CSS429 – Final Project Documentation

Team 15:

1. **Gulnaz Kuatkyzy 170107040@stu.sdu.edu.kz**
2. **Dilshat Sissimbayev 170107022@stu.sdu.edu.kz**
3. **Zhanbolat Bekmaganbetov 180107180@stu.sdu.edu.kz**
4. **Assel Tolegen 180107033@stu.sdu.edu.kz**
5. **Kaisar Phazyl 170107119@stu.sdu.edu.kz**

Business understanding

Background

As we know alcohol has a detrimental effect on our organism and our heath in general ,and especially on the health of minors(underage teenagers), children. Nowadays a problem of alcohol consumption among not only children, teenagers but also among the adults is actual and requires drastic measures and solutions. And solving the problem of consumption an alcohol may also prevent it in mature age too, and as a consequence decrease amount of alcohol consumption among the adults .

In the report of CDC(Centers for Disease Control and Prevention)[1] Youth Risk Behavior Survey for 2019 about alcohol consumption of youth(in USA) says: "*

19% of young people aged 12 to 20 years reported[1] external icon drinking

alcohol and 11% reported binge drinking in the past 30 days.4% of 8th grade students and 29% of 12th grade students reported[1] external icon drinking alcohol during the past 30 days, and 4% of 8th grade students and 14% of 12th grade students reported binge drinking during the past two weeks." In another report ,in this time - Regional Office for Europe of WHO, for 2002-2014 says: "** More than 1 in 4 15-year-olds (28%) reported that they started consuming alcohol at age 13 or younger (25% of girls and 31% of boys) in 2014...Around 1 in 10 adolescents reported first being drunk at age 13 or younger (7% of girls and 9% of boys) in 2014*"[2]. In the most researches standing out different reasons,which are in varying degree affected on influenced of cause on emergence alcohol addiction,in the generall, responders noting alcohol drinking by adults. But that's only the one of the huge amount of factors that could be influenced on minors.

Goal and business problem

The actual business problem here is more relation to society,and social issues - find out and understand the factors ,influenced on alcohol consumption among students and which type or types of effects on students it has , such as academic performance, and based on these factors ,explore the relationships between different features.

How and where it could be used? For social problems for example - teachers and parents of student,or student itself for example for determining the influence of alcohol on his academic grades

Theoretically, business could use the results of this research for example online schools or course platforms for providing their services to potential clients who ahs a problems with an material and studying in general,or private psychologists could correct their advertisement model and reach a more huge audition.

For this research we are building prediction model on student's academic grade

References:

- [1] - https://www.cdc.gov/healthyyouth/data/yrbs/index.htm
- [2] - https://www.euro.who.int/en/media-centre/sections/press-releases/2018/adolescents-drink-less,-although-levels-of-alcohol-consumption-are-still-dangerously-high

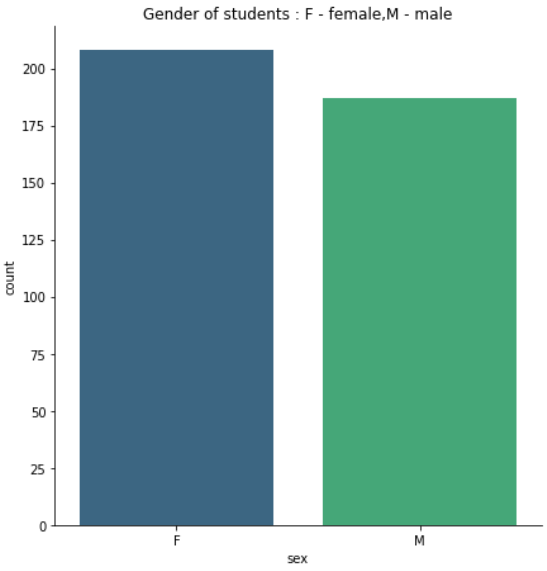
Data understanding

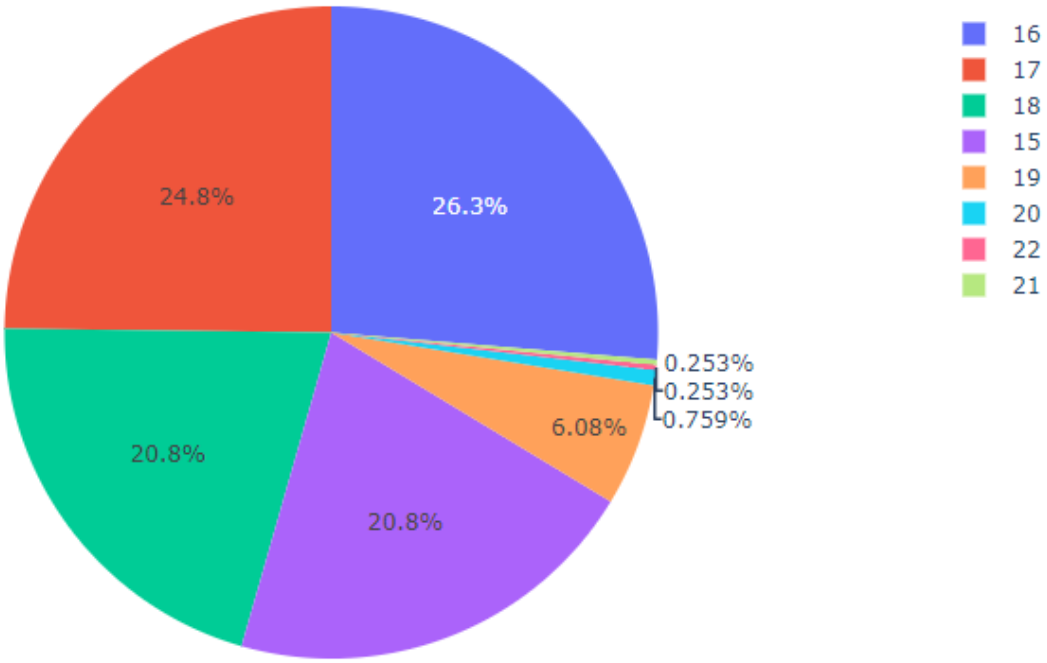
Data was taken from https://www.kaggle.com/uciml/student-alcohol-consumption.

It consists of 33 columns. Which are:

1. school - student's school
2. sex - student's sex
3. age - student's age
4. address - student's home address type
5. famsize - family size
6. Pstatus - parent's cohabitation status
7. Medu - mother's education
8. Fedu - father's education
9. Mjob - mother's job
10. Fjob - father's job
11. reason - reason to choose this school
12. guardian - student's guardian
13. traveltime - home to school travel time
14. studytime - weekly study time
15. failures - number of past class failures
16. schoolsup - extra educational support
17. famsup - family educational support
18. paid - extra paid classes within the course subject
19. activities - extracurricular activities
20. nursery - attended nursery school
21. higher - wants to take higher education
22. internet - Internet access at home
23. romantic - with a romantic relationship
24. famrel - quality of family relationships
25. freetime - free time after school
26. goout - going out with friends
27. Dalc - workday alcohol consumption
28. Walc - weekend alcohol consumption
29. health - current health status
30. absences - number of school absences
31. G1 - first period grade
32. G2 - second period grade
33. G3 - final grade

The number of girls is little more.





Basically we are dealing with students 15-18 years old.

Data preparation

So our data preparation consists of 4 parts:

1. Import library
2. Read data
3. Data Wrangling and Cleansing
4. Data Exploration

There are students taking courses in both mathematics and Portuguese. Therefore, participate in both surveys. We will not merge the two datasets yet, as this will distort the data. To solve this problem, we compare the students in both grades. We assign each student a unique number by adding a StudentID column to the math DataFrame class. Which will serve as the primary key.

We then use the Portuguese DataFrame class to map the StudentID from the math class. Using only information that would not change for the student depending on the course. For example, school, gender, address, guardian information. Next, we ask the math class to identify the same students in the Portuguese dataset. If a match is found, we assign the same StudentID to the string. If no match is found, the string is assigned a new unique StudentID.

After we have identified the same students in both courses, we delete the columns that we don't need in both DataFrames.

In data exploration we examined which factors influence alcohol consumption the most. For example, how the following factors influence alcohol cravings:

1. Whether a student's gender influences his alcohol cravings.
1. Family relationships
2. Academic progress

Modeling

We were discussing and choosed 4 data mining algorithms and built prediction models for our purposes. As alternatives we had Lasso,which we used a little bit later , and Ridge Regression. The main pros of Linear Regression - that’s the simplest type of regression, computationally efficient,calculations are fast and often well scored. Cons that it was too simple and it’t capture the whole complexity or real world project such our .It’s also affected by outliers,cause some too large outputs could make our research not effective. Then we were discussed DecisionTreeRegressor ,so the one of the main disadvantages and cons of this choice were that training of that could be expensive as the complexity and taken time, one of the and it’s also tend to overfit, as pros mentioned that it’s enough fast for inference and by the time requires little data preprocessing: no need for one-hot encoding, dummy variables etc.

Evaluation

After preprocessing the data, we tried to visualize the relationship between independent and target variables. Box plots and diagrams showed interesting results. We found variables that have good predictive power. Linear Regression, Lasso, Decision Tree and Random Forest were used during evaluation. And among them Linear Regression and Lasso have higher accuracies. We also used feature selection for determining the best of features for understanding .

Deployment

Discuss how the result of the data mining will be deployed.

With the help of the results and intellectual analysis, it is possible to determine the level of danger, factors of influence and influence on the student's studies or academic performance. Since the main goal of the analysis is to reduce alcohol consumption, this analysis will help to divide the influencing factors into several groups and, accordingly,

offer services and ways to solve them.

Discuss any issues the firm should be aware of regarding deployment.

The list of such questions is directly related to the commercial agreement of the

manufacturer and the legislation of the republic. One of their important elements is the distribution of the products of unlicensed alcohol producers and their sale. With this opportunity, any minor can buy on the domestic market. Decisions here require action from legislators and manufacturers. Restricting the sale of a product to private customers, concluding sales contracts only with those who have a full product license, and tightening requirements for the sale of alcohol on the market.

Are there important ethical considerations?

Research data cannot be limited to students only. This data is not advertising to a social audience. Professionals in the alcoholic field or pharmacists can use this data to neutralize alcohol dependence.

Identify the risks associated with your proposed plan and how you would mitigate them

To reduce the risk of the proposed plan, psychology courses can be promoted by seeking professional help and regular work on the participants in this study. In educational institutions where students study, clubs should be created to neutralize alcohol and addiction. Providing students with new opportunities in the virtual world and training of all kinds will help replace alcohol with constant work on themselves.

The contributions of each team member:

Zhanbolat Bekmaganbetov - Modeling, evaluation, documentation

Dilshat Sissimbayev - Modeling, visualization, data preparation

Kaisar Phazyl - data preparation, deployment

Gulnaz Kuatkyzy - presentation, visualization

Assel Tolegen - Visualization, presentation, documentation.