Personalized Intervention for Alcohol Consumption

1st Abdullah Aydoğan
AI and Data Engineering
Istanbul Technical University
150230731
aydogana21@itu.edu.tr

2nd Ömer Faruk Satık
AI and Data Engineering
Istanbul Technical University
150210330
satik21@itu.edu.tr

***Abstract*— This paper proposes an artificial intelligence and data-based project aimed at reducing alcohol consumption and its harmful effects. The trigger for the project was that today's redirects were ineffective on consumers, and the team thought of personalized redirects as a solution to this ineffectiveness. In this context, it was decided that a classification model that detects citizens who drink alcohol is necessary. In this way, customized guidance can be given to people by the authorities with the alcohol consumer detection output provided. This paper will explain the problem in detail and the steps to be followed in solving the problem. In addition, the work plans of the team members and how they will work in which areas of the project will be discussed.**

***Index Terms*—alcohol related problems, personalized redirection, data, data preprocessing, classification algorithms.**

I. Introduction

Today, alcohol consumption is an important factor that negatively affects human life both materially and spiritually. Health problems can be considered the most important of the problems caused by alcohol consumption. Alcohol consumption increases the risk of liver cirrhosis, chronic pancreatitis, upper gastrointestinal cancers, cardiomyopathy, polyneuropathy [1]. Moreover, diseases are not the only factors that endanger human life due to alcohol. Situations such as drunk driving and doing dangerous things that would not normally be done while drunk also pose a risk of death. On the spiritual side of the matter, it can significantly worsen family and friendship relationships and the person's quality of life. For all these reasons, alcohol consumers need to be made aware of alcohol, encouraged to quit drinking, and the harmful effects of alcohol on them need to be treated. Although information is provided through media about the harms of alcohol consumption, it is obvious that the majority of people do not care about it. We thought that this situation occurred because the information was general notifications that did not address the consumer personally. So, our main goal with this project is to be more effective on alcohol consumers by providing them with directions specifically created for them. With this project, people who consume alcohol can be identified with the help of physical and biological data obtained from people. In line with this determination, the authorities will inform each individual who consumes alcohol about the alcohol-related problems in his/her body, and will also provide referrals to doctors and therapists appropriate to the person's needs. In short, this project aims to reduce alcohol consumption and its harmful effects as much as possible. The project team consists of two members: Abdullah Aydoğan and Ömer Faruk Satık. Both members have worked on machine learning, especially prediction and classification algorithms, and have previously carried out a project together. Additionally, Aydoğan also has experience in deep learning and data visualization. Team members will carry out the pre-processing of the data together. Machine learning and deep learning will be used in the project, and Satık will mainly take part in the machine learning part. Deep learning and data visualization algorithms will be applied by Aydoğan.

II. Dataset

A. General Information

This dataset, consisting of 24 columns and 991346 rows, was collected from the National Health Insurance Service in Korea via Kaggle datasets [2]. This set, which is in 'csv' format and 181.5 MB in size, contains 22 numerical and 2 categorical features, such as people's physical characteristics, genders, some health information, smoking degree and whether they drink alcohol or not.

B. Preprocessing

This data must be pre-processed before it can be used in modelling. Since most of the features in the data set are numerical, they are suitable for many algorithms and are easy to preprocess. One-hot encoding will be used to convert categorical data into numerical data [3]. Because our goal is to predict whether a person drinks alcohol or not, unnecessary features can be eliminated by correlating the drinking status column with other columns. Thanks to the large number of rows, outliers can be easily detected and removed [4]. Furthermore, methods such as normalization and standardization will be used so that Machine Learning algorithms can process data easily and accurately.

C. Limitations

As for the limitations of this data set, the fact that the dataset was created only from Korean citizens is a limitation because the model created may not give accurate results with data from people of other races around the world. In addition, the long processing time due to the large number of rows and features can be shown as another limitation.

III. Methodology

First of all, visualization will be made using matplotlib and seaborn libraries to analyze the data. The data preprocessing phase will be started with the help of the information obtained from the visualization and analysis. During the data preprocessing phase, the steps mentioned in the data set section will be taken. Since this is a classification problem, the model will work on machine learning classification algorithms SVC (Support Vector Classifier), Decision Tree, Random Forest, KNN (K-Nearest Neighbor) and Naive Bayes algorithms. Among these algorithms, the algorithms that give the most accurate results and their optimum parameters will be determined. ANN (Artificial Neural Network) algorithm produces very effective results on large data sets [5]. For this reason, the model will be created by comparing the results obtained from other classification algorithms with the ANN results and deciding on the most appropriate algorithms.

IV. Evaluatoin Method

Project success will be assessed by measuring improvements in metrics such as accuracy, recall, precision, F1-score [6]. In order to obtain these information confusion matrix and classification report methods will be applied to each classification algorithm. The goal is to develop a model that outperforms the initial model and provides a more efficient means of identifying individuals who consume alcohol, thereby increasing awareness of its detrimental effects.

V. Time Plan & Distribution of Work

Deadline	Process	Abdullah Aydoğan	Ömer Faruk Satık
13.10.2023	Identifying the problem		
16.10.2023	Search for dataset		
20.10.2023	Data understanding		
25.10.2023	Literature review		
30.10.2023	Project proposal		
7.11.2023	Data analyzing and visualization	Data visualization (Boxplot, Scatter Plot, Heatmap, Bar Plot)	General Information (Data types, Data size, Missing values, Detection of duplicated data, Data description), Correlation
14.11.2023	Data preprocessing	Dropping outliers, duplicated data and features	Categorical to numerical (one-hot encoding), scaling
26.11.2023	Model development	Artificial Neural Network, Support Vector Classifier, KNN	Decision Tree, Random Forest, Naive Bayes
30.11.2023	Model evaluation	Classification report, Confusion matrix	
4.12.2023	Intermediate report		
20.12.2023	Model review and development		
25.12.2023	Re-evaluation of the model	Classification report, Confusion matrix	
1.01.2024	Preparing presentation		
2.01.2024	Presentation of the project		
15.01.2024	Final report		

Fig. 1. Time plan

To talk about the progress made so far, in order to determine the problem, the team members talked about the factors that negatively affect human life today and came to a common decision. Following this, detailed research was conducted to obtain the data set needed to solve the problem. After the selected data set was examined and understood in detail, a literature review on classification algorithms that could be used in creating the model was carried out in collaboration. In the next part of the project, as indicated in the table in Fig. 1, both members will have distinct areas of work in the data analysis, data preprocessing and model development stages. Reporting and presentation of the project will also be done jointly along with the evaluation of model success. Below, the areas where team members will work separately will be discussed in more detail.

A. Distribution

- Ömer Faruk Satık will start with basic analysis studies on the data. This analysis includes understanding data types, data size, and data description, as well as detecting missing values and duplicated data. Additionally, examining the correlation between features is also a part of this analysis. In the data preprocessing phase, he will deal with numerical encoding of categorical data and data scaling required for algorithms such as KNN. Finally, he will focus on Decision Tree, Random Forest and Naive Bayes, which are classification algorithms that will contribute to model formation. After researching and understanding the working logic, purpose of use, compatibility and requirements of the algorithms, he will apply them on the data.
- Data pre-processing operations such as data visualization, dropping outliers and unnecessary features will be under the responsibility of Abdullah Aydoğan. Additionally, he will prepare classification processes using leading machine learning algorithms such as support vector classifier and k-nearest neighbors. When the other algorithms are completed, he will compare the results in Artificial Neural Networks with other algorithms and determine whether the most suitable model can be found with ANN or not.

In model evaluation, Classification Report and Confusion Matrix methods will be applied with the participation of both members.

VI. Conclusion

In conclusion, this project is aimed at addressing the adverse effects of alcohol consumption on both the individual's health and their social life. Alcohol-related health problems, including cancer, cirrhosis, and nervous system disorders, continue to pose a significant threat to public health. Moreover, the potential dangers associated with alcohol consumption and the impact on interpersonal relationships underline the need for a comprehensive solution. The project is designed to be

impactful by providing personalized guidance to individuals who consume alcohol. By leveraging physical and biological data, we can identify those who may be at risk due to their alcohol consumption. This personalized approach will enable authorities to orient alcohol consumers in necessary way. The dataset, collected from the National Health Insurance Service in Korea, offers a substantial volume of data, although its applicability may be limited to the Korean population. By implementing data preprocessing techniques such as one-hot encoding, outlier detection, and normalization, data will be appropriate for the algorithms to be used. In our methodology, we will utilize various machine learning classification algorithms, such as SVC, Decision Tree, Random Forest, KNN, and Naive Bayes, to determine which provides the most accurate results. Also, ANN will be used for its effectiveness on large datasets. The success of this project will be measured through improved metrics like accuracy, recall, precision, and F1-score, ensuring that the model outperforms the initial model. In essence, this project is not only a study to help alcohol consumers, but a call to raise awareness. By creating the model, the actual aim is to improve the quality of life for individuals in need.

REFERENCES

[1] Grønbæk, M. (2009). The positive and negative health effects of alcohol-and the public health implications. *Journal of internal medicine*, *265*(4), 407-420.

[2] SooyoungHer. Smoking & Drinking Dataset. Kaggle. (https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset/)

[3] Dahouda, M. K., & Joe, I. (2021). A deep-learned embedding technique for categorical features encoding. IEEE Access, 9, 114381-114391.K. Elissa, “Title of paper if known,” unpublished.

[4] Li, C. (2019). Preprocessing methods and pipelines of data mining: An overview. arXiv preprint arXiv:1906.08510.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[5] Basu, J. K., Bhattacharyya, D., & Kim, T. H. (2010). Use of artificial neural network in pattern recognition. International journal of software engineering and its applications, 4(2).

[6] Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.