that SVM has the highest accuracy rate among all other classification algorithms for these particular datasets.

Methodology

The dataset is collected from the National Health Insurance Service in Korea. All personal information and sensitive data were excluded.

Details of Dataset:

Number of Instances : 991346
Number of attributes : 24

Column/Feature	Description
Sex	Male, Female
Age	Round up to 5 years
Height	Round up to 5 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)
Haemoglobin	haemoglobin(g/dL)
Urine_protein	Protein in urine, 1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4)
Serum_creatinnine	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	Gamma - 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

1. The "**Sex**" column indicates the gender of the individuals in the dataset, with options for "male" or "female."
2. "**Age**" represents the age of the individuals, rounded up to the nearest 5 years.
3. "**Height**" specifies the height of individuals, rounded up to the nearest 5 centimetres (cm).
4. The "**Weight**" column denotes the weight of individuals in kilograms (kg).

5. **Sight_left** / **Sight_right**: These columns describe the eyesight or visual acuity in the left and right eyes, respectively.
6. "**Hear_left**" and "**Hear_right**" indicate the hearing condition in the left and right ears, respectively.
7. "**SBP**" stands for Systolic Blood Pressure, and "**DBP**" stands for Diastolic Blood Pressure, both measured in millimetres of mercury (mmHg). These values provide critical information about an individual's cardiovascular health, helping to diagnose hypertension (high blood pressure) and assess cardiovascular risk.
8. "**BLDS**" represents Blood Sugar Level or Fasting Blood Glucose, measured in milligrams per deciliter (mg/dL). This value is important for monitoring blood sugar control and diagnosing conditions such as diabetes.
9. **Tot_chole** / **HDL_chole** / **LDL_chole** / Triglyceride: These columns pertain to cholesterol levels in the blood. "Tot_chole" represents total cholesterol, "HDL_chole" is High-Density Lipoprotein cholesterol (good cholesterol), "LDL_chole" is Low-Density Lipoprotein cholesterol (bad cholesterol), and "Triglyceride" measures a type of fat in the blood. These values are crucial for assessing cardiovascular health and risk.
10. "**Haemoglobin**" denotes the level of haemoglobin in the blood, measured in grams per deciliter (g/dL). Haemoglobin is vital for oxygen transport, and this measurement is used to assess anaemia and overall blood health.
11. "**Urine_protein**" indicates the presence of protein in urine, with values ranging from 1 (-) to 6 (+4). It is a marker for kidney function and can help diagnose kidney disease and related conditions.
12. "**Serum_creatinine**" represents the concentration of serum (blood) creatinine, measured in milligrams per deciliter (mg/dL). It is used to assess kidney function and is important for monitoring renal health.
13. **SGOT_AST** / **SGOT_ALT** / **Gamma_GTP**: These columns include enzyme measurements related to liver and metabolic health. "SGOT_AST" and "SGOT_ALT" are enzymes used to assess liver function, while "Gamma_GTP" is involved in various metabolic processes. Elevated levels can indicate liver or bile duct issues.
14. "**SMK_stat_type_cd**" indicates the smoking status of individuals, with values of 1 (never smoked), 2 (used to smoke but quit), or 3 (still smoke).
15. "**DRK_YN**" represents whether individuals are drinkers (Y for Yes) or non-drinkers (N for No).

Since we want to predict smokers and drinkers using body signal data our target variable is **SMK_stats_type_cd** and **DRK_YN**. It is a numerical variable, which suggests that it might be a numeric encoding for different health categories.

Taking references from the research papers read, we will be using the techniques which have given better accuracy for the data sets.

The data was divided into training (80%) and testing (20%) and then analysed with logistic regression, decision trees and random forest.

To compare the different classification techniques precision, recall, f1, ROC AUC and support is used.

ROC Curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

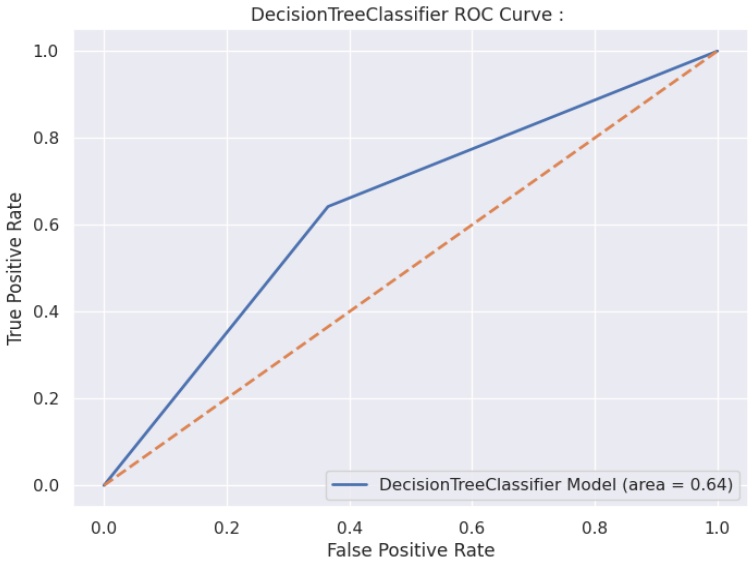
Experimental Results

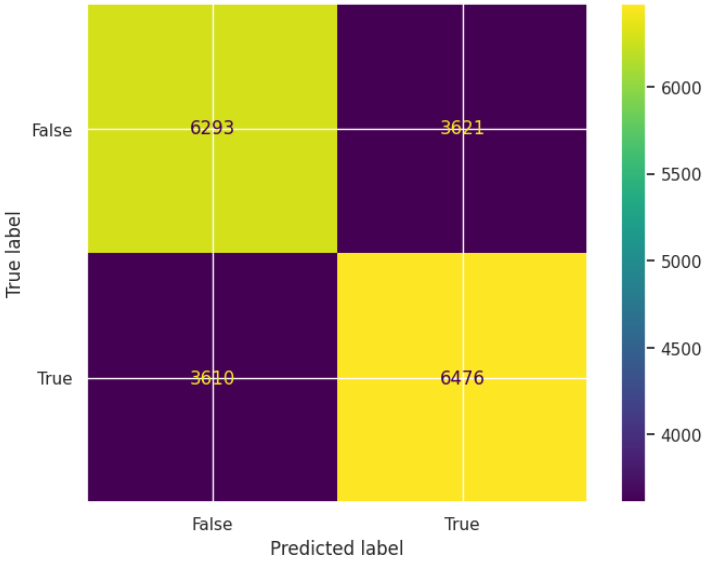
- For 2 class problem we found that the Random Forest Classifier worked best for identifying whether a person is a Drinker or a Non Drinker, Logistic Regression Classifier performed almost similar to Random Forest Classifier.
- For 3 class problem, again, Random Forest Classifier worked the best, giving the maximum accuracy, precision, recall and f-1 scores.
- Random Forest also achieved the highest ROC AUC in both the classification problems.
- No correlation was found between the Drinker and Smoker variable in the dataset.

Classification of Drinker:

Results : Decision Tree

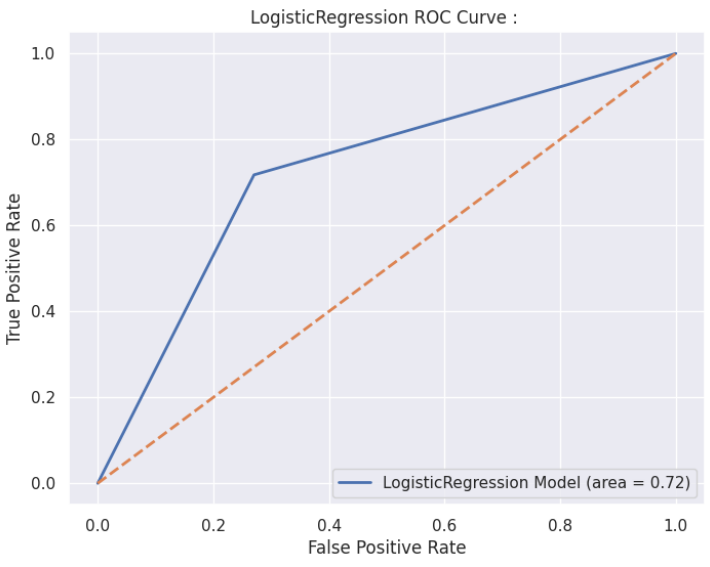
	precision	recall	f1-score	support
Yes	0.635464	0.634759	0.635111	9914.00000
No	0.641379	0.642078	0.641728	10086.00000
accuracy	0.638450	0.638450	0.638450	0.63845
macro avg	0.638421	0.638419	0.638420	20000.00000
weighted avg	0.638447	0.638450	0.638448	20000.00000

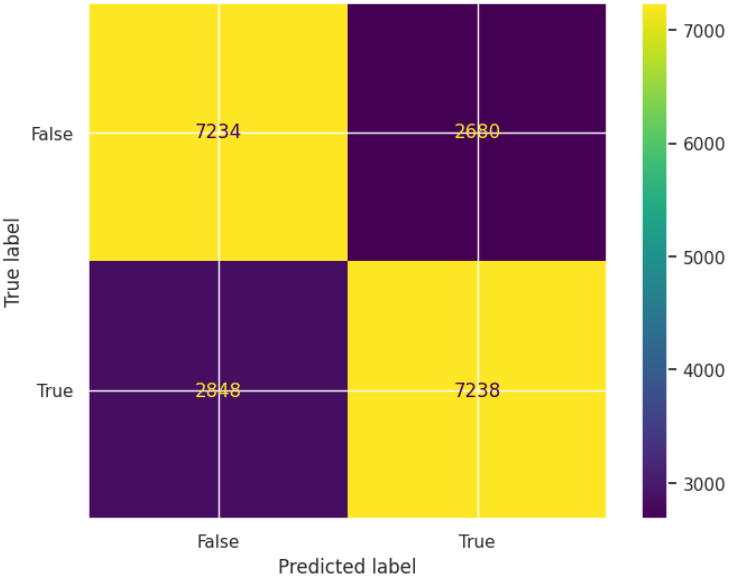




Results : Logistic Regression

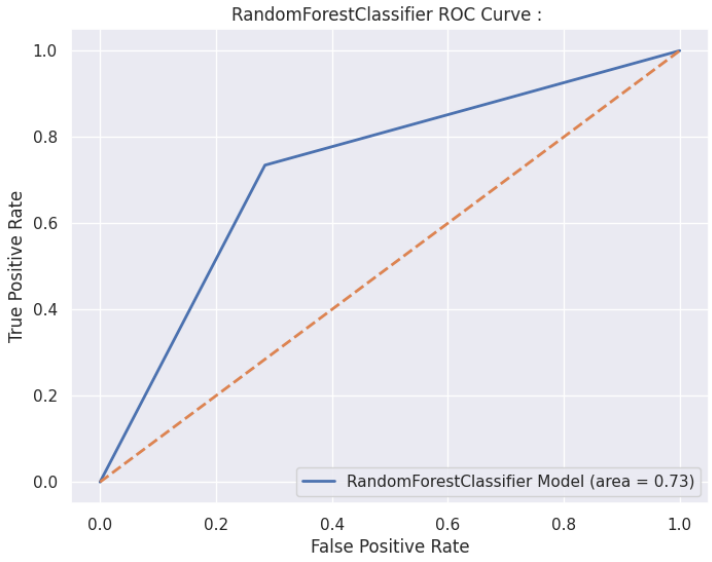
	precision	recall	f1-score	support
Yes	0.717516	0.729675	0.723545	9914.0000
No	0.729784	0.717628	0.723655	10086.0000
accuracy	0.723600	0.723600	0.723600	0.7236
macro avg	0.723650	0.723652	0.723600	20000.0000
weighted avg	0.723703	0.723600	0.723600	20000.0000

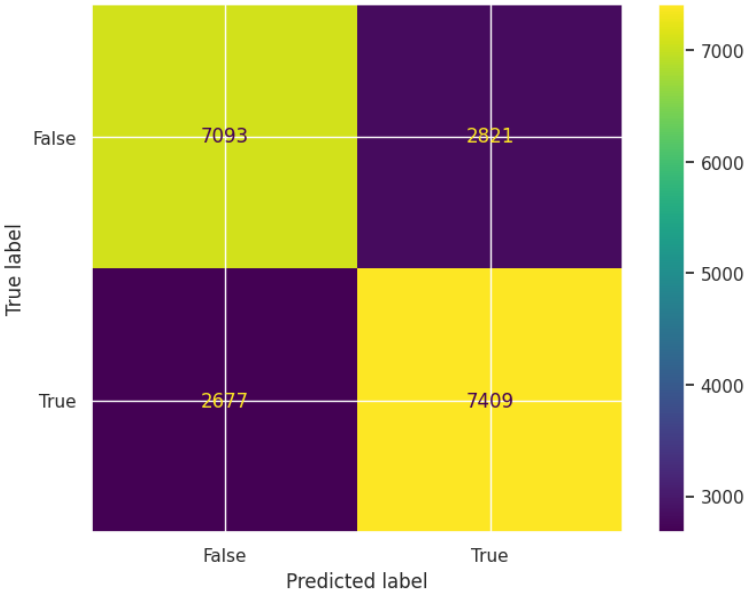




Results : Random Forest

	precision	recall	f1-score	support
Yes	0.725998	0.715453	0.720687	9914.0000
No	0.724242	0.734583	0.729376	10086.0000
accuracy	0.725100	0.725100	0.725100	0.7251
macro avg	0.725120	0.725018	0.725031	20000.0000
weighted avg	0.725113	0.725100	0.725069	20000.0000

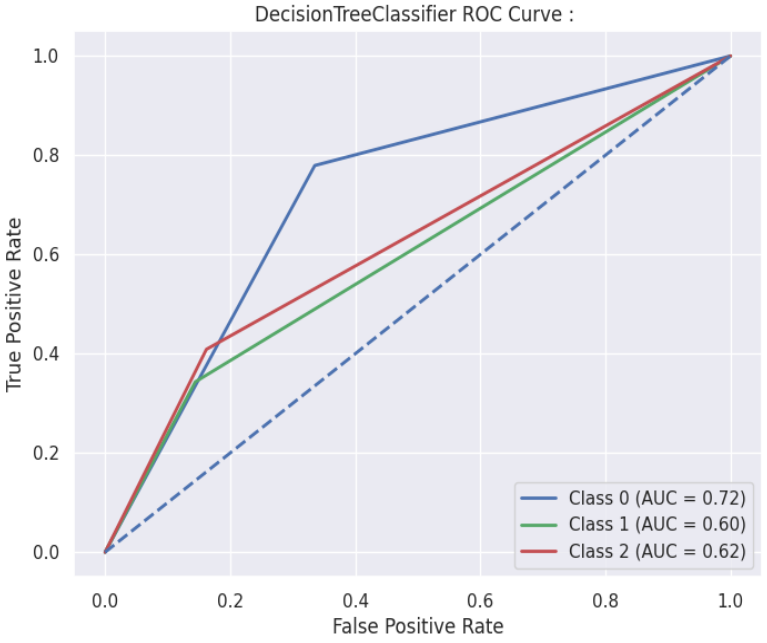




Classification of Smoker :

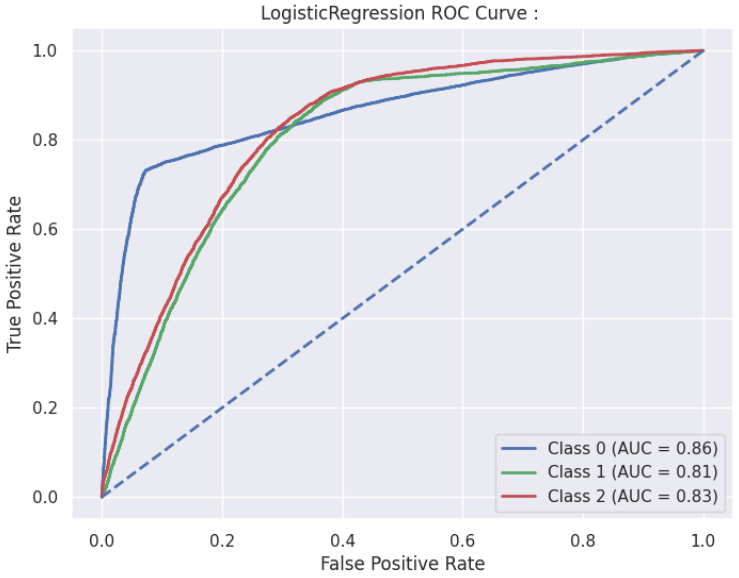
Results : Decision Tree

```
Model: DecisionTreeClassifier
ROC AUC: 0.6482924153497761
Accuracy: 0.6232
Precision: 0.6248888789025235
Recall: 0.6232
F1 Score: 0.6240341024898275
```



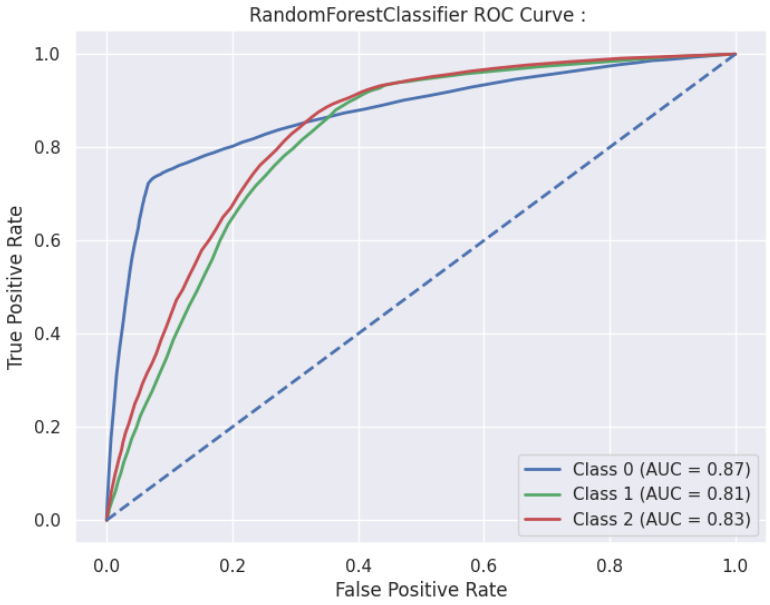
Results : Logistic Regression

```
Model: LogisticRegression
ROC AUC: 0.8322260817934648
Accuracy: 0.68245
Precision: 0.6851363669619294
Recall: 0.68245
F1 Score: 0.681551873923252
```



Results : Random Forest

```
Model: RandomForestClassifier
ROC AUC: 0.8383870916660197
Accuracy: 0.6928
Precision: 0.6878118851711239
Recall: 0.6928
F1 Score: 0.689322913005959
```



References

1. Lamprou, Sokrates, A study in alcohol: A comparison of data mining methods for identifying binge drinking risk factors in university students. (link)
2. Radhika P R, Rakhi A S Nair, Veena G, A Comparative Study of Lungs Cancer Detection using Machine Learning Algorithms.(link)