THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Final Term Project Report

Alcoholic Classification Using Body Signals

GEORGE WASHINGTON UNIVERSITY
AUTHOR: Sai Prathyusha Kanisetti
INSTRUCTOR: Prof. David W. Trott
05/08/2024
Machine Learning
CSCI 6364

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Final Term Project Report

Table of contents

Introduction

Description of the Dataset

Pre-Processing

Outlier Detection and Removal

Exploratory Data Analysis

Model Building

Conclusion

Final Term Project Report

Introduction

Alcohol consumption is a widespread concern globally, impacting public health and societal welfare significantly. The World Health Organization (WHO) highlights the profound consequences of alcohol misuse, ranging from mental and behavioral disorders to various physical ailments. Shockingly, about 13.5% of deaths among individuals aged 20–39 years are attributed to alcohol. Understanding the health patterns of individuals struggling with alcohol addiction is vital. By comprehending how alcohol affects health, frontline workers can devise more effective treatment plans tailored to individuals' needs. Analyzing these patterns is crucial, given the extensive range of health problems associated with alcohol misuse, empowering healthcare providers to address the challenges posed by alcohol-related issues more effectively.

Description of the Dataset

The dataset sourced from Kaggle comprises a total of 991346 observations with 24 features. Among these features, 22 are numerical, and 2 are categorical. This rich dataset offers a broad spectrum of variables significant in the medical and health analytics context. Key numerical features include vital health metrics such as, Hemoglobin, Glucose levels, height, weight etc.,

Source	**Kaggle**
Features	**24**
Numerical Features	**22**
Categorical Features	**2**
Observations	**991346**

Fig 1: Dataset Description

Pre-Processing:

Upon thorough examination of the dataset, 26 duplicate entries were identified and subsequently removed. This meticulous curation process ensures the integrity and cleanliness of the data for subsequent analysis and model construction. However, given the relatively small proportion of duplicates, their potential influence on analytical outcomes is deemed negligible.

Feature Engineering:

Analysis of the kernel density plot revealed substantial overlap between the data distributions of left and right eyesight measurements. This observation suggests a high correlation between the eyesight values of the left and right eyes. Furthermore, a notable peak at 9.9 on the x-axis was detected, indicating instances of extreme impairment or blindness. Leveraging these insights, a

new feature was engineered within the dataset to specifically capture cases of blindness, thereby enriching the data, and enhancing its utility for subsequent analyses and model development.

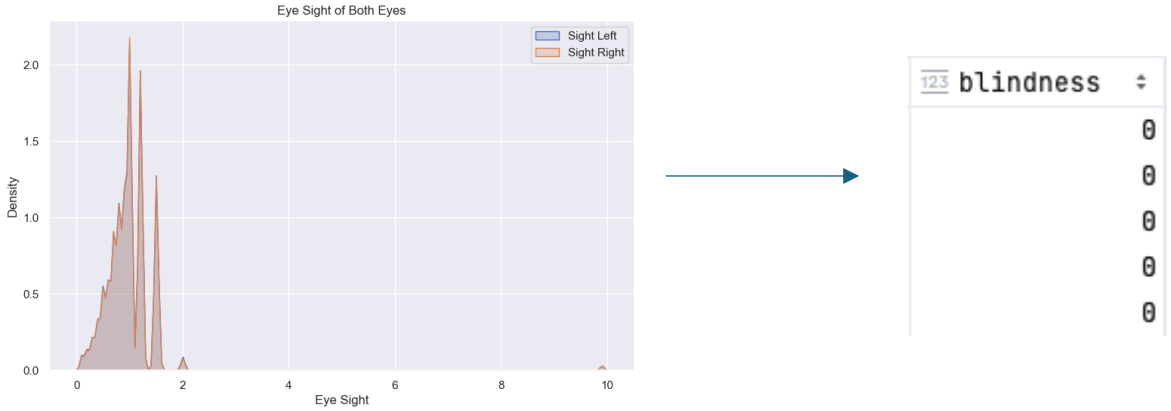

Fig: Eyesight of Both Eyes

Outlier Detection and Removal:

During the comprehensive analysis of the dataset, several outliers were identified, indicative of potential data entry errors. Notably, instances such as a recorded waistline measurement of 999.0 and systolic blood pressure readings exceeding the typical threshold of 170 signify anomalies warranting scrutiny and removal.

Additionally, irregularities were observed in cholesterol levels, where exceptionally high values were recorded. For instance, the maximum HDL cholesterol value of 8110 and LDL cholesterol value of 5119 far exceed the established upper limits, strongly suggesting data inaccuracies.

Following the removal of outliers from the dataset, the total number of observations decreased from 991,346 to 906,676. This substantial reduction underscores the prevalence and impact of outliers on the data's reliability and precision, emphasizing the necessity of their identification and elimination in ensuring the integrity of subsequent analyses and interpretations.

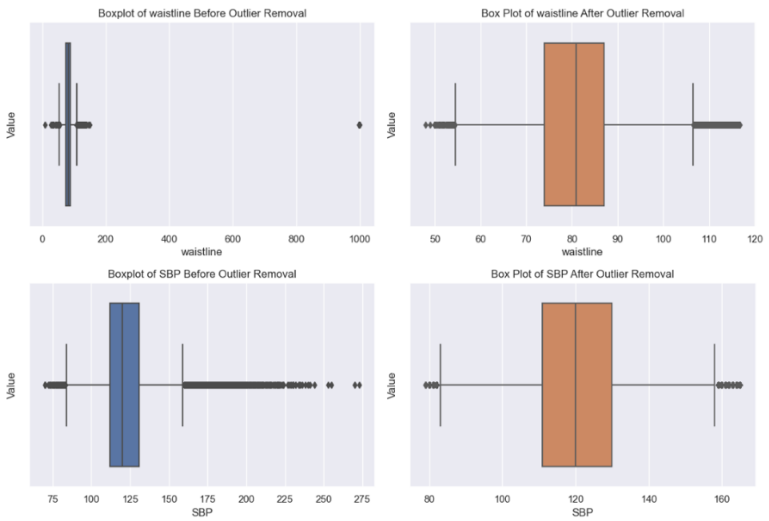

Fig 3: Outlier removal of waistline and SBP

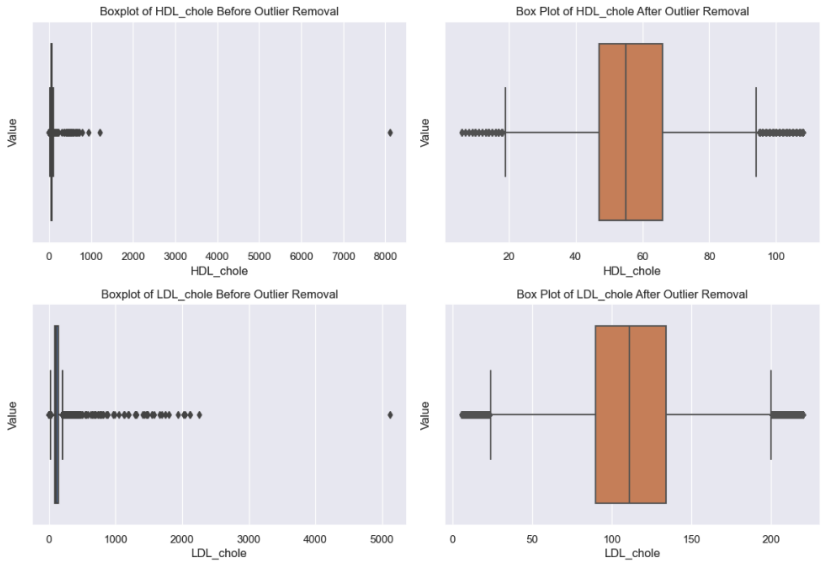

Fig 4: Outlier removal of High Density and Low-Density Cholesterol

Exploratory Data Analysis:

Exploratory Data Analysis, or EDA, is an indispensable phase in the data analysis pipeline. It entails delving into the dataset's main attributes, often through visual aids, to unearth underlying patterns, identify anomalies, test hypotheses, and validate assumptions using summary statistics and graphical representations.

Plot 1: Target Class Distribution:

Analysis of the bar plot reveals a roughly equal distribution between drinkers and non-drinkers within the dataset, indicating no significant skew towards either group.

Fig 5: Distributed of Target Class

Plot 2: Distribution of Drinkers/Non-Drinkers by Gender:

A tabulated summary showcases noticeable gender disparities among drinkers and non-drinkers. Non-drinkers exhibit a higher female count, while drinkers predominantly comprise males. This gender-based discrepancy suggests potential differences in drinking behaviors.

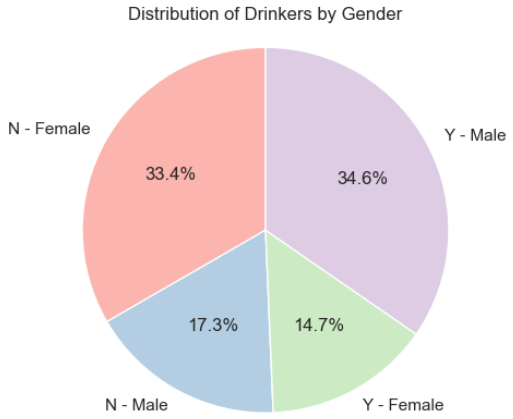

Fig 6: Pie Chart Distribution of Drinkers by Gender

Drinker (Yes/No)	Gender	Count
No	Female	302439
No	Male	156971
Yes	Female	133160
Yes	Male	314106

Table 1: Gender Count by Drinker/ Not

SMART Questions:

Exploratory Data Analysis, commonly referred to as EDA, is a crucial step in the data analysis process. It involves investigating and summarizing the main characteristics of a dataset, often using visual methods. The primary goal of EDA is to explore data to uncover underlying patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.

Final Term Project Report

Smart Question 1: Distribution of Drinker and Non-Drinkers by Age:

Exploring the data unveils a trend where younger and middle-aged individuals exhibit higher alcohol consumption rates compared to older demographics, prompting further investigation into the influence of social factors or generational disparities on drinking habits.

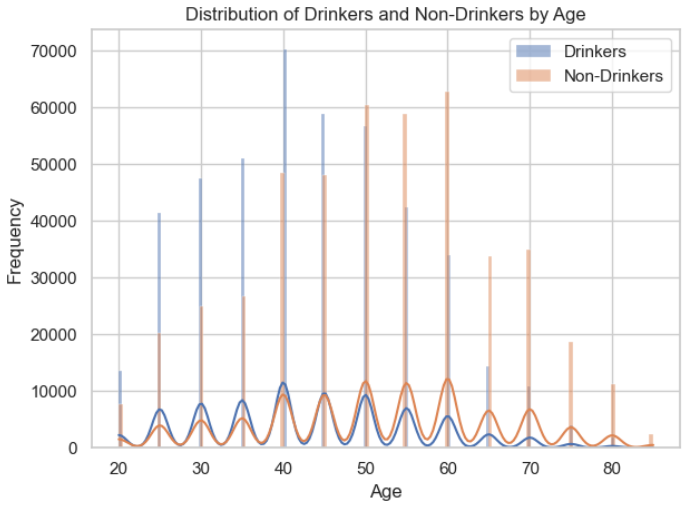

Fig 7: Distribution of Drinkers and Non-Drinkers

Smart Question 2: Does every individual who drinks follow the smoking trend?

A nuanced analysis reveals intriguing patterns between smoking and drinking behaviors, indicating a strong association between the two. Non-smokers are more inclined towards alcohol consumption, while smoking status correlates with drinking habits, underscoring the interconnectedness of these behaviors.

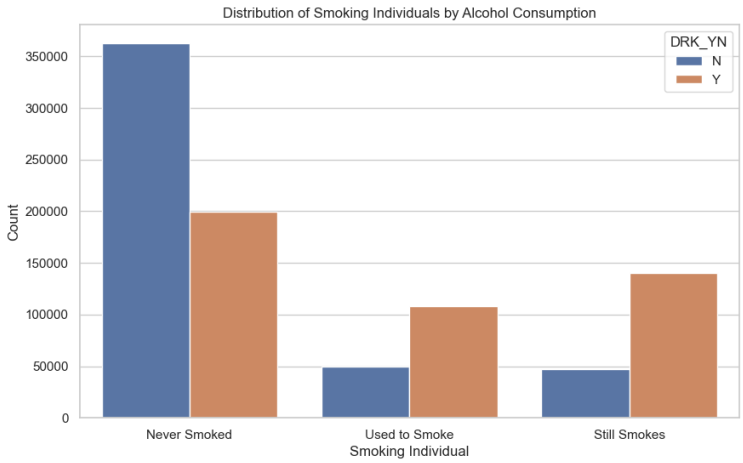

Fig 8: Distribution of Individuals by Alcohol Consumption.

Smart Question 3: Is there any significant impact of alcohol on eyesight?

Initial visual inspection suggests potential differences in eyesight between drinkers and non-drinkers. Subsequent statistical tests, specifically the Mann-Whitney U Test due to non-normal

data distribution, confirm significant disparities in eyesight scores between the two groups, warranting further investigation into the influence of alcohol consumption on eyesight.

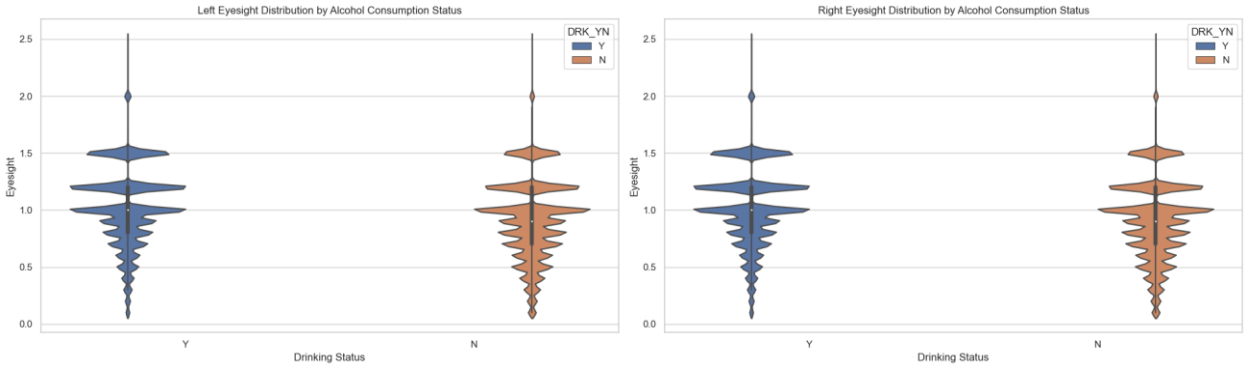

Fig 9: Left and Right Eyesight Distribution by Alcohol Consumption Status

Normality Check:

Left Eye P-Value – 0.0

Right Eye P-Value – 0.0

Mann Whitney U Test:

Left Eye P-Value – 0.0

Right Eye P-Value – 0.0

Null Hypothesis: There is difference.

Alternate

After conducting the Mann Whitney U Test, the result of P-Value < alpha value 0.05 which reject null hypothesis. There is difference in eyesight between individuals who consume/ don't consume alcohol.

Smart Question 4: Does drinking alcohol regularly affect the liver?

Elevated gamma-GTP levels among alcohol consumers hint at potential liver damage, necessitating in-depth exploration into alcohol's hepatotoxic effects. This underscores the imperative of investigating alcohol's impact on liver function for comprehensive health assessment.

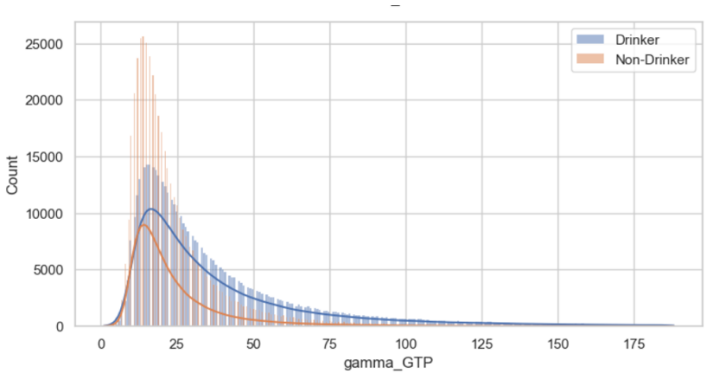

Fig 10: Gamma_GTP Distribution of Drinker and Non-Drinker

Final Term Project Report

Model Building:

Label encoding:

In the project, label encoding was employed to transform the target variable representing "drinker" and "non-drinker" categories into numerical values. Specifically, "drinker" was encoded as 1, while "non-drinker" was encoded as 0. This systematic transformation facilitated the integration of the categorical target variable into machine learning algorithms, ensuring compatibility with models that require numerical inputs.

Feature Importance:

Two distinct methods were employed to extract the importance of features within the dataset: Random Forest Classifier and Recursive Feature Elimination.

1. Random Forest Classifier: This method evaluates feature importance by assessing how random combinations of features contribute to predicting the target variable. By analyzing the impact of each feature on prediction accuracy, the Random Forest Classifier identifies those with significant predictive power.
2. Recursive Feature Elimination: In contrast, Recursive Feature Elimination systematically eliminates features to discern their individual importance in predicting the target variable. By iteratively removing features and evaluating the impact on model performance, Recursive Feature Elimination identifies the most influential features within the dataset.

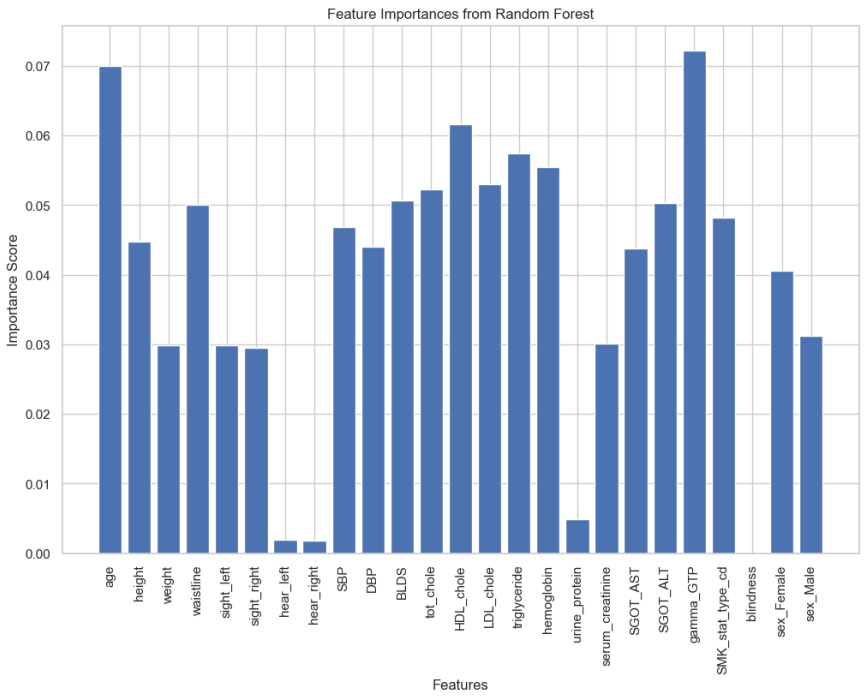

Fig 11: Feature Importance by Random Forest Classifier

Final Term Project Report

Following the application of both methods, two distinct sets of important features were delineated: Set 1: Comprising features identified as important by both the Random Forest Classifier and Recursive Feature Elimination. These features demonstrate consistent significance across multiple analytical approaches.

SET 1

1. gamma_GTP
2. HDL_chole
3. age
4. SMK_stat_type_cd
5. serum_creatinine
6. sex_Female
7. SGOT_ALT
8. SBP
9. weight
10. sex_Male
11. height
12. SGOT_AST

Set 2: Encompassing features deemed important solely by the Random Forest Classifier. This set includes features that exhibit notable predictive power according to the Random Forest Classifier but may not have been identified as crucial by Recursive Feature Elimination alone.
By delineating these distinct sets of important features, the analysis aims to provide comprehensive insights into the factors driving predictive performance within the dataset.

SET 2

1. age
2. height
3. weight
4. sight_left
5. sight_right
6. SBP
7. DBP
8. BLDS
9. tot_chole
10. HDL_chole
11. triglyceride
12. hemoglobin
13. serum_creatinine
14. SGOT_AST

15. SGOT_ALT
16. gamma_GTP
17. SMK_stat_type_cd
18. sex_Female
19. sex_Male

Models employed:

1. Random Forest Classifier.
2. Gradient Boosting.
3. XGBoost.

Random Forest Classifier:

The Random Forest Classifier stands as a stalwart in machine learning, adeptly tackling both classification and regression tasks. Its efficacy stems from its construction of numerous decision trees during training, culminating in an ensemble model that amalgamates the majority vote of individual trees for classification or the average prediction for regression. A notable attribute of this algorithm lies in its resilience, manifesting in its adept handling of outliers and noisy data, thus fortifying its applicability across diverse datasets.

SET 1

Accuracy	0.725
Precision	0.721
Recall	0.719
F1 Score	0.720
ROC AUC	0.725

SET 2

Accuracy	0.728
Precision	0.724
Recall	0.724
F1 Score	0.724
ROC AUC	0.728

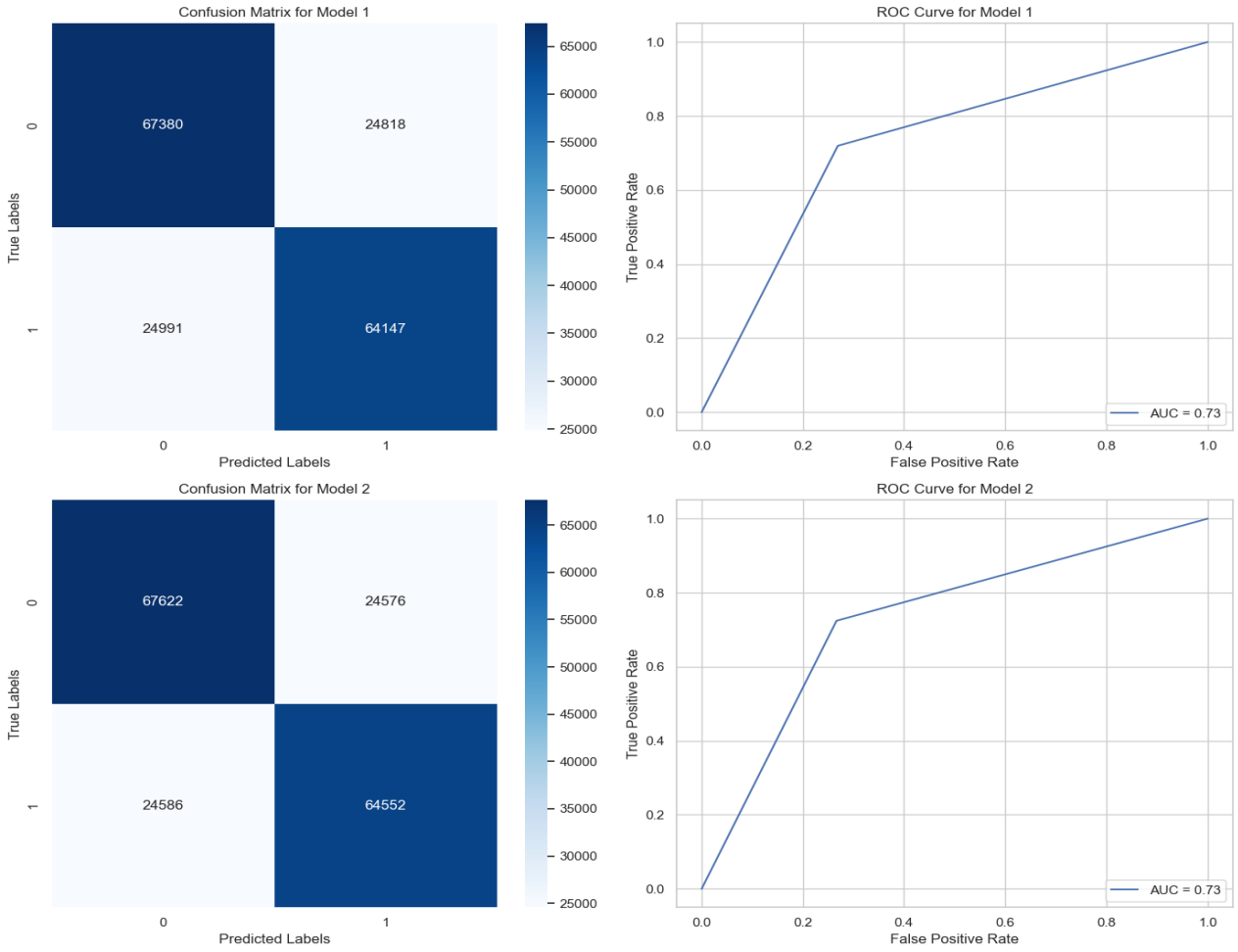

Fig 12: Confusion Matrix and AUC – ROC Curves

Gradient Boosting:

Gradient Boosting emerges as a formidable machine learning technique, leveraging the synergy of multiple weak learning models to forge a robust predictive framework. Nestled within the ensemble learning paradigm, Gradient Boosting orchestrates the training of a suite of models to address the same task, harmoniously amalgamating their insights to yield enhanced predictive prowess. Central to its methodology is the iterative refinement of predictions through an ensemble of decision trees, meticulously crafted in a stage-wise progression, thereby affording a nuanced understanding of complex data patterns.

Final Term Project Report

SET 1

Accuracy	0.729
Precision	0.718
Recall	0.740
F1 Score	0.729
ROC AUC	0.730

SET 2

Accuracy	0.730
Precision	0.718
Recall	0.741
F1 Score	0.749
ROC AUC	0.730

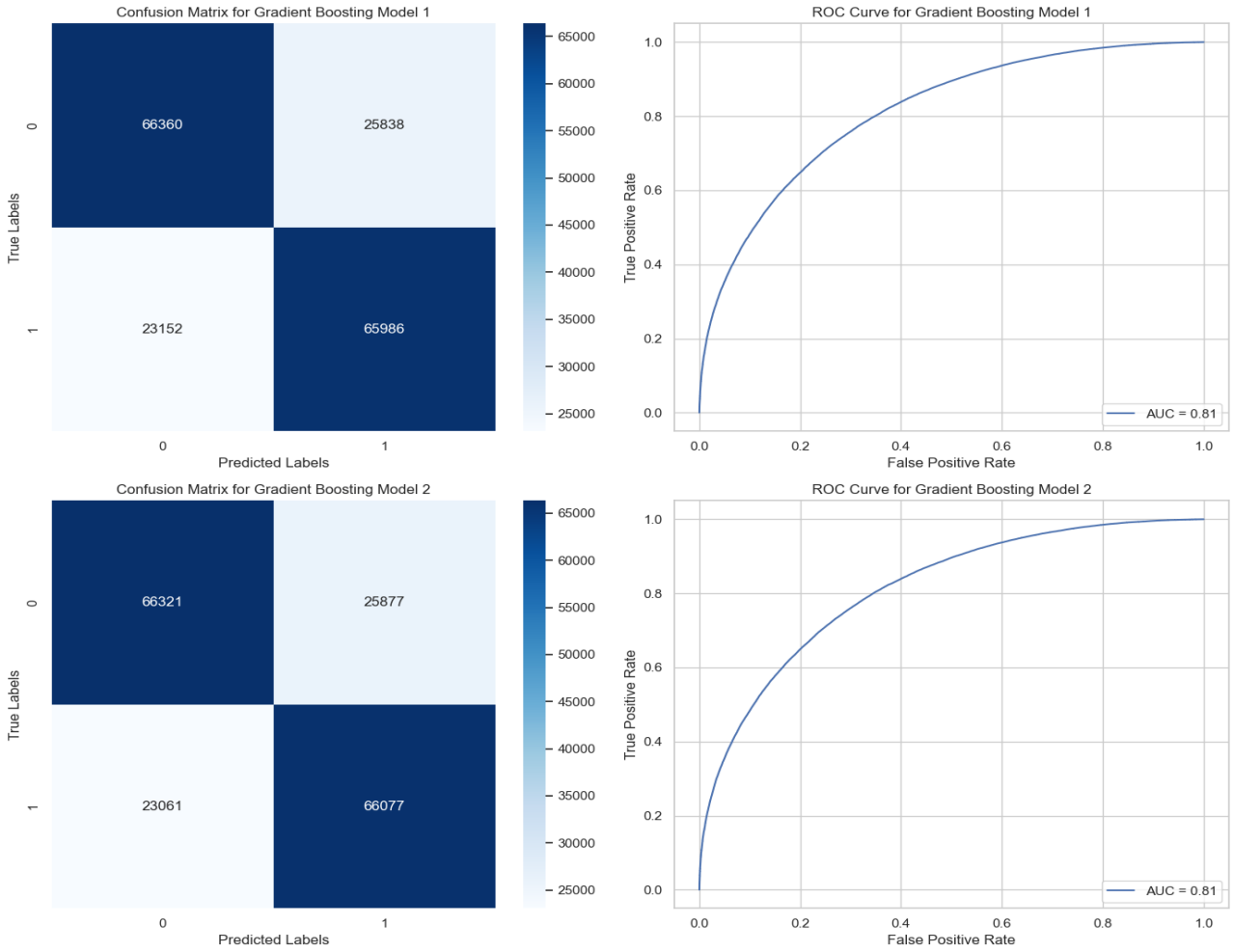

Fig 13: Confusion Matrix and AUC – ROC Curves

Final Term Project Report

XGBoost:

XGBoost, colloquially known as extreme Gradient Boosting, stands at the vanguard of machine learning libraries, redefining the contours of gradient boosting methodologies. Renowned for its unparalleled performance and efficiency, XGBoost elevates conventional gradient boosting through its adept parallel computing capabilities, expediting training processes manifold. Furthermore, its arsenal boasts an array of regularization techniques, strategically deployed to forestall overfitting and bolster model generalization. This amalgamation of speed, accuracy, and robustness renders XGBoost a preeminent choice across diverse domains, ensuring superior predictive performance and scalability.

SET 1

Accuracy	0.732
Precision	0.726
Recall	0.731
F1 Score	0.729
ROC AUC	0.732

SET 2

Accuracy	0.734
Precision	0.728
Recall	0.732
F1 Score	0.730
ROC AUC	0.734

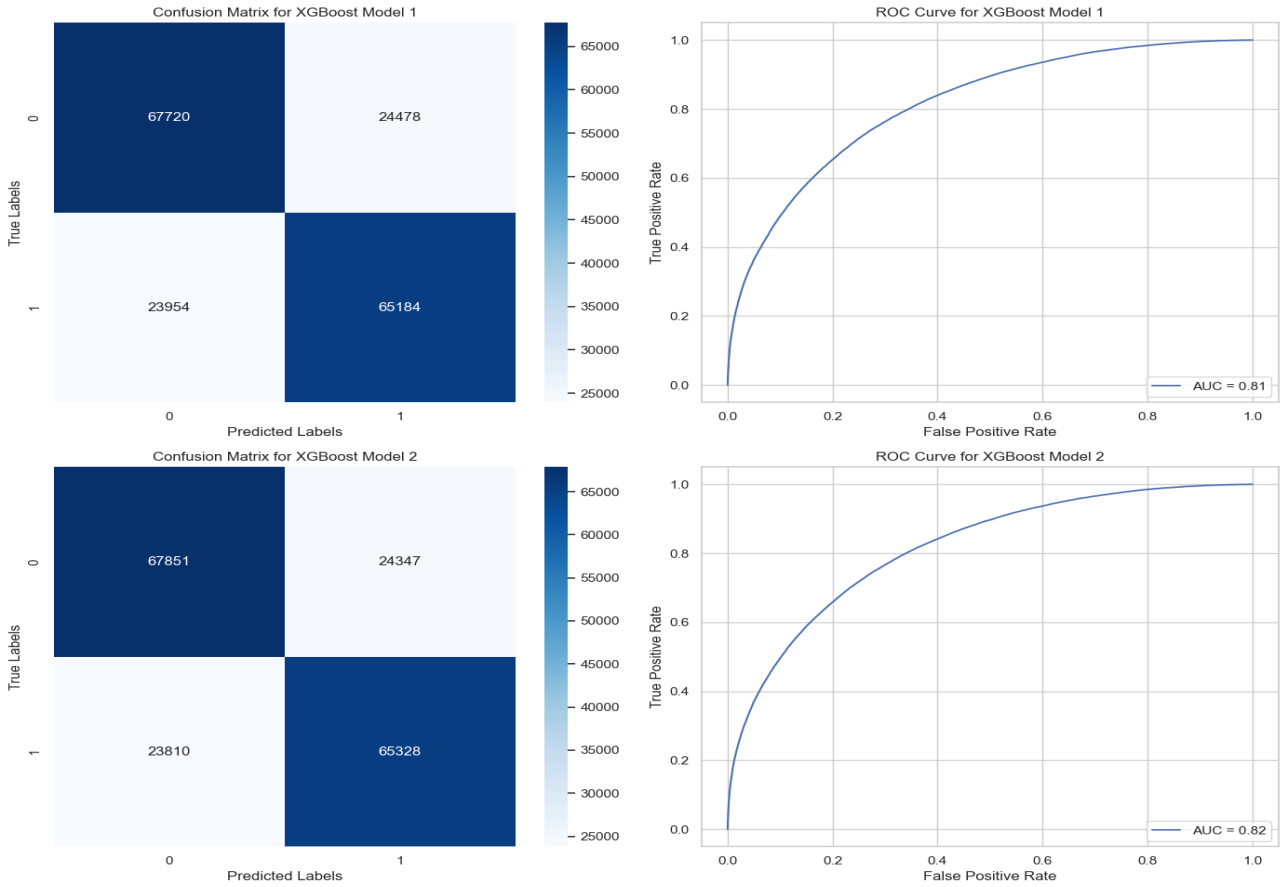

Fig 14: Confusion Matrix and AUC – ROC Curves

Upon examining the confusion matrices, we observed that all three models exhibited similar performance. XGBoost is slightly better at predicting if someone isn't drinking.

While there were slight variations in the number of true positives, true negatives, false positives, and false negatives among the models, overall, they demonstrated comparable classification accuracy. These findings suggest that all three models are effective in distinguishing between the Drinking behaviors.

K-Fold Cross Validation:

The cross-validation analysis conducted reveals a consistent performance trajectory across various subsets of the dataset, elucidating scores ranging from 0.7337 to 0.7363. This uniformity underscores the model's stable and dependable accuracy when extrapolated to novel data instances. Moreover, the mean cross-validation score, hovering at approximately 0.7348, signifies a robust and reliable predictive capability inherent in the model. Notably, the marginal standard deviation of the cross-validation scores, computed at approximately 0.00094, serves as a testament to the model's steadfast consistency. This minimal variability across different folds substantiates the

model's resilience, indicating a limited variance in its performance when confronted with unseen data from the same underlying statistical distribution.
Such consistency and predictability are pivotal attributes for fostering confidence in the model's applicability within real-world scenarios, where reliable performance is paramount for informed decision-making and actionable insights.

Fold 1	0.733
Fold 2	0.733
Fold 3	0.734
Fold 4	0.735
Fold 5	0.736

Mean CV Score	0.734
Standard Deviation	0.00094

F1 Metric Comparative Analysis:
After analyzing the bar plots comparing the F1 metric across various models in set 1 and set 2, it's evident that Gradient Boosting emerged as the top performer in set 2. Nevertheless, factoring in the time complexity involved in building models, XGBoost stands out as the preferred choice for me.

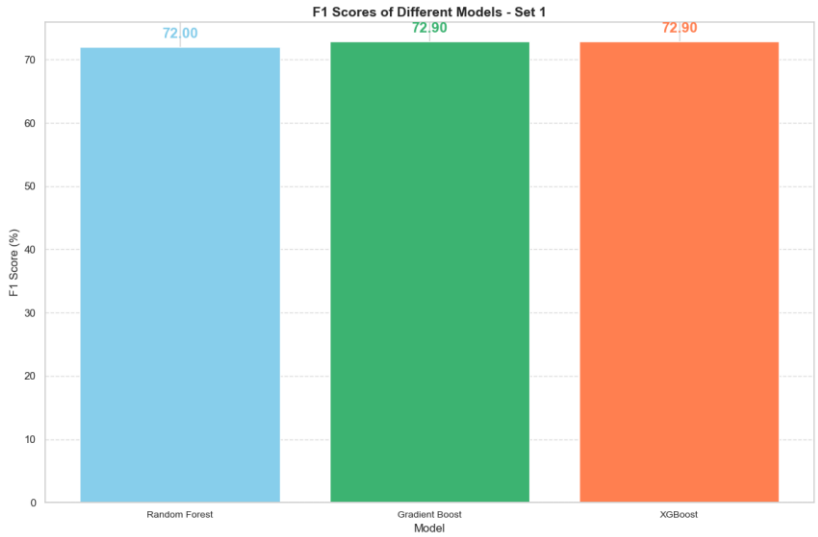

Fig 15: F1 Scores of Different Models – Set 1

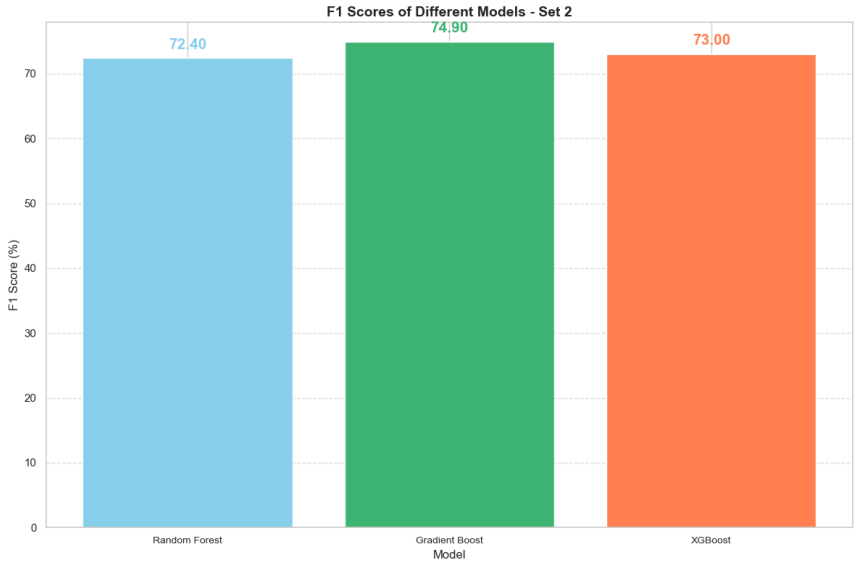

Fig 16: F1 Scores of Different Models – Set 2

Conclusion:

In summary, XGBoost emerges as a formidable contender for structured data tasks, owing to its transparency, efficiency, and adaptability. Its innate capacity to unravel intricate data patterns endows it with a distinct advantage, particularly in domains such as healthcare and finance, where insights into feature importance hold paramount significance.

Furthermore, XGBoost's resource-efficient architecture renders it a cost-effective solution for organizations grappling with datasets of varying scales, ranging from smaller to medium-sized. Beyond its prowess in classification tasks, XGBoost demonstrates remarkable proficiency in handling regression and ranking endeavors, underscoring its versatility across diverse predictive modeling domains.

Crucially, XGBoost's adept handling of noisy or incomplete datasets, coupled with its robustness against outliers and overfitting, reinforces its standing as a stalwart in the machine learning landscape. This resilience positions XGBoost as a reliable ally in navigating the complexities of real-world data scenarios, empowering organizations with the tools necessary to glean actionable insights and drive informed decision-making processes.