IMPACT OF ALCOHOL COMSUMPTION ON STUDENT GPA

Contents

EXECUTIVE SUMMARY:

School is a period that gives students their first chance to settle on their own choices and at times it is inseparable from drinking. People decide to drink liquor differs from one individual to another, yet liquor utilization has consistently been considered as a large part of the school culture. This analysis centers around the impact of liquor use on students' academic performance.

PROJECT MOTIVATION:

This project will aim to determine how alcohol consumption influences the GPA of two groups of students taking Math and Portuguese. The 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.
This project uses methods like linear regression, cluster analysis and other techniques to develop the predictions. Moreover, as the age group in this data set is between 15 to 22, this data set is interesting as being a college student.

DATASET USED:

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

DATA DESCRIPTION:

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)

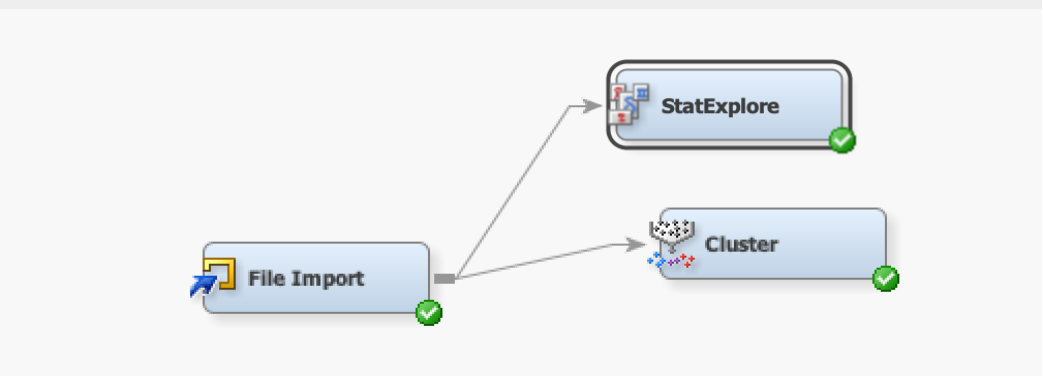
DATA PREPARATION ACTIVITIES:

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

Name	Use	Report	Role	Level
Dalc	Default	No	Input	Interval
Fedu	Default	No	Input	Interval
Fjob	Default	No	Input	Nominal
G1	No	No	Input	Interval
G2	No	No	Input	Interval
G3	Default	No	Input	Interval
Medu	Default	No	Input	Interval
Mjob	Default	No	Input	Nominal
Overall_Perc	Default	No	Input	Interval
Pstatus	Default	No	Input	Nominal
Walc	Default	No	Input	Interval
absences	Default	No	Input	Interval
activities	Default	No	Input	Nominal
address	Default	No	Input	Nominal
age	Default	No	Input	Interval
failures	Default	No	Input	Interval
famrel	Default	No	Input	Interval
famsize	Default	No	Input	Nominal
famsup	Default	No	Input	Nominal
freetime	Default	No	Input	Interval
goout	Default	No	Input	Interval
guardian	Default	No	Input	Nominal
health	Default	No	Input	Interval
higher	Default	No	Input	Nominal
internet	Default	No	Input	Nominal
nursery	Default	No	Input	Nominal
paid	Default	No	Input	Nominal

MODELS

1. CLUSTER ANALYSIS:

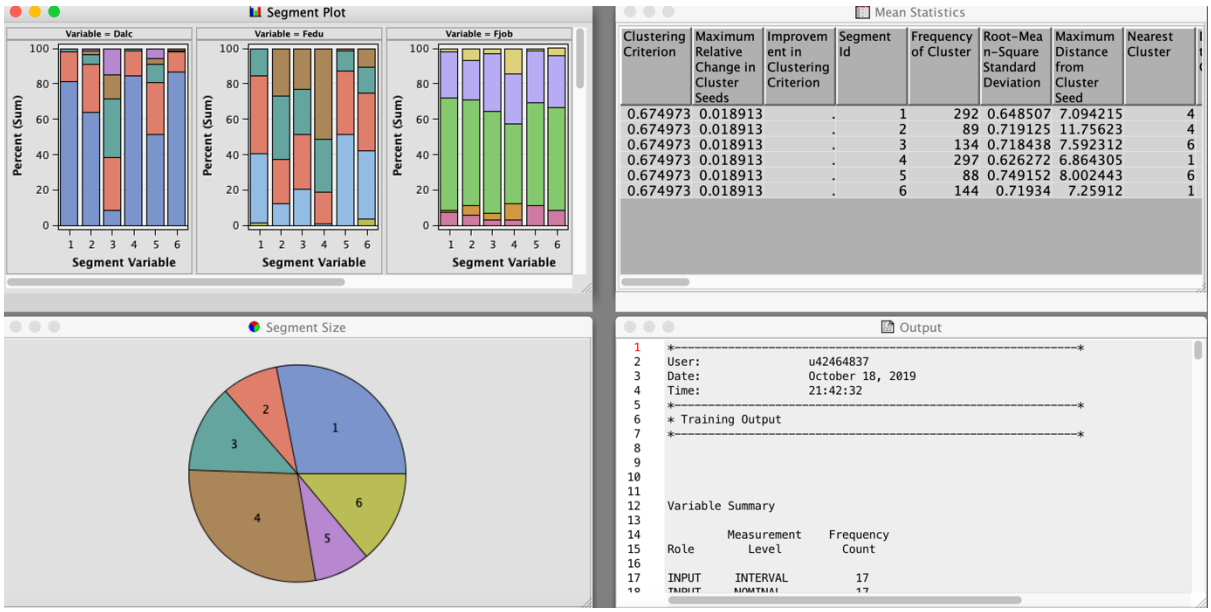


As shown in this figure we will attach the cluster node to file import and in cluster analysis we will be rejecting the G1 and G2 attributes because of them having high collinearity with G3.

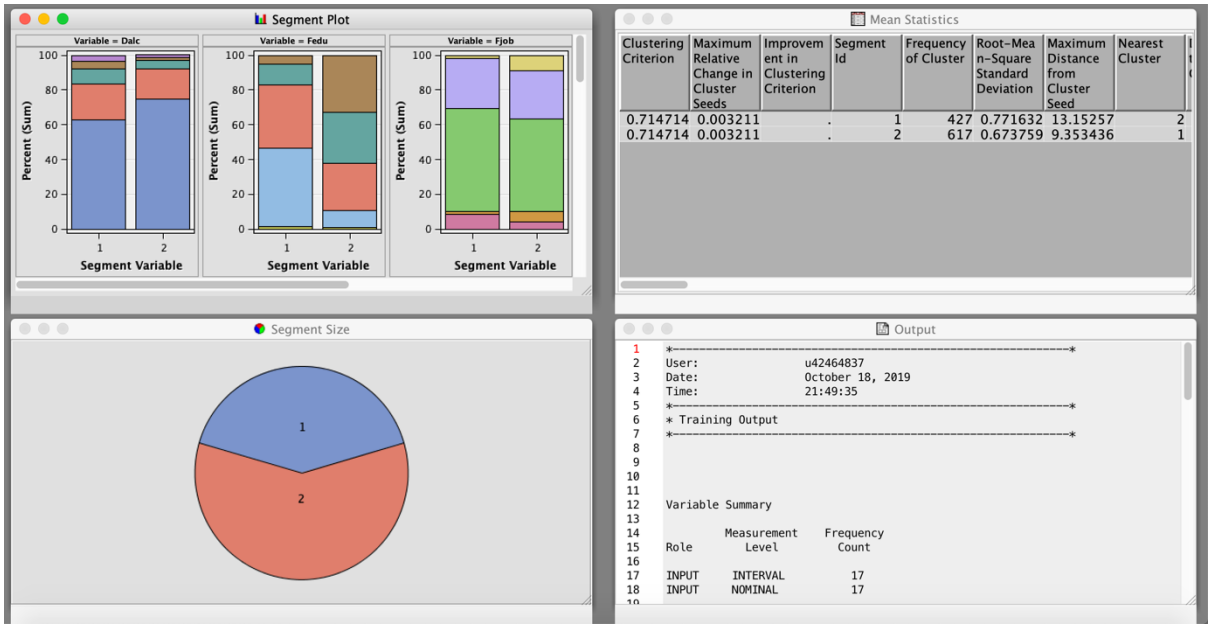
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Fjob	INPUT	5	0	other	55.94	services	27.97
TRAIN	Mjob	INPUT	5	0	other	38.22	services	22.89
TRAIN	Pstatus	INPUT	2	0	T	88.41	A	11.59
TRAIN	activities	INPUT	2	0	no	50.57	yes	49.43
TRAIN	address	INPUT	2	0	U	72.70	R	27.30
TRAIN	famsize	INPUT	2	0	GT3	70.69	LE3	29.31
TRAIN	famsup	INPUT	2	0	yes	61.30	no	38.70
TRAIN	guardian	INPUT	3	0	mother	69.73	father	23.28
TRAIN	higher	INPUT	2	0	yes	91.48	no	8.52
TRAIN	internet	INPUT	2	0	yes	79.21	no	20.79
TRAIN	nursery	INPUT	2	0	yes	79.98	no	20.02
TRAIN	paid	INPUT	2	0	no	78.93	yes	21.07
TRAIN	reason	INPUT	4	0	course	41.19	home	24.71
TRAIN	romantic	INPUT	2	0	no	64.46	yes	35.54
TRAIN	school	INPUT	2	0	GP	73.95	MS	26.05
TRAIN	schoolsup	INPUT	2	0	no	88.60	yes	11.40
TRAIN	sex	INPUT	2	0	F	56.61	M	43.39

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

- This data does not 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.
2. 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.
3. 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.
4. Absence Segment Plot:
    - In cluster 1, the students tend to be absent a greater number of times.
    - In cluster 2, we have only few people with absence in classes.
5. Sex(M/F) Segment Plot:
    - In Cluster 1, there are 72% of male students whereas, In Cluster 2 there are a greater number of female students i.e 68%.
6. 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.

7. Travel Segment Plot:
    - In Cluster 1, students travel for more time as compared to Cluster 2.

8. 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 reasons 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).

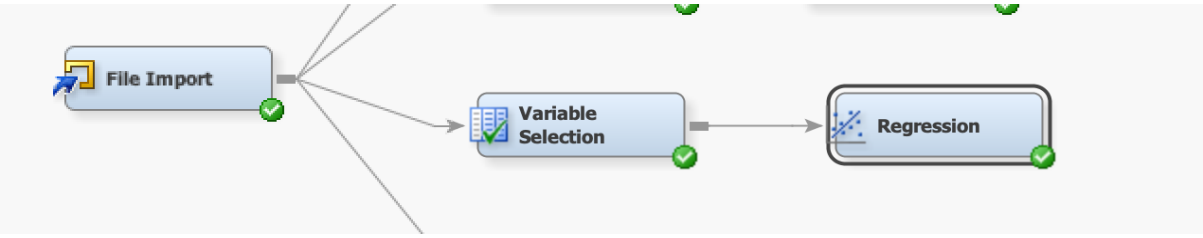
Final Conclusion from Cluster Analysis:

- 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.

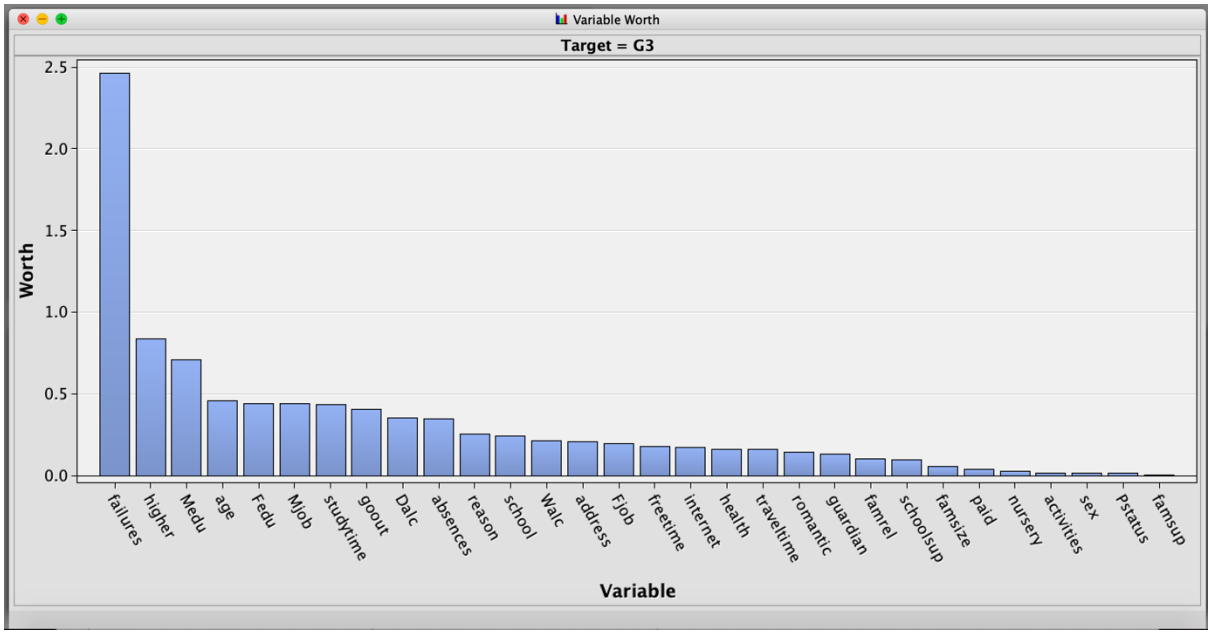
2. LINEAR REGRESSION:

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

Name	Use	Report	Role	Level
Dalc	Default	No	Input	Interval
Fedu	Default	No	Input	Interval
Fjob	Default	No	Input	Nominal
G1	No	No	Input	Interval
G2	No	No	Input	Interval
G3	Default	No	Input	Interval
Medu	Default	No	Input	Interval
Mjob	Default	No	Input	Nominal
Overall_Perc	Default	No	Input	Interval
Pstatus	Default	No	Input	Nominal
Walc	Default	No	Input	Interval
absences	Default	No	Input	Interval
activities	Default	No	Input	Nominal
address	Default	No	Input	Nominal
age	Default	No	Input	Interval
failures	Default	No	Input	Interval
famrel	Default	No	Input	Interval
famsize	Default	No	Input	Nominal
famsup	Default	No	Input	Nominal
freetime	Default	No	Input	Interval
goout	Default	No	Input	Interval
guardian	Default	No	Input	Nominal
health	Default	No	Input	Interval
higher	Default	No	Input	Nominal
internet	Default	No	Input	Nominal
nursery	Default	No	Input	Nominal
paid	Default	No	Input	Nominal

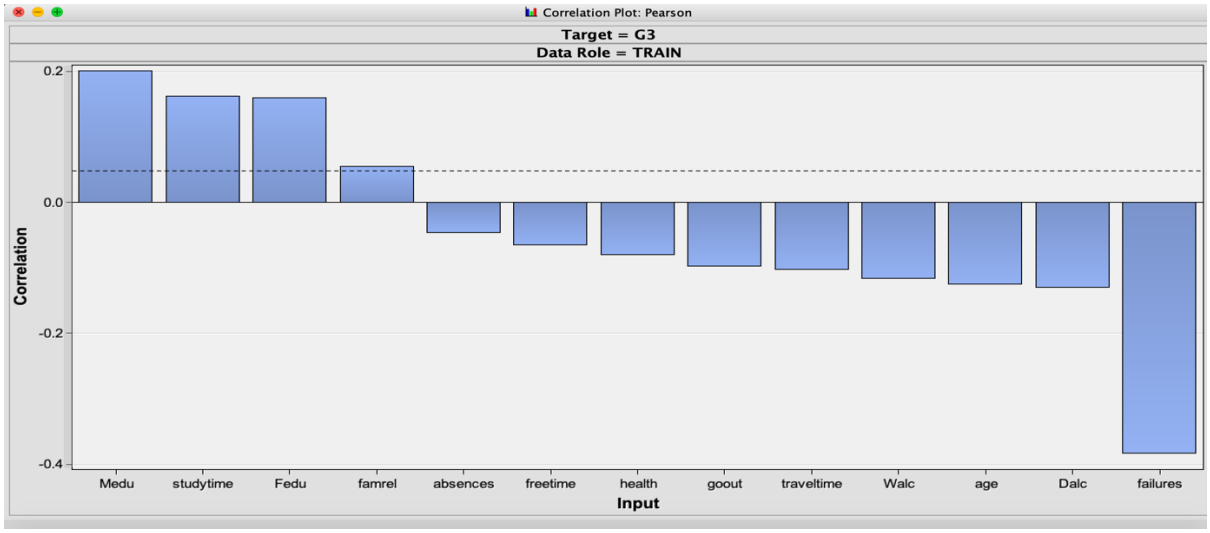


In Linear regression we used variable in selection node because it refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs to model.



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.

The DMINE Procedure

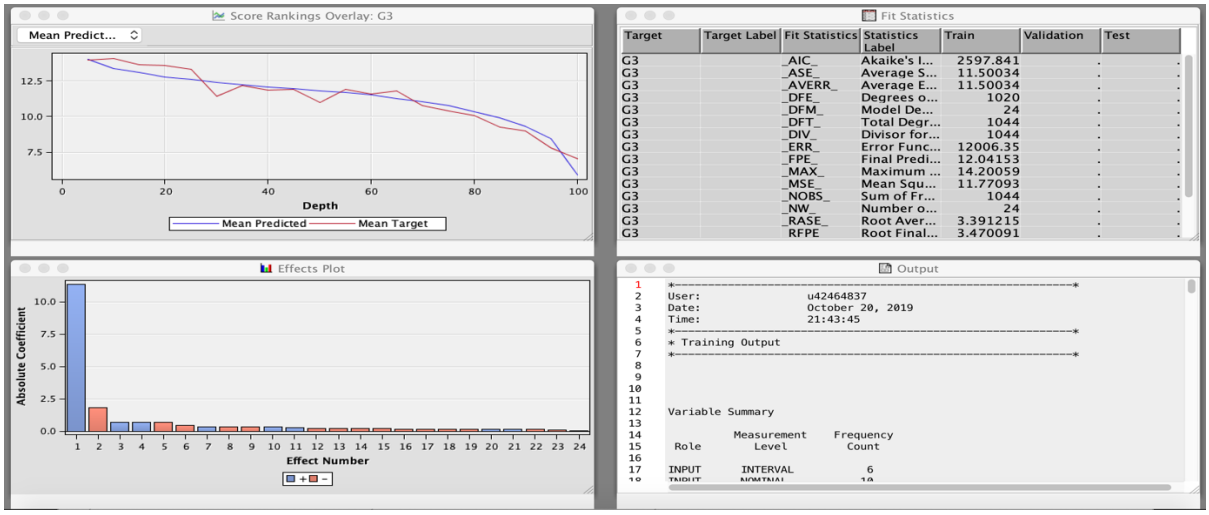
Effects Chosen for Target: G3

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Var: failures	1	0.146800	179.285013	<.0001	2286.990568	12.756173
Class: Mjob	4	0.020341	6.337653	<.0001	316.883810	12.500046
Class: higher	1	0.012666	16.014400	<.0001	197.326467	12.321814
Var: studytime	1	0.007710	9.830512	0.0018	120.106976	12.217775
Class: schoolsup	1	0.007769	9.991934	0.0016	121.028735	12.112643
Group: Fjob	3	0.006079	2.618530	0.0497	94.707657	12.056084
Var: health	1	0.005650	7.346181	0.0068	88.024876	11.982399
Var: goout	1	0.005235	6.845056	0.0090	81.557816	11.914850
Class: romantic	1	0.005201	6.838853	0.0091	81.024599	11.847688
Class: address	1	0.004492	5.934422	0.0150	69.973636	11.791145
Class: internet	1	0.001851	2.449470	0.1179	28.841385	11.774543
Group: reason	2	0.001584	1.047834	0.3511	24.673237	11.773446
Var: Dalc	1	0.001066	1.410712	0.2352	16.602293	11.768731
Var: Medu	1	0.001154	1.528541	0.2166	17.979705	11.762659
Class: guardian	2	0.000999	0.661234	0.5164	15.566047	11.770455
Class: school	1	0.000724	0.958493	0.3278	11.282362	11.770934

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.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	2802.827032	467.137839	37.92	<.0001
Error	1037	12776	12.320246		
Corrected Total	1043	15579			

Model Fit Statistics

R-Square	0.1799	Adj R-Sq	0.1752
AIC	2628.7151	BIC	2630.8095
SBC	2663.3708	C(p)	7.0000

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
failures	1	1859.7535	150.95	<.0001
goout	1	73.4168	5.96	0.0148
health	1	50.6988	4.12	0.0428
romantic	1	109.1747	8.86	0.0030
schoolsup	1	145.0962	11.78	0.0006
studytime	1	180.0490	14.61	0.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.6154	0.5383	21.58	<.0001
failures	1	-2.0727	0.1687	-12.29	<.0001
goout	1	-0.2316	0.0949	-2.44	0.0148
health	1	-0.1552	0.0765	-2.03	0.0428
romantic no	1	0.3408	0.1145	2.98	0.0030
schoolsup no	1	0.5914	0.1723	3.43	0.0006
studytime	1	0.5077	0.1328	3.82	0.0001

Fit Statistics

Target=G3 Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	1837.42	.
ASE	Average Squared Error	12.11	12.58
AVERR	Average Error Function	12.11	12.58
DFE	Degrees of Freedom for Error	724.00	.
DFM	Model Degrees of Freedom	7.00	.
DFT	Total Degrees of Freedom	731.00	.
DIV	Divisor for ASE	731.00	313.00
ERR	Error Function	8855.79	3937.22
FPE	Final Prediction Error	12.35	.
MAX	Maximum Absolute Error	13.46	12.38
MSE	Mean Square Error	12.23	12.58
NOBS	Sum of Frequencies	731.00	313.00
_NW_	Number of Estimate Weights	7.00	.
RASE	Root Average Sum of Squares	3.48	3.55
RFPE	Root Final Prediction Error	3.51	.
RMSE	Root Mean Squared Error	3.50	3.55
SBC	Schwarz's Bayesian Criterion	1869.58	.
SSE	Sum of Squared Errors	8855.79	3937.22
SUMW	Sum of Case Weights Times Freq	731.00	313.00

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).

4. 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.

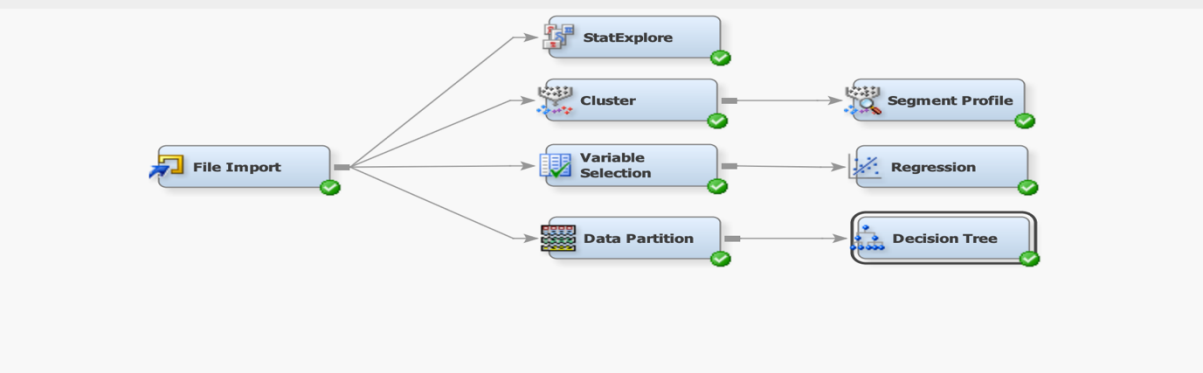
Name	Use	Report	Role	Level
Dalc	Default	No	Input	Interval
Fedu	Default	No	Input	Interval
Fjob	Default	No	Input	Nominal
G1	No	No	Input	Interval
G2	No	No	Input	Interval
G3	Default	No	Input	Interval
Medu	Default	No	Input	Interval
Mjob	Default	No	Input	Nominal
Overall_Perc	Default	No	Input	Interval
Pstatus	Default	No	Input	Nominal
Walc	Default	No	Input	Interval
absences	Default	No	Input	Interval
activities	Default	No	Input	Nominal
address	Default	No	Input	Nominal
age	Default	No	Input	Interval
failures	Default	No	Input	Interval
famrel	Default	No	Input	Interval
famsize	Default	No	Input	Nominal
famsup	Default	No	Input	Nominal
freetime	Default	No	Input	Interval
goout	Default	No	Input	Interval
guardian	Default	No	Input	Nominal
health	Default	No	Input	Interval
higher	Default	No	Input	Nominal
internet	Default	No	Input	Nominal
nursery	Default	No	Input	Nominal
paid	Default	No	Input	Nominal

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Dalc	INPUT	1.494253	0.911714	1044	0	1	1	5	2.157973	4.476565
Fedu	INPUT	2.387931	1.099938	1044	0	0	2	4	0.119447	-1.16724
Medu	INPUT	2.603448	1.124907	1044	0	0	3	4	-0.13953	-1.22795
Walc	INPUT	2.284483	1.285105	1044	0	1	2	5	0.625923	-0.78049
absences	INPUT	4.434866	6.210017	1044	0	0	2	75	3.741347	26.5962
age	INPUT	16.72605	1.239975	1044	0	15	17	22	0.434028	0.036774
failures	INPUT	0.264368	0.656142	1044	0	0	0	3	2.78366	7.49535
famrel	INPUT	3.935824	0.933401	1044	0	1	4	5	-1.05577	1.29178
freetime	INPUT	3.201149	1.031507	1044	0	1	3	5	-0.17871	-0.36034
goout	INPUT	3.15613	1.152575	1044	0	1	3	5	0.038928	-0.83549
health	INPUT	3.543103	1.424703	1044	0	1	4	5	-0.4988	-1.08155
studytime	INPUT	1.970307	0.834353	1044	0	1	2	4	0.670982	0.00662
traveltime	INPUT	1.522989	0.731727	1044	0	1	1	4	1.369314	1.475579
G3	TARGET	11.34195	3.864796	1044	0	0	11	20	-0.98596	1.744319

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

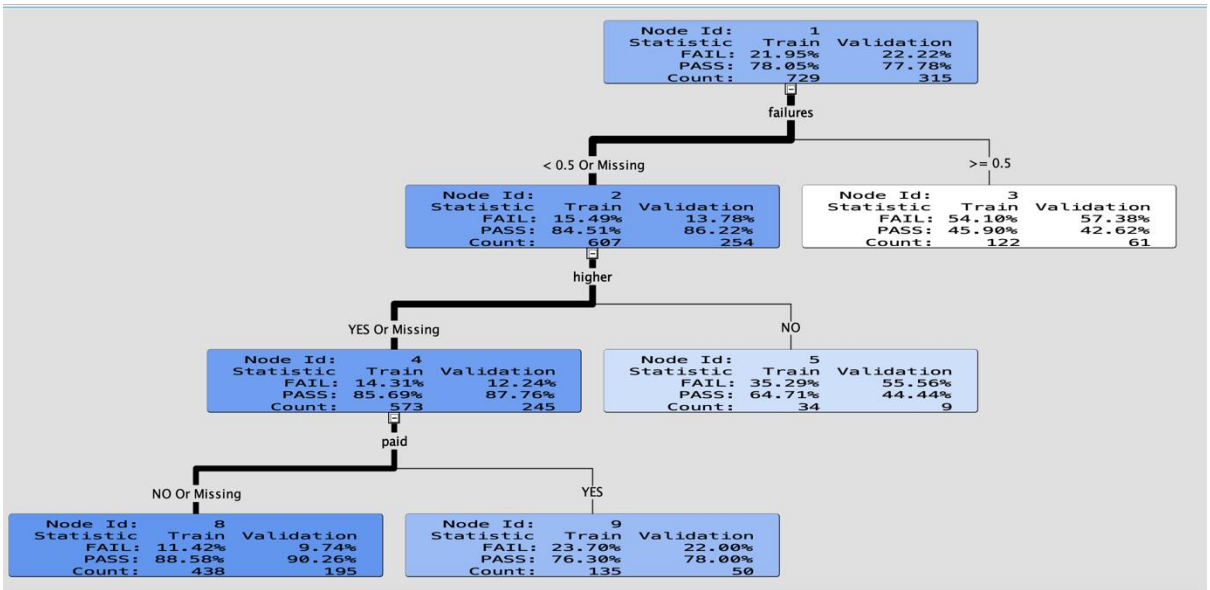


The next thing done was Adding data partition node to split the data into 70 percent training data and 30 percent validation data.

Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	12/11/19 4:24 AM
Run ID	8db124e9-bf85-374
Last Error	

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

Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validat	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Im	
Observation Based Im	No
Number Single Var Im	5

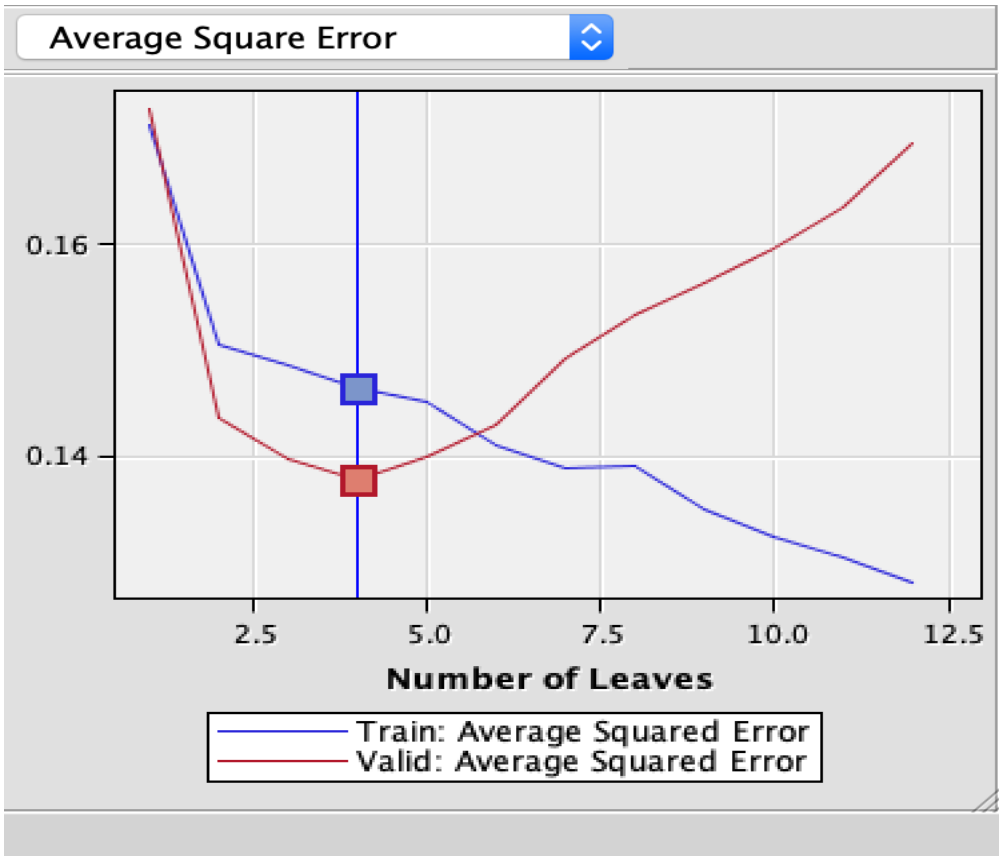


Output of the decision tree is shown in above diagram.

- 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 a greater 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 have more than 54% of chance to fail in that subject.
- In the second split higher, if the student has an interest about pursuing his/her higher studies he has probability of more than 85% to pass the subject. If he does not 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 within subjects, he/she has 77% of passing that subject.

- In the third split paid, If the students do not 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 will not 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.

Target Variable: G3Pass

Variable	Variable Description	−Log(p)	Branches
failures	failures	19.7967	2
higher	higher	4.7598	2
goout	goout	3.142	2
school	school	2.8331	2
absences	absences	2.0476	2

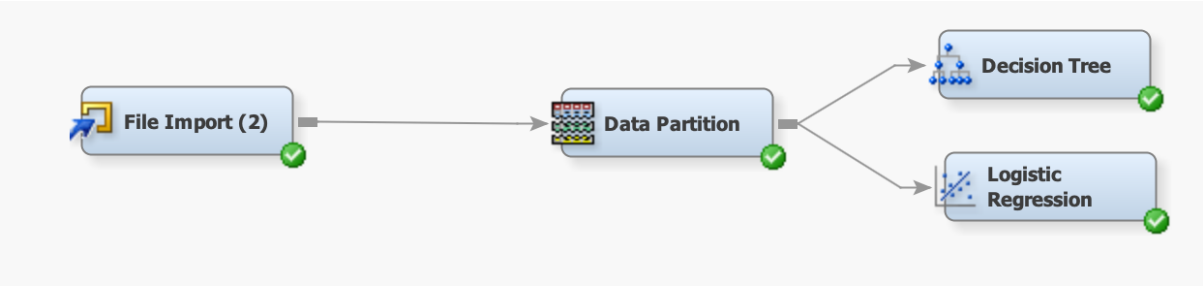
Here - Log(p) means, it is chi-square based number created which is called log worth. It shows us which variable is best to use for splitting our data to get better predictions.

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.

4: 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.



Name	Use	Report	Role	Level
Dalc	Default	No	Input	Interval
Fedu	Default	No	Input	Interval
Fjob	Default	No	Input	Nominal
G1	No	No	Input	Interval
G2	No	No	Input	Interval
G3	Default	No	Input	Interval
Medu	Default	No	Input	Interval
Mjob	Default	No	Input	Nominal
Overall_Perc	Default	No	Input	Interval
Pstatus	Default	No	Input	Nominal
Walc	Default	No	Input	Interval
absences	Default	No	Input	Interval
activities	Default	No	Input	Nominal
address	Default	No	Input	Nominal
age	Default	No	Input	Interval
failures	Default	No	Input	Interval
famrel	Default	No	Input	Interval
famsize	Default	No	Input	Nominal
famsup	Default	No	Input	Nominal
freetime	Default	No	Input	Interval
goout	Default	No	Input	Interval
guardian	Default	No	Input	Nominal
health	Default	No	Input	Interval
higher	Default	No	Input	Nominal
internet	Default	No	Input	Nominal
nursery	Default	No	Input	Nominal
paid	Default	No	Input	Nominal

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

Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	12/11/19 4:24 AM
Run ID	8db124e9-bf85-374
Last Error	

We used stepwise model. Stepwise linear regression is a method of regressing multiple variables while simultaneously removing those that are not important. Stepwise regression essentially does multiple regression several times, each time removing the weakest correlated variable.

−2 Log Likelihood Intercept Only	−2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
767.269	657.857	109.4118	7	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
absences	1	4.4602	0.0347
failures	1	47.7957	<.0001
goout	1	6.3364	0.0118
higher	1	5.6015	0.0179
paid	1	11.7724	0.0006
school	1	11.4668	0.0007
schoolsup	1	5.7893	0.0161

Analysis of Maximum Likelihood Estimates

Parameter		G3Pass	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		Pass	1	1.5165	0.3574	18.00	<.0001		4.556
absences		Pass	1	−0.0304	0.0144	4.46	0.0347	−0.1032	0.970
failures		Pass	1	−0.9786	0.1415	47.80	<.0001	−0.3449	0.376
goout		Pass	1	−0.2112	0.0839	6.34	0.0118	−0.1356	0.810
higher	no	Pass	1	−0.3748	0.1584	5.60	0.0179		0.687
paid	no	Pass	1	0.3920	0.1142	11.77	0.0006		1.480
school	GP	Pass	1	0.3711	0.1096	11.47	0.0007		1.449
schoolsup	no	Pass	1	0.3590	0.1492	5.79	0.0161		1.432

Odds Ratio:

Odds Ratio Estimates

Effect		G3Pass	Point Estimate
absences		Pass	0.970
failures		Pass	0.376
goout		Pass	0.810
higher	no vs yes	Pass	0.473
paid	no vs yes	Pass	2.190
school	GP vs MS	Pass	2.101
schoolsup	no vs yes	Pass	2.050

- **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:

Event Classification Table

Data Role=TRAIN Target=G3Pass Target Label=' '

False Negative	True Negative	False Positive	True Positive
21	36	124	548

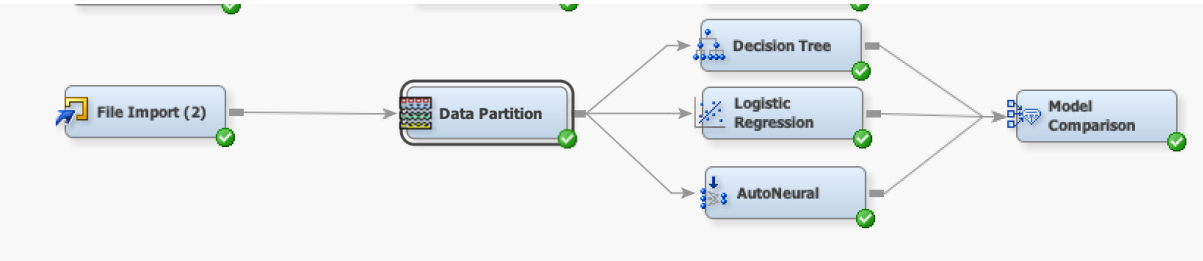
Data Role=VALIDATE Target=G3Pass Target Label=' '

False Negative	True Negative	False Positive	True Positive
7	18	52	238

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

4.NEURAL NETWORKS:



We added auto neural network node.

Results are shown below:

Event Classification Table

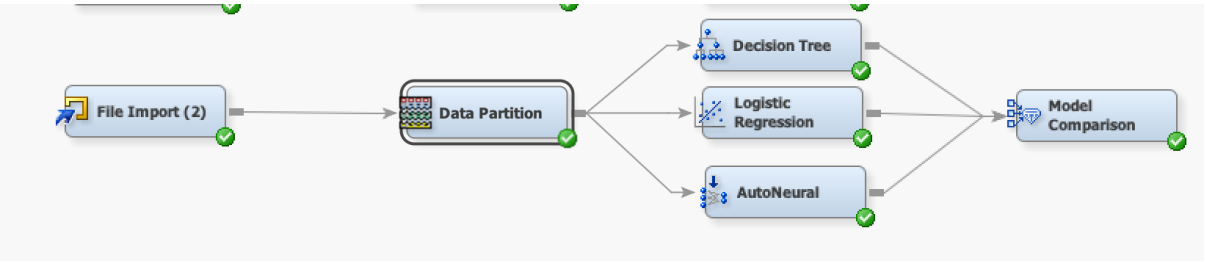
Data Role=TRAIN Target=G3Pass Target Label=' '

False Negative	True Negative	False Positive	True Positive
18	90	70	551

Data Role=VALIDATE Target=G3Pass Target Label=' '

False Negative	True Negative	False Positive	True Positive
23	26	44	222

MODEL COMPARISON NODE:



Output:

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate
Y	Reg2	Reg2	Logistic ...	G3Pass		0.187302	729	0.198903	0.942787	210.2846	0.144228	0.379774	1458	729	315	0.187
	Tree	Tree	Decision ...	G3Pass		0.193651	729	0.205761	0.885845	213.5337	0.146457	0.382696	1458	729	315	0.193
	AutoNeu...	AutoNeu...	AutoNeu...	G3Pass		0.212698	729	0.120713	0.98551	126.1234	0.086504	0.294116	1458	729	315	0.212

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 do not have school support tend to get poor overall grade. The students who do not 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 student's GPA, they consume alcohol and gets less GPA. whereas, most of the female students do consume alcohol yet they tend to get good overall GPA.

References:

- Turrisi R, Larimer ME, Mallett KA, Kilmer JR, Ray AE, Mastroleo NR, et al. A randomized clinical trial evaluating a combined alcohol intervention for high-risk college students. J Stud Alcohol Drugs. 2009;70:555–67. [PMC free article] [PubMed]

- Hingson R, Heeren T, Winter M, Wechsler H. Magnitude of alcohol-related mortality and morbidity among U.S. college students ages 18-24: Changes from1998 to 2001. Annu Rev Public Health. 2005;26:259–79. [PubMed]

- P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION