Statistics of Heavy Alcohol Consumption Amongst Student Population (between ages 15 & 22) and the Contributing Factors

Olakunle Alabede, Ashton Anderson, Jaimie Jerome, Yomiso Sanya

Overview

- ✓ Reason why topic selected
- ✓ Description of data source
- ✓ Questions to consider using this data
- ✓ Data exploration phase of the project
- ✓ Analysis phase of the project

Reason why data was selected

This topic was selected because we needed a dataset that we could use to make possible predictions to an outcome using machine learning models. ~~Meanwhile, the dataset selected is also considered simple and easy to read and majorly contains numerical values.~~ This dataset contains several features that consist of numerical data with a few features that require the nominal data to be reassigned. With over 1000 rows and 30 columns, this is a sizeable dataset ~~However, this dataset was chosen because of its size and ability to generate potential results~~ which allows our model to predict outcomes on alcohol consumption amongst student population.

Description of data source

This dataset was culled from the UCI machine learning website which shows 2 different schools in Portugal and shows students drinking consumption. This group will use different machine learning algorithms to make predictions pertaining to whether a student may be prone to being a heavy drinker (total drinks > 4) or not

Questions to consider with this data

Using the sum of the Workday Alcohol Consumption (Dalc) and Weekend Alcohol Consumption (Walc), we will generate an Alcohol Consumption column, where if values fall between 4, they will be deemed a low risk alcohol consumer while if they fall above 4, they will be deemed a high risk alcohol consumer. Factors to consider include, but aren't limited to, sex, age, parent education, parent job, number of class failures, etc.

Data exploration phase

- Renamed columns to provide clarity
- Had to create a new column that combines the Walc and Dalc (work day and weekend alcohol consumption) to then create a binary column that assigns heavy drinkers as “1” (“yes”) for total drinks greater than 4 and not heavy drinks as “0” (“no”) for total drinks less than or equal to 4
- Removed certain columns

Analysis phase of project

- Used get_dummies to encode numeric data
- After splitting encoded features and targets into train and test sets, used StandardScaler to scale data
- Used RandomForestClassifier model; generated a confusion matrix, accuracy score and classification report

Dashboard Blueprint

- Final Dashboard Tool
  - Tableau to import the CSV and visualize data relationships
- Interactive Elements
  - Scatter plot, histogram