ALCOHOL EFFECTS
ON STUDY

INTRODUCTION

THIS DATA APPROACH STUDENT ACHIEVEMENT IN SECONDARY EDUCATION OF TWO PORTUGUESE SCHOOLS. THE DATA ATTRIBUTES INCLUDE STUDENT GRADES, DEMOGRAPHIC, SOCIAL AND SCHOOL RELATED FEATURES) AND IT WAS COLLECTED BY USING SCHOOL REPORTS AND QUESTIONNAIRES. TWO DATASETS ARE PROVIDED REGARDING THE PERFORMANCE IN TWO DISTINCT SUBJECTS: MATHEMATICS (MAT) AND PORTUGUESE LANGUAGE (POR). IN [CORTEZ AND SILVA, 2008], THE TWO DATASETS WERE MODELED UNDER BINARY/FIVE-LEVEL CLASSIFICATION AND REGRESSION TASKS. IMPORTANT NOTE: THE TARGET ATTRIBUTE G3 HAS A STRONG CORRELATION WITH ATTRIBUTES G2 AND G1. THIS OCCURS BECAUSE G3 IS THE FINAL YEAR GRADE (ISSUED AT THE 3RD PERIOD), WHILE G1 AND G2 CORRESPOND TO THE 1ST AND 2ND PERIOD GRADES. IT IS MORE DIFFICULT TO PREDICT G3 WITHOUT G2 AND G1, BUT SUCH PREDICTION IS MUCH MORE USEFUL (SEE PAPER SOURCE FOR MORE DETAILS).

OUR DATASET TALK ABOUT THE EFFECTS OF ALCOHOL ON STUDY , AND OUR PROBLEM IS THAT WE WANT THE RELATIONSHIP BETWEEN STUDYTIME AND G3.

1 READ DATA

- Use the Pandas method read_csv() to load the data.
- Use the method head() to display the first five rows of the dataframe.

2 DATA PREPARATION

- identify missing values using method isnull()

- Count missing values in each column using method ".value_counts()" counts the number of "True" values.

- list the data types for each column using df.dtypes()

- Binning data :rearrange values in G3 column into three 'bins' to simplify analysis,(Low,Medium,High)

- label categories using dummy variable, column "sex" has two unique values: "female" or "male"

- Normalize data by scaling the columns "studytime", "absences" and "G3" so their value ranges from 0 to 1 .

3 DATA EXPLORATION

- remove a few unnecessary columns. using pandas drop() method
- Using methods : df.describe , df.shape , and df.tail()

4 DATA ANALYSIS

- Visualize data using matplotlib .
- Made a line graph by extract data series for G3 and plot a line plot by appending .plot() to the G3 dataframe.
- Made a histogram for grades (G1,G2,G3) .
- Then we made a line graph that describe the relationship between study time and G3 for students.

Members

1 **SHAIMA IBRAHIM**
- Visualize data using matplotlib
- Made a line graph by extract data series for G3 and plot a line plot by appending .plot() to the G3 dataframe
- Made a histogram for grades (G1,G2,G3)
- Then we made a line graph that describe the relationship between study time and G3 for students.

2 **NADA ALI**
- Use the Pandas method read_csv() to load the data.
- Use the method head() to display the first five rows of the dataframe.
- list the data types for each column using df.dtypes()
- Binning data :rearrange values in G3 column into three 'bins' to simplify analysis,(Low,Medium,High)

3 **FATIMA ALQARNI**
- remove a few unnecessary columns. using pandas drop() method
- Using methods : df.describe , df.shape , and df.tail()
- identify missing values using method isnull()
- Count missing values in each column using method ".value_counts()" counts the number of "True" values.

Finally , We found that grades not only improve by increasing study time, but are also affected by a person's health and absence from school.

REFERENCE

KAGGLE.COM