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. Predicting the Risk of Alcohol Use Disorder Using Machine Learning: A Systematic Literature Review [3]

The article is a literature review that systematically examines the work of many researchers using machine learning (ML) techniques over the past decade focusing on the prediction of alcohol use disorder (AUD). These studies between 2010 and 2021 were identified after the elimination of 3,736 studies obtained from six different academic databases and examined on the basis of technical decision analysis. These studies were analyzed on five basic dimensions such as data set characteristics, data collection methods, pre-processing and sampling techniques, feature types and selection, use of ML algorithms and performance evaluation metrics. The results highlight important findings, such as the fact that publicly available datasets are rarely used, unbalanced class distribution is considered a significant problem, and the support vector machine algorithm is the most widely used algorithm in AUD prediction.

The shortcomings of the article include the accuracy of the surveys used to measure alcohol consumption and its effects on problems and the small number of users in the data. The differences between this article and our own project are that, unlike the article, our own data set is obtained from a reliable organization such as the Korean Ministry of Health, the machine learning algorithms we use are more than those mentioned in the article, and we decided on a tree-based algorithm.

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.

This article takes an important step forward by using machine learning to classify individuals with alcohol addiction based on their treatment-seeking status. However, compared to our project, obvious differences emerge given the limitations and focus of the paper. While this study does not provide a broad perspective on general alcohol addiction by focusing on a specific subgroup, our project aims to identify individuals who consume alcohol using a wide data set. Additionally, our project adopts a more personalized approach by evaluating the physical and biological characteristics of individuals. This allows health policies and intervention strategies to be more targeted and effective. In short, our project uses data science practices like this study, but offers a more specific and customized approach to identifying individuals who consume alcohol.

C. A Deep Learning Algorithm to Predict Hazardous Drinkers and the Severity of Alcohol-Related Problems Using K-NHANES [5]

This article aims to estimate hazardous drinkers and the severity of alcohol-related problems through a large-scale survey, at a time when alcohol-related problems are on the rise. In the study conducted using the datasets of the South Korean National Health and Nutrition Examination Survey, it was determined that deep learning algorithms exhibited higher performance than traditional machine learning algorithms. Energy and carbohydrate intake have been found to have a significant impact on predicting hazardous drinkers.

Like our own project, data was taken from the South Korean National Ministry of Health in this project. A shortcoming in the article is that the number of data is approximately one tenth of our own set. This infers the entire population with a lower degree of accuracy. Our own project may allow health policies to be made more specific and effective because it contains more data.

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.

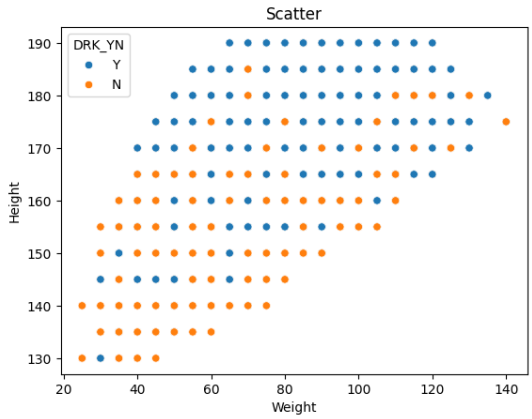

Fig. 1. Scatter Plot of Height and Weight

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

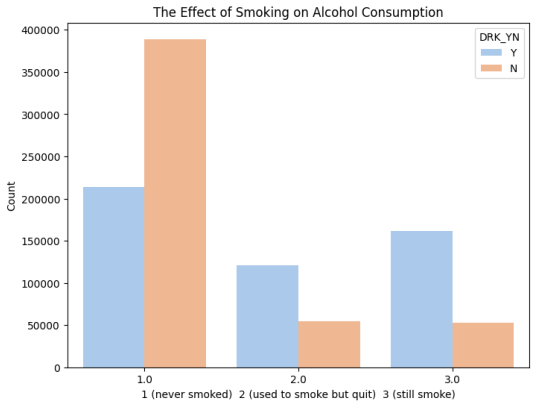

Fig. 2. Count Plot of Smoking

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

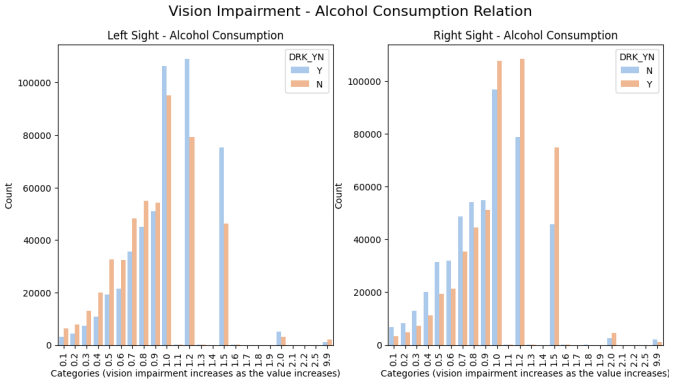

Fig. 3. Count Plot of Vision Impairment

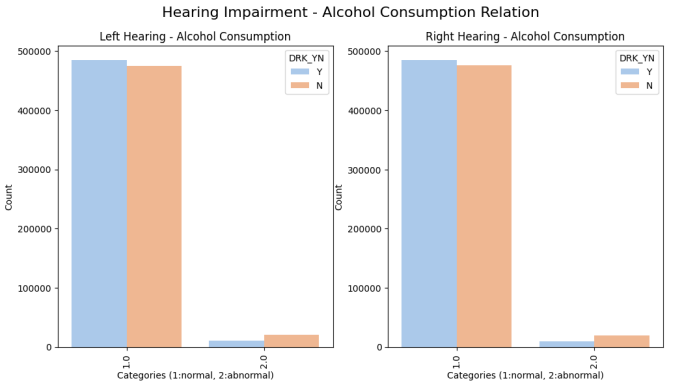

Fig. 4. Count Plot of Hearing Impairment

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

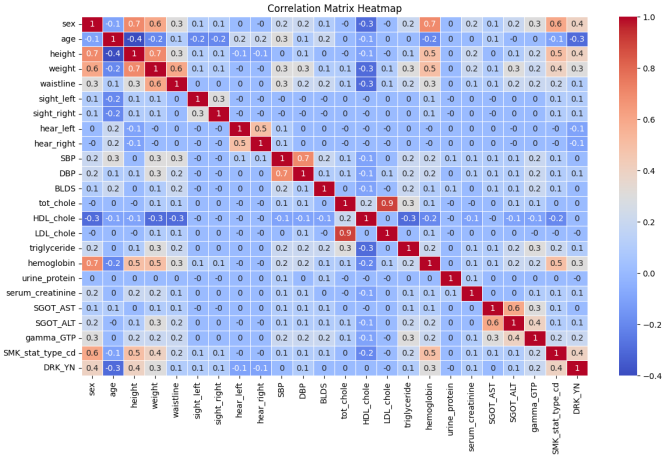

Fig. 5. Correlation Heatmap

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

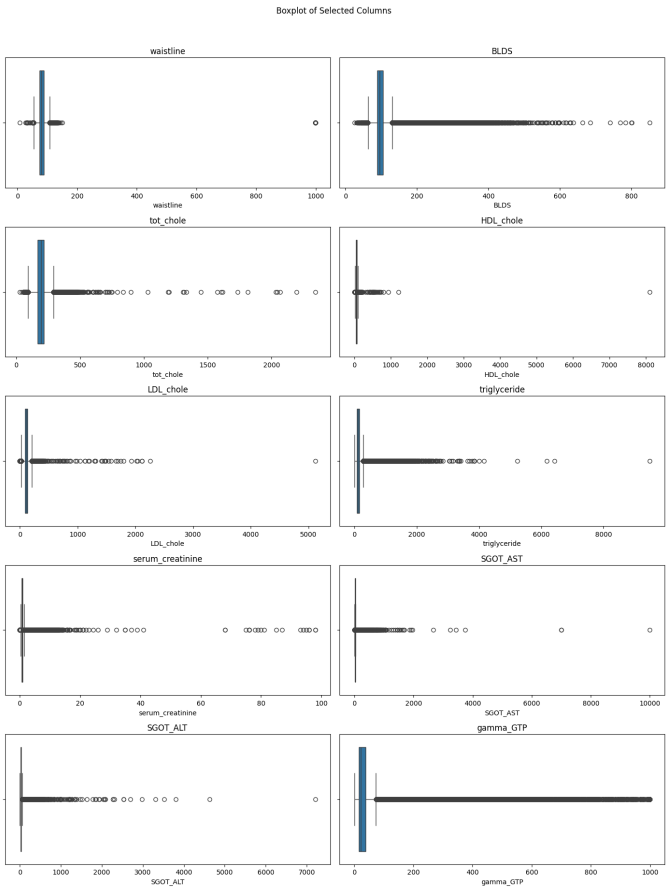

Fig. 6. Box Plots

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 [6].

- 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 [7].

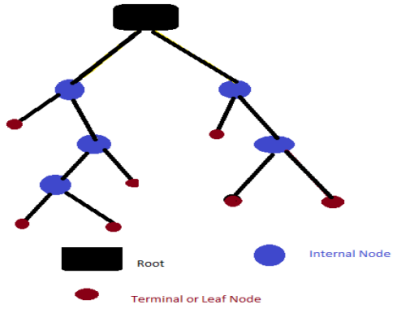

Fig. 7. Decision Tree

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 [9]:
  - 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 [10]. 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 [11].

- 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 [12].

D. Further Development

At this stage, we carried out some work to overcome some of the problems we encountered during the development of the model and to increase the success of the model. Thus, the final model was created.

- **Interquartile Range Optimization:** Interquartile Range Optimization: Detecting outlier data with high accuracy is of great importance for our project because the presence of outlier data negatively affects the model success. While detection of incomplete outlier data affects model success, deleting non-outlier data will lead to data loss. Therefore, an iterative method was followed to determine the most accurate IQR factor and was analyzed on a line plot. When Fig. 8. is examined, there

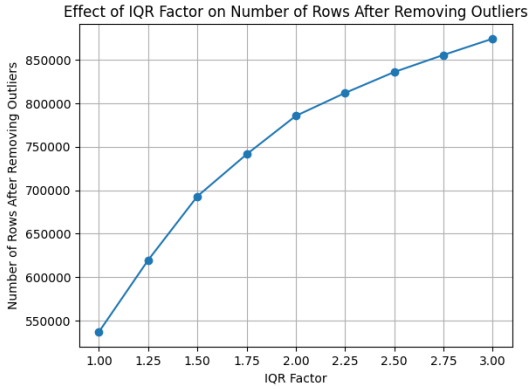

Fig. 8. IQR Factor Optimization

is a break at the point where iqr factor is 2. In addition, while there is a lot of data loss for values between 1 and 2, outlier data is removed less for values greater than 2. For these reasons, we decided that the optimum value is 2.

- **Principal Component Analysis:** Dimensional reduction was deemed necessary to reduce unnecessary variance while preserving the most important information in the data set. Thus, it was aimed to increase the success of the model. However, it turned out that training the model was quite costly after this method was applied. For this reason, this method was not applied.

- **Random Sampling:** As stated under the Preliminary Findings heading of Part IV of this paper, training the Support Vector Classifier model was very costly due to the large data set and even the model could not be run. To solve this problem, a sample containing approximately one-tenth of the rows of the data set was selected by the Random Sampling method. The SVC model was trained with this data.

- **Grid Search CV:** Finally, the Grid Search CV algorithm was used to determine the most suitable parameters of the decided model.

IV. Experimental Results

A. Preliminary Findings

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.

- 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
```

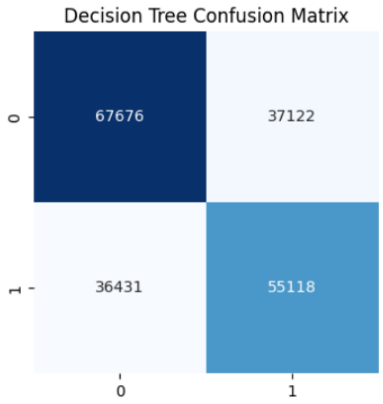

Fig. 9. Decision Tree Evaluation (Standard Scaled))

```
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
```

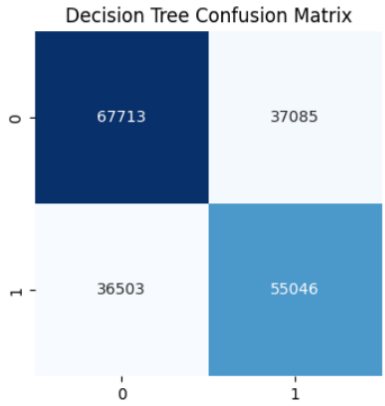

Fig. 10. Decision Tree Evaluation (Robust Scaled)

- 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
```

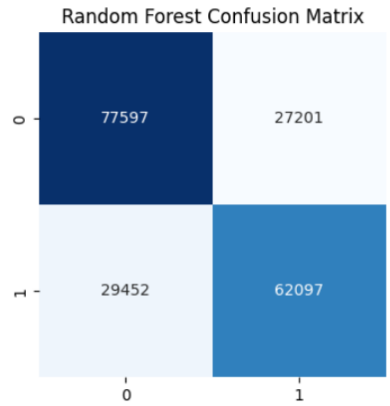

Fig. 11. Random Forest Evaluation (Standard Scaled))

```
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
```

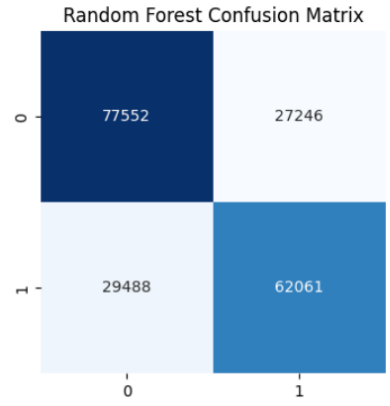

Fig. 12. Random Forest Evaluation (Robust Scaled)

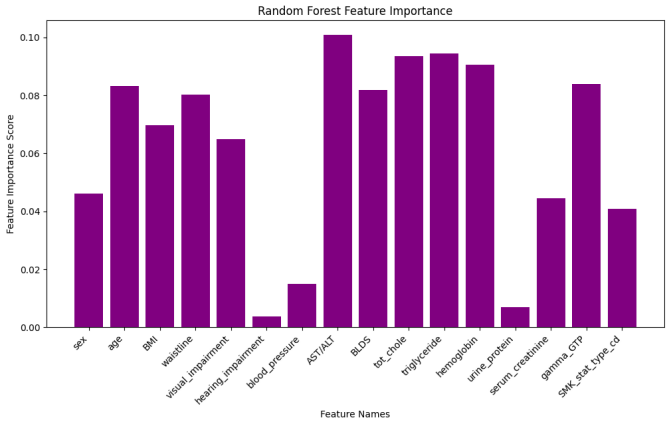

Fig. 13. Random Forest Feature Importances

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
```

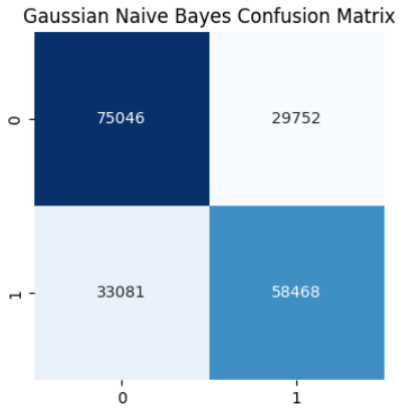

Fig. 14. Gaussian Naive Bayes Evaluation (Standard Scaled)

```
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
```

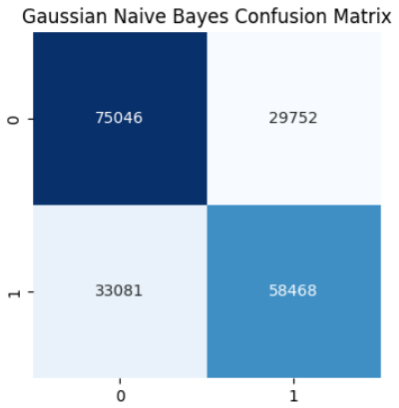

Fig. 15. Gaussian Naive Bayes Evaluation (Robust Scaled)

- 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
```

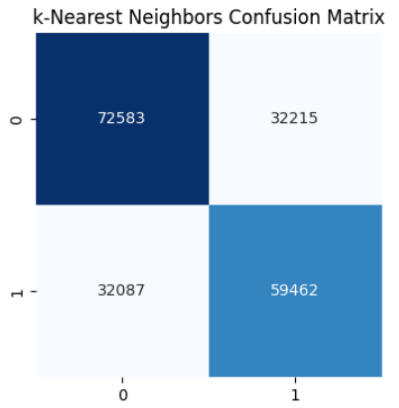

Fig. 16. K-Nearest Neighbors Evaluation (Standard Scaled)

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

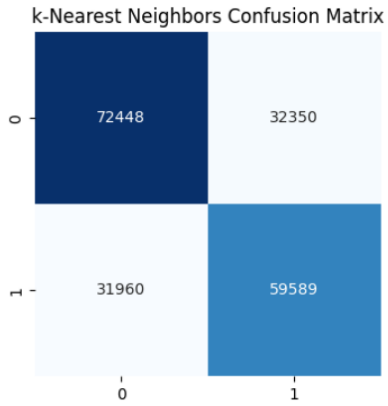

Fig. 17. K-Nearest Neighbors Evaluation (Robust Scaled)

Support Vector
Classifier:
Accuracy Score: 0.7103333333333334

	precision	recall	f1-score	support
N	0.73	0.73	0.73	11237
Y	0.69	0.69	0.69	9763
accuracy			0.71	21000
macro avg	0.71	0.71	0.71	21000
weighted avg	0.71	0.71	0.71	21000

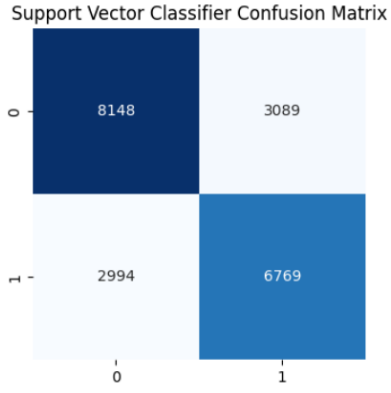

Fig. 18. SVC Evaluation

- 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. Latest Findings

As noticed in the preliminary findings section, there is no evaluation of the SVC model because it could not be run due to its cost in the initial development phase. However, in the further development phase, this problem was overcome with Random Sampling and model evaluation was carried out.

As a result of all analyzes and evaluations, it was decided that the most suitable model was Random Forest. After selecting the most accurate parameters for the Random Forest model with the Grid Search CV algorithm, the final model was created. Final evaluation metrics are shown in Fig. 19.

After determining the best parameters(Random Forest Classifier)
Accuracy Score: 0.7123307206119778

	precision	recall	f1-score	support
N	0.73	0.73	0.73	104798
Y	0.69	0.69	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

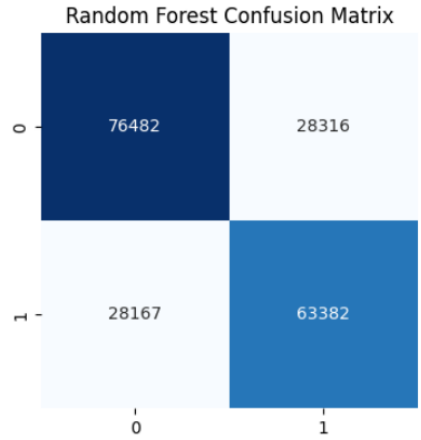

Fig. 19. Final Model Evaluation

V. Conclusion

In conclusion, this paper presented a project aimed at providing a modern solution to the problem of alcohol consumption. The project focused on building a classification model using artificial intelligence techniques to predict whether an individual consumes alcohol based on various features obtained from a dataset provided by the Korean Ministry of Health.

The initial findings revealed that Random Forest consistently outperformed other models in terms of accuracy, precision, recall, and F1-score. Feature importance analysis indicated that the AST/ALT (De Ritis) ratio, created during the feature engineering step, played a crucial role in the Random Forest model's success.

Further development efforts included optimizing the Interquartile Range factor for outlier detection, attempting Principal Component Analysis for dimensional reduction, random sampling to overcome computational costs, and employing Grid Search CV for parameter tuning. The final model, based on Random Forest, achieved an accuracy of 71.23%.

The implications of this work are significant, as it provides a personalized intervention approach for addressing alcohol consumption. The classification model can be utilized to identify individuals at risk, allowing for targeted interventions and guidance. However, it is essential to acknowledge some limitations of the study, such as the reliance on a specific dataset and potential biases in the data.

In the future, data from other countries of world may be obtained and integrated to further advance this study. In addition, the developed model can be tested and used in real-life applications by making it available to ministries of health that have a lot of data. Finally, a mobile application that provides personalized guidance can be developed.

In summary, the developed classification model presents a promising step towards addressing the challenge of alcohol consumption, and further research and collaboration are crucial to refining and implementing this solution effectively in real-world scenarios

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. Ebrahimi, A., Wiil, U. K., Schmidt, T., Naemi, A., Nielsen, A. S., Shaikh, G. M., & Mansourvar, M. (2021). Predicting the risk of alcohol use disorder using machine learning: a systematic literature review. IEEE Access, 9, 151697-151712.

[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. Kim, S. Y., Park, T., Kim, K., Oh, J., Park, Y., & Kim, D. J. (2021). A deep learning algorithm to predict hazardous drinkers and the severity of alcohol-related problems using K-NHANES. Frontiers in psychiatry, 12, 684406.

[6] 6. 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.

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

[8] 8. 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.

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

[10] 10. 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

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

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