**Data Mining**

**ISCG 7426**

**Semester 2**

**November 2020**

****

**Assignment-2**

**Effect on academic performance of secondary students**

**consuming alcohol**

**Prepared By: Priya Payyadi Surendran**

**ID:1533656**

**Supervisor: Mr. Hamid Sharifzadeh**

**Declaration:” Wherever this assignment draws on the work of others, such sources are clearly acknowledged.”**

Table of Contents

[**1.** **Introduction** 2](#_Toc55742688)

[**a.** **Problem Statement** 2](#_Toc55742689)

[**b.** **Proposed methods to solve the problem** 2](#_Toc55742690)

[**c.** **Criteria used to evaluate the performance** 2](#_Toc55742691)

[**2.** **Methodology** 3](#_Toc55742692)

[**a.** **Dataset Description** 3](#_Toc55742693)

[**b.** **Data Pre-processing and Transformation** 4](#_Toc55742694)

[**c.** **Data Mining Techniques chosen** 5](#_Toc55742695)

[**Unsupervised Techniques** 5](#_Toc55742696)

[**•** **K Means Clustering** 5](#_Toc55742697)

[**•** **Filtered Clustering** 6](#_Toc55742698)

[**Classification** 6](#_Toc55742699)

[**•** **Naïve Bayesian** 6](#_Toc55742700)

[**•** **K Nearest Neighbour** 6](#_Toc55742701)

[**d.** **Tuning Parameters.** 6](#_Toc55742702)

[**3.** **Evaluation** 6](#_Toc55742703)

[**a.** **Quantitative Metrics Definition** 6](#_Toc55742704)

[**b.** **Reason for the selections** 7](#_Toc55742705)

[**c.** **Evaluation Process** 7](#_Toc55742706)

[**4.** **Discussions** 7](#_Toc55742707)

[**a.** **Review of Results** 7](#_Toc55742708)

[**b.** **Analysis of findings and best technique** 16](#_Toc55742709)

[**5.** **Conclusion** 18](#_Toc55742710)

[References 19](#_Toc55742711)

## **Introduction**

### **Problem Statement**

Secondary school students have the target of graduating school and getting passing marks. In any case, in every life there are various elements that impact these plans and their outcomes, both decidedly and contrarily. Here, I have decided to explore information relating to secondary school students from Porto in Portugal, to get a comprehension of the connections between way of life factors, for example, familial business and liquor utilization, and the evaluations acquired by the students upon whom these components hold impact.

Students are put resources into the administration of way of life through order as a method for acquiring positive results from our schooling. In any case, there are continually alleviating conditions outside our ability to control. Having an individual knowledge into the aftereffects of different students and having the option to dissect any relationships with their ways of life makes this venture expressly intriguing. With comprehension, the information acquired can be material to countless students .

There has been research directed into this subject on a few events, strikingly in the University of Gloucestershire, UK . It was seen that boys spent more hours during the latest drinking event than girls. Most students (74%) consumed liquor in any event once week, and just 3% detailed no liquor consumption. Guys devoured liquor more regularly than females. A huge share of the example (54%) affirmed considerable rambling drinking in any event multiple periods during the recent month, and just 20% of respondents announced no weighty wordy drinking. Recurrence of hefty roundabout drinking was altogether higher among guys than females (El Ansari, Stock, & Mills, 2013). As this exploration was directed against information from the United Kingdom, I decided to investigate information from different areas to check whether the results were comparable or different.

### **Proposed methods to solve the problem**

My proposed strategy is to dissect the information utilizing various data mining methods to decide if the Portuguese models share any similar results with the prior investigations in this area. The information accumulated and worked on originates from a study of students taking Mathematics and Portuguese language courses in high school. My strategy is to utilize the characteristics of a student response to anticipate their last grades with an end goal to comprehend if liquor utilization is a critical factor in academic performance.

The data mining techniques used are:

• K Nearest Neighbour • Filtered Clustering

• Naïve Bayesian Classification • K Means Clustering

### **Criteria used to evaluate the performance**

The performance of classification techniques will be assessed using confusion matrix, Root Mean Squared Error and the Relative Absolute Error and ROC Curve approach so that correct optimal value for each case can be chosen and clustering techniques can be evaluated using Sum of Squared Errors.

## **Methodology**

### **Dataset Description**

In a study led from the University of Minho, students from two higher secondary schools in Porto, Portugal were addressed on various components that are recommended to affect academic performance of the students addicted to alcohol.

The dataset is obtained from Kaggle site and consists of 2 datasets. One pertaining to the students who opted Mathematics and second dataset belongs to students pertaining to Portuguese subject. Each of the instances belonged to details of a student.

I merged the datasets and the new dataset comprises of 33 attributes and 1044 instances,

The attributes of the dataset are:

Text

Description automatically generated

### **Data Pre-processing and Transformation**

The initial step was to remove the outliers from the merged dataset. 9 outliers were detected along with 186 extreme values.

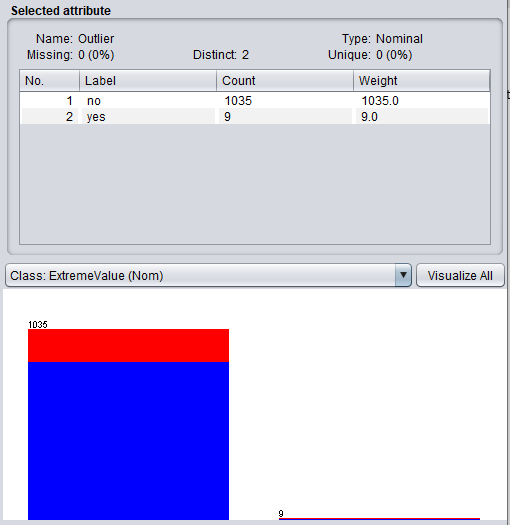


Figure 1: Outliers and Extreme Values Detected

After removing the 9 outliers, the extreme values reduced to 180 implying, 6 of the outliers were extreme values.

As a part of pre-processing, unwanted attributes were removed, and the retained attributes are as follows:

Text, letter

Description automatically generated

The main attributes are sex, age, Dalc, Walc, absences, G1, G2 and G3. The secondary supporting attributes for the conclusion of research are Pstatus, studytime, schoolsup and famsup, paid and higher, health. These attributes are considered to clarify if it’s really alcohol consumption that is affecting the academic performance or other factors like inability to learn, lack of parental or family support or other health issues.

After removal of outliers and unwanted attributes, the dataset contains 17 attributes and 1035 instances.

As part of normalization, I have converted the attributes grades G1, G2, G3 in to percentage (0-100) .I added 2 new attributes Average(numeric) and Result(nominal), which was calculated based on the grades G1, G2 and G3. Final dataset consisted of 19 attributes and 1035 instances.

### **Data Mining Techniques chosen**

#### 

#### **Unsupervised Techniques**

These models don't anticipate a target value, yet centre around the fundamental structure, interconnectedness and relations of the data (Cortez & Silva, 2008). Two unsupervised techniques I chose are:

#### **• K Means Clustering**

Quite effective as unsupervised method and works well for large scale data. Compared to other clustering techniques, it is easy to implement and no distributional assumptions about data.

#### **• Filtered Clustering**

The clustering algorithm ,based on a particular filter to be applied to the instances, is used to devise associated clusters of instances.

#### **Classification**

It predicts categorical class labels either discrete or normal. It constructs a model based on training sets and the class labels in classifying attributes and uses it in classifying new data (Trivedi & Kotak, 2019). The classification techniques I used are:

#### **• Naïve Bayesian**

It uses a simplified assumption that attributes are conditionally independent. It greatly reduce the computation cost by counting the class distribution.

#### **• K Nearest Neighbour**

One of the most used and simplest algorithms. Classification of new sample point is predicted by this algorithm by using database where earlier data points are stored in different groups.

### **Tuning Parameters.**

I used classification techniques to divide the students in to groups (group which consumes alcohol and groups which do not) . I chose Clustering which is unsupervised techniques as it provides a better graphic illustration of alcohol consumption and academic performance. I believed that there is a strong connection between amount of alcohol consumed and academic performance.

For classification, tuning parameter includes tuning the classifiers, search algorithms , batch size which can improve performance and cross validations. For Clustering tuning parameter includes value of k in case of K mean clustering and clustering distance function to identify the inter cluster distances and maximize it (El Ansari, Stock, & Mills, 2013).

## **Evaluation**

### **Quantitative Metrics Definition**

Evaluating the performance is the fundamental aspect of data mining. Determining the efficiency and performance of data mining model is hard. The data mining model is used for prediction and to check if the model is reliable, choose the right evaluation measure. Evaluation measure can differ from model to model.

ZeroR selects major classes of the dataset and makes all predictions. This will be the baseline for the dataset and always predicts the most frequent category. It predicts the mean for a numeric class and mode for a nominal class, hence is very useful for my dataset ( Pagnotta & Amran, 2016).

To quantify the classifier performance, Area Under Curve value method is used. ROC curve plots the true positives against false positive, it is useful to manipulate the dataset. The two criteria used is “Pass” and “Fail” which is used to find the true positive and false positive rates (Murphy, McDevitt-Murphy, & Barnett, 2005).

### **Reason for the selections**

I chose ZeroR rule because it is the simplest classification method which ignores all other predictors and relies on the target. It is mainly used to predict the baseline performance which acts as a benchmark for other classification techniques.

I chose AUC ROC Curve analysis because the area under ROC curve gives a picture about the impact of daily or weekly alcohol consumption on the student’s academic performance.

### **Evaluation Process**

I did not split the data but used test methods like cross validation .Many attributes were removed from the initial dataset and 2 new attributes were added to reduce overfitting of data. I never anticipated underfitting of data. Unfortunately, I faced underfitting of data as there depicted a weak correlation between alcohol consumption and student’s academic performance.

## **Discussions**

### **Review of Results**

**K Means Clustering**

I chose the K value to be 2(2 clusters) and is evaluated using the sum of squared errors. Entire 17 attributes were chosen to begin with which resulted in a high value of sum of squared errors. When the attributes Pstatus, schoolsup, famsup, paid, higher were ignored and the SSE value reduced.

Table

Description automatically generated

Figure 2: Summary of K means clustering



Figure 3:KMC: Daily alcohol consumption Vs Result output

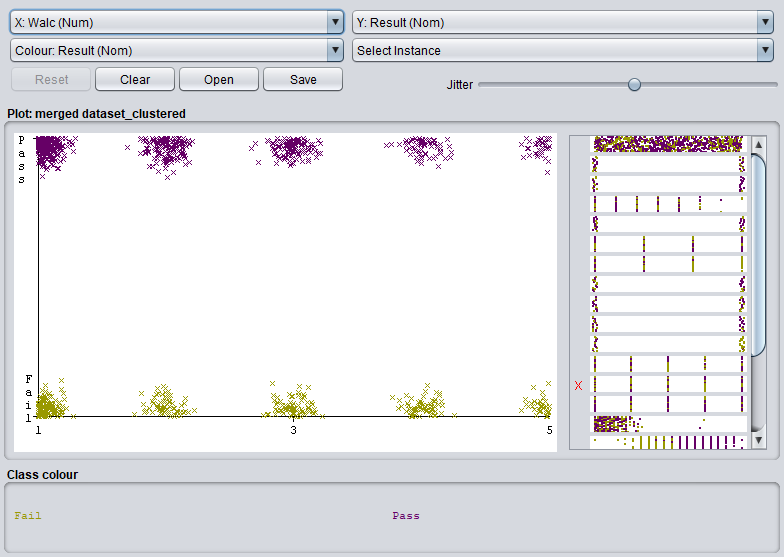


Figure 4:KMC: Weekly alcohol consumption Vs Result output

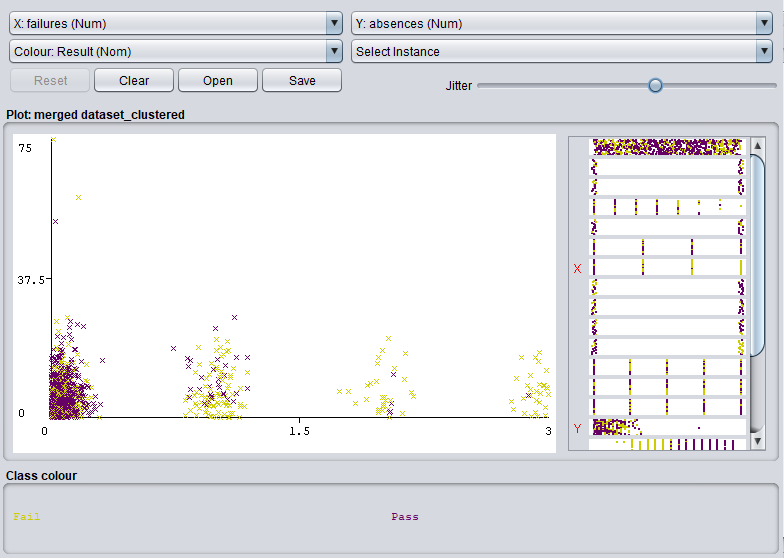


Figure 5: KMC : Absence of students Vs failures output

As shown in figure 3 and figure 4, the cluster of students passed and failed is distributed equally across al the measures of Dalc and Walc. In figure 5, it is clear that students with less absence from class have passed more(pass: purple, Fail: Yellow).

**Filtered Clustering**

Filtered clustering was performed using all attributes except the average and the result attributes and it resulted in a lesser SSE value than K means clustering.

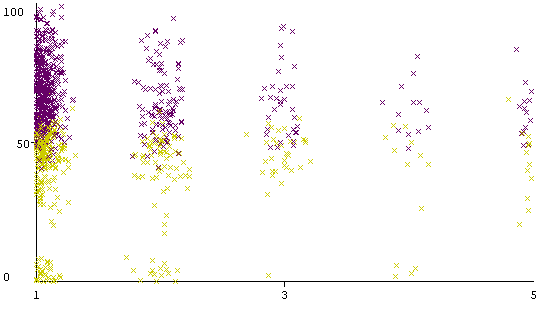


Figure 6:Filtered clustering: Dalc vs G3 output

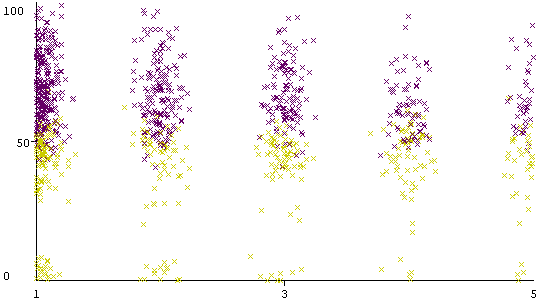


Figure 7:Filtered clustering: Walc vs G3 output



Figure 8:Filtered clustering: Dalc vs sex output

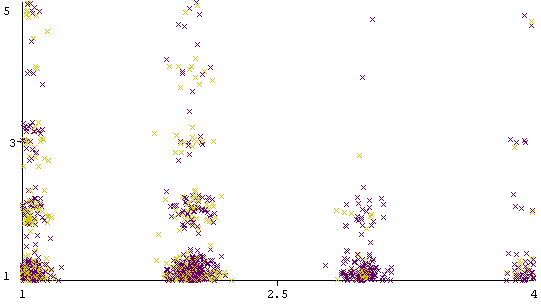


Figure 9:Filtered clustering: Studytime vs Dalc output

Table

Description automatically generated

Figure 10:Summary of Filtered Clustering

**Naïve Bayes Classification**

Naive Bayesian Classification was executed using a ZeroR algorithm which is a class for Naïve Bayes classifier using estimator classes.

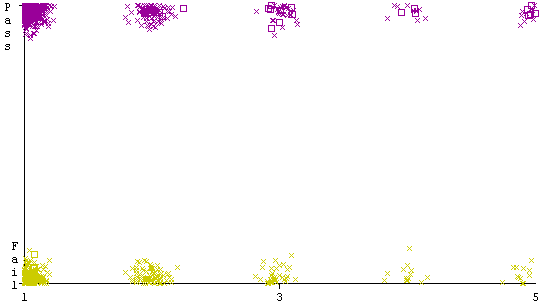


Figure 11:NBC: Dalc vs Result

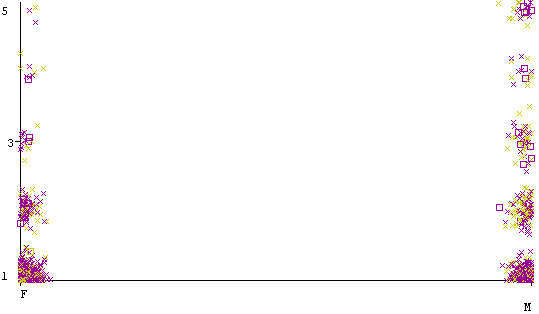


Figure 12: NBC: Sex vs Dalc

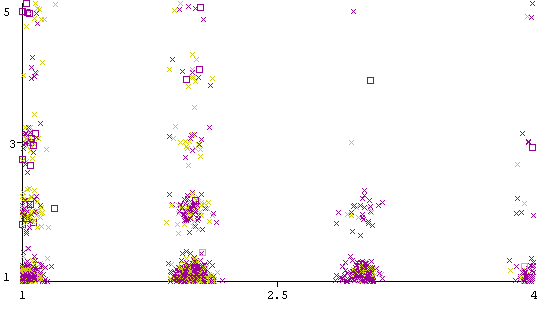


Figure 13:NBC: Studytime vs Dalc

A picture containing table

Description automatically generated

Figure 14: Summary of ZeroR

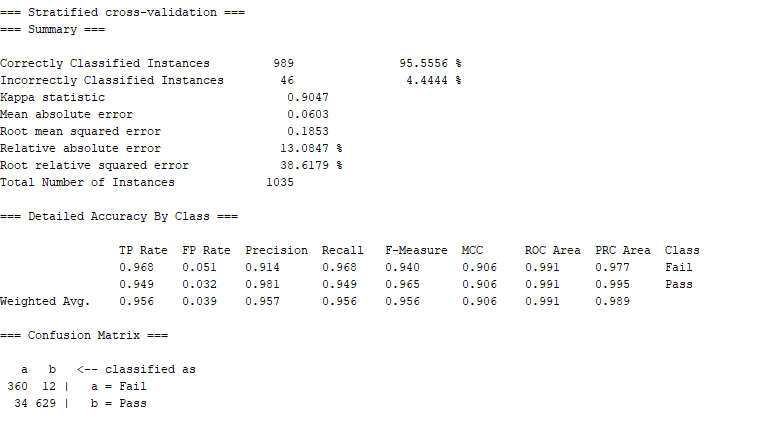


Figure 15:Summary of Naive Bayes Classification

The performance of Naïve Bayes was better than ZeroR. The accuracy of Naïve Bayes was 96% compared to 64% of ZeroR.

**K Nearest Neighbour**

When it was applied on training set, the accuracy showed 100%. Again, cross validation of 10 folds were used as test option to execute k nearest neighbour and the accuracy obtained was 80%.

A picture containing table

Description automatically generated

Figure 16: Summary of K Nearest Neighbour

When major attributes were ignored and only the gender, daily and weekly alcohol consumptions were measured against the Result attribute, the accuracy increased to 98% as shown in figure 17.

From the summary of KNN, it is obvious that student’s performance is not solely dependent on the alcohol consumption. Other factors also effects the academic performance.

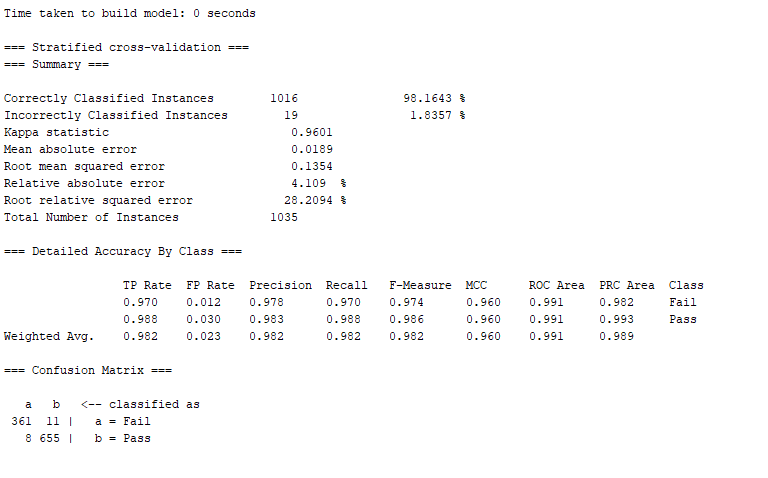


Figure 17: Summary of KNN without few attributes

### **Analysis of findings and best technique**

|  |  |  |  |
| --- | --- | --- | --- |
| **Technique Used** | **Correctly Classified Instances** | **SSE/Cross Validation** | **Accuracy (%)** |
| K Means Clustering | 663 | 2170.32 | 64% |
| Filtered Clustering | 661 | 1747.83 | 64% |
| Naïve Bayesian Classification | 989 | 10-fold cross validation | 95.56% |
| K Nearest Neighbor | 833 | 10-fold cross validation | 80.48% |

I found that the Naïve Bayes Classification proved to be the superlative at classifying and predicting the academic performance especially the success rate of students and the accuracy was almost 96%. In unison, K Nearest Neighbour Classification gave an 80% accuracy. Hence these algorithms can be organized to create a very accurate model with further tuning.

From the data mining techniques used, I learned that the correlation between student’s academic performance and alcohol consumption is less.

Boys consumed more alcohol daily compared to girls whereas during weekends girls consumed as much as boys. But girl’s success rate was more compared to boys proving that girl’s study more than boys and have high tolerance towards alcohol’s effect.

It was also observed that, student whose parents are still living together consumed more alcohol compared to the students whose parents were apart. Also, more students who consumed alcohol daily and weekly have more failures.

The major problem I identified is the less or weak correlation between the alcohol consumption and the academic performance of the student. Hence, I’m not sure if I have applied the correct techniques or not. I found it difficult to infer something from the dataset based on the daily and weekly alcohol consumption. I strongly feel that many other factors like family, lifestyle etc also effects the academic performance of the student.

## **Conclusion**

With liquor consumption progressively normal among teenagers, the effect on student’s academic performance must be constantly considered. By studying these, a help can be handed if needed to these teens to come out of alcohol. My decision was to explore the relation between the academic performance and alcohol consumption among the students of higher secondary school. Many literature reviews gave way to a strong relation between these two stated parameters. But I could not find much correlation between them by the techniques I chose . There is by all accounts no precise strategy to decide the pass pace of a student by essentially estimating how regularly they drink liquor. Fortunately, many other correlations were found. Boys consumed more alcohol daily compared to girls whereas during weekends girls consumed as much as boys. But girl’s success rate was more compared to boys proving that girl’s study more than boys and have high tolerance towards alcohol’s effect. It was also observed that, student whose parents are still living together consumed more alcohol compared to the students whose parents were apart. Also, more students who consumed alcohol daily and weekly have more failures. I conclude that, alcohol is not the only culprit that effects the student’s academic performance but family background, financial status, lifestyle etc also highly impacts the performance.

# References

Pagnotta, F., & Amran, H. (2016). *USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT ALCOHOL CONSUMPTION.*

Assanangkornchai , S., Mukthong , A., & Intanont, T. (n.d.). Prevalence and Patterns of Alcohol Consumption and Health‐Risk Behaviors Among High School Students in Thailand. *Alcoholism clinical and experimental research*, pp. 2037-2046.

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *Proceedings of 5th Annual Future Business Technology Conference* (pp. 5-12). porto: EUROSIS-ETI.

Cox, W., Steven G, Hosier, S., Crossley, S., Kendall, B., & L.Roberts, K. (2006). Motives for drinking, alcohol consumption, and alcohol-related problems among British secondary-school and university students. *Addictive Behaviors*, 2147-2157.

El Ansari, W., Stock, C., & Mills, C. (2013). Is Alcohol Consumption Associated with Poor Academic Achievement in University Students? *International Journal of preventive medicine*, 1175–1188.

Gervilla García , E., & Jiménez López, R. (2009). The methodology of Data Mining. An application to alcohol consumption in teenagers. 65-80.

Murphy, J., McDevitt-Murphy, M., & Barnett, N. (2005). Drink and Be Merry? Gender, Life Satisfaction, and Alcohol Consumption Among College Students. *Psychology of Addictive Behaviors*, 184-191.

Trivedi, T., & Kotak, D. (2019). Exploring Prediction Modeling of Students Alcohol and Drug Addiction Affecting Performance using Data Mining Approach. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY*.

W. Miller, J., S. Naimi, T., D. Brewer , R., & Everett Jones, S. (2007). Binge Drinking and Associated Health Risk Behaviors Among High School Students. *Pediatrics*, 76-85.

Yamada , T., Kendix , M., & Yamada, T. (1996). The impact of alcohol consumption and marijuana use on high school graduation. *Health Economics*, 77-92.