# Personalized Intervention for Alcohol Consumption

1st Ömer Faruk Satık
*Artificial Intelligence and Data Engineering*
*Istanbul Technical University*
Istanbul, Turkey
satik21@itu.edu.tr

2nd Abdullah Aydoğan
*Artificial Intelligence and Data Engineering*
*Istanbul Technical University*
Istanbul, Turkey
aydogana21@itu.edu.tr

***Abstract*—This paper is an intermediate report of a project that aims to produce a modern solution to the problem of alcohol consumption. This document focuses on the current status of the project and the findings obtained.**

***Index Terms*—alcohol consumption, data, classification, visualization, evaluation, preprocessing**

## I. Introduction

It is an undeniable fact that alcohol consumption causes many material and spiritual problems. Alcohol directly or indirectly reduces people's quality of life, both physically and psychologically, due to the effects it has on the human body. So much so that it even causes death. As an example, among people who die by suicide, AUD (Alcohol Use Disorder) is the second most common mental disorder and involved in roughly 1 in 4 deaths by suicide [1]. In this context, it is essential to prevent alcohol consumption and the negative effects that occur when consumed. Even though many authorities, especially healthcare professionals, carry out studies in this field, it can be seen with simple observation that societies do not respond to these studies. Therefore, the project team set out with the desire to produce a modern solution to this problem. Today, artificial intelligence not only facilitates work in all areas of life, but can also be very useful in overcoming such unresolved problems. If every individual who consumes alcohol can be personally identified, direct guidance to individuals can prevent them from consuming alcohol. This is a problem that concerns the classification field of data mining. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. The data analysis task classification is where a model or classifier is constructed to predict categorical labels (the class label attributes). Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data [2]. In this project, a classification model will be built using artificial intelligence techniques on a data set obtained from the Korean Ministry of Health, containing approximately one million samples, showing some body signals (cholesterol, blood pressure) of people. And with the model created, it is aimed to determine whether each new person whose data is obtained consumes alcohol.

Identify applicable funding agency here. If none, delete this.

## II. Related Work

### *A. Health Risks and Benefits of Alcohol Consumption (National Library of Medicine)* [3]

This article is a review examining the health risks and benefits of alcohol consumption. Alcohol consumption can affect the health of individuals and the lives of those around them. The article covers extensive research to reduce negative effects, guide healthcare professionals, and help individuals make informed decisions. He states that 44% of the adult population in America drinks alcohol, the majority drinks safely, but the minority who drink heavily has a large impact. It is emphasized that there are 14 million Americans diagnosed with alcoholism, and more than half of Americans who have or have had alcoholism have a close relative. The article documents the effects of alcohol consumption on physical health and highlights that it contributes to injuries, violence, workplace productivity, and deaths. Overall, it evaluates the relationship between alcohol consumption and mortality and emphasizes that such research is important for understanding the effects on individuals and society.

The current article is focused on evaluating the health effects of alcohol consumption in general. Although this article examines the consequences of alcohol consumption at the societal and individual levels, it does not focus on data science applications on a specific data set. On the other hand, our project aims to detect individuals who consume alcohol using a large dataset created by the Korean Ministry of Health. The main difference between the two studies is that the current paper provides an overview and evaluates the general effects of alcohol consumption, whereas our project focuses on a specific goal, namely identifying individuals who consume alcohol. Furthermore, while the current paper presents general results regarding alcohol consumption on the general population, our project offers a more personalized approach by assessing the physical and biological characteristics of individuals. This could contribute to making health policies and intervention strategies more specific and effective. In conclusion, the contribution of our own project to the above article may be that it offers a specific approach to detecting individuals who consume alcohol by focusing on data science applications on a specific data set, thereby further specificizing the general conclusions of the current article.

### B. *Using Machine Learning to Classify Individuals With Alcohol Use Disorder Based on Treatment Seeking Status (National Library of Medicine)* [4]

This article discusses a study that developed a decision tree classifier using cognitive, behavioral and laboratory measures, based on the treatment-seeking status of individuals with alcohol addiction. The primary objective focuses on the set of measures that best predict treatment-seeking among individuals with alcohol dependence. The study included data from 778 alcohol-dependent individuals using 178 clinical measurements. The developed decision tree classifier accurately classified individuals as treatment seekers and non-seekers using 10 important measurements such as drinking habits, depression, drinking-related psychological problems, intelligence, race, body mass index (BMI) and substance abuse. The study evaluated the validity of this classification on both cross-validation and an independent data set. The results show that the decision tree is effective in identifying individuals seeking alcohol addiction treatment and can make predictions with similar accuracy with fewer measurements.

## III. Proposed Work

### A. *Data Understanding and Analysis*

In this section, the focuse is understanding and analyzing the dataset. The dataset consists of various features related to individuals, such as gender, age, height, weight, health metrics, and lifestyle choices like smoking and drinking status. Drinking status is the dependent variable that aimed to classify. The data set contains only 3 categorical features, while each of the other columns consists of numerical values.

- Checked data types and information to understand the structure of the dataset.
- Examined missing values and found that the dataset has no NaN values, facilitating data analysis and model training.
- Removed duplicate rows to prevent misleading analysis and errors in statistical calculations and model building.
- Utilized descriptive statistics to gain insights into the distribution of numerical features.
- Identified potential outliers in certain columns and planned to address them.
- Created visualizations to explore relationships between features and drinking status.
  - Illustrated the relationship between weight and height, with points colored by drinking status. This visualization hinted at a potential correlation between height and drinking rates.

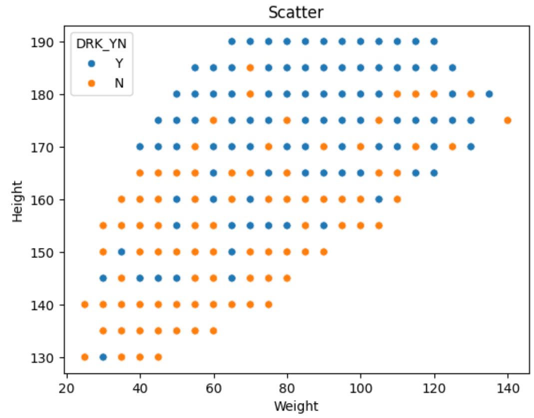


Fig. 1. Scatter Plot of Height and Weight

  - Examined the impact of smoking status on alcohol consumption, providing insights into the relationship between smoking habits and drinking tendencies.

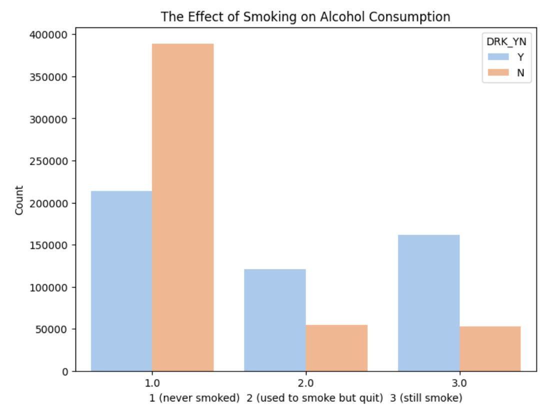


Fig. 2. Count Plot of Smoking

  - Count plots for sight and hearing features to analyze their relationship between alcohol consumption.

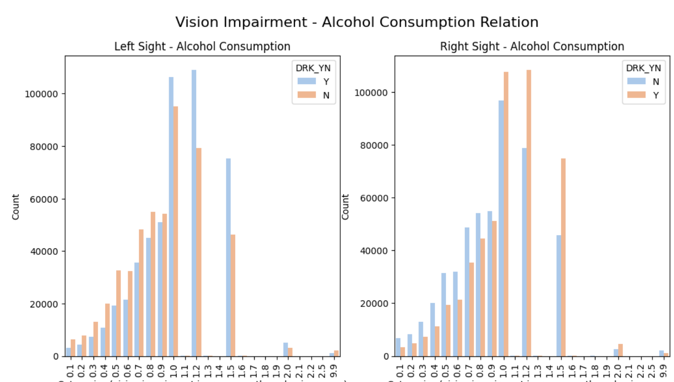


Fig. 3. Count Plot of Vision Impairment

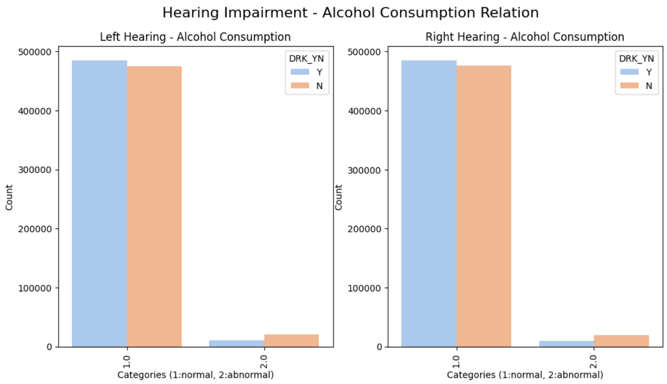


Fig. 4. Count Plot of Hearing Impairment

– Box plots for selected columns to identify and visualize outliers.

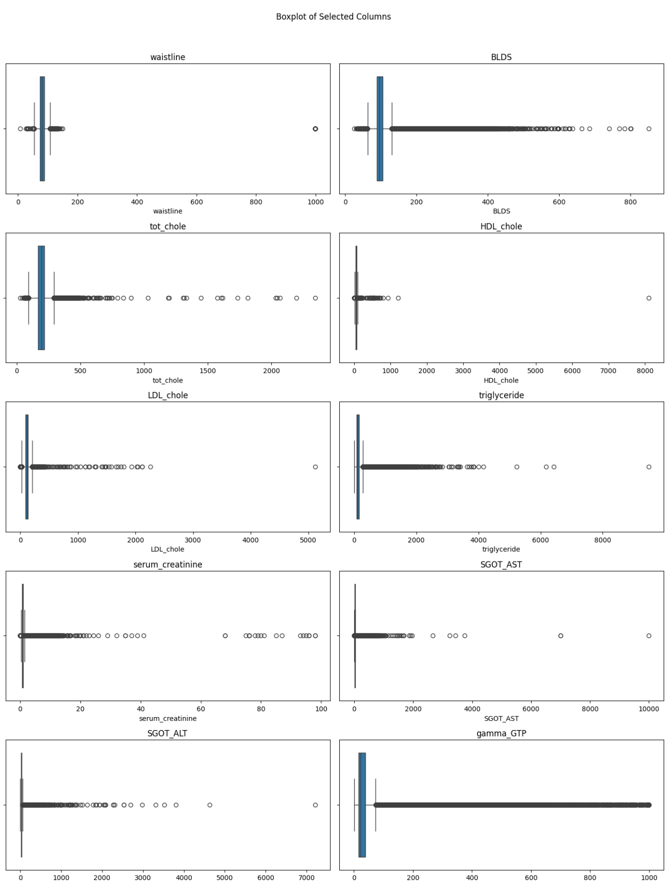


Fig. 5. Box Plots

– Display of the correlation matrix with a heat map to detect positive or negative relationships between numerical features

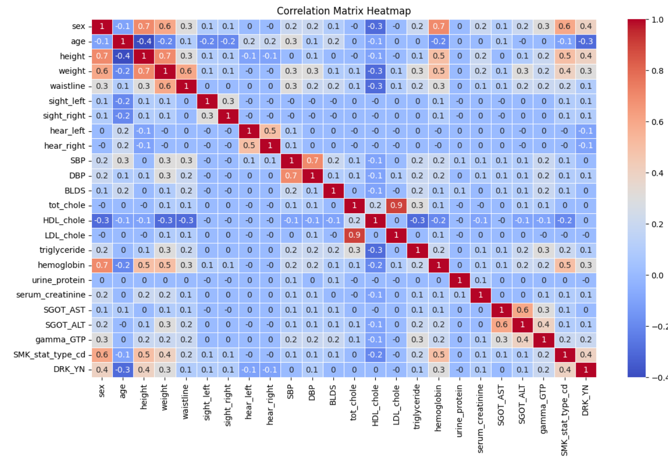


Fig. 6. Box Plots

## *B. Data Preprocessing*

In this phase, data preprocessing steps were performed to enhance the quality of the dataset.

- Detected and removed outliers using the Interquartile Range (IQR) method for selected columns.
- New features were created to extract additional information from existing ones.
  - Body Mass Index (BMI): $\frac{Weight(kg)}{(Height(m))^2}$
  - Visual impairment: Average of defects in left and right eye.
  - Hearing health: Indicates whether there is hearing impairment in any of the ears
  - Blood pressure categories: Divides continuous blood pressure values into 3 categories: Normal, High-Normal, Hypertension
  - AST/ALT (De Ritis) ratio: In the evaluation of elevated liver enzymes; Diagnostic approaches using some parameters such as the rate and increase rate of transaminases have also been described. One of these is the "De Ritis ratio": In 1957, Fernando De Ritis defined the ratio between serum AST and ALT levels as the De Ritis ratio. In this definition, the reflections of the relationship between AST and ALT on diseases are analyzed and some important proportional findings that can guide physicians in terms of etiology are defined [5].
- HDL cholesterol and LDL cholesterol columns were dropped because they are already used in calculation of total cholesterol and the total cholesterol column was included in the data set.
- Categorical features were encoded using one-hot encoding for nominal variables and label encoding for ordinal variables, facilitating the inclusion of these features in machine learning models.

- Two separate data sets were created with applying standard and robust scaling methods on numerical features. The reason for choosing robust scaling is that this method can tolerate outliers. The machine learning model will be trained on these two data sets and the results will be evaluated to decide on the appropriate scaling method.

### C. Model Development and Evaluation

Implemented machine learning classification algorithms to predict drinking status based on the preprocessed dataset and accuracy of the model evaluated. Selected classification algorithms are Decision Tree, Random Forest, Gaussian Naive Bayes, K-Nearest Neighbors and Support Vector Classifier. We trained these models on both standard scaled and robust scaled datasets and then evaluated model performance using classification report and confusion matrix.

- Decision Trees embody a supervised classification approach. The idea came from the ordinary tree structure which is made-up of a root and nodes (the positions where places branches divides), branches and leaves. In a similar manner, a Decision Tree is constructed from nodes which represent circles and the branches are represented by the segments that connect the nodes. A Decision Tree starts from the root, moves downward and generally are drawn from left to right. The node from where the tree starts is called a root node. The node where the chain ends is known as the "leaf" node. Two or more branches can be extended from each internal node i.e. a node that is not leaf node. A node represents a certain characteristic while the branches represent a range of values. These ranges of values act as a partition points for the set of values of the given characteristic. Figure 7 describes the structure of a tree [6].

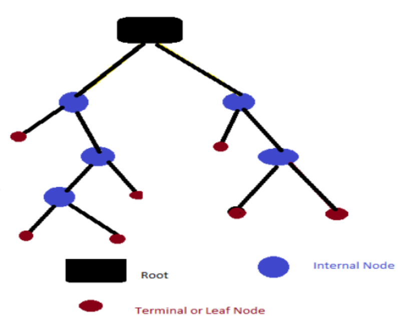


Fig. 7. Decision Tree

  Decision tree classifiers obtain similar or better accuracy when compared with other classification methods [7].
- Random Forest developed by Leo Breiman is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Each tree is grown as described in [8]:
  - By Sampling $N$ randomly, If the number of cases in the training set is $N$ but with replacement, from the original data. This sample will be used as the training set for growing the tree.
  - For $M$ number of input variables, the variable $m$ is selected such that $m << M$ is specified at each node, $m$ variables are selected at random out of the $M$ and the best split on these $m$ is used for splitting the node. During the forest growing, the value of $m$ is held constant.
  - Each tree is grown to the largest possible extent. No pruning is used.

  The advantages of Random Forest are [6]:
  - Overcoming the problem of over fitting
  - In training data, they are less sensitive to outlier data
  - Parameters can be set easily and therefore, eliminates the need for pruning the trees
  - Variable importance and accuracy is generated automatically
- Naive Bayes classifier is an easy and simple probabilistic classifier dependent on applying Bayes theorem. NB considers each attribute variable as independent variable. This classifier can be trained very effectively in supervised learning and can be also utilized in complex real world situations. The major advantage of NB is that it requires little measure of training data which are vital for characterization and necessary for classification [9]. The reason why Gaussian Naive Bayes was chosen is that most columns in the data set consist of continuous values.
- K-Nearest Neighbor (KNN) algorithm is a classification and regression method belonging to the lazy learning category. The basic idea is to determine the class or value of new data points using a similarity measure. KNN, which is widely used especially for classification problems, uses the class of its nearest neighbors to determine the class of a data point. The basic step is to determine the K nearest neighbors of a new data point. This neighborhood is usually calculated using the Euclidean distance or other similar distance measures. The Euclidean distance formula is:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

  Here $x$ and $y$ represent the feature vectors of two points. Then, after the K nearest neighbors are determined, the majority class is used as the prediction class of the algorithm. This forms the basic logic of KNN for classification. The flexibility of KNN does not occur during the training phase of the model; It only keeps the training data in memory. This causes KNN to be called a "lazy" algorithm. The advantages of KNN include simplicity and the ability to obtain effective results. However, in large data sets and high-dimensional feature spaces, the computational cost may increase [10].

- Support Vector Classification (SVC) is a powerful learning algorithm designed specifically for classification problems. The ability to handle two-class and multi-class classification problems is based on working with support vectors and hyperplane principles. The marginal separation principle aims at better classification of future examples by maximizing the margin between classes. The soft margin approach and the C penalty parameter allow error tolerance and the construction of more general models. Various kernel options provide the flexibility to create models suitable for different data structures. The user-friendly interface and wide range of applications make SVC effective. By using it in our own algorithm, we can achieve successful results, especially in high-dimensional and complex data sets, and integrate powerful solutions to classification problems [11].

### D. Future Plan

- Visualizations will be made for the new structure of the data set created after the feature engineering phase.
- Dimensional reduction methods will be studied.
- Removal of outlier data will be optimized
- Sampling will be applied in data preprocessing
- Grid search algorithm will be applied to the models with the highest accuracy value.
- If the desired level of accuracy cannot be achieved, Artificial Neural Networks model will be worked on.

## IV. Experimental Results

Two separate data sets were created by scaling the analyzed and preprocessed data with 2 different methods (Standard, Robust). Each of these data sets was divided into two: train and test. Decision Tree, Random Forest, Gaussian Naive Bayes, K-Nearest Neighbors and Support Vector Classifier models were trained separately with the train sets and the model success was examined on the test sets. When testing the model success, the evaluation metrics, such as precision, recall, and F1-score, were calculated for each model, and confusion matrices were visualized. Support Vector Classifier could not be printed because it was a very costly method.

### A. Preliminary Findings

- Decision Tree

Standard Scaled
---------------
Decision Tree:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| N | 0.65 | 0.65 | 0.65 | 104798 |
| Y | 0.60 | 0.60 | 0.60 | 91549 |
| accuracy | | | 0.63 | 196347 |
| macro avg | 0.62 | 0.62 | 0.62 | 196347 |
| weighted avg | 0.63 | 0.63 | 0.63 | 196347 |

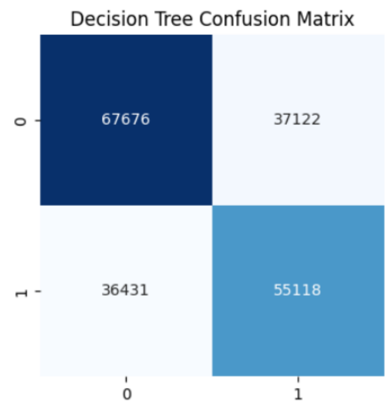


Fig. 8. Decision Tree Evaluation (Standard Scaled))

Robust Scaled
---------------
Decision Tree:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| N | 0.65 | 0.65 | 0.65 | 104798 |
| Y | 0.60 | 0.60 | 0.60 | 91549 |
| accuracy | | | 0.63 | 196347 |
| macro avg | 0.62 | 0.62 | 0.62 | 196347 |
| weighted avg | 0.63 | 0.63 | 0.63 | 196347 |

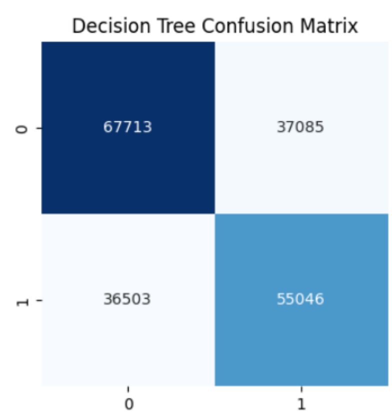


Fig. 9. Decision Tree Evaluation (Robust Scaled)

- Random Forest

Standard Scaled
---------------
Random Forest Classifier:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| N | 0.72 | 0.74 | 0.73 | 104798 |
| Y | 0.70 | 0.68 | 0.69 | 91549 |
| accuracy | | | 0.71 | 196347 |
| macro avg | 0.71 | 0.71 | 0.71 | 196347 |
| weighted avg | 0.71 | 0.71 | 0.71 | 196347 |

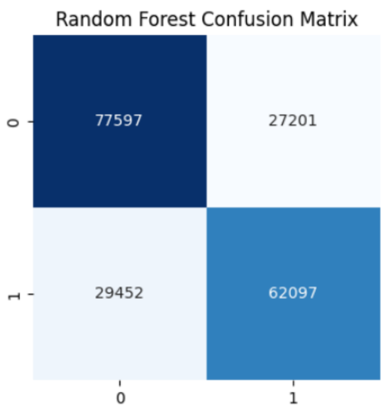


Fig. 10. Random Forest Evaluation (Standard Scaled))

Robust Scaled
---------------
Random Forest Classifier:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| N | 0.72 | 0.74 | 0.73 | 104798 |
| Y | 0.69 | 0.68 | 0.69 | 91549 |
| accuracy | | | 0.71 | 196347 |
| macro avg | 0.71 | 0.71 | 0.71 | 196347 |
| weighted avg | 0.71 | 0.71 | 0.71 | 196347 |

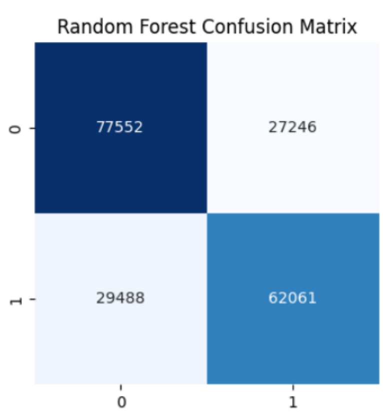


Fig. 11. Random Forest Evaluation (Robust Scaled)

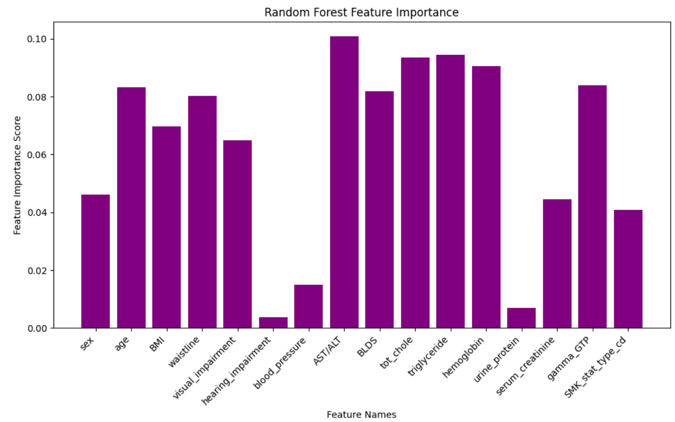


Fig. 12. Random Forest Feature Importances

AST/ALT (De Ritis) ratio which was created in feature engineering step is the most important feature for the random forest. This means we produced a useful feature beside reducing dimension.

- Gaussian Naive Bayes

Standard Scaled
---------------
Gaussian Naive Bayes:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| N | 0.69 | 0.72 | 0.70 | 104798 |
| Y | 0.66 | 0.64 | 0.65 | 91549 |
| accuracy | | | 0.68 | 196347 |
| macro avg | 0.68 | 0.68 | 0.68 | 196347 |
| weighted avg | 0.68 | 0.68 | 0.68 | 196347 |

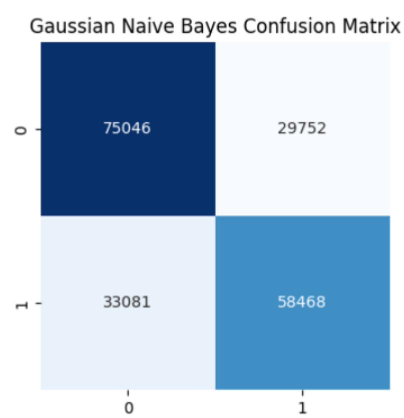


Fig. 13. Gaussian Naive Bayes Evaluation (Standard Scaled)

Robust Scaled
--------------
Gaussian Naive Bayes:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| N | 0.69 | 0.72 | 0.70 | 104798 |
| Y | 0.66 | 0.64 | 0.65 | 91549 |
| accuracy | | | 0.68 | 196347 |
| macro avg | 0.68 | 0.68 | 0.68 | 196347 |
| weighted avg | 0.68 | 0.68 | 0.68 | 196347 |

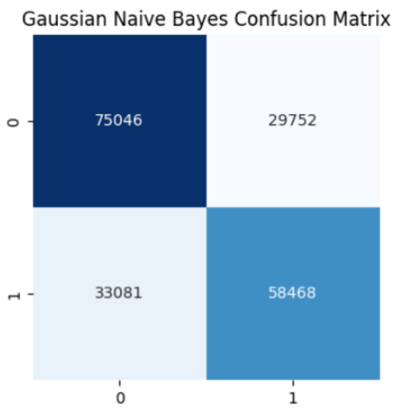


Fig. 14. Gaussian Naive Bayes Evaluation (Robust Scaled)

- K-Nearest Neighbors

Standard Scaled
--------------
k-Nearest Neighbors:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| N | 0.69 | 0.69 | 0.69 | 104798 |
| Y | 0.65 | 0.65 | 0.65 | 91549 |
| accuracy | | | 0.67 | 196347 |
| macro avg | 0.67 | 0.67 | 0.67 | 196347 |
| weighted avg | 0.67 | 0.67 | 0.67 | 196347 |

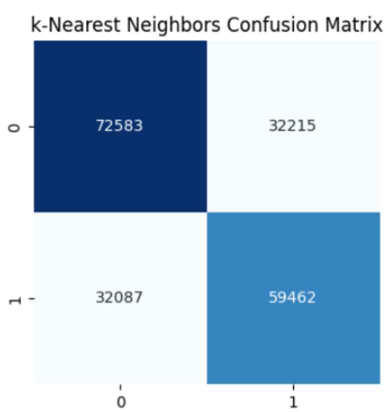


Fig. 15. K-Nearest Neighbors Evaluation (Standard Scaled)

Robust Scaled
--------------
k-Nearest Neighbors:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| N | 0.69 | 0.69 | 0.69 | 104798 |
| Y | 0.65 | 0.65 | 0.65 | 91549 |
| accuracy | | | 0.67 | 196347 |
| macro avg | 0.67 | 0.67 | 0.67 | 196347 |
| weighted avg | 0.67 | 0.67 | 0.67 | 196347 |

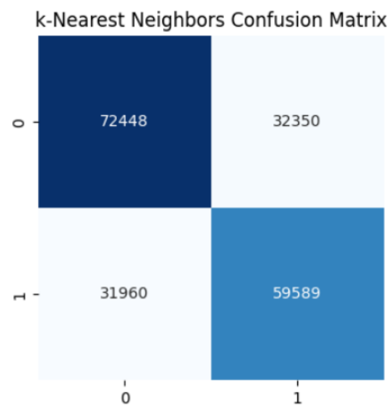


Fig. 16. K-Nearest Neighbors Evaluation (Robust Scaled)

- Random Forest consistently outperforms other models in terms of accuracy, precision, recall, and F1-score for both scaling techniques.
- Gaussian Naive Bayes and k-Nearest Neighbors demonstrate similar performance.
- Decision Tree shows decent performance but is slightly less accurate compared to Random Forest.
- The choice of scaling technique (Standard or Robust) does not significantly impact the model performance in this context.

### B. *Inferences*

Although the accuracy value obtained especially in the Random Forest model is good enough, it is possible to further increase the performance of the model. The following methods can be followed for this:

- Selecting the parameters that provide the highest performance for each model using the Grid Search algorithm.
- Dimension reduction using PCA.
- To find the most appropriate multiplier for IQR in the step of removing outlier data. This item is important because the current code destroys approximately two hundred thousand lines, which means it causes serious data loss.

On the other hand, the Support Vector Classifier model also needs to be made operational. As a solution to this, sampling methods of data mining can be used to reduce costs.

## V. Conclusion

As a result, within the scope of this project, it is aimed to develop a personalized intervention system for alcohol consumption using artificial intelligence techniques. Dataset analysis and preprocessing stages provided important information regarding the relationships between various characteristics and drinking status. Classification models developed include Decision Tree, Random Forest, Gaussian Naive Bayes, K-Nearest Neighbors and Support Vector Classifier, and these models have been evaluated on standard and robust scale datasets.

The Random Forest model has demonstrated its success in predicting alcohol consumption by consistently performing more effectively than other models in terms of accuracy, precision, recall and F1 score. Additionally, the AST/ALT (De Ritis) ratio was determined to be an important factor in the classification process, as highlighted by the Random Forest model.

However, it should be taken into account that the performance of the model can be further improved. Further work may include performing parameter adjustments using Grid Search, integrating the PCA method to reduce dimensionality, and minimizing data loss to reduce outliers. Additionally, applying data mining sampling methods to enable the Support Vector Classifier can improve the overall efficiency of the model.

Despite the promising results, it is important to acknowledge the limitations of this study, the need for more detailed parameter settings, and potential biases in the dataset. However, considering that the project has not been completed yet, it can be predicted that it will not make a valuable contribution to the field of personalized interventions for alcohol consumption.

## References


[1] 1. Berglund M, Ojehagen A. The influence of alcohol drinking and alcohol use disorders on psychiatric disorders and suicidal behavior. Alcohol Clin Exp Res. 1998;22(7 Suppl):333S-345S. PubMed PMID: 9799958.

[2] 2. G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, India, 2013, p. 1, doi: 10.1109/ICCCNT.2013.6726842.

[3] 3. Health risks and benefits of alcohol consumption. (2000). Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism, 24(1), 5–11.

[4] 4. Lee, M. R., Sankar, V., Hammer, A., Kennedy, W. G., Barb, J. J., McQueen, P. G., & Leggio, L. (2019). Using Machine Learning to Classify Individuals With Alcohol Use Disorder Based on Treatment Seeking Status. EClinicalMedicine, 12, 70–78. https://doi.org/10.1016/j.eclinm.2019.05.008

[5] 5. Botros M, Sikaris KA. The de ritis ratio: the test of time. Clin Biochem Rev. 2013 Nov;34(3):117-30. PMID: 24353357; PMCID: PMC3866949.

[6] 6. Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), 273.

[7] 7. Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR), 5(4), 2094-2097.

[8] 8. Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), 274.

[9] 9. Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019, June). Cancer classification using gaussian naive bayes algorithm. In 2019 international engineering conference (IEC) (pp. 165-166). IEEE

[10] 10. Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 40(7), 2038-2048.

[11] 11. Lee, D., & Lee, J. (2007). Domain described support vector classifier for multi-classification problems. Pattern Recognition, 40(1), 41-51.