PROJECT - TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE USING ALCOHOL CONSUMPTION DATA

COURSE - DATA ANALYTICS
Prof: DR. SREEJA SR

Submission Date - 14.04.2021

TEAM DETAILS:

G.MADHAN (S20180020210)
K.MANOHAR (S20180020215)
P.HEMASAI (S20180020236)
ROHITH ALLA(S20180020238)

1 Introduction

We all know that excessive alcohol consumption of teenagers has been a serious issue in many countries. Drinking alcohol is often assumed to effect their academic performance and also some social and economic factors may also be related to one's grades. We are interested in the relationship between alcohol consumption and GPA. And if drinking does not significantly predict academic success, then what other factors should be considered will be known by this project.

2 Problem Statement

The primary goal of this project is to explore whether students' habits, especially alcohol consumption is a good classifier of student's grade and build a model that classify one's GPA based on variables including alcohol consumption.

3 Dataset description

The student performance dataset records the response to a questionnaire of students from two public high school in Portuguese during the 2005-2006 school year. The data is collected from two classes, Maths and Portuguese.
The data set has a total of 33 variables. The variables are listed as follow:

- school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- sex - student's sex (binary: 'F' - female or 'M' - male)
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: 'U' - urban or 'R' - rural)
- famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

- Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services')
- Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services')
- reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- traveltime - home to school travel time (numeric: numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour)
- studytime - weekly study time (numeric: 1 –< 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 –> 10 hours)
- failures - number of past class failures (numeric: n if 1¡=n¡3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (binary: yes or no)
- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93)
- G1 - first period grade (numeric: from 0 to 20)
- G2 - second period grade (numeric: from 0 to 20)
- G3 - final grade (numeric: from 0 to 20, output target)

4 Methodology

Build a decision tree based classifier model using ID3 algorithm.

4.1 Decision Tree:

- A Supervised Machine Learning Algorithm, used to build classification and regression models in the form of a tree structure.
  - Node - a feature
  - Branch - a decision
  - leaf – an outcome.
- Each node corresponds to a splitting attribute and each arc is a possible value of that attribute.
- At each node,the splitting attribute is selected to be the most informative among the attributes in the path starting from the root.

4.2 Entropy

- Entropy is a measure of the amount of uncertainty in the training data due to the presence of more than one class.
- Formula for entropy calculation:

$$E = \sum_{i=1}^{c} -pi \log_2(pi)$$

4.3 ID3 algorithm:

- Iterative Dichotomiser 3 is a classification algorithm that follows a greedy approach of building a decision tree by selecting a best attribute that yields maximum Information Gain (IG) or minimum Entropy (H).
- This algorithm defines a measurement of a spilitting called Information Gain to determine the quality of a spilt.
  - The attribute with the largest value of information gain is chosen as the splitting attribute and it partitions into a number of smaller trainings sets based on the distinct values of attribute under split.
- Steps in ID3 Algorithm:
  - Calculate entropy for dataset.
  - For each attribute/feature
    - Calculate entropy for all its categorical values.

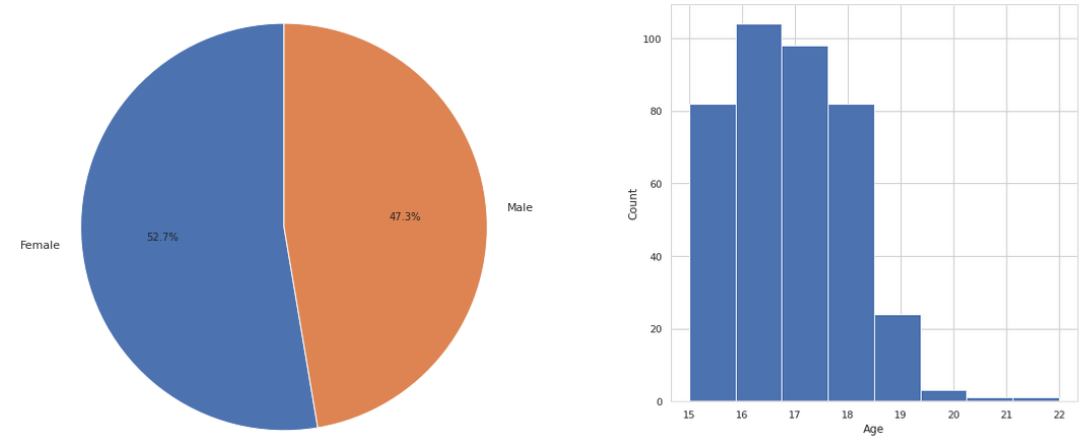
* Calculate information gain for the feature.

– Find the feature with maximum information gain.

– Repeat it until we get the desired tree.

5 Exploratory data analysis

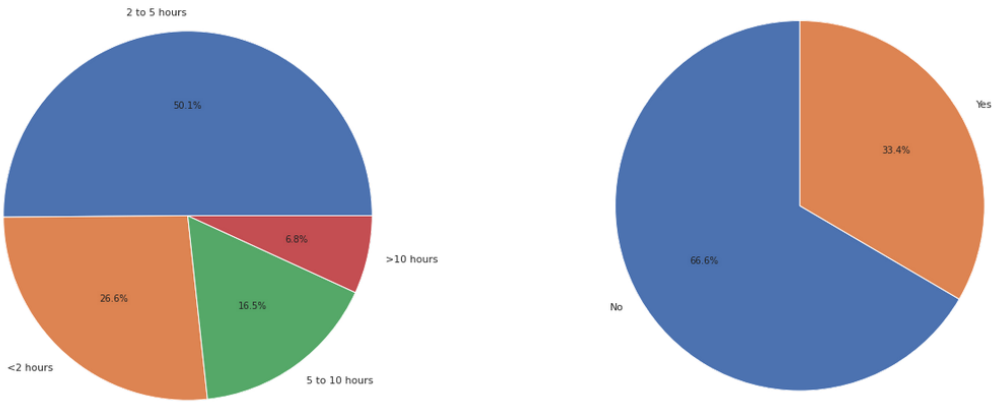
From the given dataset, we have applied exploratory data analytics on certain attributes and below are the observations.



(a) Gender (b) Age

Figure 1: Gender and age split in the dataset

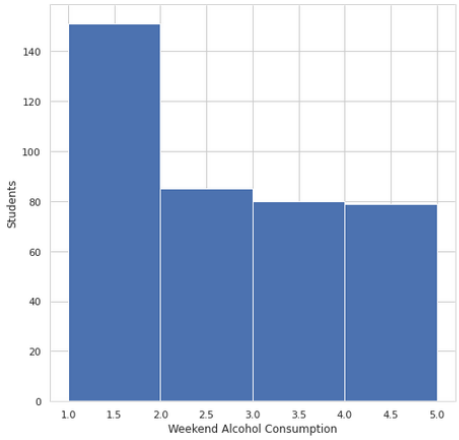
- We can observe that almost boys and girls are equal in number.
- Most of the students age is in between 15 and 18.



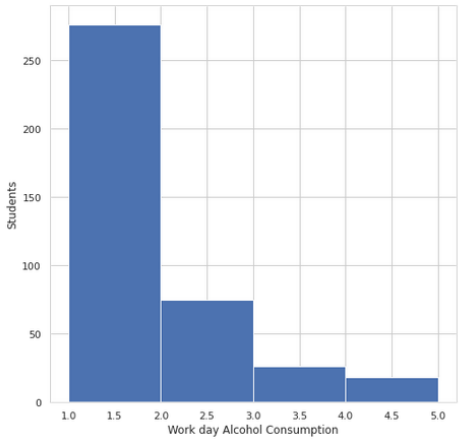
(a) Study time in hours (b) Romantic relationship

Figure 2: Study time and Relationship split in the dataset

- Nearly 50% of the students study 2-5 hours per week and approximately 20% of the students study greater than 5 and 10 hrs per week.
- 33.3 % of the students are in romantic relationship.



(a) Weekend alcoholic consumption

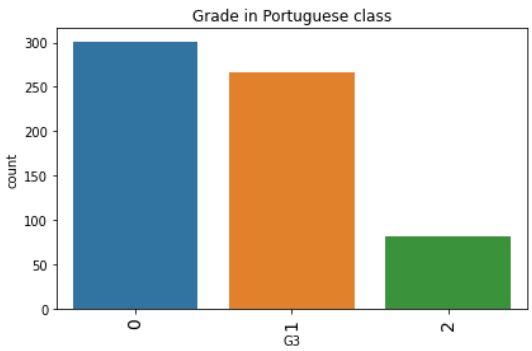


(b) Workday alcohol consumption

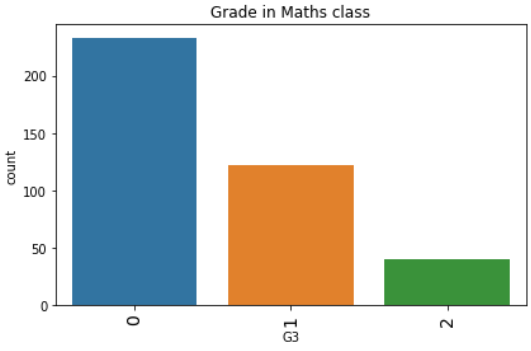
Figure 3: Alcohol consumption

- Nearly 250+ students don't consume alcohol during working days. Whereas around 20+ students consume high amount of alcohol.
- Nearly 80 students consume high amount of alcohol during weekends, and around 160 students consume moderately and 140+ students don't consume at all even in the weekends.

The grades in the dataset is in a range of 0 to 20. Hence we have defined three ranges which will be considered low grade (0-7), moderate grade (8-12) and high grade (13 - 20).



(a) Portuguese



(b) Maths

Figure 4: Final Grade splits

6 Implementation

- First identification of training attributes and class is made clear from the dataset. The attribute G3 is considered as the class attribute.
- Data cleaning is performed where any null values, repeated values are removed.
- All the categorical data is encoded into numerical using Label encoder.
- Now the dataset is divided randomly into training and validation data in 2:1 ratio using Bootstrap sampling method.
- The training data set attributes are used for training the decision tree.
- With the help of constructed decision tree, now we have predicted the grade using the validation dataset.
- Now using the validation data and the predicted data values we have constructed the confusion matrix.
- Using confusion matrix, we have calculated all the performance metrics such as accuracy, precision, recall and F1 score.

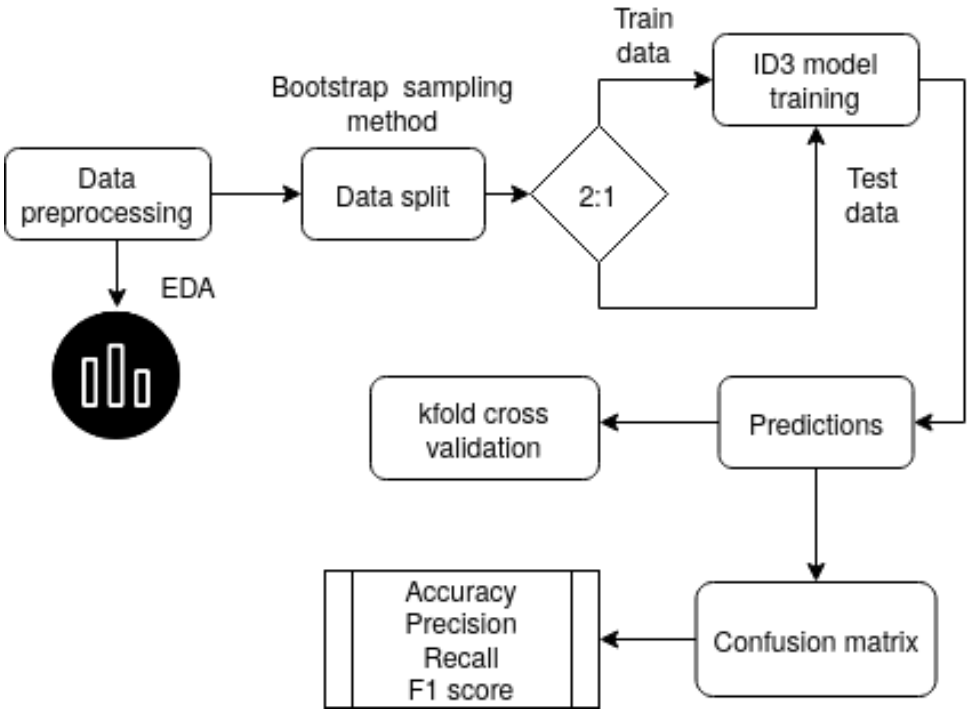


Figure 5: Work Flow of the entire project

7 Tree structure

```
{'G2': {0: 0.0,
        2: 0.0,
        3: 0.0,
        4: 0.0,
        5: 0.0,
        6: 0.0,
        7: {'G1': {4.0: 0.0,
                   5.0: {'sex': {0.0: 0.0, 1.0: 1.0}},
                   6.0: 0.0,
                   7.0: 0.0,
                   8.0: 0.0,
                   9.0: {'age': {2.0: 1.0, 4.0: 0.0}},
                   10.0: 0.0}},
        8: {'absences': {0.0: {'famsup': {0.0: 0.0, 1.0: 1.0}},
                         2.0: {'sex': {0.0: 0.0, 1.0: 1.0}},
                         3.0: 0.0,
                         4.0: 0.0,
                         6.0: 0.0,
                         8.0: 0.0,
                         11.0: 0.0,
                         16.0: 0.0}},
        9: {'absences': {0.0: {'school': {0.0: 1.0, 1.0: 0.0}},
                         1.0: 1.0,
                         2.0: 0.0,
                         3.0: {'school': {0.0: 0.0, 1.0: 1.0}},
                         4.0: 0.0,
                         6.0: 1.0,
                         8.0: 1.0,
                         10.0: {'school': {0.0: 1.0, 1.0: 0.0}},
                         13.0: 1.0,
                         14.0: 1.0,
                         20.0: 1.0,
                         31.0: 0.0}},
        10: {'nursery': {0.0: 0.0, 1.0: 1.0}},
        11: 1.0,
        12: {'age': {0.0: 1.0,
                     1.0: {'address': {0.0: 1.0, 1.0: 2.0}},
                     2.0: {'sex': {0.0: 1.0, 1.0: 2.0}},
                     3.0: 2.0}},
        13: {'Pstatus': {0.0: 1.0, 1.0: 2.0}},
        14: 2.0,
        15: 2.0,
        16: 2.0}}
```

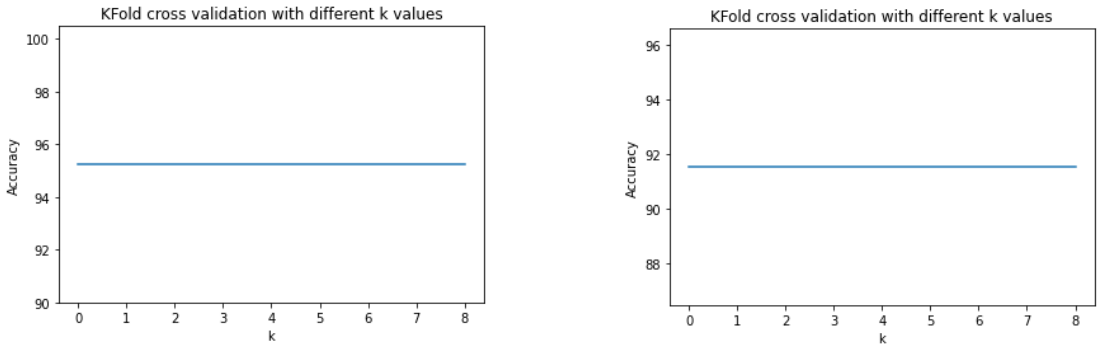
Figure 6: Tree constructed for maths class

```
{'G2': {0: 0.0,
        1: 0.0,
        2: 0.0,
        3: 0.0,
        4: 0.0,
        5: 0.0,
        6: {'G1': {5.0: 0.0,
                   6.0: 0.0,
                   7.0: 0.0,
                   8.0: 0.0,
                   9.0: 0.0,
                   10.0: 1.0}},
        7: {'absences': {0.0: {'age': {0.0: {'address': {0.0: 1.0, 1.0: 0.0}},
                                       1.0: 0.0,
                                       2.0: 0.0,
                                       3.0: 1.0}},
                         1.0: 1.0,
                         2.0: {'age': {0.0: 0.0,
                                       1.0: {'school': {0.0: 0.0, 1.0: 1.0}},
                                       2.0: {'Fedu': {1.0: 0.0,
                                                      3.0: 1.0,
                                                      4.0: 1.0}},
                                       3.0: 1.0}},
                         4.0: {'age': {1.0: 0.0, 2.0: 1.0, 3.0: 0.0, 4.0: 0.0}},
                         5.0: 0.0,
                         6.0: {'reason': {0.0: 0.0, 1.0: 1.0, 3.0: 0.0}},
                         8.0: {'Fjob': {0.0: 1.0, 2.0: 1.0, 3.0: 0.0}},
                         10.0: 0.0,
                         11.0: 0.0,
                         12.0: 0.0,
                         14.0: 0.0,
                         16.0: {'sex': {0.0: 0.0, 1.0: 1.0}},
                         19.0: 0.0}},
        8: {'Medu': {0.0: 0.0,
                     1.0: {'age': {1.0: 1.0, 2.0: 1.0, 3.0: 0.0}},
                     2.0: 1.0,
                     3.0: 1.0,
                     4.0: 1.0}},
        9: {'higher': {0.0: 0.0, 1.0: 1.0}},
        10: {'G1': {10.0: 1.0,
                    11.0: 1.0,
                    12.0: {'Fedu': {2.0: 1.0, 3.0: 2.0, 4.0: 1.0}},
                    13.0: {'sex': {0.0: 2.0, 1.0: 1.0}}}},
        11: {'freetime': {0.0: 1.0,
                          1.0: {'guardian': {0.0: 1.0, 1.0: 2.0}},
                          2.0: {'Fedu': {1.0: 2.0,
                                         2.0: 1.0,
                                         3.0: 2.0,
                                         4.0: {'sex': {0.0: 2.0, 1.0: 1.0}}}},
                          3.0: 1.0,
                          4.0: 1.0}},
        12: {'famsize': {0.0: 2.0, 1.0: {'sex': {0.0: 1.0, 1.0: 2.0}}}},
        13: 2.0,
        14: 2.0,
        15: 2.0}}
```

Figure 7: Tree constructed for Portuguese class

8 k-Fold Cross-Validation

- Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.
- The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into.
- The general procedure is as follows:
  - Shuffle the dataset randomly.
  - Split the dataset into k groups.
  - For each unique group:
    * Take the group as a hold out or test data set
    * Take the remaining groups as a training data set
    * Fit a model on the training set and evaluate it on the test set
    * Retain the evaluation score and discard the model
  - Summarize the skill of the model using the sample of model evaluation scores
- After applying kfold cross validation many times, the accuracy remains same for different values of k, for both the datasets.



(a) Portuguese class kfold cross validation

(b) Maths class kfold cross validation

Figure 8: kFold cross validation

- We can choose k=5 as the optmised k value as this values has been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance in various tests.

9 Performance metrics

- A confusion matrix is a table that is used to describe the performance of a classification model on a set of test data for which the true values are known.

- Accuracy is the fraction of predictions of the model that are true.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: Precision tells us how many, out of all instances that were predicted to belong to class X, actually belonged to class X.
- The precision for class X is calculated as:

$$Precision = \frac{TP}{TP + FP}$$

- Recall: expresses how many instances of class X were predicted correctly.
- The recall is calculated as:

$$Recall = \frac{TP}{TP + FN}$$

  where TP = the number of true positives for class X.
  TN = the number of true negatives for class X.
  FN = the number of false negatives for class X.
  FP = the number of false positives for class X

- F1 Score: is a function of Precision and Recall. It can be calculated by:

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

10 Results

- Confusion matrix

	0	1	2
0	63	2	0
1	6	41	3
2	0	0	16

(a) Maths class

	0	1	2
0	95	3	0
1	4	86	1
2	1	1	25

(b) Portuguese class

Figure 9: Confusion matrix (Column is prediction class and Row is actual class)

- Accuracy

S.NO	Data	k value	Accuracy
1	Maths class	5	91.6%
2	Portuguese class	5	95.4%

Table 1: Accuracy that is finalised after kfold cross validation

- Precision, Recall, F1 score

S.NO	class label	Precision	Recall	F1 score
1	0	0.913	0.969	0.940
2	1	0.953	0.820	0.882
3	2	0.842	1.000	0.914

Table 2: performance metrics for maths class data

S.NO	class label	Precision	Recall	F1 score
1	0	0.950	0.969	0.960
2	1	0.956	0.945	0.950
3	2	0.962	0.926	0.943

Table 3: performance metrics for maths Portuguese data

- In the table 1, we can observe the accuracy is 91.6% for maths class data and 95.4% for Portuguese class.
- In table 2 and table 3, precision, recall and F1 score is mentioned for each label.

11 conclusion

- If we observe the decision trees that are constructed, we can easily say that the attributes "G2" and "G1" are the main parameter for deciding the class. But those grade are not a satisfactory support for final grade.
- If we see the tree, the parameters such as absences, freetime also placed an important role to classify the students grade.
- And the lower end of the decision tree, parameters such as school, age and gender placed an important role.
- We can say that other than alcohol consumption, the social and person life of a student effects their academic score.