**Group Members:**

- **Burak Topçu 283078027**
- **İsmail Cebeci 272001030**
- **Güliz Akkaya 283078008**

# CENG542 – KNOWLEDGE DISCOVERY

# TERM PROJECT PROPOSAL REPORT

## ABSTRACT

The continued high level of alcohol consumption in the majority of university students highlights the need for effective preventive and treatment interventions for both genders. The purpose of our research on this issue is that binge drinking is common among university students. While the sale and use of alcohol is legally free at the age of 18 and over, those under the age of 18 also consume alcohol a lot. Therefore, there is no concentration on university students only. In addition, drinking habits are not dependent on marital status; and most students don't drink alone. It is more well known that they consume alcohol for entertainment rather than melancholy. In this proposal, we will investigate the alcohol consumption among students. It will be determined what is the alcohol consumption tendency for a given student. To find prediction pattern used for predicting alcohol tendencies, we think that we will use classification and regression approaches.

**Keywords:** classification approach, regression approach.

## INTRODUCTION

In this project, we will investigate the relation between some student features such as age, family size, education status of the parents and alcohol consumption tendency of students. Even if selling alcohol is forbidden for an age limit such that it is 18 in Turkey and can be changed from one country to another, it can be consumed by students in a lot of various way. The point is not evaluating effects of alcohol consumption in student period. We will focus on whether there are some relations between alcohol consumption and student features or not. If we can realize a relation, we will try to implement a model such that it will operates on data in a mining manner to try to a pattern which enables us while determining the alcohol tendency of any student.

As a result, we think that we will use both classification and regression approaches to predict alcohol tendency and consumption amount of the students. Maybe, our findings can be beneficial for phycological works to solve problematic tendencies especially for those who have a huge tendency to consume alcohol in teenage duration without taking care both physical and phycological effects.

## PLANS AND METHODS

First of all, we are planning to look for correlation between features and alcohol tendencies. To investigate relations, we will use Spearman and Pearson correlation methods. The reason behind why we

are doing this is that if there are no relations between some features and alcohol consumptions, the irrelevant features can corrupt the regression and classification mechanisms. To prevent this, we will first look at the correlation results. Second, we will try to figure out meaningful patterns from our dataset for predicting alcohol consumption amount and tendencies among the students. We are planning to use 2 approaches. One of them is classification that will classify a student's alcohol tendency by rating from 1 to 5 where 1 for no tendency and 5 for high tendency. Apart from that, we will try to predict alcohol consumption amount for a given student. To do so, we will use regression approach.

We are planning to use k-NN, SVM and SGD algorithms to test classify students with respect to their tendencies. Also, we are planning to implement SVM, GB, and RF regression algorithms to find out meaningful patterns between alcohol consumption of the students and their features. Those algorithms have already implemented in sci-kit library of the python. Furthermore, we will use a dataset taken from Paulo Cortez and Alice Silva [1] works. There are lots of attributes that can be effective for our prediction and classification mechanisms.

## RELATED RESEARCH

There are several studies that analyze student's alcohol consumption with different algorithms. For instance, decision tree and random forest algorithms are utilized to perform the classification of student alcohol consumption in [2]. The performance results of these algorithms are compared, and it is stated that the random forest algorithm performs better than the decision tree algorithm for this classification problem. Also, important features related to alcohol consumption such as gender, go-out level and study time are mentioned in the study. Also, another study [3] analyzes the factors that affect student's alcohol consumption using five different alcohol consumption measures and three different educational achievement indicators. Another study [4] uses four different algorithms, Decision Tree, k-Nearest Neighbour (k-NN), Random Forest and Naïve Bayes to predict alcohol consumptions among secondary school students. As a result of the study, it is stated that decision tree algorithm produces highest accuracy values.

## REFERENCES

[1] Cortez, Paulo & Silva, Alice. (2008). Using data mining to predict secondary school student performance. EUROSIS.

[2] A. Pisutaporn, B. Chonvirachkul and D. Sutivong, "Relevant factors and classification of student alcohol consumption," *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, 2018, pp. 1-6, doi: 10.1109/ICIRD.2018.8376297.

[3] El Ansari, W., Stock, C., & Mills, C. (2013). Is alcohol consumption associated with poor academic achievement in university students?. *International journal of preventive medicine*, *4*(10), 1175–1188.

[4] S. Ismail, N. I. A. N. Azlan and A. Mustapha, "Prediction of alcohol consumption among Portuguese secondary school students: A data mining approach," *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2018, pp. 383-387, doi: 10.1109/ISCAIE.2018.8405503.