BUS 462 Business Analytics (D100)

Professor Michael Brydon

Final Analytics Project

December 10, 2023

Austin Yu (301372579)

Azil Rasyani (301393062)

Carmen Chen (301360068)

Harveer Dhadda (301379347)

Ramiz Sabahat (301290512)

Executive Summary

We discovered a data set on Kaggle that produced the final grades of students in a high school. The final grades were the dependent variable in this scenario as the grade was contingent upon numerous independent variables like study time, free time, absences, social, etc. As schools begin to gauge and assess the factors it will help them understand their students' habits and allow them to allocate resources and programs that will assist the students in their educational journey. Our goal was to utilize these variables to make assumptions and determine what factors specifically impact final grade performances. To assist us in our decision-making process we plan to go through the process of root cause analysis, to help quantify the existence and strength of the relationship between each explanatory variable and the response variable. We plan to use multiple business analytics tools like SAS and KNIME to perform regressions, graphical outputs, and classification trees, etc. to help us achieve our expected results and support our analysis to conclude which explanatory variable has a significant impact on final grades (our response variable) of students.

The findings of this report are meant to guide decision-making processes within schools, facilitating the formulation of targeted interventions and educational strategies aimed at nurturing student success.

Moving forward, our subsequent report sections will delve into a detailed exploration of influential factors impacting final grades, offering a systematic breakdown of our analytical approach, findings, and their implications for educational institutions.

1) **SAS Multiple Linear Regression Analysis:**

To analyze the impact of various variables in the school dataset, a linear regression was performed using SAS Enterprise Guide. The resulting model demonstrated a reasonable, albeit moderately weak, R-squared value of 0.3428. This value indicates the proportion of variance explained in the log-transformed final grade by the variables included in the model.

Upon examination of the output below, the standardized estimates offer insights into the relative impact of variables on the log-transformed final grade. Out of the 18 significant variables in our linear regression model, class_failures emerges as the variable with the highest standardized estimate, indicating its substantial impact on the log-transformed final grade. The interpretations of several key variables, accounting for the log transformation, are detailed below:

- **Class_failures (Standardized Estimate = -0.25738, Parameter Estimate = -0.10843):** For every one-unit increase in class_failures, it corresponds to an approximate 10.8% decrease in the final grade, after log transformation.

- **WantsHigherEd (Standardized Estimate = 0.17326, Parameter Estimate = 0.14028):** When students aspire to pursue higher education, the final grade, after log transformation, is expected to be approximately 14.0% higher compared to those without such aspirations.
- **StudyTime<2hours (Standardized Estimate = -0.14064, Parameter Estimate = -0.07343):** Students studying less than 2 hours, after log transformation, are anticipated to achieve a final grade approximately 7.3% lower than students who study longer.
- **MotherEducationHigher (Standardized Estimate = 0.13092, Parameter Estimate = 0.07195):** Students with mothers who pursued higher education are predicted to attain a final grade approximately 7.2% higher than students without this educational background, considering the log transformation.
- **HasSchoolSupport (Standardized Estimate = -0.12056, Parameter Estimate = -0.09565):** Students receiving support from the school are expected to have a final grade approximately 9.6% lower, given the log transformation, compared to those without such support.
- **SchoolChoiceReputation (Standardized Estimate = 0.11192, Parameter Estimate = 0.06565):** Students selecting a school based on its reputation are projected to achieve a final grade about 6.6% higher, considering the log transformation, compared to students without this consideration.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: LogFinalGrade

Number of Observations Read	649
Number of Observations Used	634
Number of Observations with Missing Values	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	12.93138	0.71841	17.83	<.0001
Error	615	24.78627	0.04030		
Corrected Total	633	37.71765			

Root MSE	0.20076	R-Square	0.3428
Dependent Mean	2.47362	Adj R-Sq	0.3236
Coeff Var	8.11587		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	95% Confidence Limits	
Intercept	1	2.50405	0.05116	48.95	<.0001	0	2.40358	2.60451
IsFemale	1	0.04829	0.01818	2.66	0.0081	0.09726	0.01260	0.08398
class_failures	1	-0.10843	0.01479	-7.33	<.0001	-0.25738	-0.13748	-0.07938
social	1	-0.01847	0.00720	-2.57	0.0105	-0.08804	-0.03262	-0.00433
weekday_alcohol	1	-0.02077	0.00971	-2.14	0.0329	-0.07802	-0.03984	-0.00169
health	1	-0.00954	0.00569	-1.68	0.0940	-0.05660	-0.02071	0.00163
absences	1	-0.00428	0.00181	-2.37	0.0181	-0.08181	-0.00784	-0.00073482
MotherEducationHigher	1	0.07195	0.01986	3.62	0.0003	0.13092	0.03295	0.11095
FatherJobServices	1	-0.03508	0.01834	-1.91	0.0562	-0.06440	-0.07109	0.00093578
FatherJobTeacher	1	0.07632	0.03756	2.03	0.0426	0.07146	0.00255	0.15008
SchoolChoiceHome	1	0.05610	0.02035	2.76	0.0060	0.09684	0.01614	0.09607
SchoolChoiceReputation	1	0.06565	0.02101	3.12	0.0019	0.11192	0.02438	0.10691
StudyTime<2hours	1	-0.07343	0.02411	-3.05	0.0024	-0.14064	-0.12079	-0.02608
StudyTime2to5hours	1	-0.04371	0.02139	-2.04	0.0415	-0.08944	-0.08572	-0.00170
HasSchoolSupport	1	-0.09565	0.02676	-3.57	0.0004	-0.12056	-0.14821	-0.04309
HasFamilySupport	1	-0.02898	0.01691	-1.71	0.0870	-0.05764	-0.06219	0.00422
InExtraCurricular	1	0.03823	0.01655	2.31	0.0213	0.07833	0.00572	0.07074
WantsHigherEd	1	0.14028	0.02882	4.87	<.0001	0.17326	0.08368	0.19688

Figure 1. Linear Regression Model

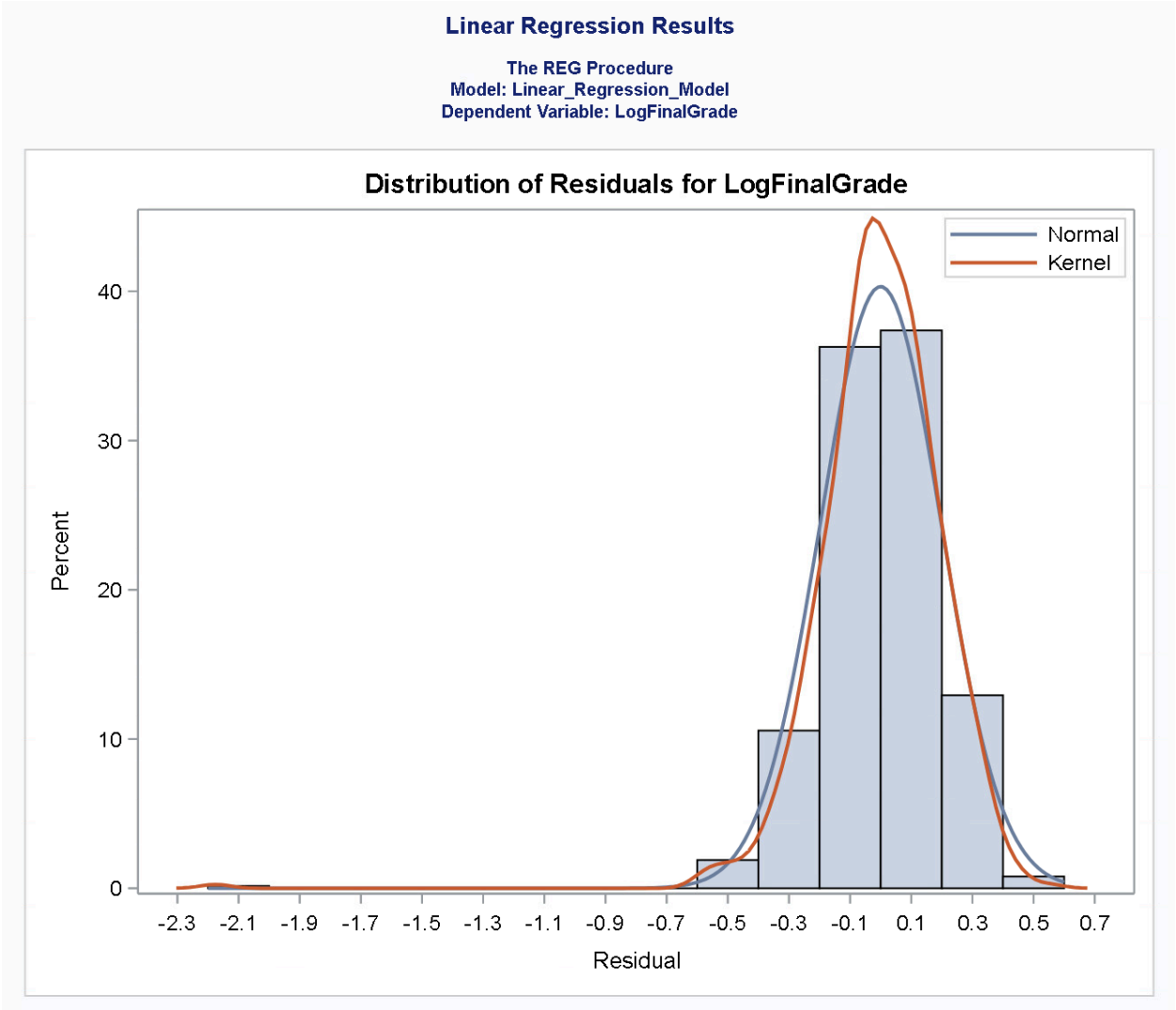


Figure 2. Log Final Grade Histogram

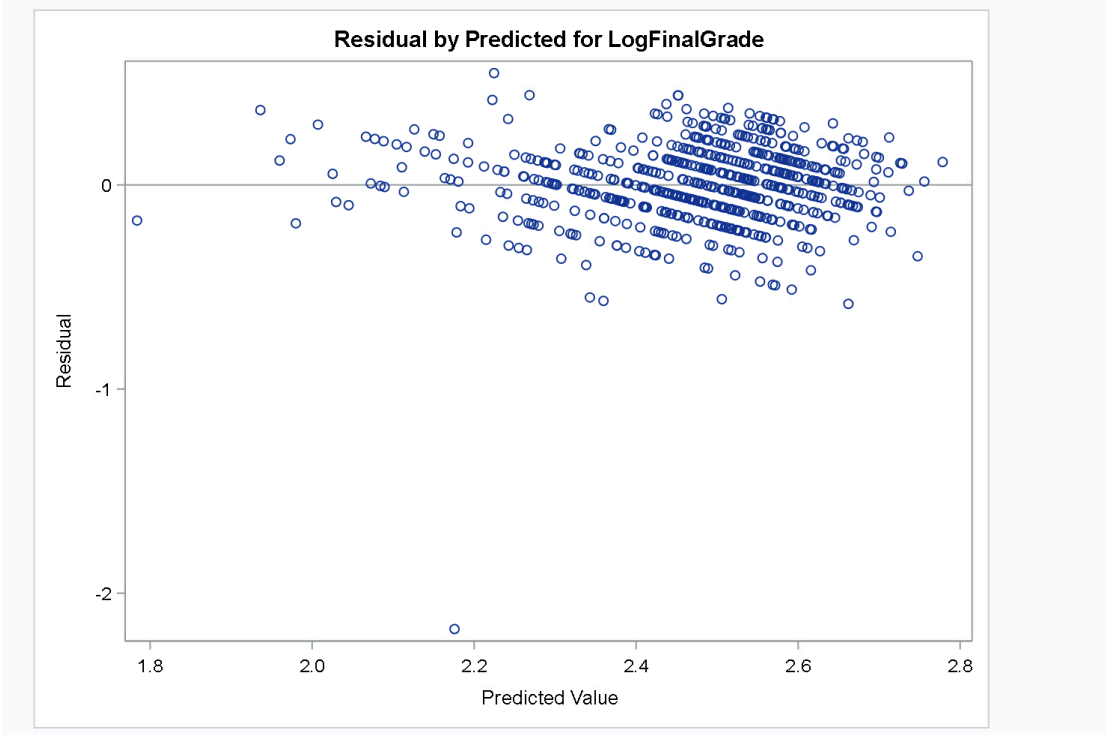


Figure 3. Log Final Grade Residual Scatter Plot

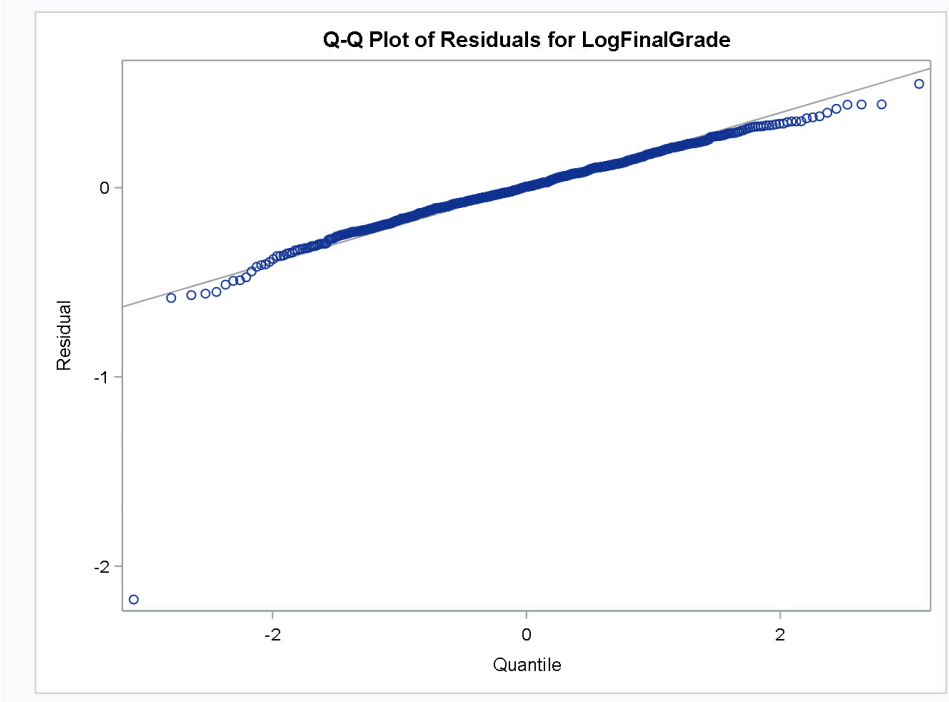
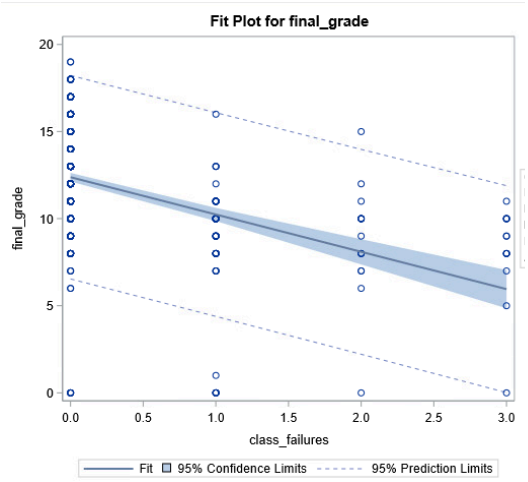


Figure 4. Log Final Quantile Plot

Interpretation of Graphics for SAS

For our analysis, we used multiple explanatory variables to predict students' final grade (response variable). First, we ran a multiple linear regression on SAS and used backward elimination to get the linear regression results shown in the above. However, we had to recode multiple explanatory variables and log the final grade to give us a more normalized distribution. In this case, our model (Figure 1) depicts that we have a R-square of 0.3428, so, our model explains 34% of the observed variance in the response variable. Next, we examine the residuals and it depicts a fairly normal distribution shown in Figure 2, Figure 3, and Figure 4. However, the histogram and scatter plot shows that there are some large negative residuals. This may be due to the data points in the model that underestimates the observed values, but a further investigation is required to identify the limitations and improvements for this model.

Regarding the most impactful variable, which is the amount of previous classes the student has failed, the following plot is a fit plot for the variables class_failures vs the variable final_grade to provide an isolated view of its impact on final grade:



While there are outliers present, the causes may perhaps be explained by other factors in the model or even factors not considered in the model.

Regarding the 2nd most impactful variable within the context of this model, which is whether the student aspires to pursue higher education or not, the two-sample t test output below highlights the differences between these two categories of students. With a difference in means of 0.2683, the students with higher education aspirations have on average a 30.77% higher final grade compared to those who do not.

WantsHigherEd	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		64	2.2324	0.1920	0.0240	1.6094	2.6391
1		570	2.5007	0.2344	0.00982	0	2.9444
Diff (1-2)	Pooled		-0.2683	0.2305	0.0304		
Diff (1-2)	Satterthwaite		-0.2683		0.0259		

Regarding the 3rd most impactful variable within the model, which is whether students studied less than 2 hours or more than 2 hours per week, the two-sample t-test output below highlights the difference between these two categories of students. With a difference in means of 0.1109, students who studied more than 2 hours every week had on average a 11.72% higher final grade than students who studied less than 2 hours every week.

StudyTime<2hours	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		430	2.5093	0.2469	0.0119	0	2.9444
1		204	2.3984	0.2205	0.0154	1.6094	2.8904
Diff (1-2)	Pooled		0.1109	0.2387	0.0203		
Diff (1-2)	Satterthwaite		0.1109		0.0195		

2) **KNIME Regression Tree Analysis:**

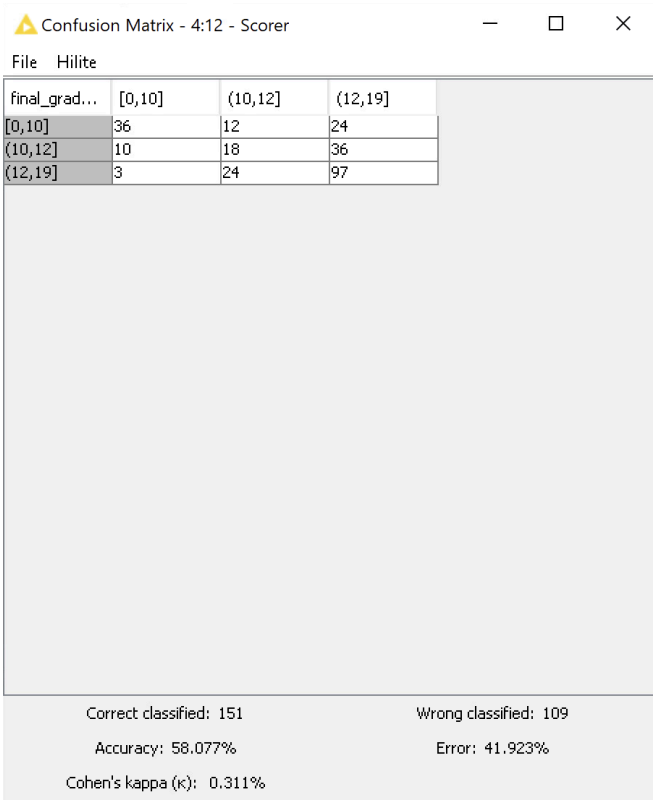
Classification Tree Analysis:

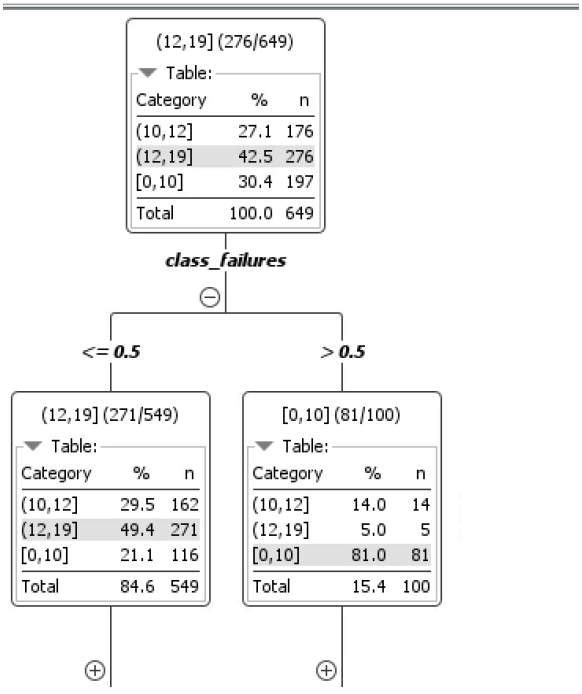
We initiated the classification tree analysis using Knime by categorizing final grades into three distinct bins: low, medium, and high. These categories were formed to represent the bottom third, middle third, and top third of students based on their final grades, each holding equal frequency, enabling a balanced set of classes for improved model information.

Utilizing an auto binning method, the data was segmented into three categories: 0-10 for low grades, 10-12 for medium grades, and 12-19 for high grades. We refined the model by employing an auto binner node, a decision tree learner, and a column filter, focusing solely on relevant explanatory variables used in our prior SAS analysis (shown in Classification Tree Workflow in Appendix) .

A partitioning approach was also adopted, dividing the data into training (60%) and testing (40%) sets for model refinement and accuracy assessment. Through hold-back sampling and subsequent pruning, we achieved approximately 58% accuracy, an important metric for evaluating the model's predictive capacity (as shown in the confusion matrix below). This procedure was applied to the entire dataset (“All Data” node shown in Classification workflow in Appendix), ensuring consistency and accuracy across the complete dataset. This allowed us to

get the best possible prediction using all 649 pieces of data instead of a subset we used in the hold back partitioning sample.

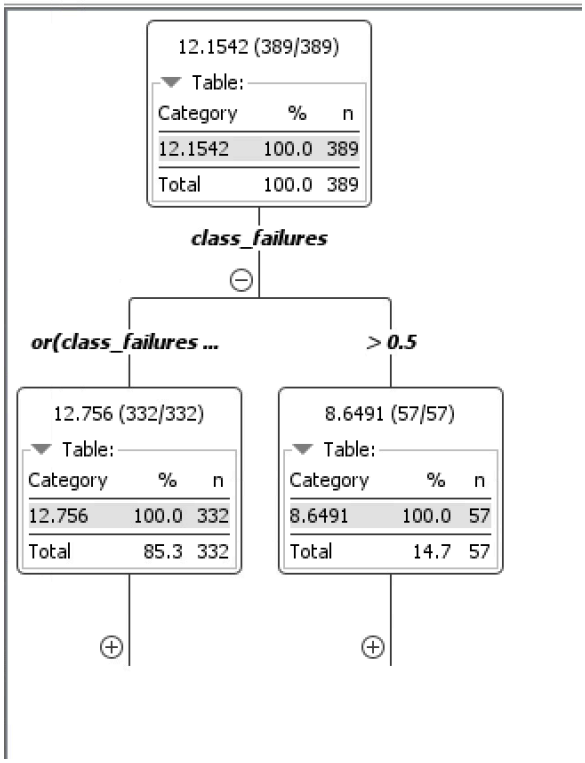




As shown in the graphic above which shows the first node of our classification tree, it shows that class failures, just as it was in our SAS analysis, is one of the significant predictor variables.

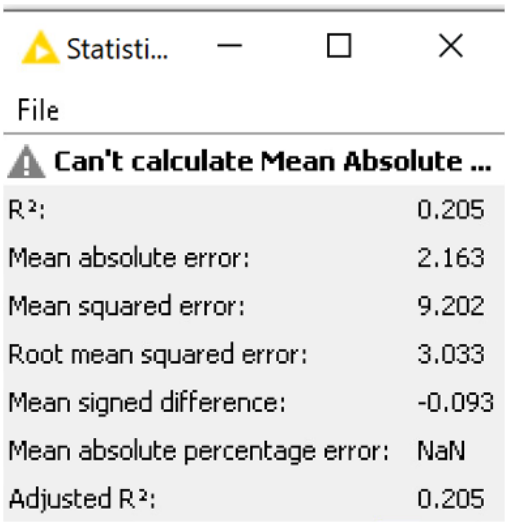
Regression Tree Analysis:

The regression tree analysis furthered our understanding of influential variables affecting final grades. Notably, class_failures emerged as a significant predictor, mirroring its importance revealed in the multiple linear regression via SAS. As shown on the regression tree below, class_failures is one of the more important discriminating variables.

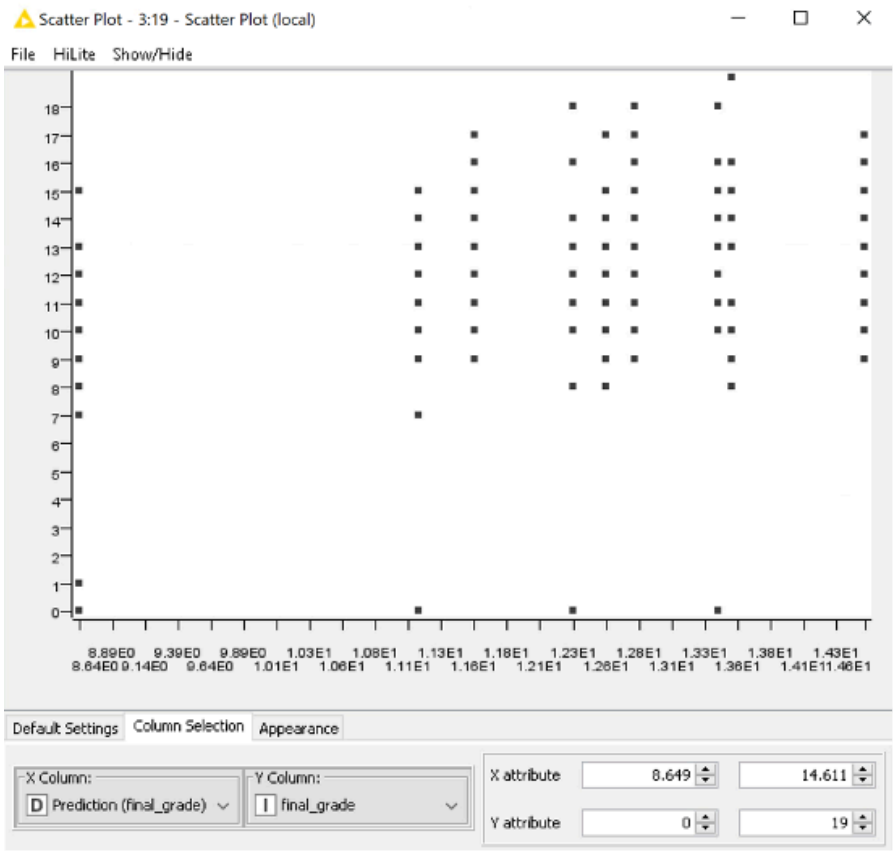


The root of the regression tree emphasized class_failures indicates that students with more than 0.5 class failures had an estimated mean final grade of 8.6, with 57 students who satisfy this condition; while those with fewer than 0.5 class failures had an expected average of 12.7, with 332 students who satisfy this condition. This reaffirmed the variable's significance in predicting final grades.

Evaluating the model's accuracy using a testing loop revealed an R-square value of 0.205 as seen below.



Additionally, a scatter plot depicting predicted versus actual final grades showcased a discernible yet slightly variable trend, revealing areas where the model's predictive power could be further enhanced. As seen below, the values shown on the far left show that within the students that match the criteria for the prediction, the range of values are quite large, hence why we only achieved a 0.205 r square.



Conclusion - Summary of Findings/Business Implications

The SAS analysis applied multiple linear regression techniques to explore the relationship between various explanatory variables and final grades. Among the factors investigated, class_failures emerged as a significant predictor negatively impacting final grades. Contrary to expectations, an increase in class_failures was associated with a decrease in final grades. The model highlighted several key variables impacting final grades, including students' aspirations for higher education, study time, parental education levels, school support, and school choice based on reputation.

The application of both classification and regression tree analyses through Knime offered valuable insights and furthered our understanding into the factors influencing final grades for students. While the classification tree provided a segmented understanding, the regression tree underscored class_failures as a pivotal variable affecting final grades.

Both methodologies consistently highlighted the pivotal role of class_failures in predicting final grades. While the multiple linear regression in SAS revealed the individual impact of various factors, the classification and regression trees in Knime underscored class_failures as a key discriminator affecting final grades. The analyses collectively stress the importance of identifying and addressing class_failures as a crucial factor influencing student academic performance.

Understanding these variables' impact on final grades can aid educational institutions in devising targeted interventions. Strategies focusing on reducing class_failures, fostering a desire for higher education, optimizing study time, and enhancing school support systems can significantly improve student outcomes. These insights enable educational administrators to allocate resources effectively, tailor support mechanisms, and design proactive measures to elevate student academic achievements.

Brief Description of the Data Source

The dataset in question is sourced from Kaggle, specifically from the "High School Student Performance and Demographics" dataset available at the link. This dataset is a comprehensive collection of data points covering various aspects of high school students' lives, including demographic details, family backgrounds, school-related information, and personal behaviors, with the aim of analyzing their impact on academic performance of portuguese subject.

Mini Data Dictionary

1. **Response Variable**: 'LogFinalGrade' - Logarithm of the final grade, representing the academic performance of a student.
2. **Demographic Variables**: 'Age', 'Sex', 'Address Type' (Urban/Rural).
3. **Family Background Variables**: 'Family Size', 'Parent Status' (living together/apart), 'Mother's Education', 'Father's Education', 'Mother's Job', 'Father's Job'.
4. **School-Related Variables**: 'School Support' (indicates extra educational support), 'Study Time', 'Failures' (number of past class failures).
5. **Behavioral Variables**: 'Internet Access' (yes/no), 'In a Relationship' (yes/no), 'StudyTime2to5hours' (a recoded variable indicating if study time is between 2 to 5 hours).
6. **New Variables Created: (All Dummy Variables)**

   StudyTime2to5hours indicates whether a student studies between 2 to 5 hours per week, a factor often directly correlated with academic performance. **HasSchoolSupport** reflects whether the student receives extra educational support from the school, which could significantly impact learning outcomes. **HasFamilySupport** is included to determine if the student has educational support from their family, a key aspect in academic success. The variable **InExtraPaidClasses** shows whether a student is enrolled in extra paid classes within the course subject, potentially indicating additional learning opportunities or academic pressure. **WantsHigherEd** represents the student's aspiration for higher education, a variable that could be a strong motivator for academic performance. **HasInternetAccess** is crucial in the modern educational context, indicating whether the student has home internet access for resource and information access. Lastly, **InRelationship** provides insight into whether the student is in a romantic relationship, which can influence their academic focus and performance.

Granularity and Reference

- **Granularity**: The dataset is granular at the individual student level, meaning each row represents a unique student and their specific characteristics.
- **Reference**: It refers to high school students who are taking the Portuguese course, capturing a wide range of factors from personal demographics to family and school-related variables.

Appendix

Main data dictionary:

Variable Name	Description	Data Type
school	Student's school (Gabriel Pereira or Mousinho da Silveira)	Binary
sex	Student's sex (female or male)	Binary
age	Student's age	Numeric
address_type	Student's home address type (Urban or Rural)	Binary
family_size	Family size (≤3 or >3)	Binary
parent_status	Parent's cohabitation status (Living together or Apart)	Binary
mother_education	Mother's education level	Ordinal
father_education	Father's education level	Ordinal
mother_job	Mother's job type	Nominal
father_job	Father's job type	Nominal
reason	Reason to choose this school	Nominal
guardian	Student's guardian	Nominal
travel_time	Home to school travel time	Ordinal
study_time	Weekly study time	Ordinal
class_failures	Number of past class failures	Numeric
school_support	Extra educational support (yes or no)	Binary
family_support	Family educational support (yes or no)	Binary
extra_paid_classes	Extra paid classes within the course subject	Binary
activities	Extra-curricular activities (yes or no)	Binary
nursery	Attended nursery school (yes or no)	Binary
higher_ed	Wants to take higher education (yes or no)	Binary
internet	Internet access at home (yes or no)	Binary

romantic_relationship	With a romantic relationship (yes or no)	Binary
family_relationship	Quality of family relationships	Numeric
free_time	Free time after school	Numeric
social	Going out with friends	Numeric
weekday_alcohol	Workday alcohol consumption	Numeric
weekend_alcohol	Weekend alcohol consumption	Numeric
health	Current health status	Numeric
absences	Number of school absences	Numeric
grade_1	First period grade (Math or Portuguese)	Numeric
grade_2	Second period grade (Math or Portuguese)	Numeric
final_grade	Final grade (Math or Portuguese, output target)	Numeric

New/ Dummy Variables:

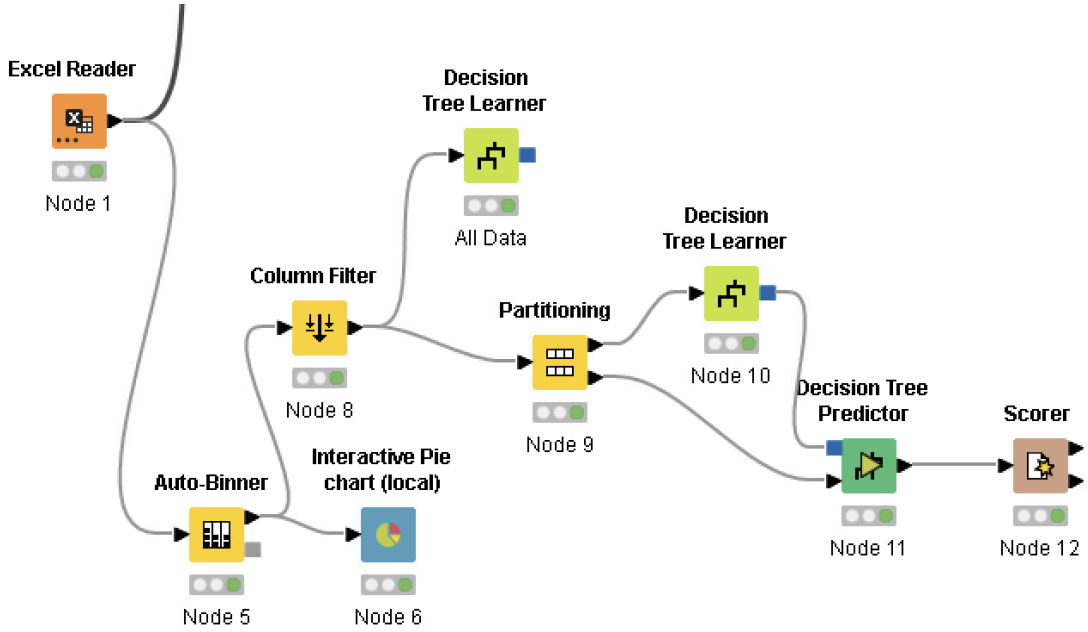
Variable Name	Description	Data Type
IsUrban	Student lives in an urban area (1) or not (0)	Binary
FamilySize>3	Family size is greater than 3 (1) or not (0)	Binary
ParentsLiveTogether	Parents are living together (1) or apart (0)	Binary
MotherEducationNone	Mother has no education (1) or not (0)	Binary
MotherEducationSecondary	Mother has secondary education (1) or not (0)	Binary
MotherEducationHigher	Mother has higher education (1) or not (0)	Binary
FatherEducationNone	Father has no education (1) or not (0)	Binary
FatherEducationSecondary	Father has secondary education (1) or not (0)	Binary
FatherEducationHigher	Father has higher education (1) or not (0)	Binary
MotherJobHome	Mother's job is at home (1) or not (0)	Binary

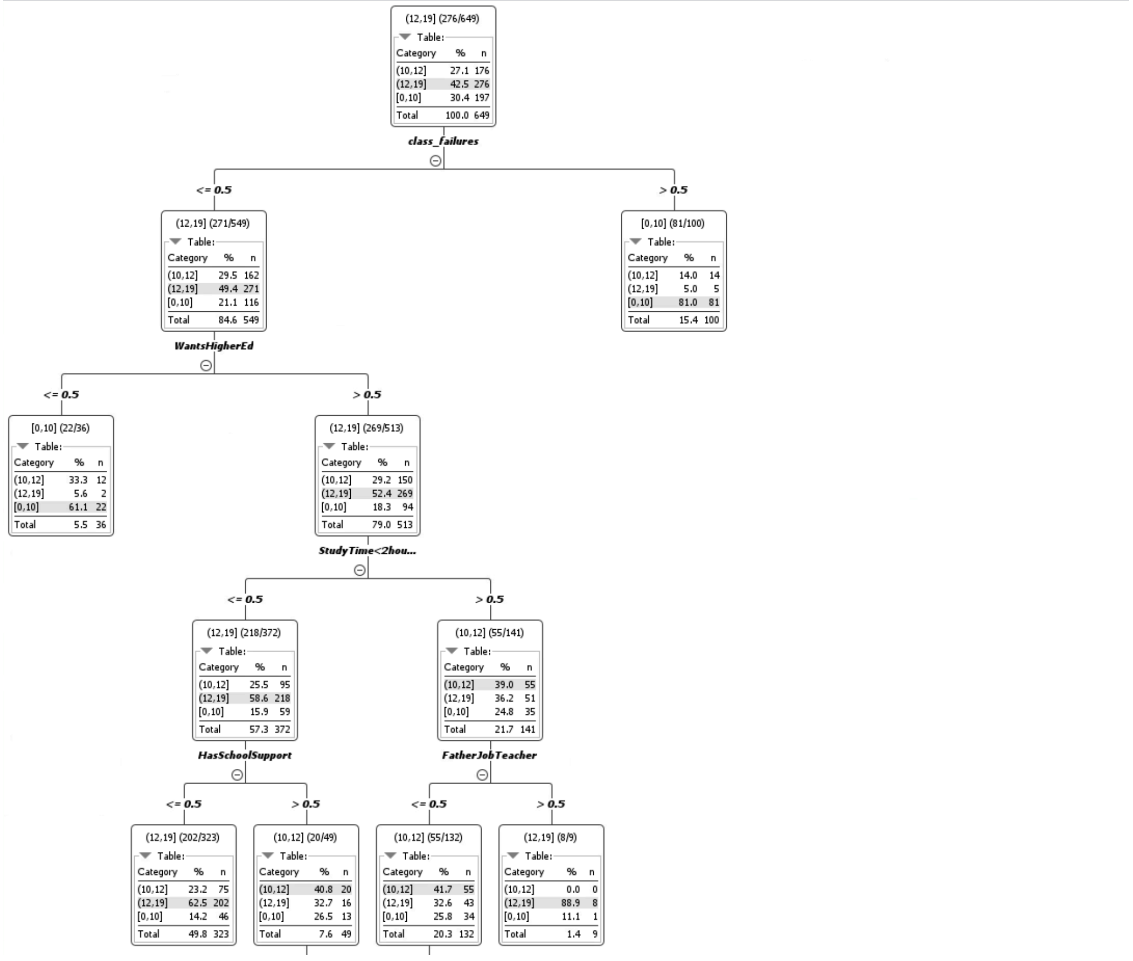
MotherJobHealth	Mother's job is health related (1) or not (0)	Binary
MotherJobServices	Mother's job is in services sector (1) or not (0)	Binary
MotherJobTeacher	Mother is a teacher (1) or not (0)	Binary
FatherJobHome	Father's job is at home (1) or not (0)	Binary
FatherJobHealth	Father's job is health related (1) or not (0)	Binary
FatherJobServices	Father's job is in services sector (1) or not (0)	Binary
FatherJobTeacher	Father is a teacher (1) or not (0)	Binary
SchoolChoiceCourse	School chosen for course preference (1) or not (0)	Binary
SchoolChoiceHome	School chosen for being close to home (1) or not (0)	Binary
SchoolChoiceReputation	School chosen for its reputation (1) or not (0)	Binary
GuardianIsFather	Guardian is father (1) or not (0)	Binary
GuardianIsMother	Guardian is mother (1) or not (0)	Binary
TravelTime>1hour	Travel time to school is more than 1 hour (1) or not (0)	Binary
TravelTime<15min	Travel time to school is less than 15 minutes (1) or not (0)	Binary
TravelTime30minto1hour	Travel time to school is 30 minutes to 1 hour (1) or not (0)	Binary
StudyTime<2hours	Study time is less than 2 hours per week (1) or not (0)	Binary
StudyTime>10hours	Study time is more than 10 hours per week (1) or not (0)	Binary
StudyTime2to5hours	Study time is 2 to 5 hours per week (1) or not (0)	Binary
HasSchoolSupport	Has extra educational support from school (1) or not (0)	Binary
HasFamilySupport	Has family educational support (1) or not (0)	Binary
InExtraPaidClasses	Enrolled in extra paid classes (1) or not (0)	Binary
InExtraCurricular	Participates in extra-curricular activities (1) or not (0)	Binary
AttendedNurserySchool	Attended nursery school (1) or not (0)	Binary

WantsHigherEd	Wants to pursue higher education (1) or not (0)	Binary
HasInternetAccess	Has Internet access at home (1) or not (0)	Binary
InRelationship	Is in a romantic relationship (1) or not (0)	Binary
LogFinalGrade	Logarithm of the final grade	Numeric

Boxplot from SAS:

Classification Tree Workflow & Best Classification Tree from All Data Learner





Decision Tree Workflow & Best Decision Tree from All Data Learner