Angel V. Rivera Galaviz

Data Mining and Knowledge Discovery

Christopher Harris

May 02, 2021

What factors impact final grades in Portuguese math students?

I chose this dataset because I had a high interest in seeing how external factors regarding the Portuguese students' personal lives could affect their final grades in math. I wished to see if my assumptions of how those factors could affect their grades were correct. For the analysis, I used regression and statistical analysis. I tried using classification, association, and clustering techniques but none of them provided meaningful results. I felt that regression and statistical analysis could help me better to analyze the relationships among the different attributes on the dataset.

Research questions and hypotheses:

For simplicity, each time I refer to students I will be referring to Portuguese students, and when I refer to their final grades, I will be referring to final grades on math.

- Will students that study more time get better final grades?

Hypothesis - A student that studies more hours will be better prepared for its math assessments, allowing that student to earn higher final grades than students who study less.

- Will the parent's cohabitation status affect the students' final grades?

Hypothesis- Normally I could assume that this situation could affect the student's emotional stability, therefore, affecting their performance in math but, in this case, since the students are in the age range 15-22, I believe that they will have a better emotional maturity. Meaning that they will not be affected.

- Will the level of alcohol consumption on workdays affect the student's final grades?

Hypothesis- Students who have a higher alcohol consumption level during workdays should get lower final grades than students that have a lower alcohol consumption level during workdays. That is since those who have a higher alcohol consumption level during workdays will not be as focused on their academic duties.

-Will the level of alcohol consumption on weekends affect the students' final grades?

Hypothesis- Students that consume alcohol on weekends should be able to get good final grades regardless of the level of consumption since it will not affect their academic duties as much since the consumption will not happen during school days.

- Will students who get higher final grades be more interested in pursuing higher education?

Hypothesis- Students who are interested in pursuing higher education should get higher final grades since those grades will affect their education possibilities in the future.

- Will the parent's education level impact the student's interest in pursuing higher education?

Hypothesis- Students whose parents have a higher level of education will be more interested in pursuing higher education since they will be influenced by their parent's example or expectations.

- Will the student's access to the internet at home be a factor that will affect their final grades?

Hypothesis- Students that have access to the internet at home will get higher final grades than those who do not since they will have access to more resources that will help them to be successful.

- Will students' absences be a factor that influences students' final grades?

Hypothesis- Students who have a low number of absences will have higher final grades than those who have a higher number of absences since the latter missed more lectures, meaning that they could be less prepared for their assessments.

- Will the level of time students go out with friends affect the students' final grades?

Hypothesis- I believe that students who have a higher level of going out with friends will have lower final grades than students with a lower level of going out with friends since they will have less time to focus on their academic duties than the latter.

Data set Information:

The dataset contains real-world data collected from two Portuguese schools; Gabriel Pereira, and Mousinho da Silveira. According to Paulo Cortez, author of "Using data mining to predict secondary school student performance" and the dataset owner, the dataset includes information about students' grades, demographics, social and school-related features, among others. Cortez mentions that the data was collected by using school reports and questionnaires (Cortez & Silva).

Data source:

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+Performance. File Name: student-mat.csv

The data set I chose to use had 395 rows and 19 attributes. Originally it had 32 attributes, but I believe the attributes I chose to keep provided the best results for my research. Also, the database was divided into two different databases, they had the same attributes, but each one was for a different subject (Math or Portuguese) I worked with the math database.

Attribute descriptions:

• School – Student's school (GP-Gabriel Pereira, MS-Mousinho da Silveira)

• Sex - Student sex (F- Female, M-Male)

• Age - Student's age (15-22)

• Pstatus – Parent's cohabitation status (T- Living together, A-Living Apart)-

• Medu – Mother's education (0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3-secondary education (High school), 4- Higher education)

• Fedu – Father's education (0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3-secondary education (High school), 4- Higher education)

• Mjob – Mother's job (Teacher, Health care related, civil services (administrative or police), At

home, other)

• Fjob – Father's job (Mother's job (Teacher, Health care related, civil services (administrative or police), At home, other)

• Studytime – Student's weekly study time (1- <2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours, 4 > 10 hours)

• Failures – Number of past class failures

• Higher – Student wants to take higher education (Yes, No)

• Internet – Student has internet access at home (Yes, No)

• Romantic – Student is in a romantic relationship (Yes, No)

• Famrel – family relationships quality (from 1 – very bad to 5- excellent)

• Goout – Going out with friends (from 1- very low to 5 very-high)

• Dalc – Workday alcohol consumption (from 1- very low to 5 very-high)

• Walc- Weekend alcohol consumption (from 1- very low to 5 very-high)

• Absences – Student's number of school absences

• FinalGrade (G3 originally) – Final grade

(P. Cortez and A. Silva)

Data transformations:

No data type transformations were necessary to be able to use the Weka features I desired to use. I was required to discretize the attribute absences in Excel to be able to do my statistical analysis. To do so I used Excel's VLOOKUP function to bin all absence values on the dataset.

Metrics:

The metric I evaluated was the mean of each of the different values of each attribute of interest.

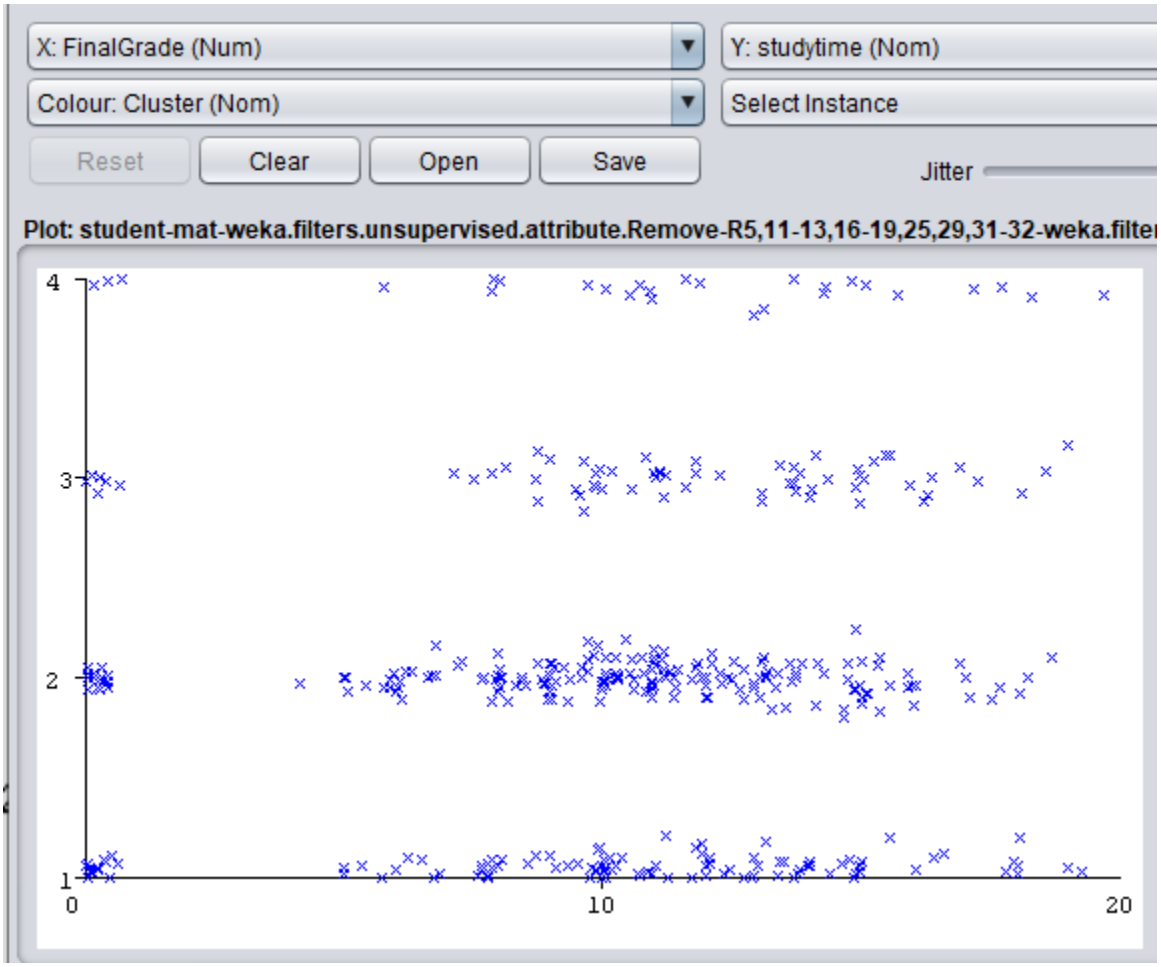
Techniques:

For my regression analysis, I used Weka's visualization feature. More precisely I used the clustering visualization feature, but it is necessary to mention that clustering was not involved in the analysis. I used clustering visualizations over Weka's standard visualizations since they allowed me to keep a single color for the instances. Having instances of different colors complicated the analysis of the patterns especially since some colors were barely visible. I did my regression analysis by observing the patterns on those visualizations and then compared them to what I could expect to see based on my hypothesis. For my statistical analysis, I used Excel. I used Excel functions; VLOOKUP, AVERAGEIF, and COUNTIFS. VLOOKUP allowed me to discretize the attribute absences into bins, AVERAGEIF allowed me to calculate the average of specific values based on a criterion, and COUNTIFS allowed me to count the number of instances that met a double criterion (one per column). Using these techniques, I was able to get the means of all the relationships I wished to analyze.

Results:

- Will students that study more time get better final grades?

Hypothesis - A student that studies more hours will be better prepared for its math assessments, allowing that student to earn higher final grades than students who study less.



Based on the visualization I got using Weka I do not believe there is enough evidence to support this hypothesis. As we can see in the visualization the students that studied less than two hours (1) were able to score similarly to those who studied more, and students who studied more than 10 hours (4) did not score higher than those that studied fewer hours. If the hypothesis was supported, we could expect to see most of the lower grades on 1 and most of the higher grades on 4. The grade distribution could resemble a diagonal line starting a bottom-left 1 and ending at top-right 4.

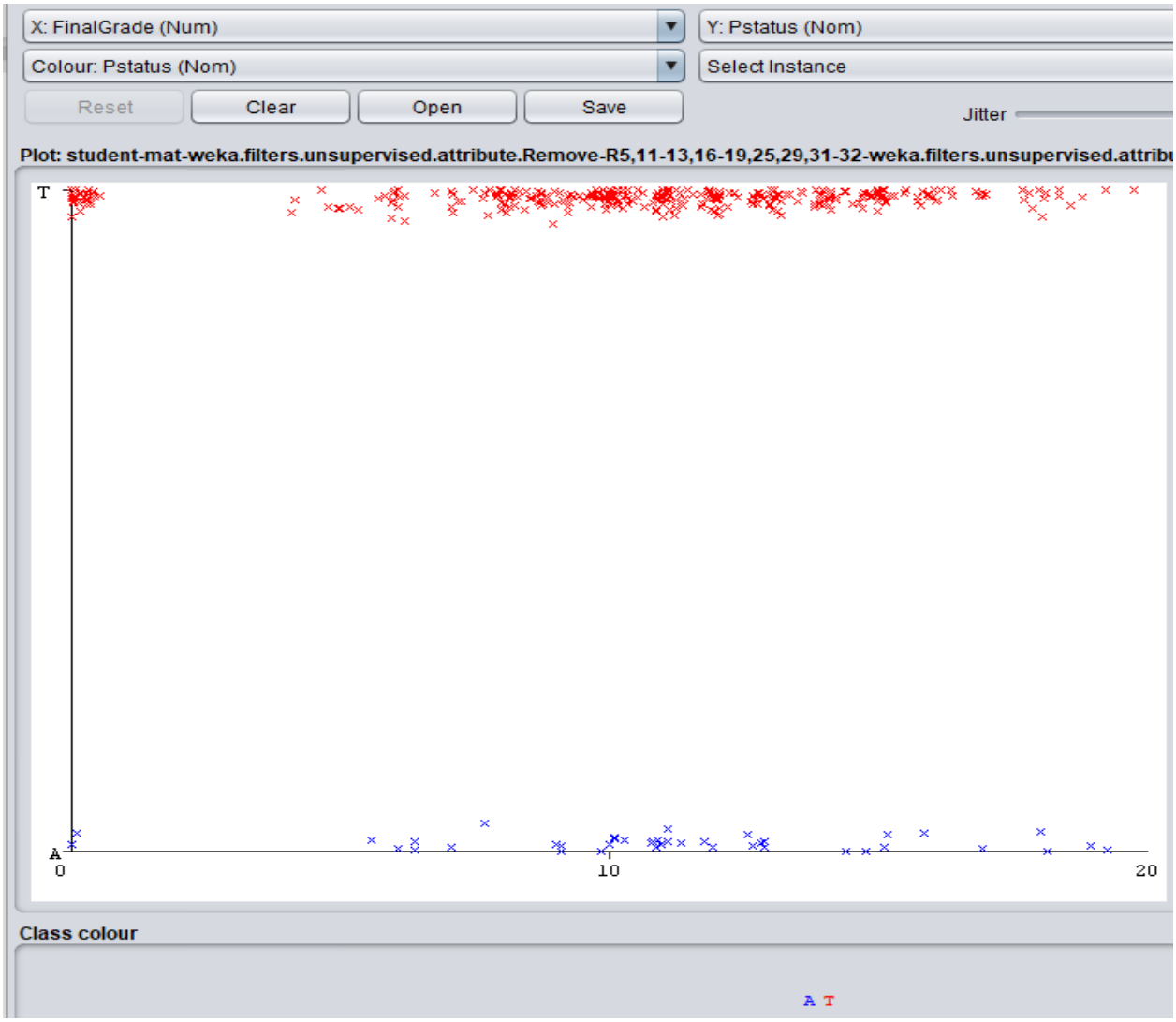
This can also be observed by looking at the grade average of each group of studied hours:

Study time (Hours)	Final Grade average	Change
1 (<2)	10.05	x
2 (2-5)	10.17	0.12
3 (5-10)	11.4	1.23
4 (>10)	11.26	-0.14

We can indeed see increasing grades averages from 1 to 3, but I believe the change is not significant enough to give support to the hypothesis, especially if we take into consideration the difference in studied hours between each group, a big increase in studied hours, low increase in grade. Also, the grade average went down at 4, which could contradict the hypothesis.

- Will the parent's cohabitation status affect the students' final grades?

Hypothesis- Normally I could assume that this situation could affect the student's emotional stability, therefore, affecting their performance in math but, in this case, since the students are in the age range 15-22, I believe that they will have a better emotional maturity. Meaning that they will not be affected.



Based on the visualization I got using Weka I do believe there is enough evidence to support this hypothesis. As we can see in the visualizations the students whose parents live apart and students whose parents live together have similar grade distributions. For example, if we assumed that students whose parents lived apart could get lower final grades since they could be affected emotionally, we would expect to see more blue (A) instances to the left of the visualization and

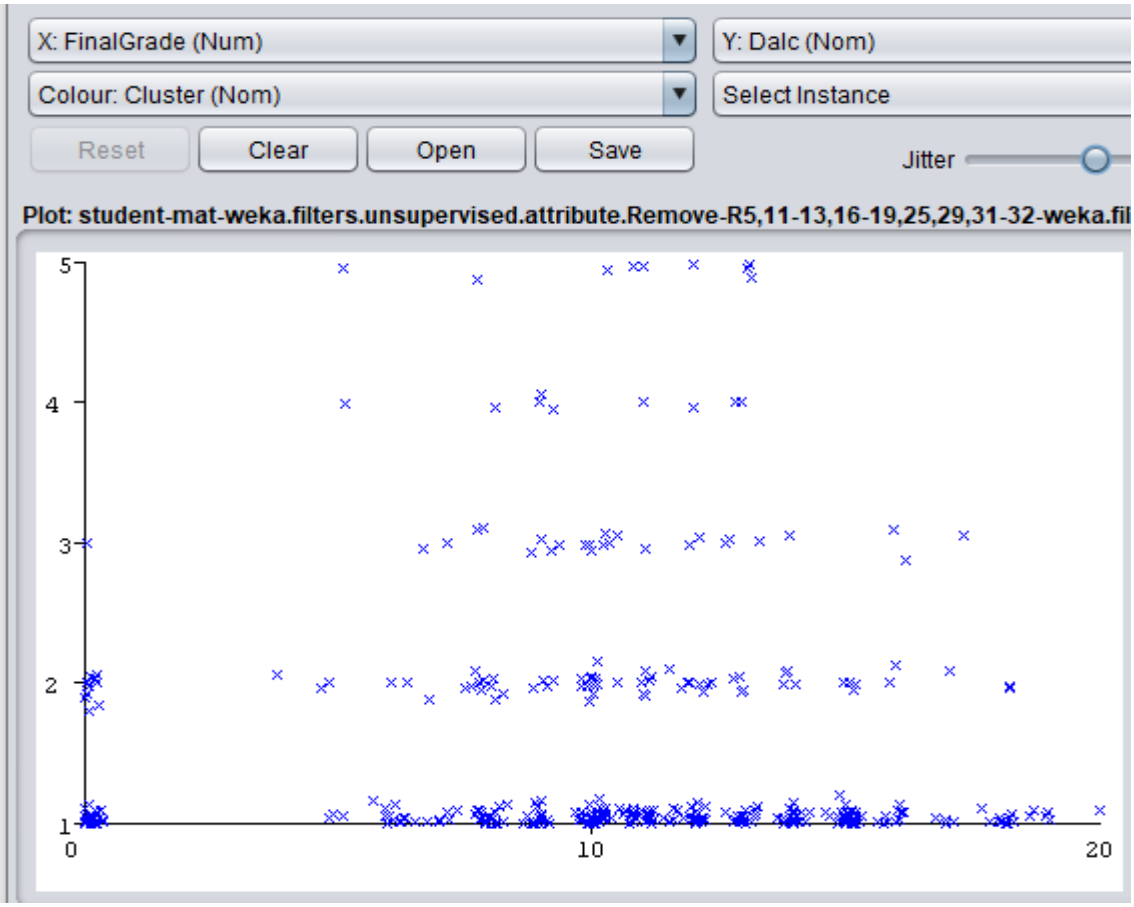
more red (T) instances to the right of the visualization. Instead, we can see a similar distribution between the two.

We can observe this situation better by looking at the averages of each group. There is indeed a difference between the averages, but I do not consider the difference big enough to indicate there is a relationship between students' final grades and their parents' cohabitation status.

Pstatus	Final Grade average	Change
T-Together	10.32	x
A-Apart	11.20	0.88

- Will the level of alcohol consumption on workdays affect the student's final grades?

Hypothesis- Students who have a higher alcohol consumption level during workdays should get lower final grades than students that have a lower alcohol consumption level during workdays. That is since those who have a higher alcohol consumption level during workdays will not be as focused on their academic duties.



Based on the visualization I got using Weka I do not believe there is enough evidence to support this hypothesis. As we can see in the visualization, regardless of the student's level of alcohol consumption during workdays, we can see a similar grade distribution between the groups. If the hypothesis was supported, we could expect to see most of the higher grades on 1 and most of the lower grades on 5. The grade distribution could resemble a diagonal line starting at a top-left 5 and ending at bottom-right 1. It is true that in the visualization we can observe that students with higher levels of alcohol consumption on workdays have less high grades than those with lower levels of alcohol consumption on workdays, but at the same time those with lower levels of

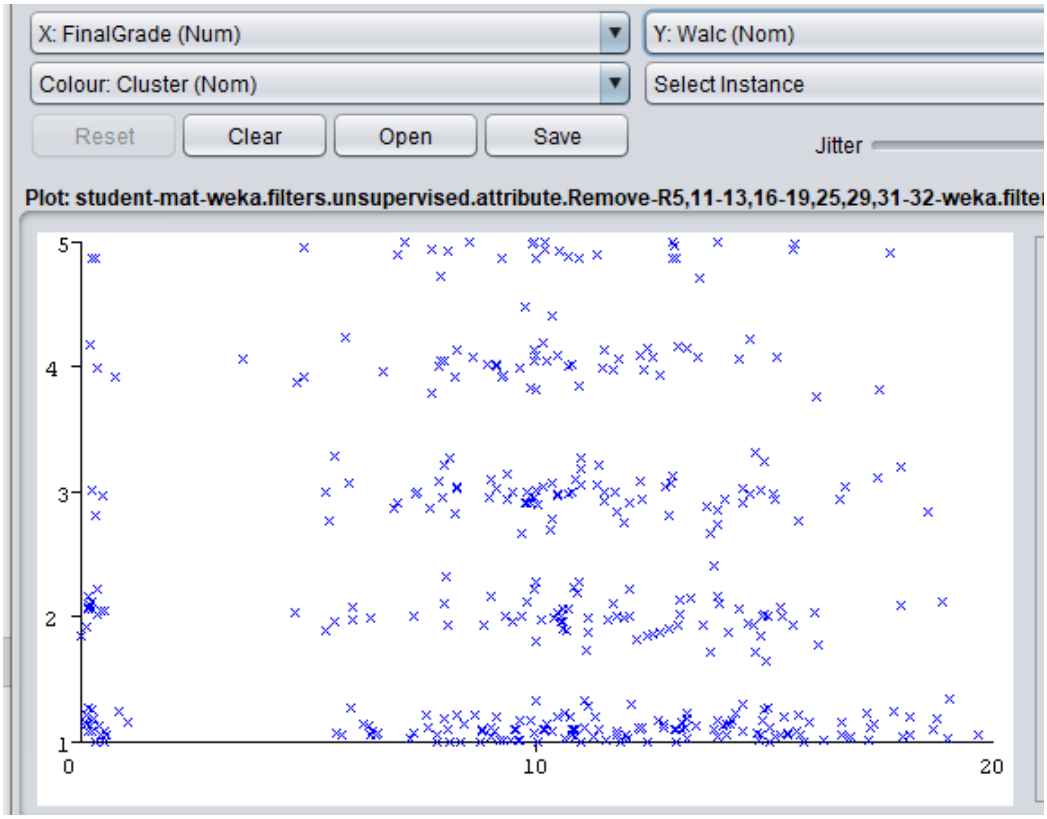
alcohol consumption on workdays have more lower grades than those with higher levels of alcohol consumption on workdays. We can observe this relationship better by looking at the averages of the different groups.

Workday LAC	Final Grade average	Change
1-Very low	10.73	x
2	9.25	-1.48
3	10.5	1.25
4	9.89	-0.61
5- Very high	10.67	0.78

If the hypothesis was supported, we could see a decreasing final grade average as the level of consumption increased, but instead of getting decreasing averages, we get similar averages with a small change. Not only that but we get an alternating pattern of grade decreases and increases as the level of consumption increases. Also, on average those with a level 5 of alcohol consumption in average scored higher than all the other levels of alcohol consumption except for the level of alcohol consumption 1. This could contradict the hypothesis.

-Will the level of alcohol consumption on weekends affect the students' final grades?

Hypothesis- Students that consume alcohol on weekends should be able to get good final grades regardless of the level of consumption since it will not affect their academic duties as much since the consumption will not happen during school days.



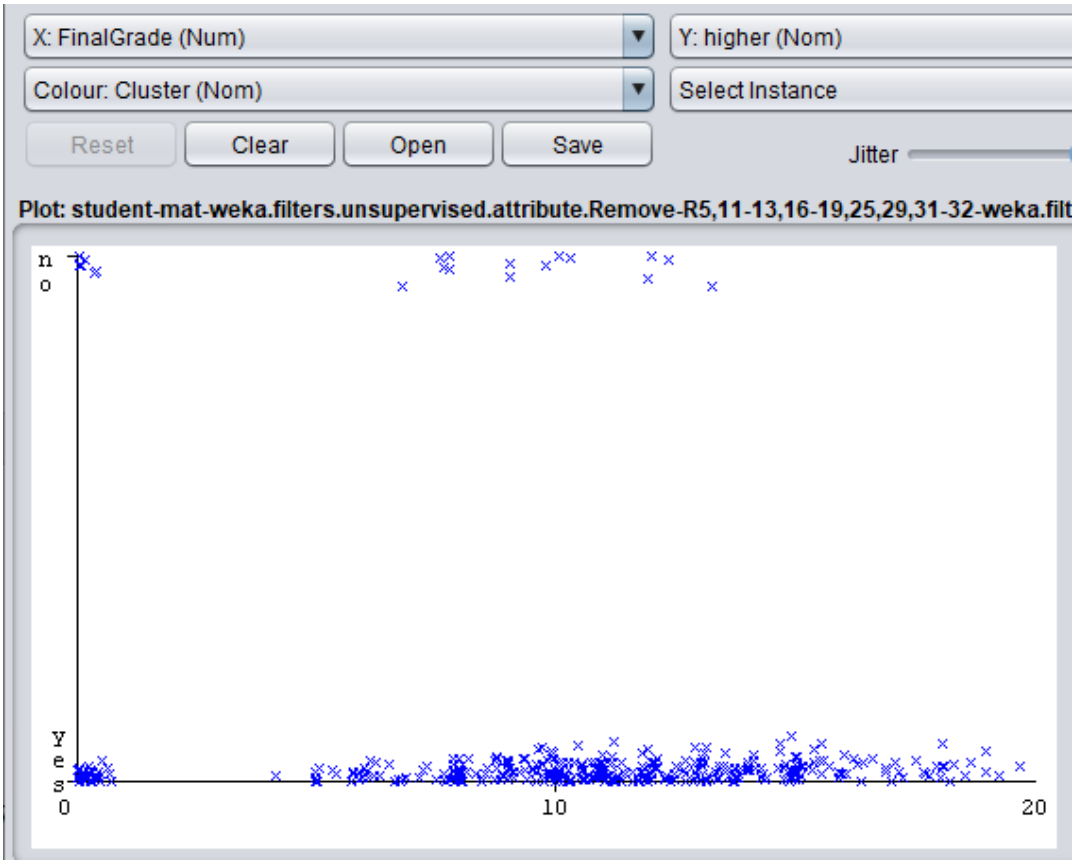
Based on the visualization I got using Weka I do believe there is enough evidence to support this hypothesis. As we can see in the visualization regardless of the students' level of alcohol consumption during the weekends the final grade's distribution is similar between the different levels. For example, if we assumed that a higher level of alcohol consumption during weekends could negatively impact the students' final grades, we could expect to see most of the higher grades on 1 and most of the lower grades on 5. The grade distribution could resemble a diagonal line starting at a top-left 5 and ending at bottom-right 1, but that is not the case. Instead, we can observe a uniform grade distribution among the different levels of consumption. We can observe this situation better by looking at the averages of each level of consumption.

Weekend LAC	Final Grade average	Change
1-Very low	10.74	x
2	10.08	-0.66
3	10.73	0.65
4	9.69	-1.04
5- Very high	10.14	0.45

If the hypothesis was not supported, we could expect to see a decreasing final grade average as the level of alcohol consumption increased, but it is not the case. Instead, we get an alternating pattern of grade increases and decreases as the level of alcohol consumption during the weekend increases. I do not consider the change between the grades significant enough to make other assumptions.

- Will students who get higher final grades be more interested in pursuing higher education?

Hypothesis- Students who are interested in pursuing higher education should get higher final grades since those grades will affect their education possibilities in the future.



Based on the visualization I got using Weka I do believe there is enough evidence to support this hypothesis. As we can see in the visualization even though students who want to pursue higher and students who do not want to pursue higher education have similar grade distributions on the low-grade area and the medium grade area, the latter do not have instances on the high-grade area while students that want to pursue higher education have a considerable number of high grades on that area. This situation can be observed better by looking at the averages of each group.

Interest in higher education	Final Grade average	Change
yes	10.61	x
no	6.8	-3.81

As we can see students who are interested in pursuing higher education have a higher final grade average. I believe that the change between the averages is significant enough to give support to the hypothesis.

- Will the parent's education level impact the student's interest in pursuing higher education?

Hypothesis- Students whose parents have a higher level of education will be more interested in pursuing higher education since they will be influenced by their parent's example or expectations.

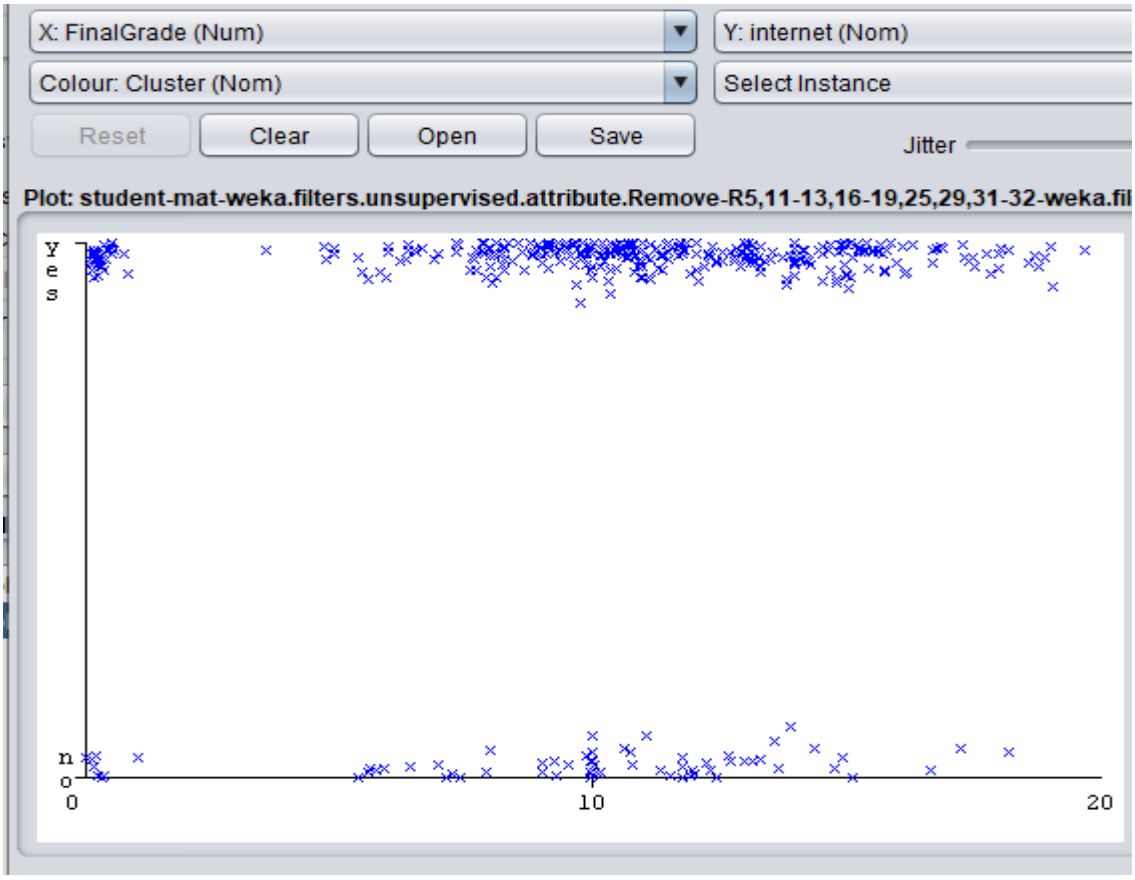
Parent's education	Mother - yes	Mother - no	Father - yes	Father-no
0 - None	0.67	0.33	Not enough data	Not enough data
1- Primary education	0.90	0.10	0.88	0.12
2 - 5^{th}-9^{th} grade	0.93	0.07	0.94	0.06
3 - High school	0.94	0.05	0.98	0.02
4 - Higher education	0.99	0.01	0.99	0.01

Averages of interest on higher education taking into consideration each parent education level.

Based on the information collected I believe there is enough evidence to support this hypothesis. Regardless of which parent we are looking at we can see the same trend. The average of interest (yes) increases as the parent's level of education increases, and the average of interest (no) decreases as the parent's level of education increases. That means that the interest in pursuing higher education increases for students as their parent's level of education increases and the interest in not pursuing higher education decreases as their parent's level of education increases.

- Will the student's access to the internet at home be a factor that will affect their final grades?

Hypothesis- Students that have access to the internet at home will get higher final grades than those who do not since they will have access to more resources that will help them to be successful.

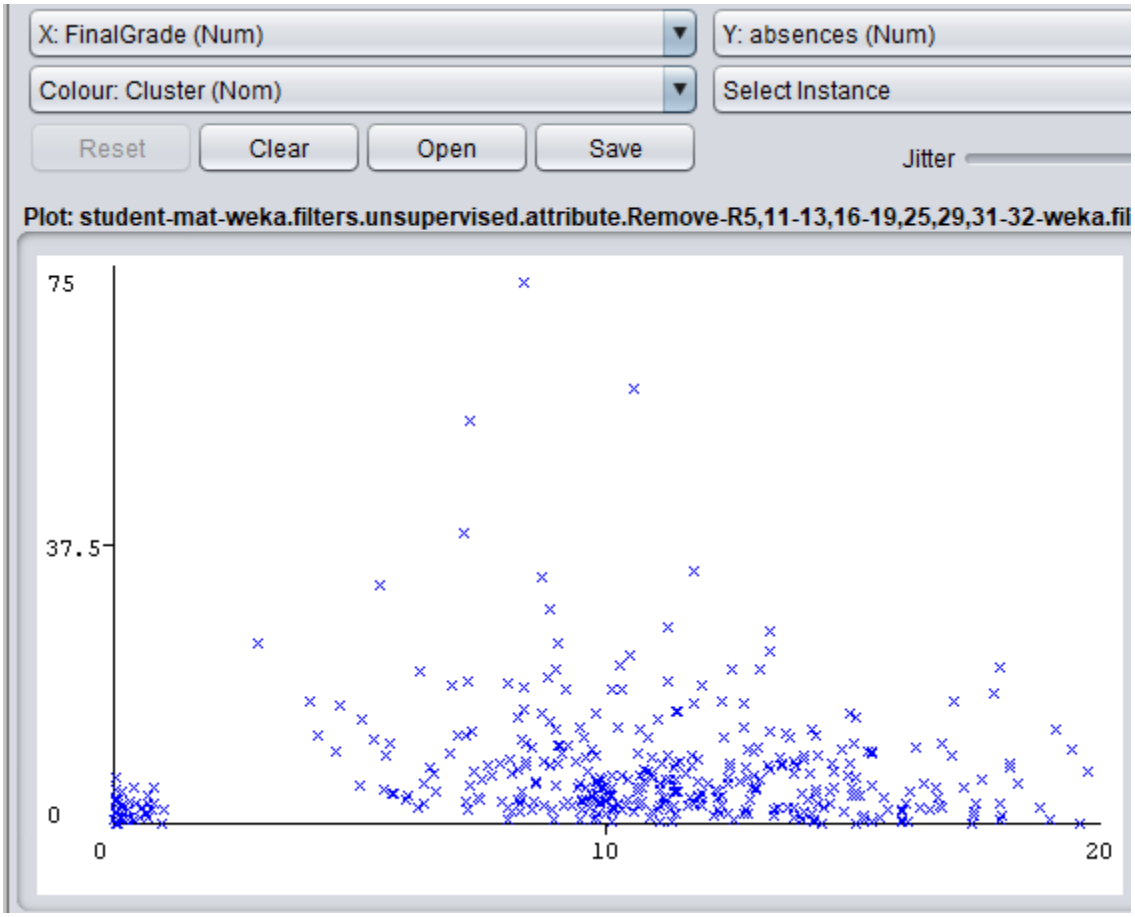


Based on the visualization I got using Weka I do not believe there is enough evidence to support this hypothesis. As we can see in the visualization regardless of the student's access to the internet at home the grade distribution is similar. If the access to the internet at home played a major role in the final grades of the students by giving those with access a major advantage, we could expect to see most of the low grades on the bottom-left of the visualization and most higher grades on the top-right of the visualization. Instead, we see a similar final grade distribution between the groups. We can also see this situation by looking at the averages between the groups. It is true that on average the students who have internet access at home scored higher, I do not believe that the change in grade is significant enough to give enough support to the hypothesis.

Internet access	Final Grade average	Change
yes	10.62	x
no	9.41	-1.21

- Will students' absences be a factor that influences students' final grades?

Hypothesis- Students who have a low number of absences will have higher final grades than those who have a higher number of absences since the latter missed more lectures, meaning that they could be less prepared for their assessments.



Based on the visualization I got using Weka I do not believe there is enough evidence to support this hypothesis. If we assumed that students with a large number of absences could earn lower grades since they could be less prepared for their assessments due to missed lectures, we could expect to see more lower values on the top-left of the visualization and more higher values on the bottom-right of the visualization. The grade distribution could resemble a diagonal line starting

at a top-left and ending at bottom-right, but that is not the case. Is true that most of the higher grades are on the bottom-right, but at the same time most lower values are on the bottom left, and bottom-middle. To observe this situation better we can look at the averages.

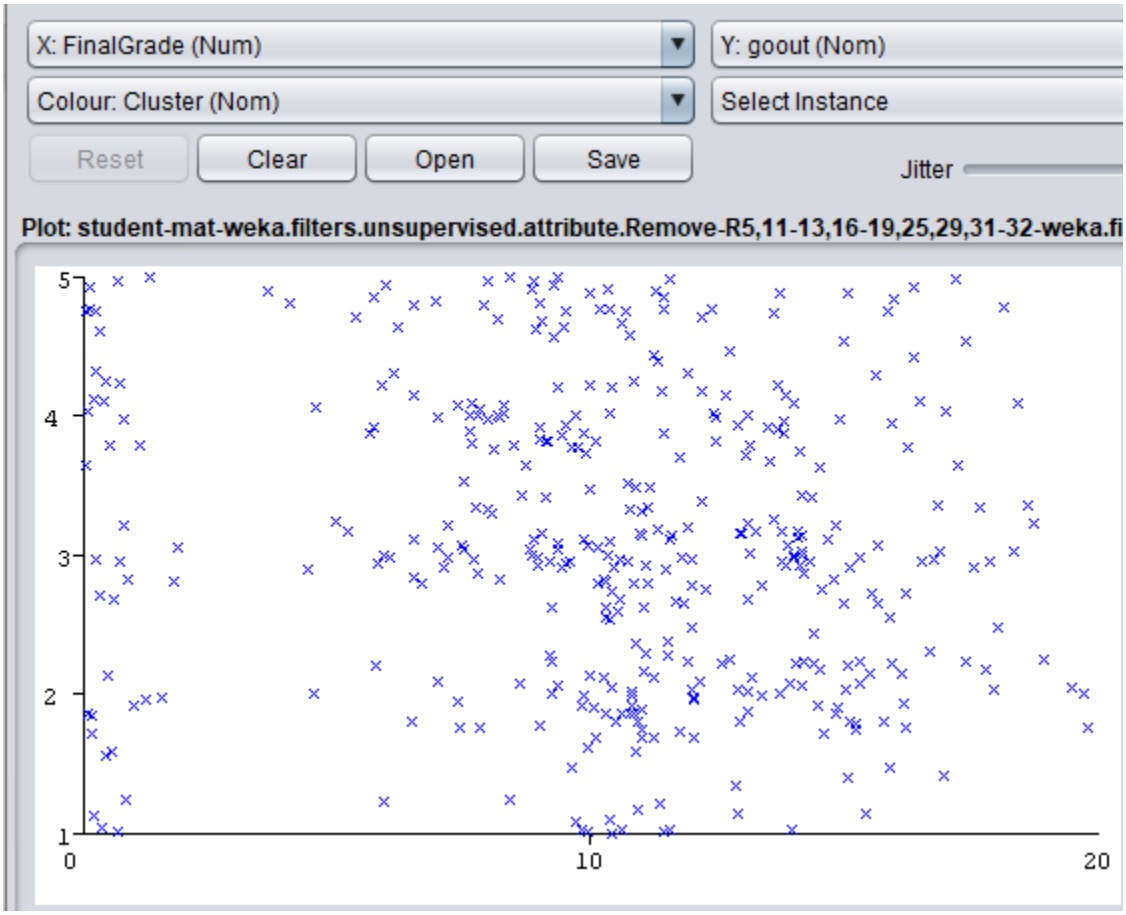
Absences range	Final Grade average	Change
0 > 5	10.17	x
6 > 10	11.4	1.23
11 > 15	10.27	-1.13
16 > 20	9.83	-0.44
21 > 25	12	2.17
26 > 30	10.17	-1.83
36 > 40	9.5	-0.67
51 > 55	11	1.5
56 > 60	8	-3

Some ranges were omitted for lack of data.

If the hypothesis was supported, we could expect to see a decreasing final grade average as the number of absences increased. Instead, we can observe an alternating pattern of increases and decreases as the number of absences increases. Not only that but the highest final grade average is at the center of the table when if the hypothesis was supported, we could expect to see it at the beginning of the table.

- Will the level of time students go out with friends affect the students' final grades?

Hypothesis- I believe that students who have a higher level of going out with friends will have lower final grades than students with a lower level of going out with friends since they will have less time to focus on their academic duties than the latter.



Based on the visualization I got using Weka I do not believe there is enough evidence to support this hypothesis. If we assumed that students who have a higher level of going out with friends could earn lower grades than those with a lower level since they could have less time to focus on their academic duties, we could expect to see most lower grades on the top-left of the visualization and most higher grades on the bottom-right of the visualization. The data distribution could resemble a diagonal line starting at top-left 5 and ending at lower-right 1. Instead, we can observe a very similar distribution between the final grades regardless of the level of going out with friends. We can observe this situation better by looking at the averages.

Go Out w/t friends	Final Grade average	Change
1-Very low	9.87	x
2	11.19	1.32
3	10.96	-0.23
4	9.65	-1.31
5- Very high	9.04	-0.61

If we assumed that the hypothesis was supported, we could expect to see a decreasing final grade average as the level of going out with friends increased, but that is not the case. We can observe that pattern after level 2, but even then, I could not consider the change significant enough to give support to the hypothesis. If anything, it seems that going out with friends can be beneficial at a lower level and prejudicial at a higher level. More precisely it could look that level 2 is the most beneficial for the students.

Conclusion:

Of my nine hypotheses, 4 of them were supported by the data analyzed, while 5 of them were not. Strictly referring to Portuguese math students, we learned that as suspected; students' parent's cohabitation status did not have a major effect on their final grades, students' level of alcohol consumption on weekends did not have a major effect on their final grades, students that got higher final grades were more interested in pursuing higher education, and students whose parents had a higher level of education were more interested in pursuing higher education. Opposed as suspected, more study hours did not have a major effect on the students' final grades, students' level of alcohol consumption on workdays did not have a major effect on their final grades, access to the internet at home did not have a major effect on their final grades, the number of absences did not have a major effect on their final grades, and the level of going out with friends did not have a major effect on their final grades.

From this, I learned that just because some students live in more favorable conditions or have more resources, that does not mean they will earn higher grades than those who live under

less favorable conditions or have fewer resources. I believe that students that want to earn high grades, will make the most of what they have.

Work Cited

Cortez, Paulo & Silva, Alice. (2008). Using data mining to predict secondary school student performance. EUROSIS. https://www.researchgate.net/publication/228780408_Using_data_mining_to_predict_secondary_school_student_performance

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+Performance