SDSU | San Diego State University

SAKSHI PRADHAN HIMALI SHEWALE VEDANT GHAVATE

CLASSIFYING DRINKERS AND SMOKERS BY BODY SIGNALS

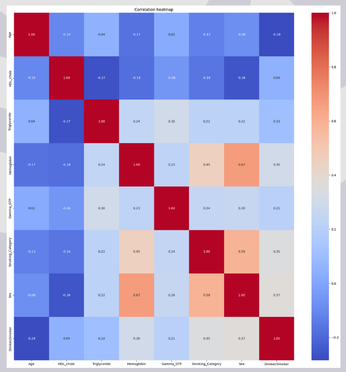
Introduction

Through this dataset we seek to find Drinking and smoking habits have that been extensively proven to have severe long-term effects on health, however, these habits are still on the rise. Analyzing the data, which contains demographic information and specific body signals and blood tests, we seek to understand any demographic-specific body signals that could help us identify smokers and drinkers.

Problem Motivation

- Despite established health risks, smoking and drinking habits persist.
- There is a critical need to identify specific physiological signals indicative of these behaviors.
- Diverse dataset (nominal, ordinal, ratio data) to bridge the knowledge gap for targeted interventions and public health strategies.

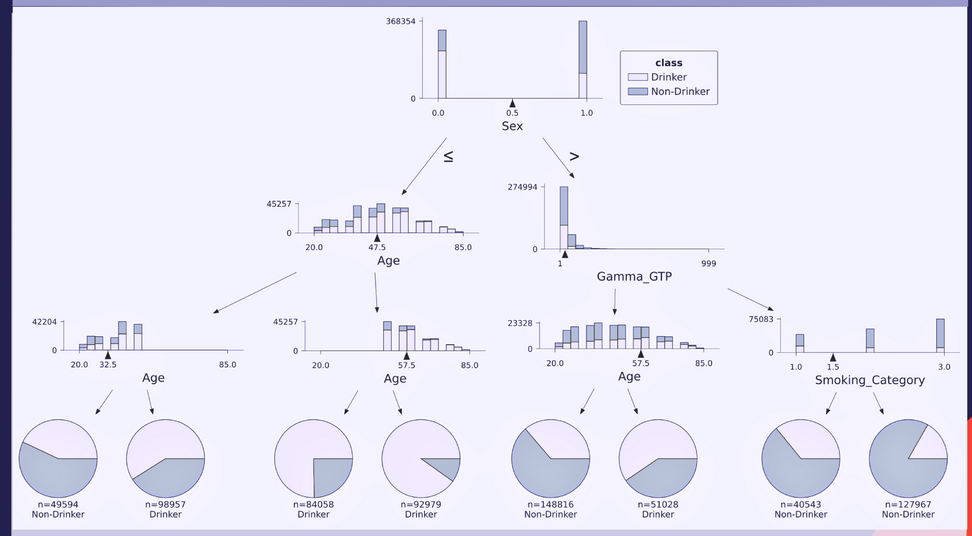
Background



- Multiple features like height weight and hearing are closely related to age however, do not have a significant impact on classifying drinkers and smokers.
- Non-medical factors like age and sex prove to be a strong indicators of establishing an individual as a smoker or drinker.

- The data is obtained from a government-based survey and sample collection drive in Korea
- The dataset contains 24 columns, features like cholesterol levels, urine protein hemoglobin, and so on. Many of the features prove to be insufficient to identify drinkers and smokers, however, Gamma GTP and HDL Cholesterol show a correlation.

Approach



- The diagram above shows the decision tree with Age, Gamma GTP, and Sex as the most important features.
- The Table below shows features and their importance on a fined-tuned XGBoost classifier.

Note: Smoking Category Survey - including this feature in training causes significantly varying accuracy for all classifiers.

Feature	Importance
Sex	0.2786290533
Smoking_Category	0.2114359532
Age	0.1571511594
Gamma_GTP	0.09824124812
Height	0.09013679029
Hemoglobin	0.07338954002
Sight_Hearing_Age	0.05621313467
HDL_chole	0.03277808121
SGOT_ALT	0.0008673337827
Weight	0.0007307413619
Triglyceride	0.0002893329898
Waistline	7.10E-05
SBP	6.67E-05

Evaluation & Results

Model	*Accuracy*	*Recall*
Logistic Regression	0.714	0.705
Naive Bayes	0.693	0.651
Decision Tree	0.724	0.769
Random Forest	0.718	0.733
AdaBoost	0.726	0.720
QDA	0.687	0.614
XGBoost	0.736	0.739
XGBoost-Tuned	0.779	0.728

Decision Tree had the highest recall rate while a fine-tuned XGBoost had the highest accuracy.

Conclusion

- From the EDA and Feature Importance Analysis, Body Signals can provide an accuracy of 71% to identify drinkers and smokers, and the most significant metrics apart from Age, Sex and Smoking survey are Gamma GTP and HDL Cholesterol levels. Thus, additional metrics will be required to identify smokers and drinkers apart from body signals.
- It is established that out of all the classification models, the Decision Tree is the most interpretable, a fine-tuned XGBoost model fits that data effectively.
- For the Decision Tree, pruning yielded a decrease in accuracy rate while others using more computation-expensive hyperparameters resulted in no significant increase in accuracy.

References

- https://www.kaggle.com/code/sidde95/roc-auc-of-82-14-with-xgb-classification#Modelling-with-XGBoost-Classifier
- https://www.data.go.kr/data/15007122/fileData.do