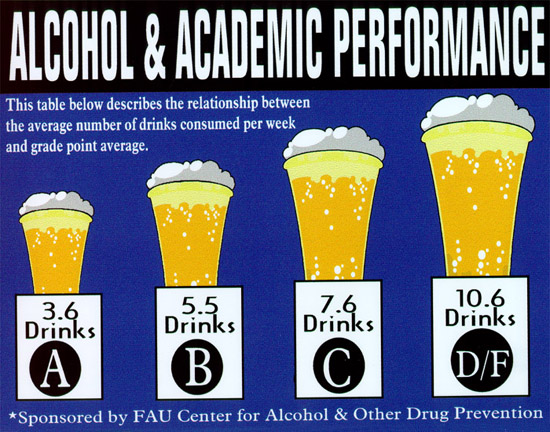
DATA MINING

****

SUBMITTED BY :

Asish Pabbisetty

Hamza Siddique Ahmed

Shweta Batra

Lakshmi Sunanda Nagareddygari

**Contents**

* [OBJECTIVE AND PROPOSAL : 2](#_Toc25094321)
* [DATASET USED : 2](#_Toc25094322)
* [Attributes of Merged1.csv dataset: 2](#_Toc25094323)
* [MODELS 4](#_Toc25094324)

[**1.** **CLUSTER ANALYSIS:** 4](#_Toc25094325)

[Final Conclusion from Cluster Analysis : 7](#_Toc25094326)

[2. LINEAR REGRESSION : 7](#_Toc25094327)

[Results from linear regression model: 10](#_Toc25094328)

[3. DECISION TREE: 11](#_Toc25094329)

[Output of the decision tree is shown in above diagram. 14](#_Toc25094330)

[CONCLUSION FROM DECISION TREE: 15](#_Toc25094331)

[4: LOGISTIC REGRESSION: 16](#_Toc25094332)

[Odds Ratio: 18](#_Toc25094333)

[Confusion matrix: 18](#_Toc25094334)

[Accuracy of the model 18](#_Toc25094335)

[4.NEURAL NETWORKS: 19](#_Toc25094336)

[Managerial Implications and conclusions 20](#_Toc25094337)

**IMPACT OF LIQUOR COMSUMPTION ON GPA**

# OBJECTIVE AND PROPOSAL :

Our project will aim to determine how alcohol consumption influences the GPA of two groups of students taking Math and Portuguese. Our target variable will be GPA and the predictors can be workday alcohol consumption and weekend alcohol consumption, as these are the only two attributes involved with alcohol consumption. Other variables could affect GPA as well, but we will be focusing on alcohol consumption.

We aim to use methods like linear regression, cluster analysis and other techniques we learn in and out of class, to develop models. Moreover, as the age group in this data set is between 15 to 22, this data set is interesting to us being college students as well!

# DATASET USED :

For this Prediction we are using second hand data obtained from the below source <https://www.kaggle.com/uciml/student-alcohol-consumption>

# Attributes of Merged1.csv dataset:

1. School - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

2. Sex - student's sex (binary: 'F' - female or 'M' - male)

3. Age - student's age (numeric: from 15 to 22)

4. Address - student's home address type (binary: 'U' - urban or 'R' - rural)

5. Famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

11. Reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12. Guardian - student's guardian (nominal: 'mother', 'father' or 'other')

13. Traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14. Studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15. Failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16. Schoolsup - extra educational support (binary: yes or no)

17. Famsup - family educational support (binary: yes or no)

18. Paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19. Activities - extra-curricular activities (binary: yes or no)

20. Nursery - attended nursery school (binary: yes or no)

21. Higher - wants to take higher education (binary: yes or no)

22. Internet - Internet access at home (binary: yes or no)

23. Romantic - with a romantic relationship (binary: yes or no)

24. Famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25. Freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

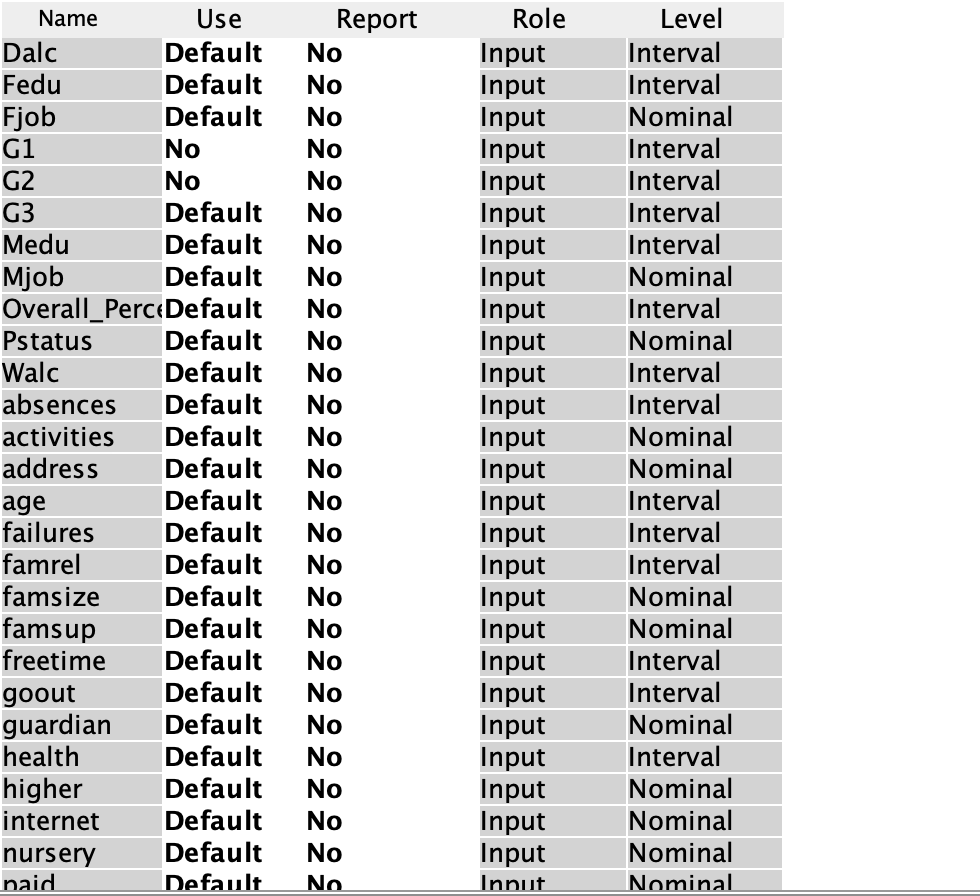
29. Health - current health status (numeric: from 1 - very bad to 5 - very good)

30. Absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese:

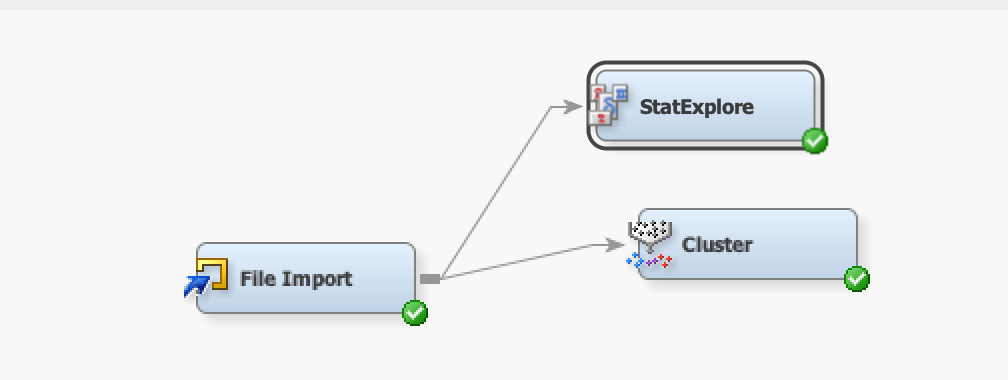
1. G1 - first period grade (numeric: from 0 to 20)
2. G2 - second period grade (numeric: from 0 to 20)
3. G3 - final grade (numeric: from 0 to 20, Target Variable)

**NOTE : IN THE PROJECT WE WILL BE NEGLECTING G1 AND G2 ATTRIBUTES BECAUSE THESE TWO VARIABLES HAVE HIGH COLLINEARITY WITH THE OUTPUT VARIABLE G3.**

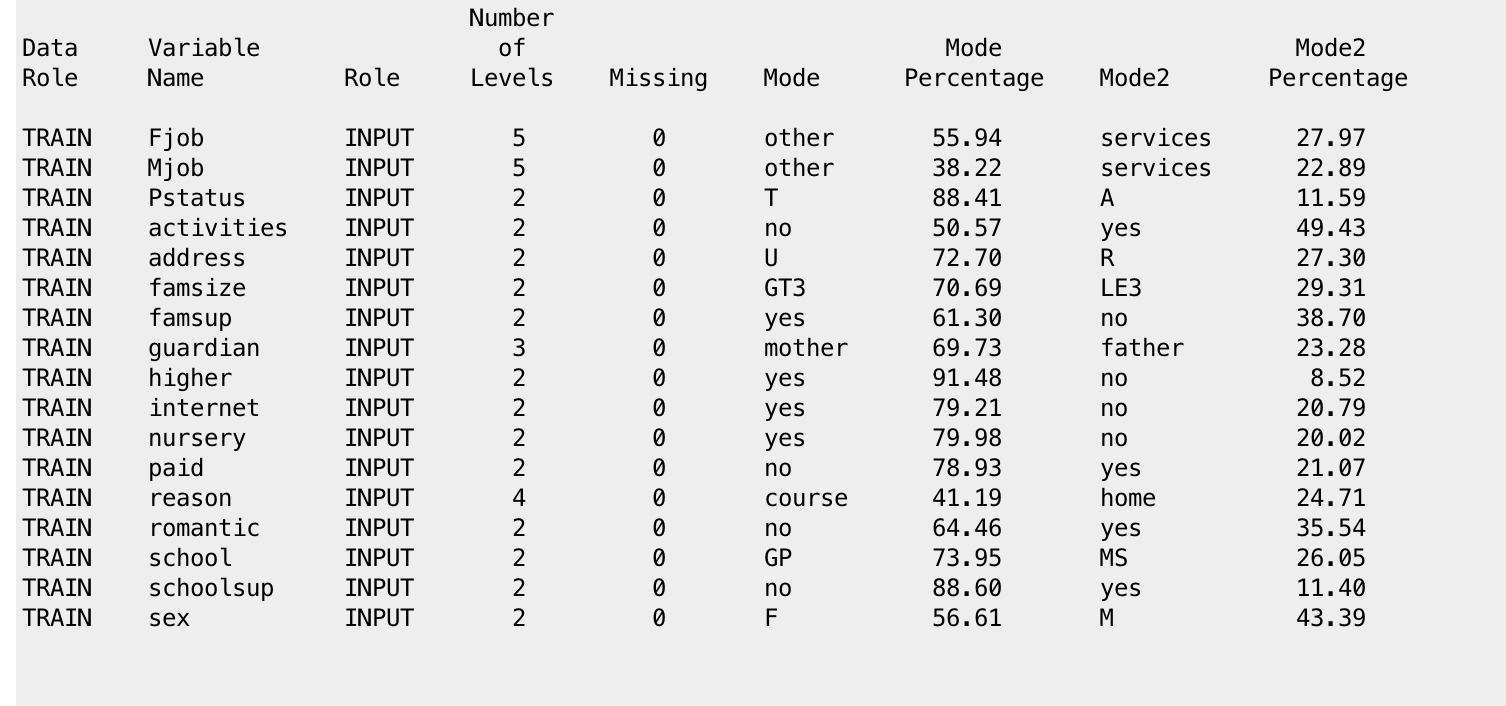


# MODELS

## **1. CLUSTER ANALYSIS:**

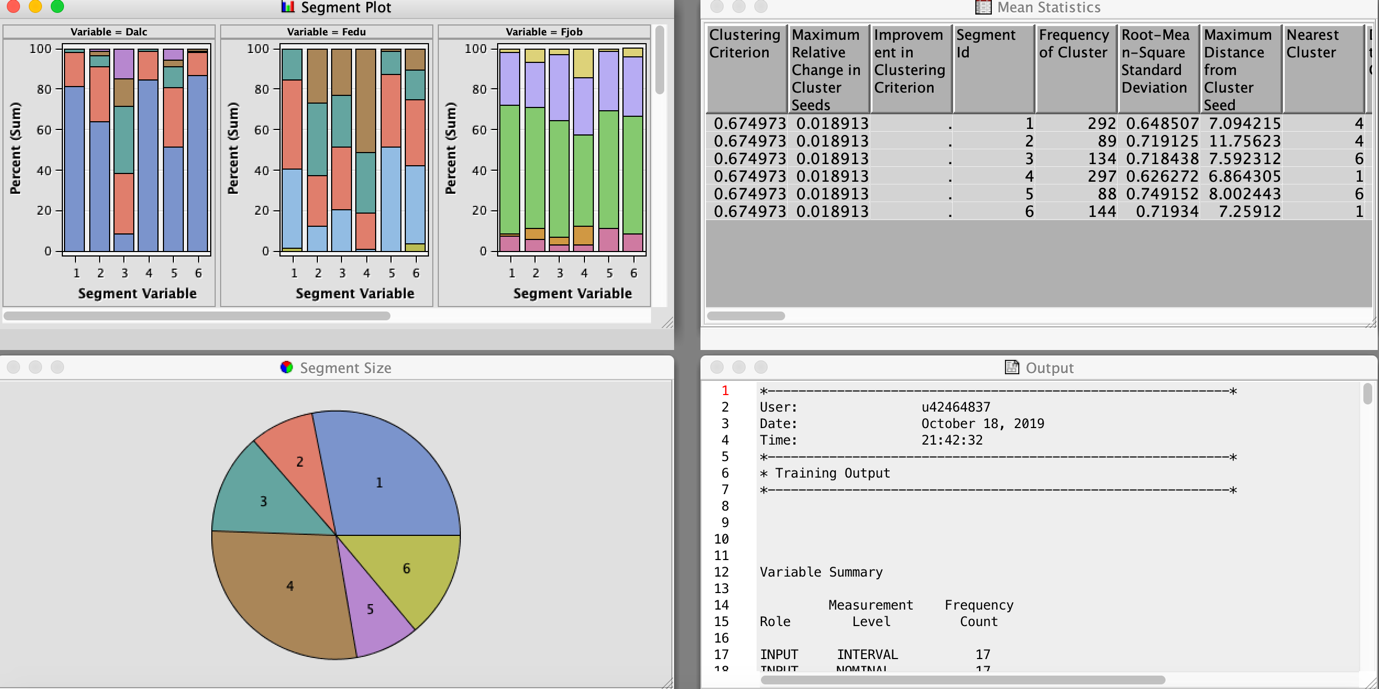


As shown in this figure we will attach the cluster node to file import and in cluster analysis we will be rejecting the G1 and G2 attributes because of them having high collinearity with G3.

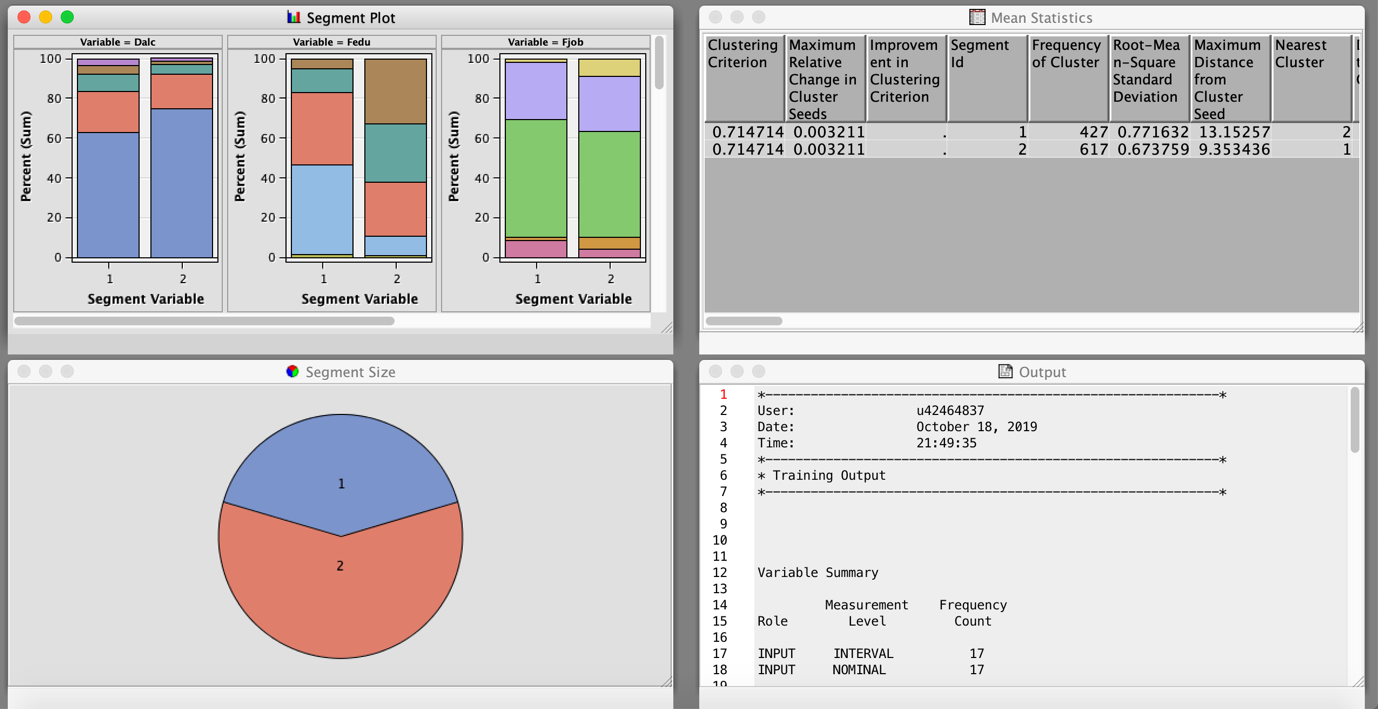
****

The above figure is the output from StatExplore node and data is trained data:

* This data doesn’t have missing values.
* Most of student’s father jobs and Mother’s job is mentioned as ‘other’ and the second highest kind of jobs is mentioned as ‘services’.
* Most of the students are shown to have family support
* Most of the students are from the school GP (Gabriel Pereira).
* Most of the students are not having school support.
* Most of students in trained data are Females

****

Initially we performed cluster analysis by specifying the number of clusters to be automatic. We found out that there are 6 clusters in it. The interpretation becomes difficult by using 6 clusters.



After specifying number of clusters = 2, we got to know information from the above diagram,

**From Segment plot:**

1. Failure Segment Plot:

* In Cluster 1, most of the students have 1 to 3 failure subjects
* In Cluster 2, the students have no subject failures until now but, some of the students might have failed in one subject.

1. DALC (Weekday alcohol consumption) Segment Plot:

* In Cluster 1, it shows the rating of student’s alcohol consumption on weekdays is mostly 2 to 5. This means that the students are consuming alcohol regularly.
* In Cluster 2, the rating of the student’s alcohol consumption is a bit low if compared to Cluster 1. This tells you that, the students in the Cluster 2 consume alcohol rarely during weekdays.

1. WALC (Weekend alcohol consumption ) Segment Plot:

* The weekend alcohol consumption of Cluster 1 is very high. The consumption rating of 57% of students is mostly greater than 3(out of 5).
* In Cluster 2, 92% of the student’s alcohol consumption is similarly high.

1. Absence Segment Plot:

* In cluster 1, the students tend to be absent more number of times.
* In cluster 2, we have only few people with absence in classes.

1. Sex(M/F) Segment Plot:

* In Cluster 1, there are 72% of male students whereas, In Cluster 2 there are more number of female students i.e 68%.

1. Study time Segment Plot:

* In Cluster 1, 93% of students have their study time between 1-2 hours and remaining students have study time more than 2 hours.
* In cluster 2, 30% of students have their study time more than 2 hours and remaining students have 1-2 hours.
* This shows that, Cluster 2 students study more than the cluster 1 students

1. Travel Segment Plot:

* In Cluster 1, students travel for more time as compared to Cluster 2.

1. G3 Segment Plot:

* In Cluster 1, 63% of students got the final grade below 10 (out of 20)
* In Cluster 2, 83% students got final grade more than 10(out of 20)

**According to Mean Statistics :**

By comparing both the clusters,

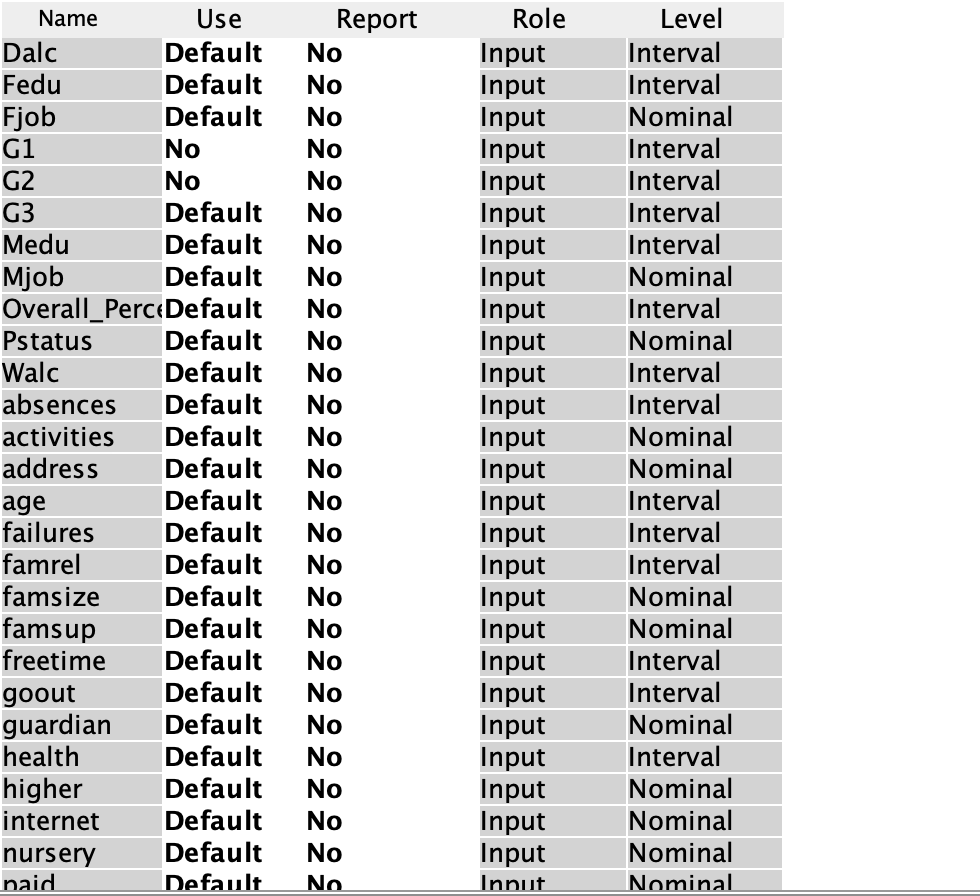
* DALC of cluster 1 is greater than the DALC of Cluster 2.
* G3 of cluster 1 is lesser than G3 of cluster 2.
* WALC of Cluster 1 is greater than that of cluster 2.
* Absence of students in cluster 1 is greater, compared to of cluster 2.
* Cluster 1 has more chances of failing in 1 subject compared to cluster 2.
* Cluster 1 students go out to enjoy more than the cluster 2 students.
* Cluster 1 has an average of 1.5 hours of study time and cluster 2 has an average of 2.2 hours of study time.
* Travel time for students in cluster 1 is more than cluster 2 students. It may also be the one of the reason reduced grades.
* Cluster 2 has more interest in Higher Studies than cluster 2.
* Cluster 1 has more free time than cluster 2. This can influence on overall grade(The students who have more leisure time or free time tend to get less grade).

### Final Conclusion from Cluster Analysis :

* CLUSTER 1 - ‘WEAKER STUDENTS’ : The students who have more free time, who drink regularly on weekdays as well as on weekends, who have more travel time, who have no internet, who don’t have school support tend to get poor overall grade.
* CLUSTER 2 – ‘BRIGHTER STUDENTS’ : The students who don’t drink regularly on weekdays as well as weekends, who have less travel time, who have internet, who have school support tend to get good overall grade.

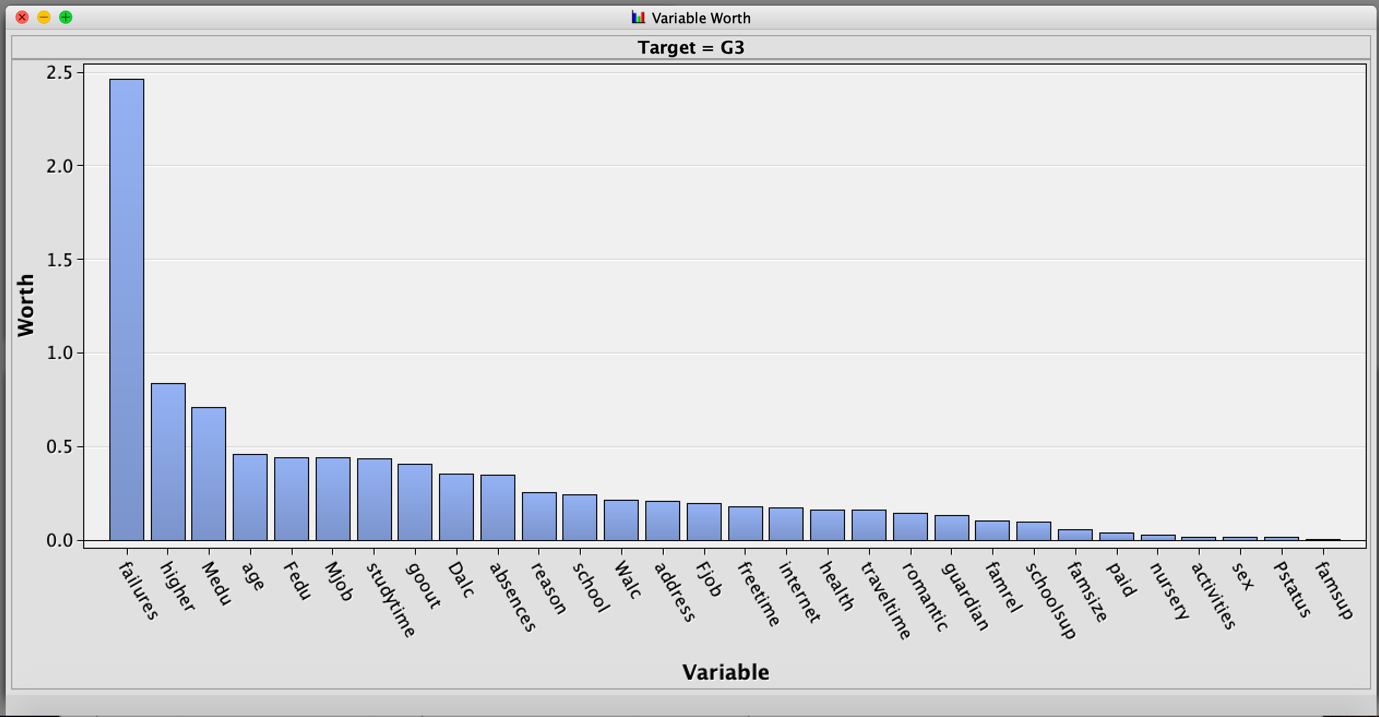
## 2. LINEAR REGRESSION :

**NOTE :** **IN THE PROJECT WE WILL BE NEGLECTING G1 AND G2 ATRRIBUITES BECAUSE THESE TWO VARIABLES HAVE HIGH COLINEARITY WITH OUTPUT VARIABLE G3.**



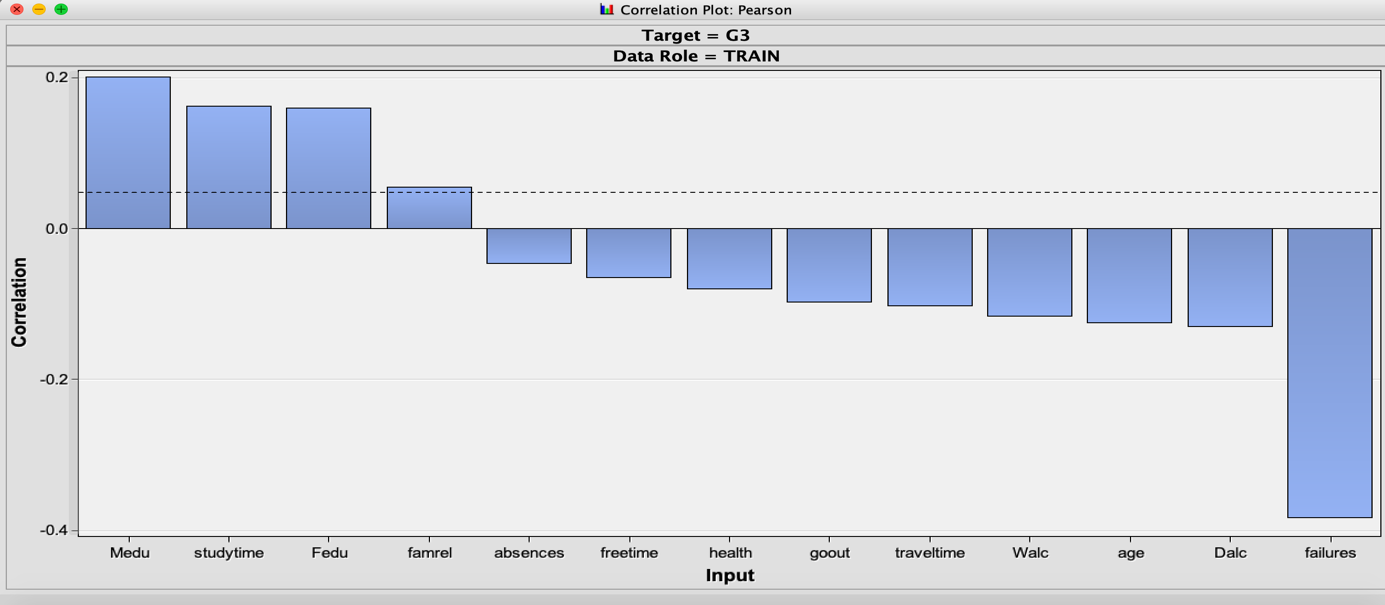
A close up of a device

Description automatically generated In Linear regression we used variable in selection node because it refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs to model.



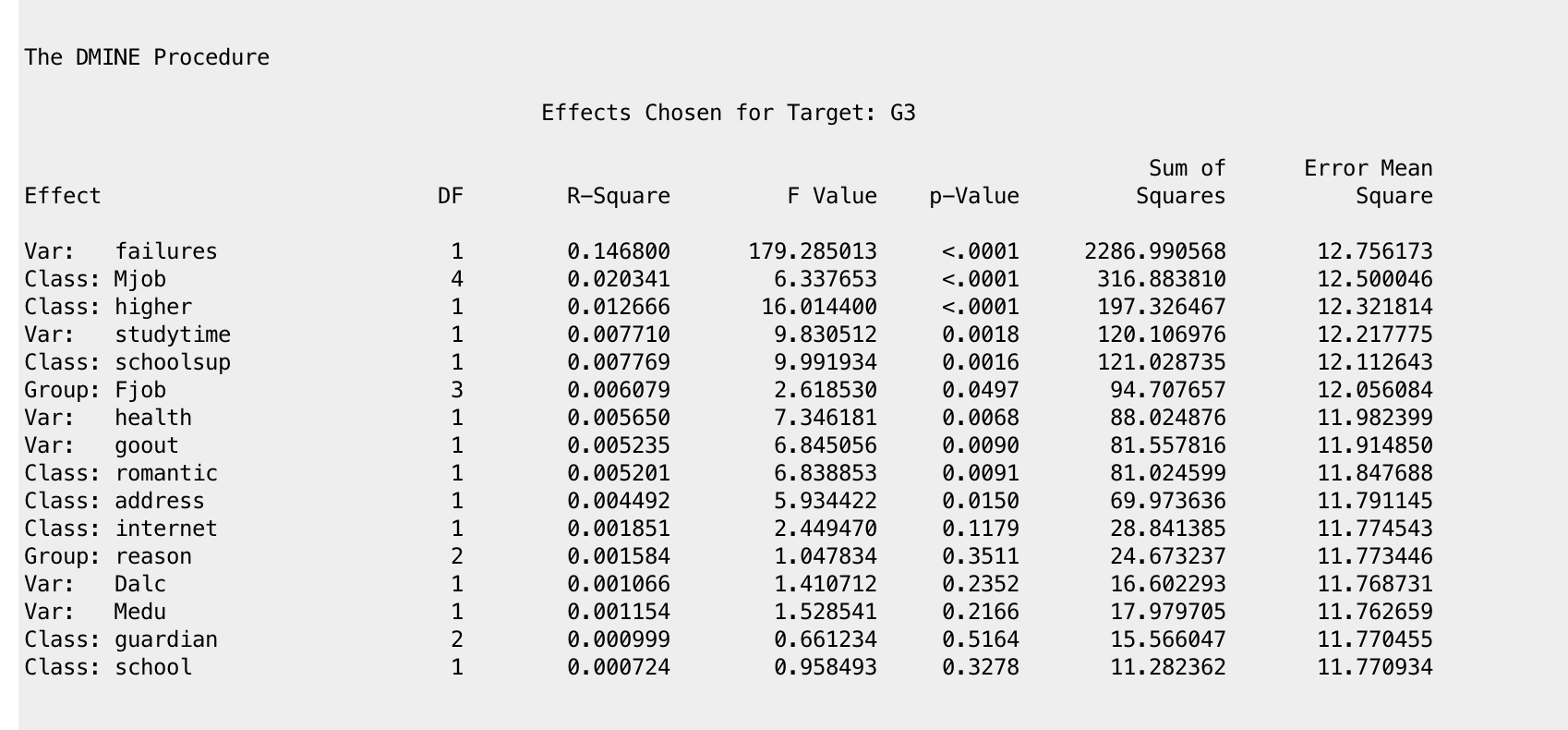
According to above figure,

* Failure is the most important variable among all the variables. Whereas the other variables such as paid, nursery, activities, sex, Pstatus, famsup are the lesser important variables.



According to Correlation plot of the dataset,

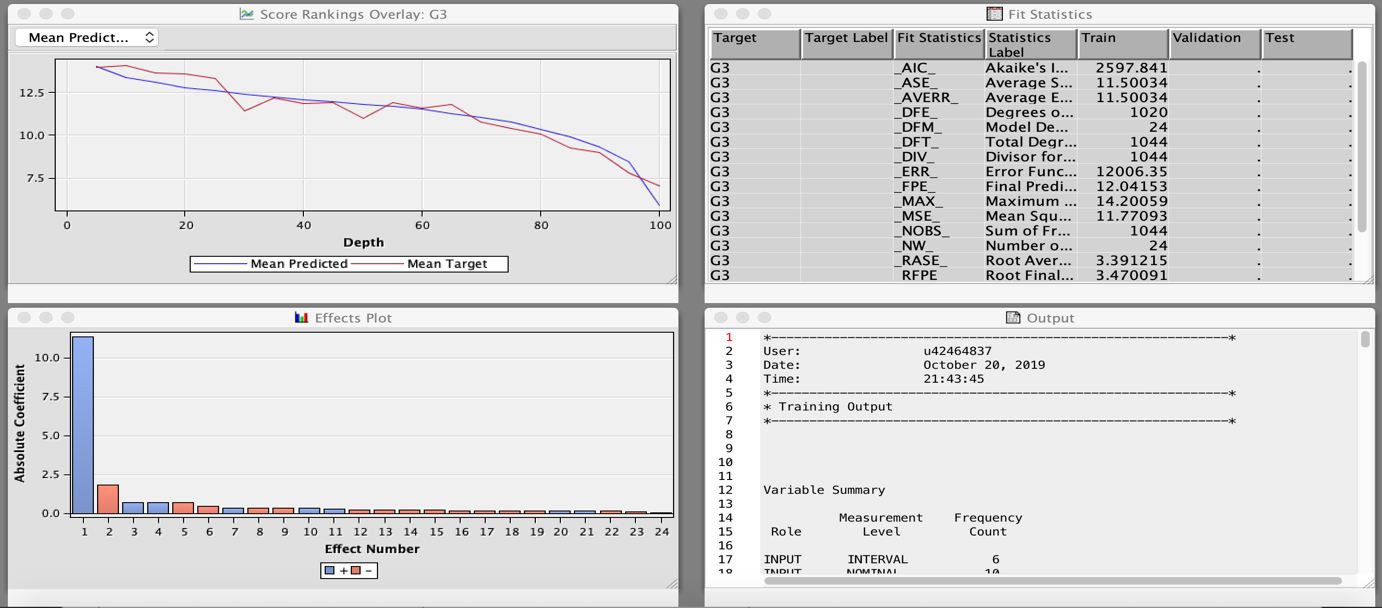
* Mother’s education (Medu) is highly correlated with G3(Target variable). Which means, if the student’s mother is education good then the student tends to get good knowledge and can get good overall marks(G3).
* The study time is also positively correlated with G3. Therefore, if we study for more time it can give good overall score.
* The WALC (weekend alcohol consumption) variable, which is negatively correlated, means if you have more alcohol consumption on weekends student’s marks tend to decrease.
* DALC (Weekday alcohol consumption) variable, which is negatively correlated. Which means, if you have more alcohol consumptions on weekdays, the student’s marks tend to decrease. If you compare WALC and DALC, Weekday consumption of alcohol(DALC) affects more than Weekday alcohol consumption.
* Failures are highly negatively correlated with the target variable G3. This means, the less the number of failures, more will be the overall marks and vice versa

****

**These are the most important variables for the model, which resulted from variable selection node.**

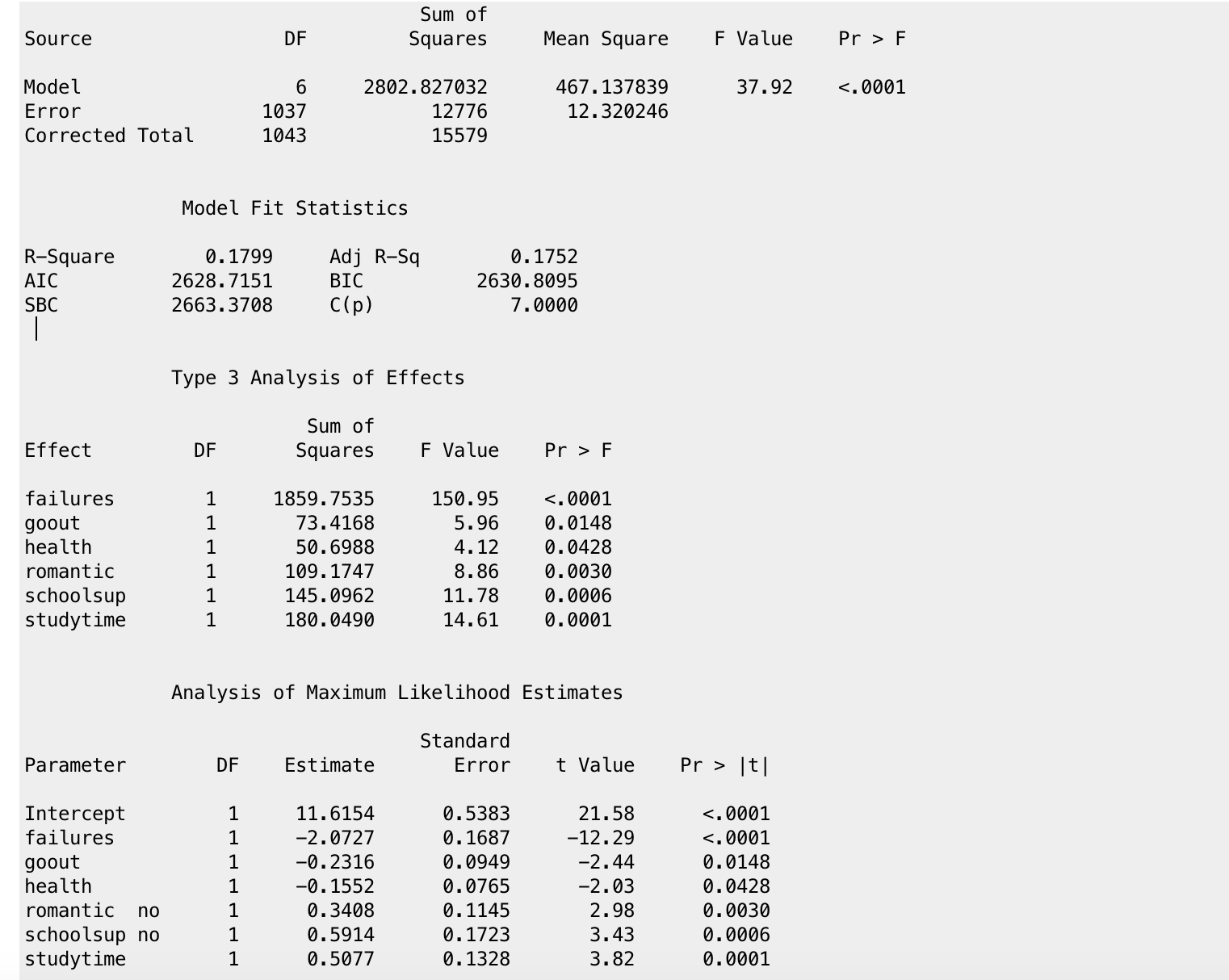
* Failures
* Mjob
* Higher
* Studytime
* School sup
* Fjob
* Health
* Go out
* Romantic
* Address
* Internet
* Reason
* Dalc
* Guardian
* School

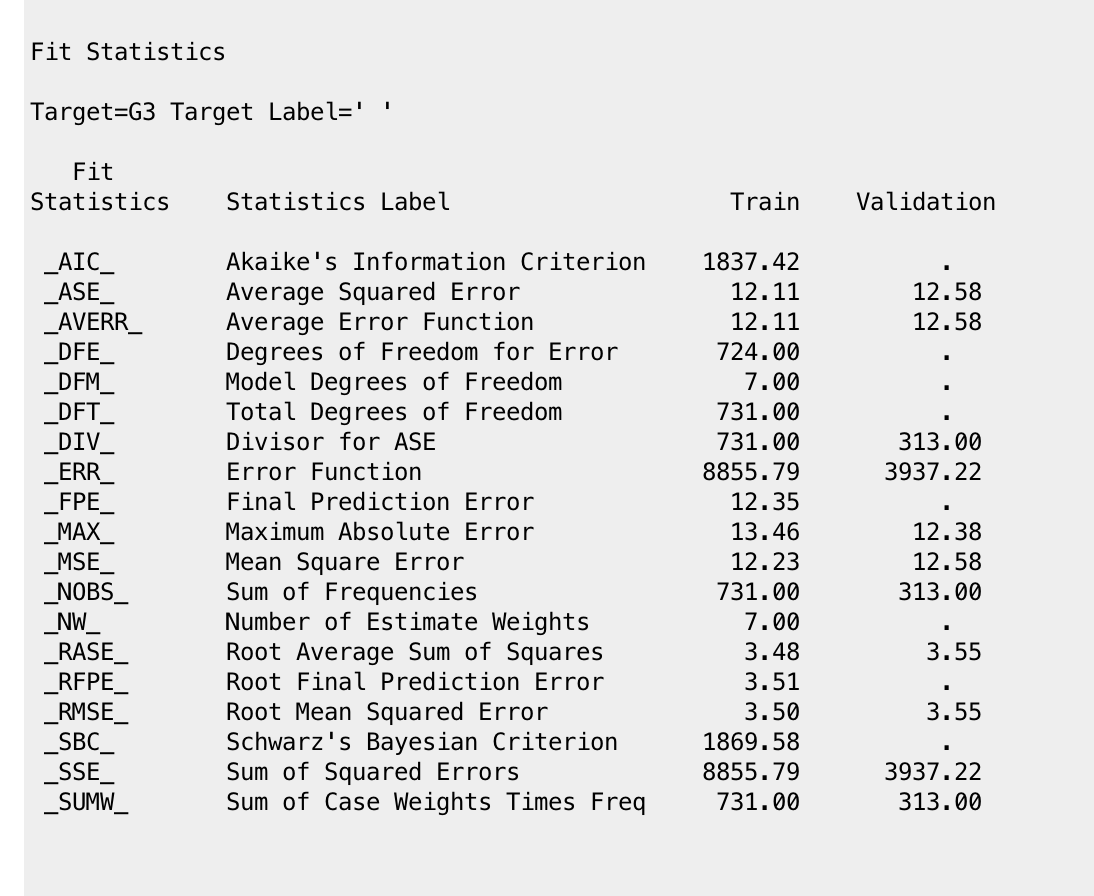
These variables are fed into linear regression model as inputs.

****

### Results from linear regression model:

After removing the insignificant variables from model. These are the significant variables shown in below figure.





Here, the F-test value is less than 0.05. Therefore, model is significant.

From these results, the significant variables are **failures, go out, health, romantic, schoolsup (school support), studytime.**

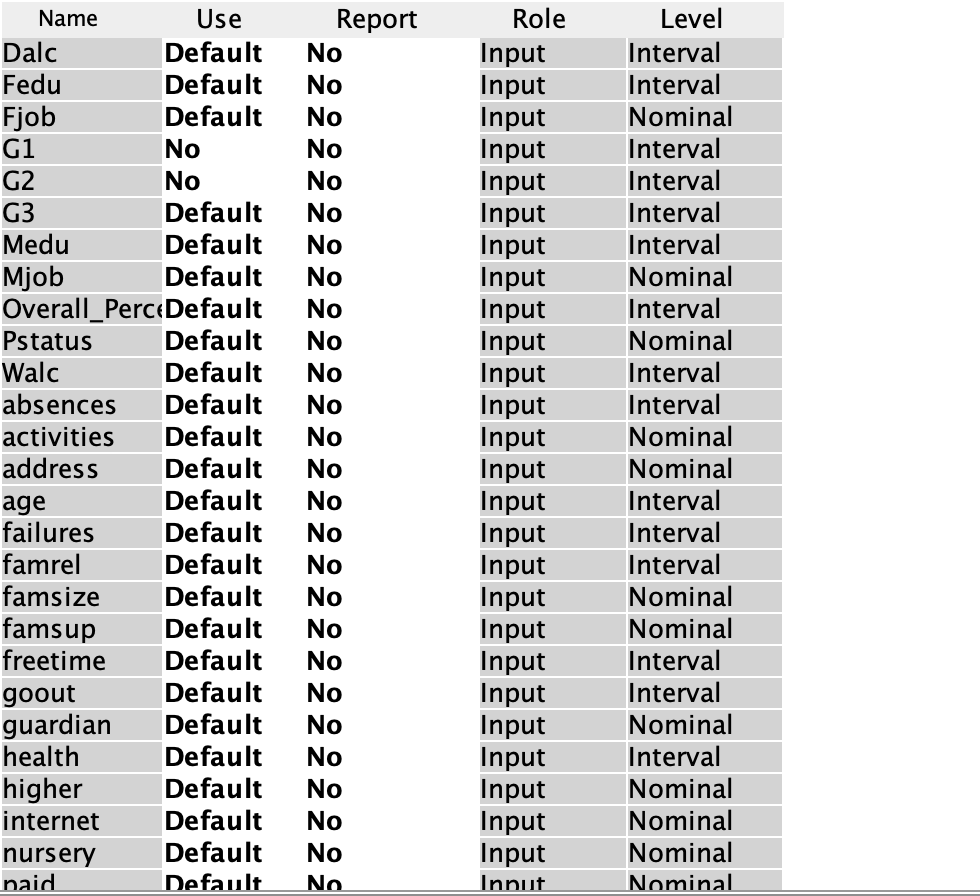
R square- 0.1799 Adjusted R square – 0.1752.

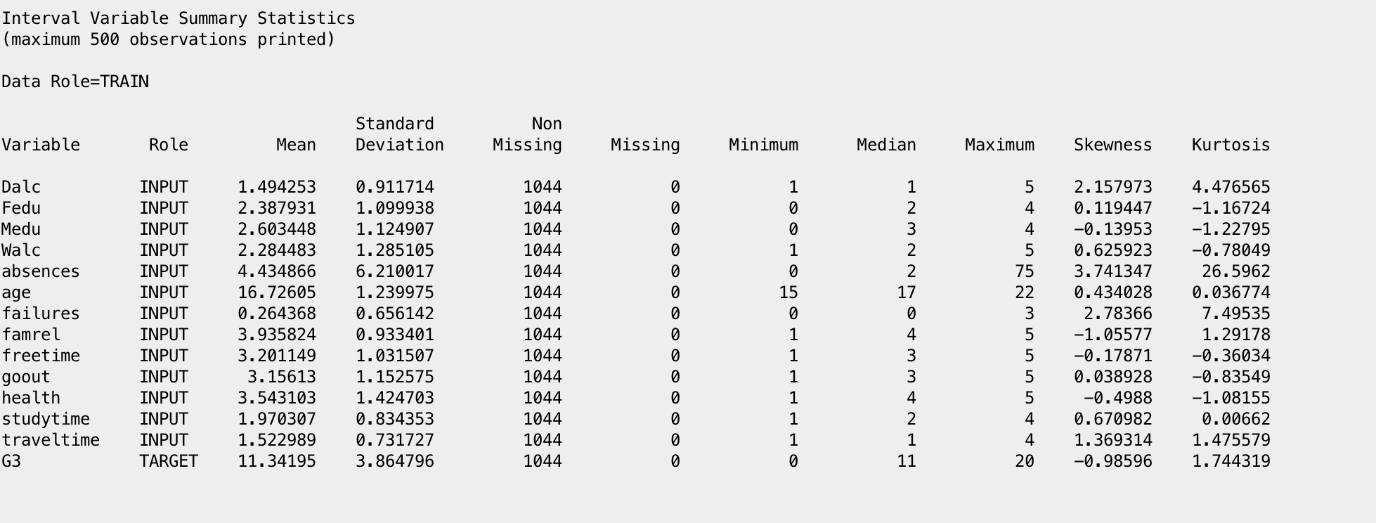
**Multi linear regression line equation is:**

G3= Estimate(Intercept)+failures\* Estimate(failures)+go out\* Estimate(go out)+health \* Estimate(health)+romantic \* Estimate(romantic)+ school sup\* Estimate(school sup)+ studytime\*Estimate(studytime).

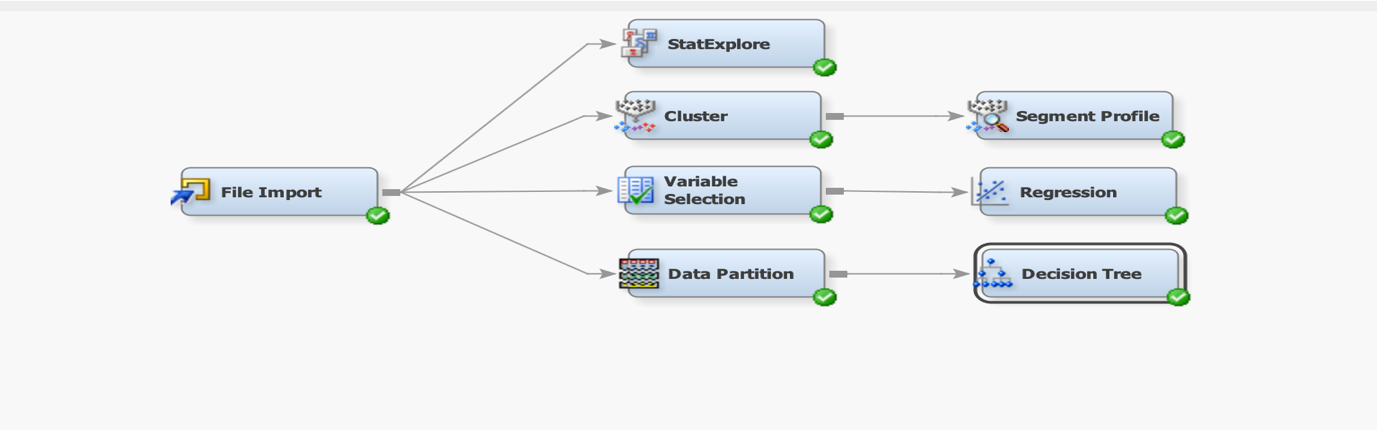
## DECISION TREE:

**NOTE**: **IN THE PROJECT WE WILL BE NEGLECTING G1 AND G2 ATRRIBUITES BECAUSE THESE TWO VARIABLES HAVE HIGH COLINEARITY WITH OUTPUT VARIABLE G3.**

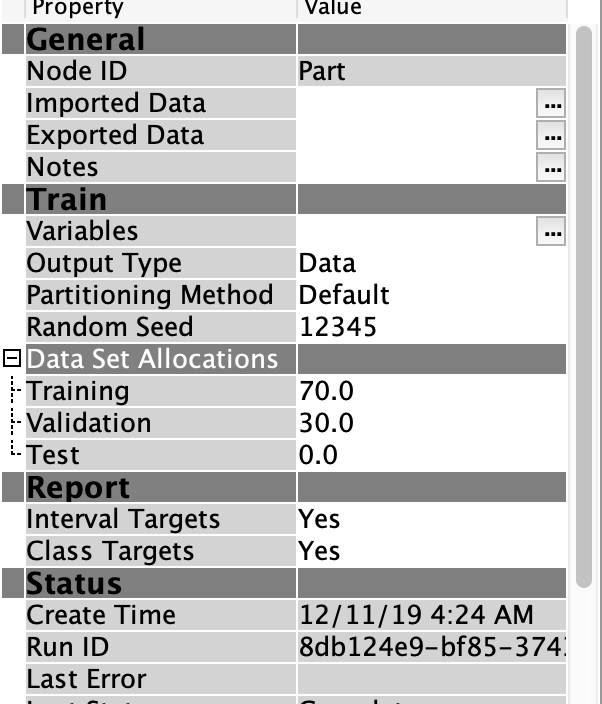


****

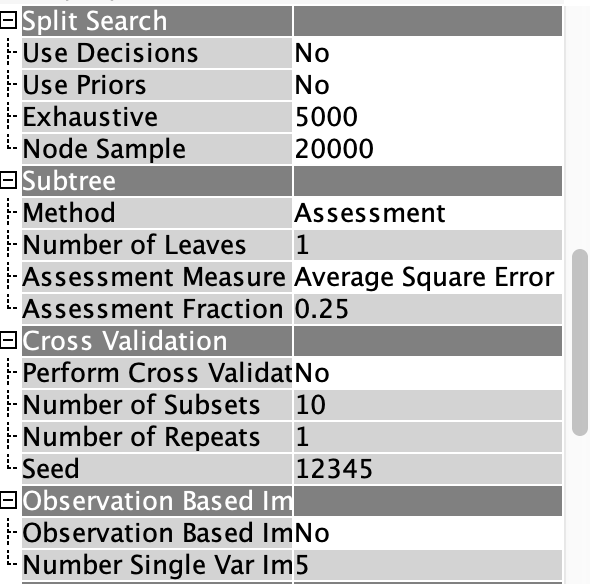
Since there are no missing values in the data, we need not use replacement node or impute to clean the data.

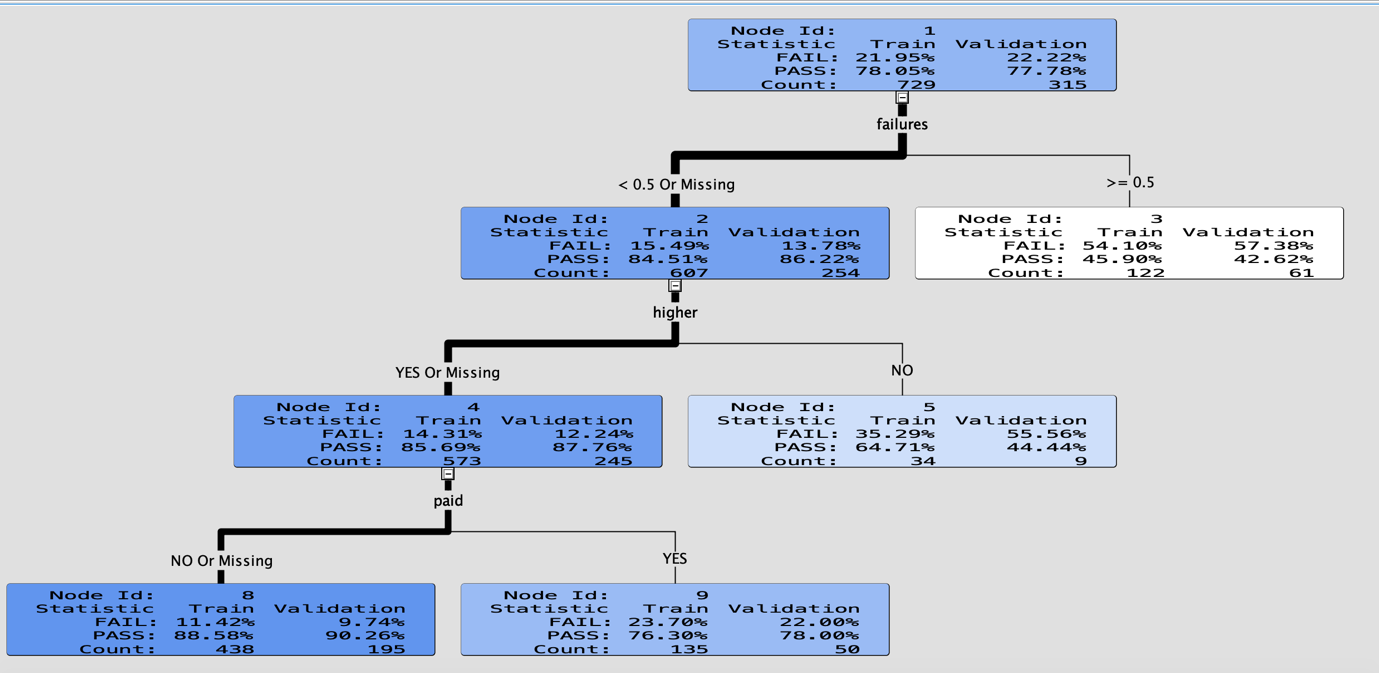
****

The next thing done was Adding data partition node to split the data into 70 percent training data and 30 percent validation data.



Next, the Decision tree was created using average square error as the model assessment statistic:





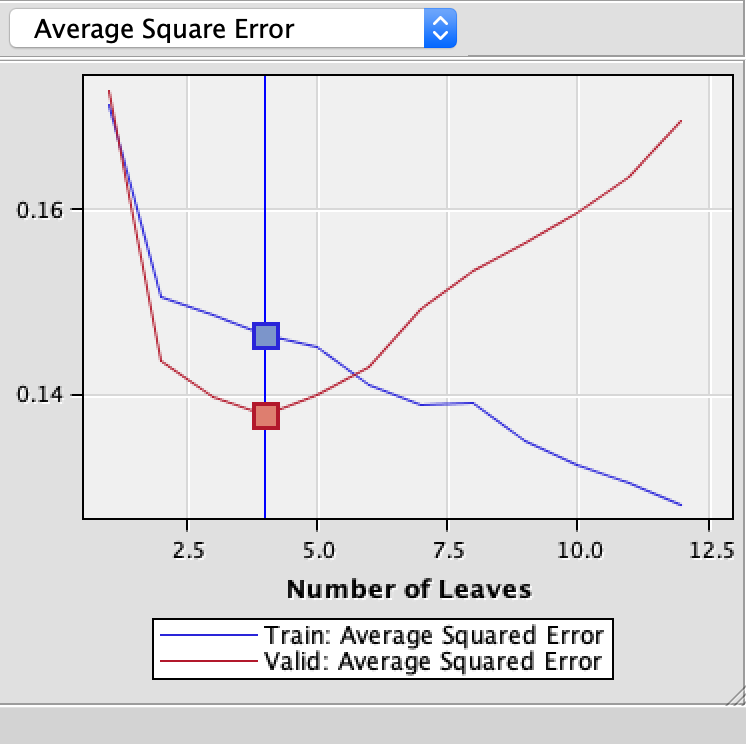
### Output of the decision tree is shown in above diagram.

* The weight of the line is heavier for the node where most observations are located.
* Darker the node, more the purity of node. Which means, the node which is darker tells you that the node has number of observations with “Yes”
* If the node is white, there are more number of observations with a “No”
* In failures split, if it is less than 0.5 student has more probability of passing the subject. The pass percentage is 84.51% in validation data.

If the failures are greater than 0.5, the students has more than 54% of chance to fail in that subject

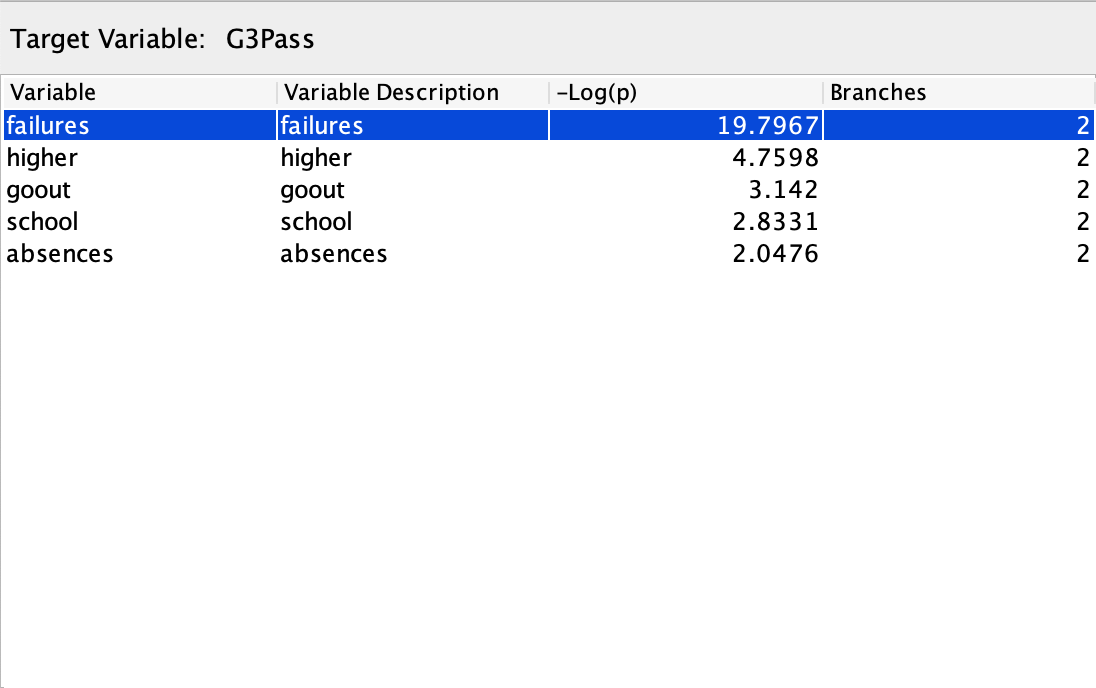
* In the second split higher, if the student has a interest about pursuing his/her higher studies he has probability of more than 85% to pass the subject. If he doesn’t have any thought of doing higher studies he has 65% of pass rate.
* In the third split paid, if the student has paid for extra hours with in subjects, he/she has 77% of passing that subject.
* In the third split paid, If the students doesn’t require any extra studying hours for that subject they have 89% of passing the subject. Since because they could have good knowledge with that subject. So they won’t require to pay extra amount.

**Sub-Tree Assessment Plot :**



The line represents minimum miss-classification rate at node 4, it has average square error of 0.1377.

The competing nodes for node1 are in the figure below.



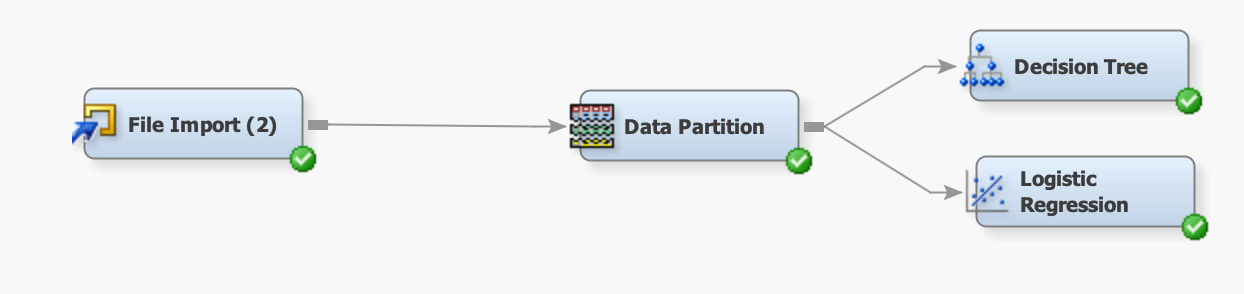
Here - Log(p) means, It is chi-square based number created which is called log worth. It shows us which variable is best to use for splitting our data to get better predictions.

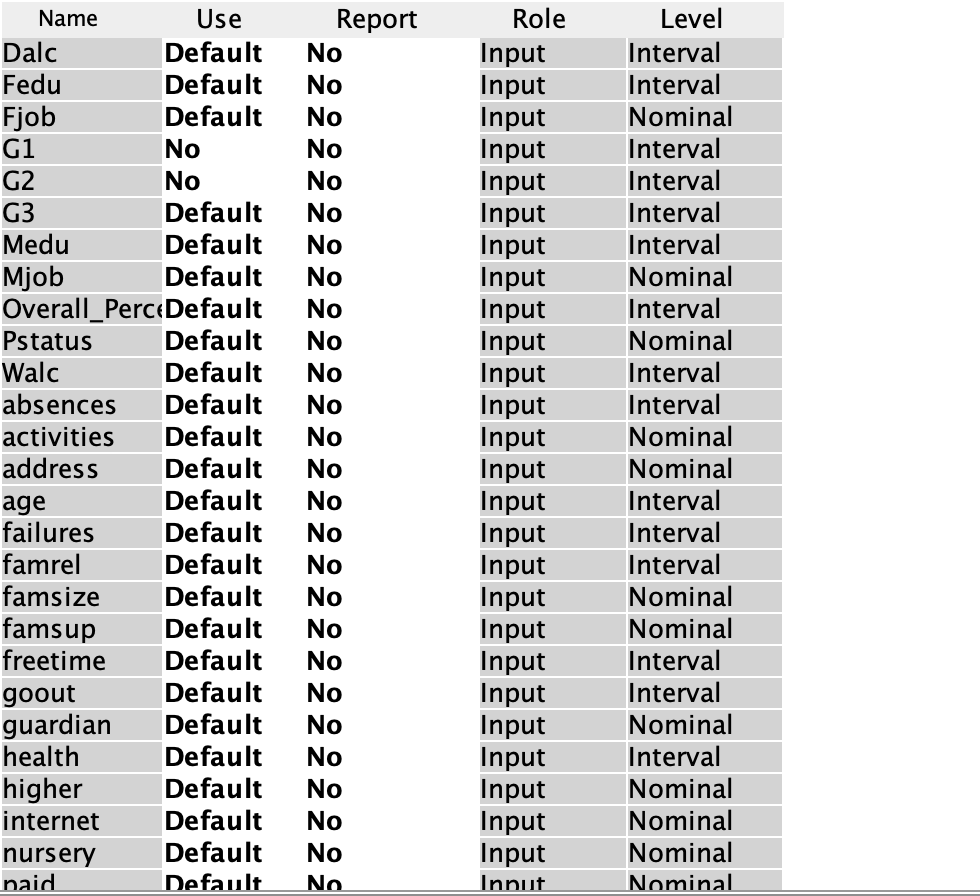
### CONCLUSION FROM DECISION TREE:

From this decision tree model, we cannot directly derive that Alcohol consumption has any impact the overall grades of the students. This model only tells you whether the students pass or fail the subject or students get good grade or not. Because, this is classification type model, it has the target variable as categorical variables.

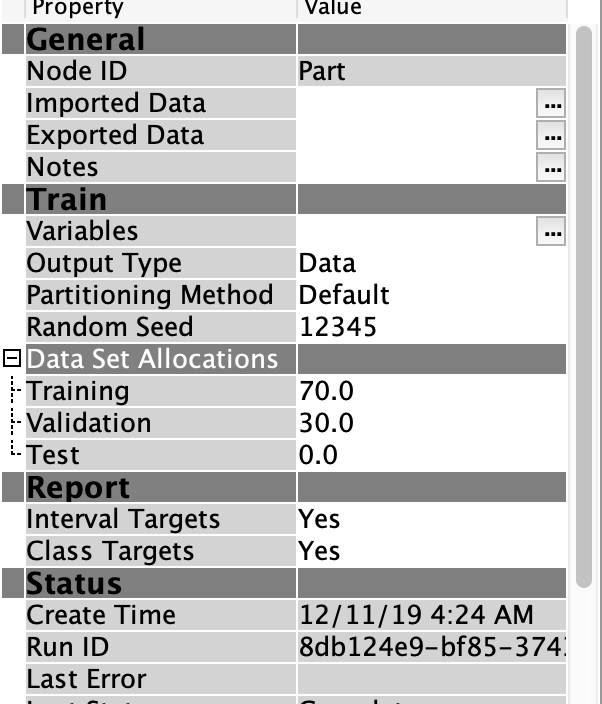
## LOGISTIC REGRESSION:

**NOTE : IN THE PROJECT WE WILL BE NEGLECTING G1 AND G2 ATRRIBUITES BECAUSE THESE TWO VARIABLES HAVE HIGH COLINEARITY WITH OUTPUT VARIABLE G3.**

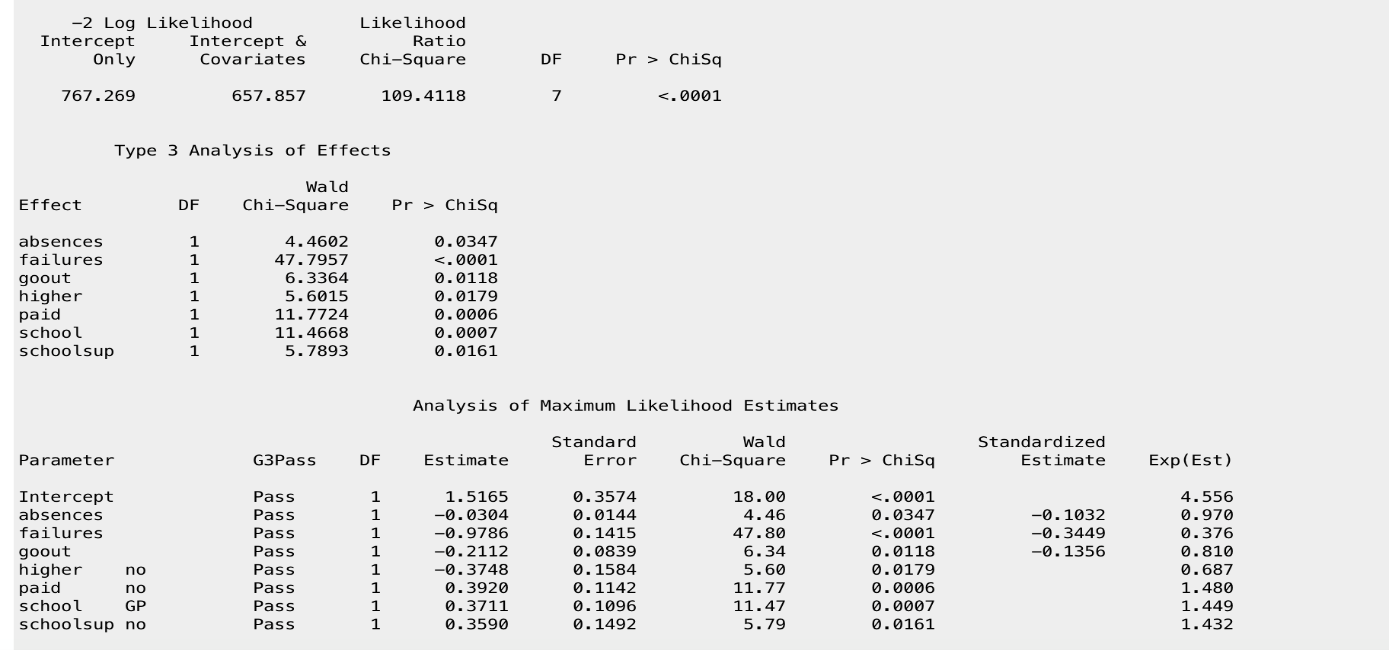




Add data partition node to split the data into 70 percent training data and 30 percent validation data.



We used step-wise model.Stepwise linear regression is a method of regressing multiple variables while simultaneously removing those that aren't important. Stepwise regression essentially does multiple regression a number of times, each time removing the weakest correlated variable.



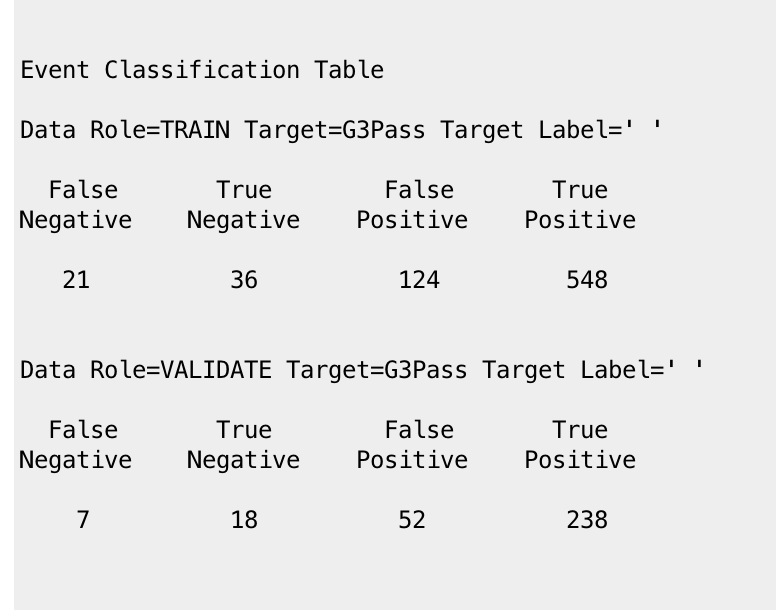
### Odds Ratio:

**A screenshot of a cell phone

Description automatically generated**

* The odds ratio got decreased by 0.3 for absences
* Failures, go out, higher odds ratio is less than 1
* Odds ratio Paid, school, schoolsup is greater than 1.

### Confusion matrix:

****

Accuracy of the model: can be determined by using validating data.

Accuracy= TP+ TN/(FN+TN+FP+TP)=238+18/7+18+52+238 = 0.81 = 81% accuracy

## NEURAL NETWORKS:

**A picture containing screenshot

Description automatically generated**

We added auto neural network node.

Results are shown below :

A screenshot of text

Description automatically generated

**MODEL COMPARISON NODE**:

**A picture containing screenshot

Description automatically generated**

Output:

A screenshot of a social media post

Description automatically generated

According to model comparison, it shows logistic regression is best model for this data.

Managerial Implications and conclusions **:**

* Linear regression and regression tree models establishes the level of impact of alcohol consumption on GPA, it was only derivable for either one of the groups
* From Cluster analysis, The students who have more free time, who drink on weekdays as well as on weekends, who travel more, who have no internet, who don’t have school support tend to get poor overall grade. The students who don’t drink regularly on weekdays as well as weekends, who have less travel time, who have internet, who have school support tend to get good overall grade
* The alcohol impact is high on male students GPA, they consume alcohol and gets less GPA. where as, most of the female students do consume alcohol yet they tend to get good overall GPA.

References:

<https://www.kaggle.com/uciml/student-alcohol-consumption>