This course requires a major project, which you will complete in your assigned groups.  For this project, you will **be assigned** a useful real-world dataset and **conduct** your own data mining activities as prompted through 10 weekly questions.

The report must be 8-15 pages, excluding source code but including figures, and should be well written. **It must have the following 9 sections (please copy the section headings into your final writeup):**

**Student Alcohol Consumption Analysis Report**

**Report:   
  
  
  
  
  
Introduction:**

1. What does the dataset describe?  -- 10 pts

In our group assignment of alcohol consumption, there are two datasets of student-mat.csv (Math course) and student-por.csv (Portuguese language course), where we had currently considered one dataset of student-por.csv (Portuguese language course) and we had identified 5 important attributes for alcohol consumption from which identify the inference and conclusion of the ratio of alcohol consumption. The attributes we identified are **age, famrel, dalc, walc, health**. Let we display the content in tabular format to make it more explanatory.

|  |  |  |
| --- | --- | --- |
| S.no | Attributes | Datatypes |
| 1 | **age** - student's age | Integer(Numeric) |
| 2 | **famrel** - quality of family relationships | Integer(Categorical) |
| 3 | **Dalc** - workday alcohol consumption | Integer(Categorical) |
| 4 | **Walc** - weekend alcohol consumption | Integer(Categorical) |
| 5 | **health** - current health status | Integer(Categorical) |

Fig: Selected Attributes with Datatypes  
  
Every Attribute which we had selected would be creating a greater impact on inferencing a good result where many people like government agencies , schools get benefited which looking with results, and make a precautionary steps to decrease alcohol consumption.   
  
Age- This is one the place where we can keep restriction, and identify from what age the people had started alcohol consumption, and increase the bar of age.

Dalc- On being workday how much alcohol consumption would be taken place and where irrespective of age, and where the revenue generated to the government can be identified.

famrel- Quality of a family relation would also impact a greater alcohol consumption in which a child would be taken from parents behavior.

Walc- keeping same rules of weekdays how much it is said to be for weekends.

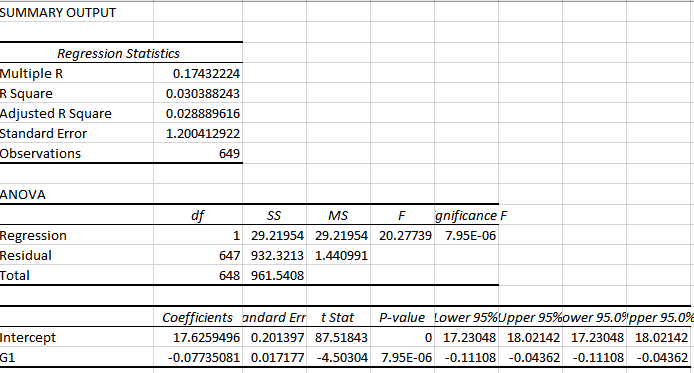
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S.no** | **Attribute** | **Mean** | **Median** | **Mode** | **Standard deviation** | **Interquartile Range** |
| 1 | Age | 16.74 | 17 | 17 | 1.218138 | 2 |
| 2 | Famrel | 3.931 | 4 | 4 | 0.9557169 | 1 |
| 3 | Dalc | 1.502 | 1 | 1 | 0.9248344 | 1 |
| 4 | Walc | 2.28 | 2 | 1 | 1.28438 | 2 |
| 5 | Health | 3.536 | 4 | 5 | 1.446259 | 3 |

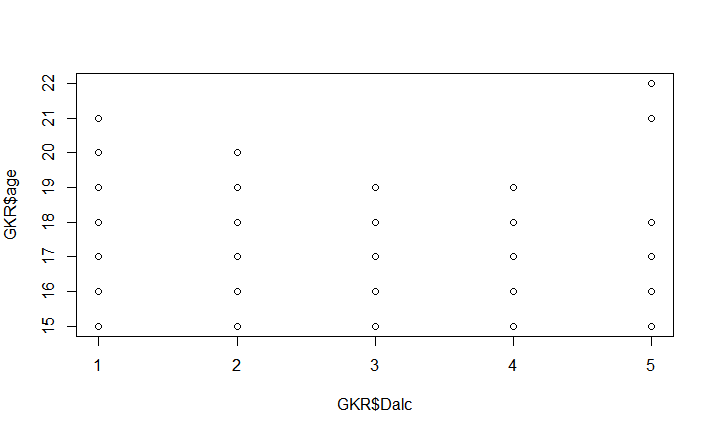
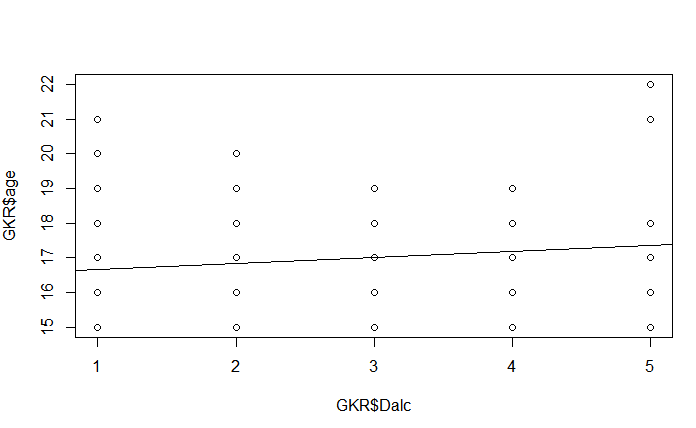
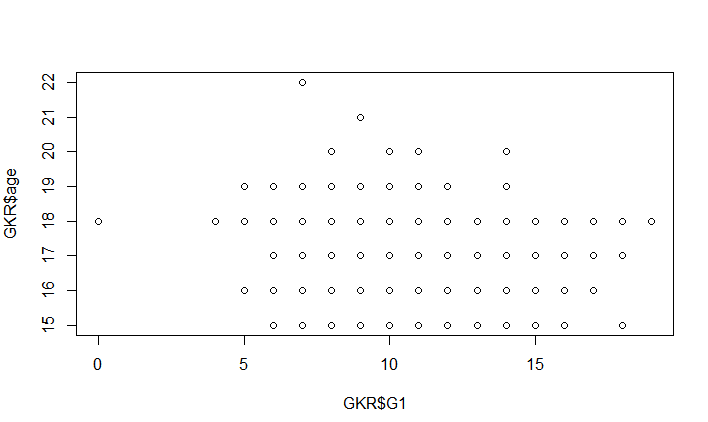
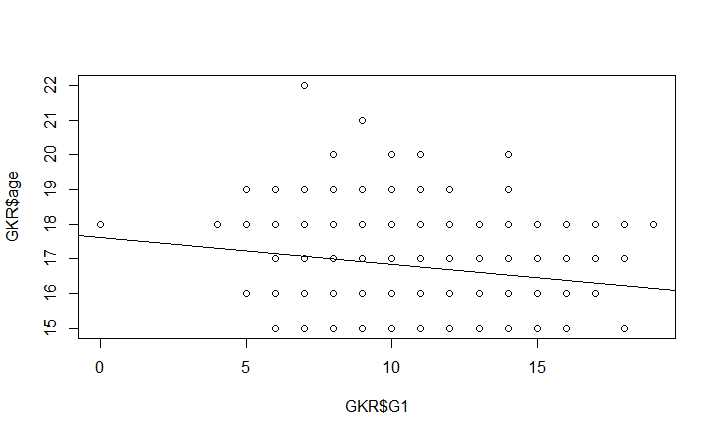
Health- Seeing alcohol consumption rate, what can be the health status of the people can be inferred, so that we can have a good society can be build from this.

As per the attributes, we find that there are no data value missing, but taken in consideration, if we think age value is missing, we would like numeric value to be added with median, and if it is a nominal value of education, we would be adding a grade of graduation to 16 because as per graduation the minimum requirement is 16, so like that we can make additional amendments to it and change nominal to numeric value. These Measurements are been important for to get all more clear results with inference of data , and make a more clear output with it.

**Prepossessing Analysis:**

1. What preprocessing was needed for what attributes?  What replacement strategies for missing values were used and why?  --10 pts

Taking student grade on basis, and what best features relates to fluctuate the grade of the student is making us to identify the best three attributes in the dataset. For this we are identifying that DALC, G1, age are the best three attributes which we consider as the independent attributes because by using these attributes, it derives to a conclusion of the grade of the student how it is getting fluctuating. As a matter of fact, In Nut shell, it is been observed using common sense that student how is having age less than mandatory age required, it is effecting the grade value, absence of the student. Students those who are drinking on daily basis are related to Grade down grade in fact health of the student also decreases which means by using one attribute we can anticipate a relation with other attribute to identify the accurate results from the data. We identify the age as a primary attribute where it get inter relates to other attribute using common sense. Also not only using common sense, we can identify the relation using R Code and MS Excel to identify how best the attributes get related.   
  
  
Example:  
  
  


  
  
  
  
  
  
  
  
In the Group Data Set , We try to identify if there are any missing values in the student alcohol distribution, and we identify there are no missing values, and there are no duplication of values. However, there said to be some nominal values where we managed to make it with numeric values, which we considered to identify the summary of all the fields, and find there is no duplication of the content provide. As a matter of fact , in nut shell , trusting the data completely is wrong, so we ideloized some points with the results which we found in this, and we found some in consistency in the data, and where we try to make over or fill the data using the mean and mode value to get the accurate results which help the person to take a decision.

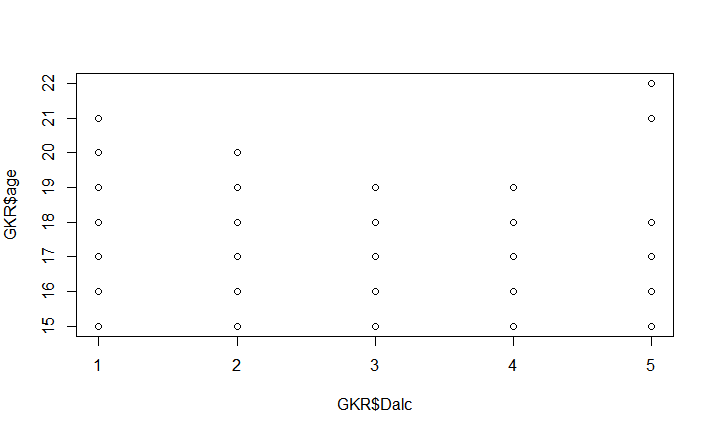
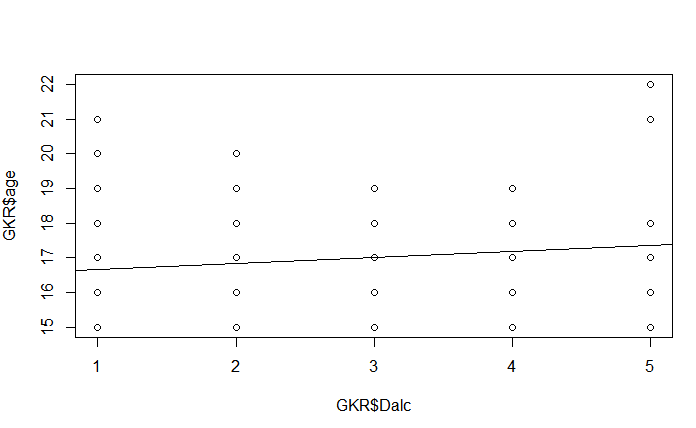
**Attributes Selection:**

1. Which are the three most important attributes?  Provide their basic statistical measures (see weeks 1 and 2) and explain what these measures tell you about the entire dataset. – 10 pts

We had first identified and made a respective identification of attributes to make a effective analysis in terms of and here are some of the classification we made to get to a idea of how did we processed our data, to get to a conclusion to the dataset.

In the Group Data Set , We try to identify if there are any missing values in the student alcohol distribution, and we identify there are no missing values, and there are no duplication of values. However, there said to be some nominal values where we managed to make it with numeric values, which we considered to identify the summary of all the fields, and find there is no duplication of the content provide. As a matter of fact , in nut shell , trusting the data completely is wrong, so we ideloized some points with the results which we found in this, and we found some in consistency in the data, and where we try to make over or fill the data using the mean and mode value to get the accurate results which help the person to take a decision.

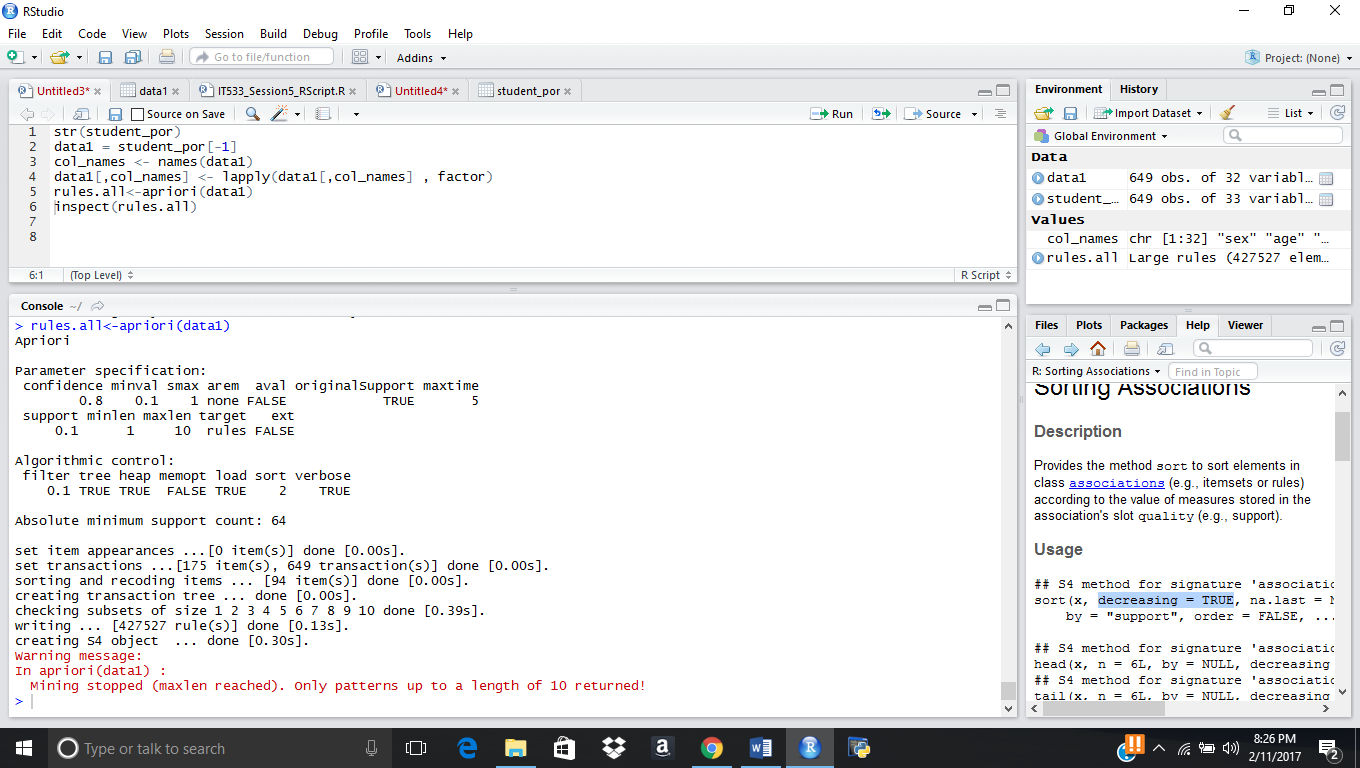
By this results, we inferred this that we can bin the results to categories, to make a best understanding about of data, and make them meaning ful even if the values are missing, so that it becomes easy to make over with the data. Taking three attributes which considered I nominal value so we moved the dat to numeric to get a meaning results which is easy in understanding, and the code is been added below.  
  
  
Show how the three most important attributes relate to each other and to the class attribute (if it exists in your dataset) using correlation and regression analysis.  What does this analysis tell you about the entire dataset? – 10 pt

  
  
  
  
  
  
Taking it with above specified results, we can identify that as on the graph display that as on the value as on the both the value increases, where they become directly proportional to each other and the value identifies they are dependent with each other with class attribute, and in order to disprove H0 and make an alternative hypothesis to make and near to results derived, and keep the confidence value of the data to be 95% , and the variance of the data should be less 5% which resembles with less than 0.05 value.   
  
In order to predict the future values in the selected attributes , where each attribute would be considered as interdependent in this data set, we consider the Df Sum Sq Mean Sq F value Pr(>F) values in consider where it can identify the trust of the data if the p value is than 0.05 , and we can have r square value to be used to predict the future values.  
  
  
Taking the R square value of the desired attributes we can have future value can be predicted.

**Association Algorithm:**

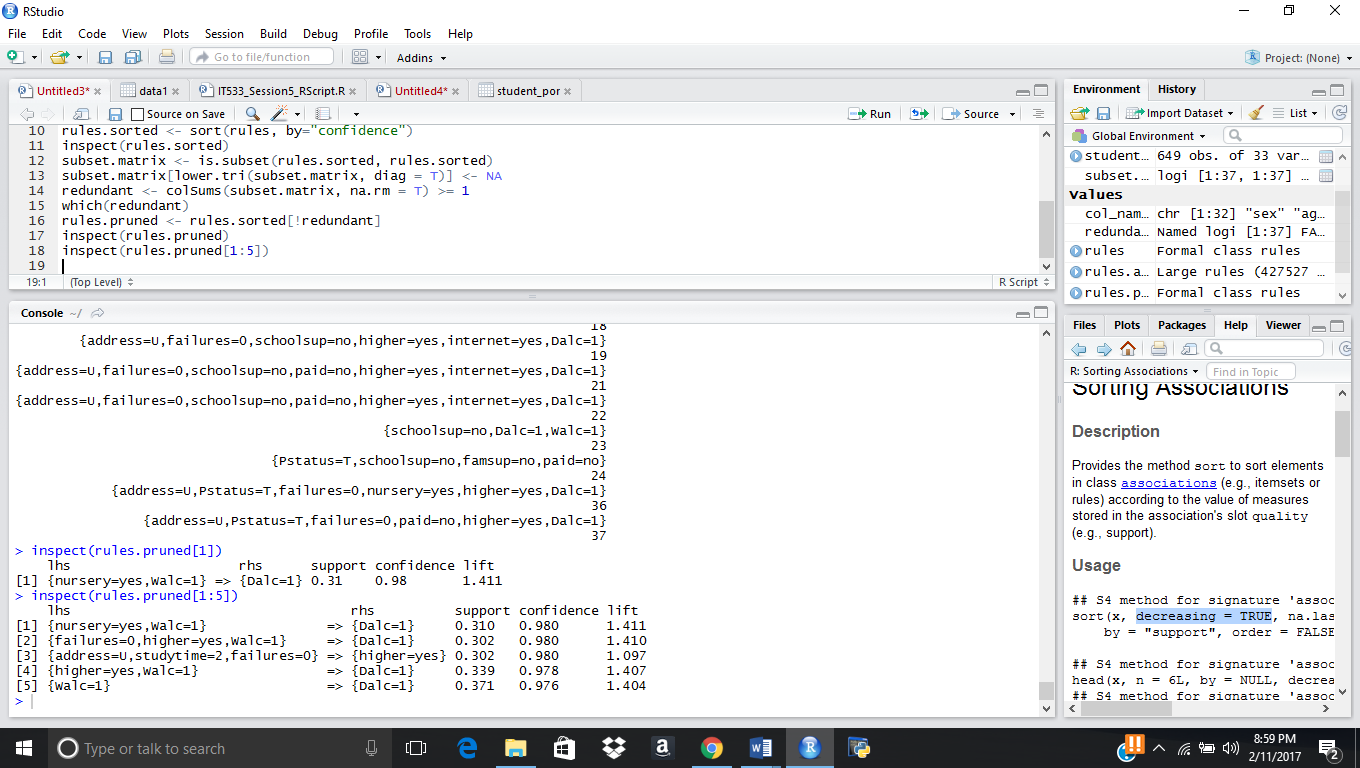
1. Which association algorithms and tuning techniques did you use and what information did they give you about your dataset?  – 10 pts

Student Alcohol Consumption is consisting with 649 observations and with 33 variables, where the data set was related to student education and grading, which factors are impacting the grade of the student. In fact, to make a closer relation and define aa best relations and minimize the vision of the dataset, we identified the cor and cov of the dataset and identified the best related attributes which helps in identifying the best output helps the students to improve the society to make a student and make government to implement best practices to make student educated. We used Chisq test , and provided the ab line to get the best relations and which what factors relates to best matters relates to result. Now as the dataset, being a good minner, it is always important to reduce the dataset to increase the dataset and observation, to make the dataset most effective, to get effective results and in order to identify which value collides and interprets best value in resolving best results, to make them most effective, and make them identified with apriori algorithm and generating rules based on confidence and support we can identify the best attributes as a group best related and make them results with best results regarding the question asked by people and government. In this Dataset, we had taken min len as 2, and support to 0.3 and confidence s 0.97 because to resolve with minimal results reason was to get as much as less results and make them effective to output and if the results are less collecting more attributes then we consider that your are making on right track for results. We made couple of combination with length, support and confidence and identified this identifies the best minimal results which would suffice the requirement.

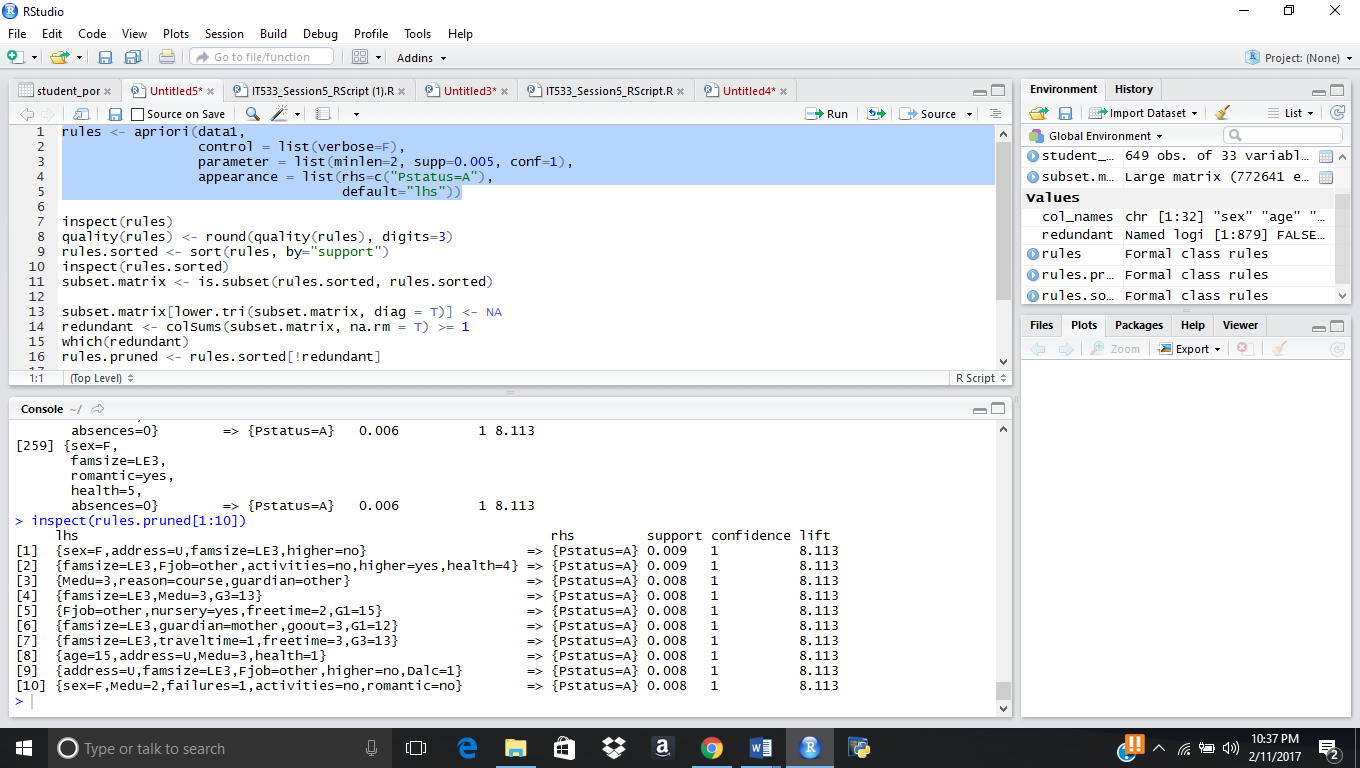


After applying the rules of apriori algorithm it was resolved with this output which was identified with screenshot when condition is added.

After the redundancy it was identified that it was decreased to 22 attributes and where when cor and cov was been identified we where able to interpret the same output what we are able to do with this algorithm and results with length of 2 and redundant value was been removed.



In order to identify best results for students we had identify with health factor of student how it is impact with education and grade of the student, if the health is 5 and then taking with 5 values of the output , we identified with confidence is 1 which mean student with high health is having to have good grade, and now we can use common sense, which factor defact the health to make grade reduced. Attaching the screenshot and script we visualized.



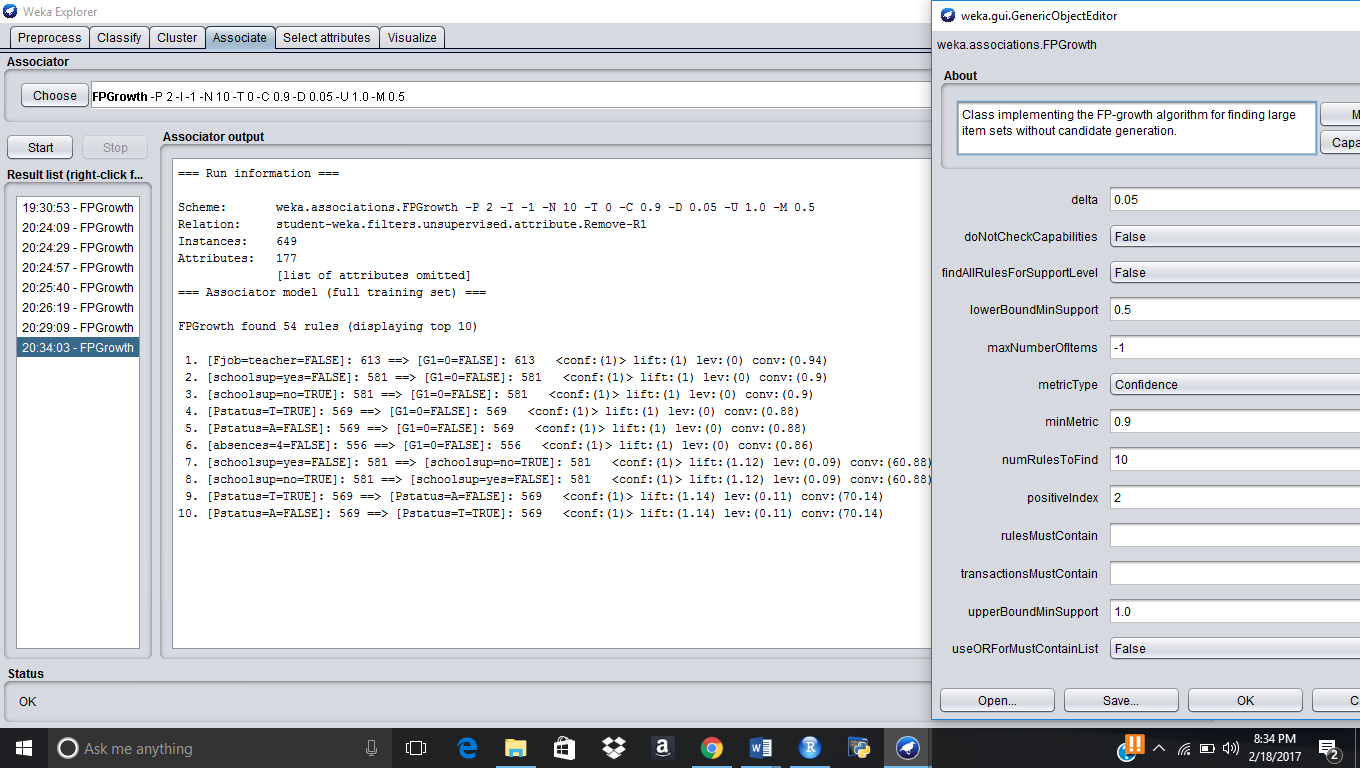
Above are the results with pruned and make them with closer values and used for interpreter by changing with different attributes confidence and support with best association rules.

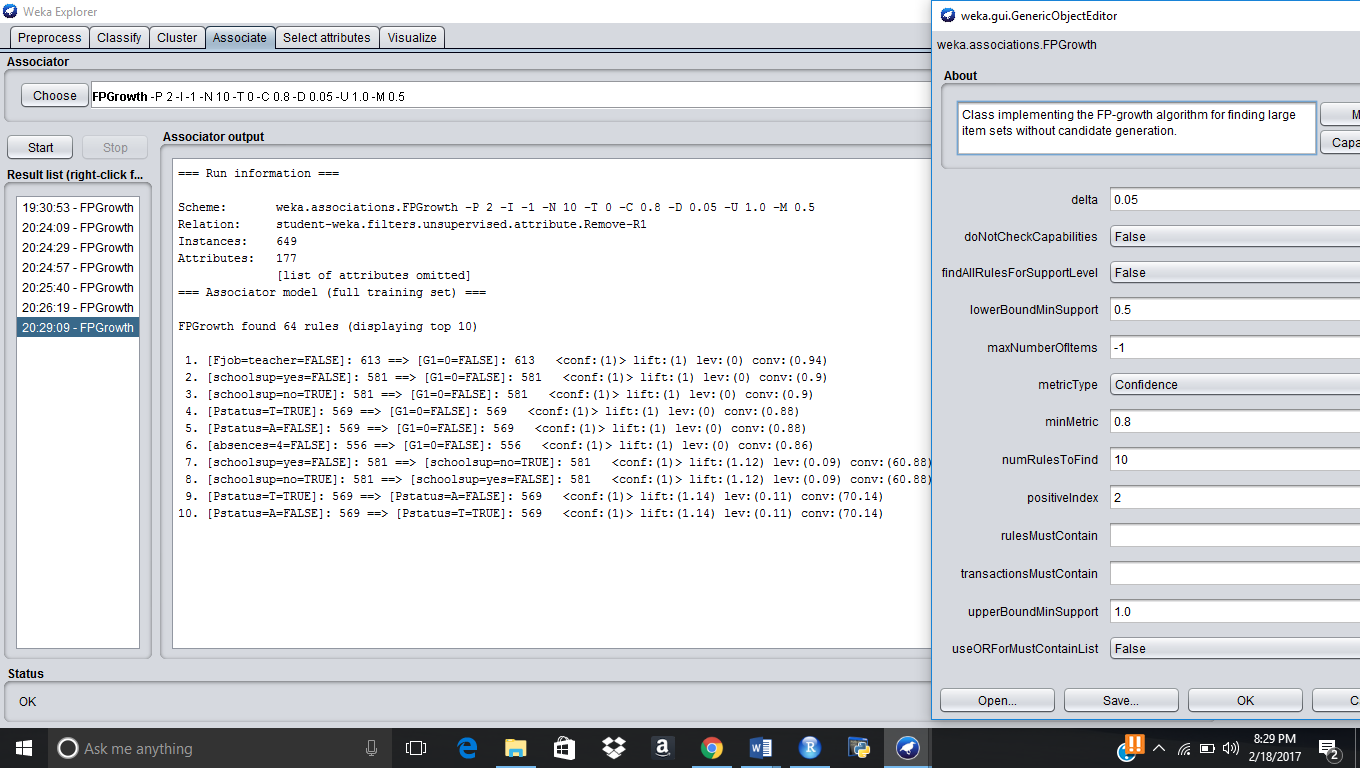
1. Which classification algorithms and tuning techniques did you use and what information did they give you about your dataset?  – 10 pts

Apriori Algorithm is a very well astonished used association learning which is highly considered on a dataset with its data values and association with its each eqvivalance provided to each value associated. It would have a greater support and confidence between each value of heavier dataset internally dependent of each attributes and for that we used Apriori for it’s effective utilization and for that we would reach for effective results and also effective output analysis based on the results.

So, for our dataset, Student alcohol Consumption and then we had used Dalc and walc as a effective attribute and which help in identifying with how it would impact a student with regards of student grade, and where what other more factors relates with the best analysis in correlation with the attributes. In order to identify it varies with how best the we vary the confidence and support such best we can identify the best related analysis regarding dataset with each attribute.

By using FP Growth , we are trying to identify the other different attributes to have them better related, and where FP Growth would be considered for smaller Datasets, in fact which would not be helpful most interactive databases.  
  
  
**Fp growth Analysis:**



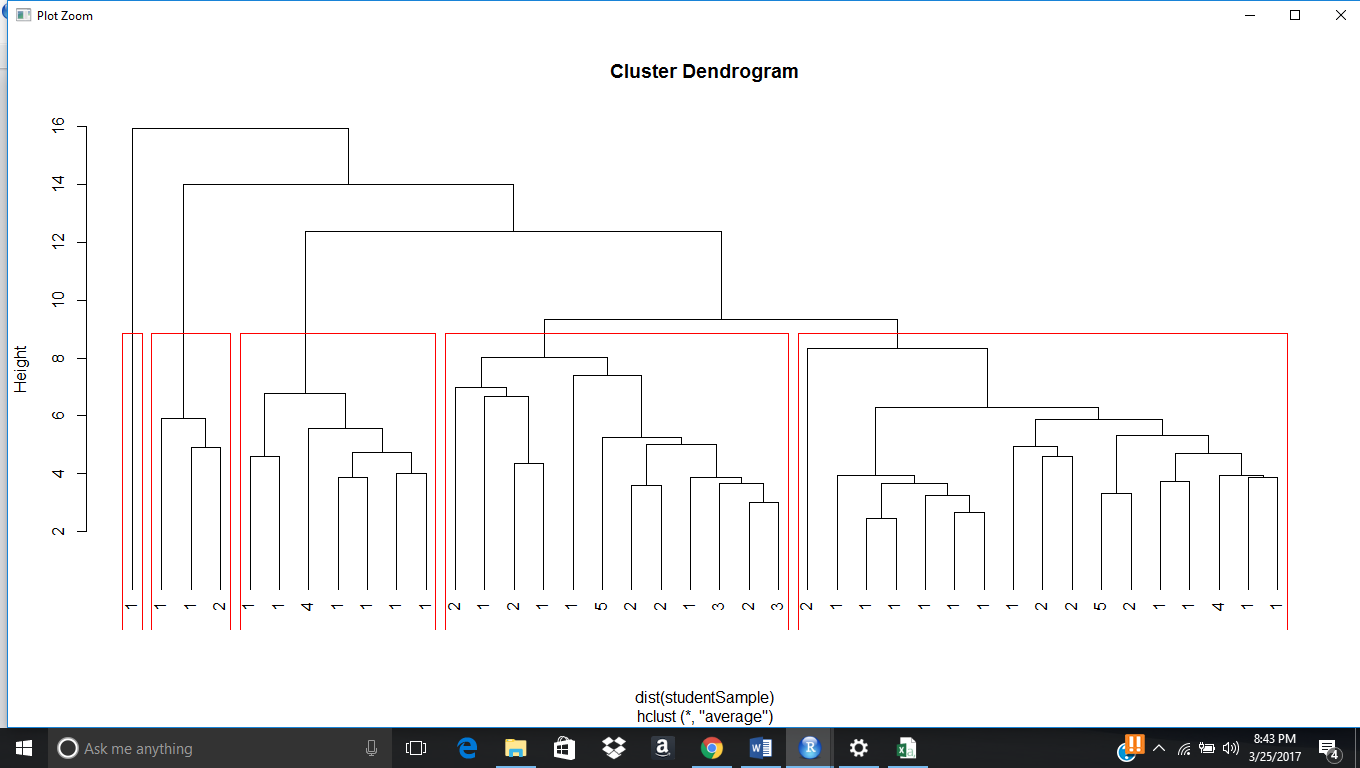


**Clustering Algorithm Analysis:**

1. Which clustering algorithms and tuning techniques did you use and what information did they give you about your dataset?  – 10 pts

Student Alcohol Consumption is our dataset , which is regarding student alcohol consumption, and where which could interrupt the education with grade of education , what factors would interrupt the education would interrupt the education and grade of the student. In the dataset, and for the dataset, we would like to have a different cluster to maintain what does a machine try to learn when I give the dataset.

Number of clusters which was created with the dataset by the machine by using different algorithms as given below, and it was 5, and different values are been provided by different clusters.



1. Which algorithm produces the best results?

Generally density algorithm would generally best results for dataset, and then in our results, we find hierarchical density is providing most effective results, and then the results in terms of graph, it would be a effective results for our dataset,

idx <- sample(1:dim(student4)[1], 40)

studentSample <- student4[idx,]

studentSample$Alc <- NULL

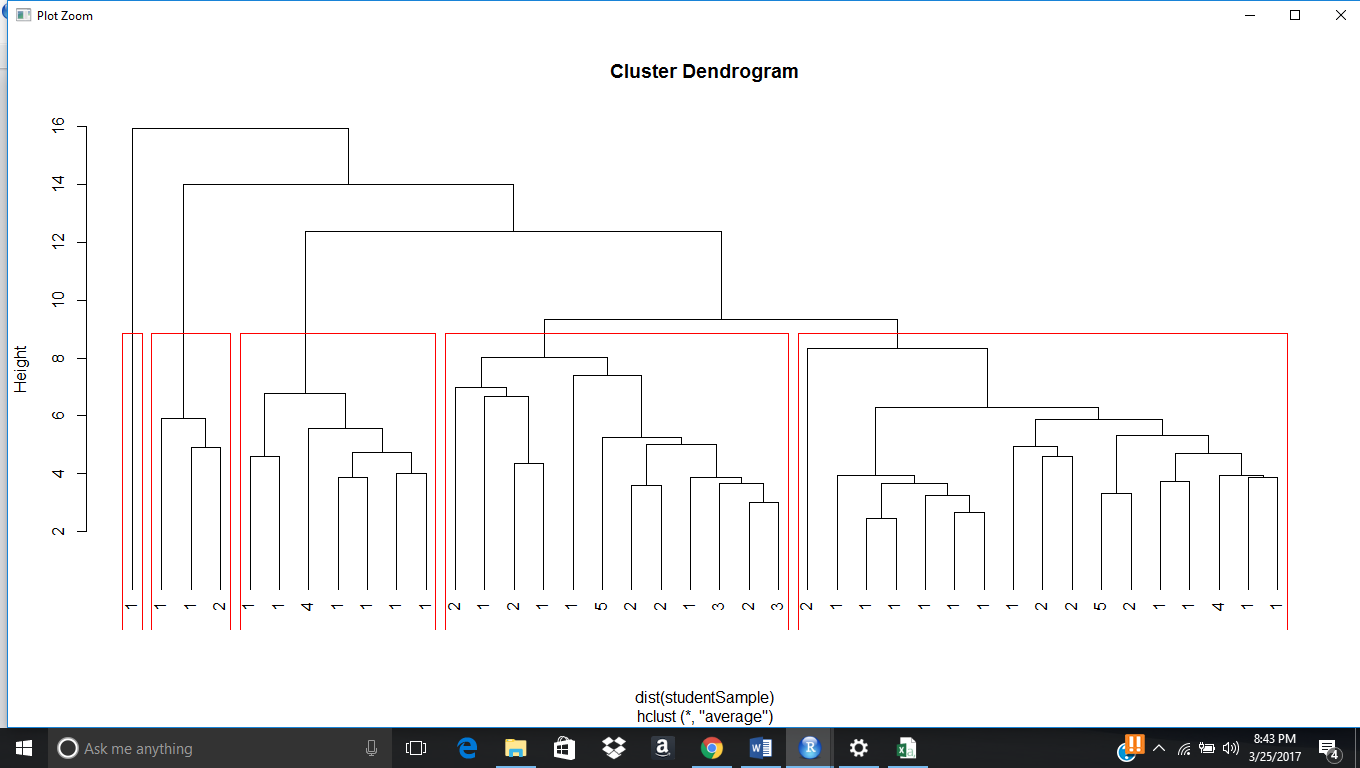
hc <- hclust(dist(studentSample), method="ave")

plot(hc, hang = -1, labels=student4$Alc[idx])

## Let's cut the tree into 3 clusters

rect.hclust(hc, k=5)

groups <- cutree(hc, k=5)



Our Dataset was regarding with student alcohol where how much each factor creates impact for student grades by taking drink, and in our dataset we are having 32 attributes, and by having drinking by weekdays and weekends how much famrel and health would impact he student education in terms of grade, so in order to get a forecast of what attributes best relates to the ideology of the dataset, and then taking attributes in consideration, we are having,   
  
school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason guardian traveltime studytime failures schoolsup famsup paid activities nursery higher internet romantic famrel freetime goout Dalc Walc health absences G1 G2 G3  
  
these are the attributes of our dataset, and for this we made Dalc and Walc are the basic class attribute which we thought and for we made the close analysis what best can lead and best decision for the dataset to provide best results, and for this we made a realistic analysis using forecast analysis to best replicate with clustering, and association analysis, and for this we got results in this way.   
  
**Data Interpretation:**

1. Interpretation of results:  Which data mining method used in points 3-7 provides the most useful and important information about your dataset? —10 pts

Identifying the best data mining method for our analysis was very difficult as because in each sector it was helping us in many ways, but in all the above data methods we provided we are been helped with clustering, Kmeans, Apriori , Naïve bayes, which has helped us in identifying the best analysis on student alcohol dataset. But, it was one situation where Apriori and clustering was such algorithm which has provided accurate with the random input taken. We used to use FP Growth to get 95% trust over the dataset, and make it used for larger dataset, and increase the trust over the dataset.

**Overall analysis and design:**

1. What action would you recommend management to take as a result of your data analysis, and where are the limitations of your analysis?  Use at least 5 sentences to explain how you arrived at your recommendation, which managerial conclusions from your analysis would be useful, and which ones would be erroneous. – 20 pts

* Our results decide that we would get good understanding that machine learning helped us in identifying what best clusters takes in consideration, and how much data points will have best and near distance, where it creates a impact in terms of making student education, and where a government can take a effective decision in terms of analysis which it brought. Our results relies what our forecast we made like what made that sex, health Dalc and Walc are created with best impact with student grade and it has got provided with the latest algorithm test which we made to identify and best interpret what factors impact student grade. By using this attributes and results a political can be made like   
    
    
    
  First, student grade is impact people are drinking more at weekends, so government can have them stop selling drinks in the weekend.
* Clusters which has given that student health is also impacting the student grade in terms of student grade.
* People with good famrel is having good education and number of datapoints which is there in there third clusters best implies how much relation is given.
* There are considered to be many more relations, but we find many more points what best reactions can be taking government and political, but as per business we have like if we stop drinking that would impact government and people investing in drinking and liquor factory get away, so we can have one decision as none of the drinking shop should be nearby school, so that atleast this can stop people coming to shop to by drinks
* Some of the decisions said to be as without having time given to student for education, it helps student to drink more, so they can start time in the schooling as study time to make them avigated with some other work for education which make student to decrease intke of alcohol. Family time is something very important, so to build more parks for students so that they can spend more time with family, and events for students so that they can time with family.
* . Health is one factor, to make a regular check up in school to check student health checkup done on monthly basis. Providing more jobs for women would also make student to get inspired from child hood and make them start working.