Journal of Educational Analytics (JEDA)
Vol.1, No.1, 2022: 41-60

Exploring Secondary School Performance by Using Machine Learning Algorithms

Felicia[1]*, Ferren[2]
Universitas Pelita Harapan

ABSTRACT: Education is an important factor to achieve a better life and to help the economy. There are lots of levels of education and the education level that we will analyze is secondary education. Secondary education provides lots of benefits, starting from knowledge and skills, training in attitudes, instincts, and ensuring students will get a job after graduating. Not only that, but Portuguese secondary education also guides the development of the students so they will be well prepared for work and real-life situations. The educational level of Portuguese has also improved from last decades because in the past, lots of students failed and this causing failure rates is increasing. The failures are caused by Mathematics which are the core subjects. Because of this, Portuguese schools are still monitoring students who didn't pass yet by using the data. We will analyze using 3 operators (i.e. Generalized Linear Model , Random Forest, Naive Bayes) andfound out that past grades, demographic, and several attributes play a role in education(Cortez & Silva, 2008). We also found that Naive Bayes method has a high accuracy. The goal of these projects is to identify what makes education successful and fail and to aim for any new prediction.

Keyword: Education; secondary education; descriptive; predictive; prescriptive.

Submitted: ***04-05-2022****; Revised:* ***13-05-2022****; Accepted:* ***24-05-2022***

INTRODUCTION

Education is a key factor for achieving long-term economic progress. In the present era of the knowledge economy, the students are the key element for socio-economic growth of any country, so keeping their performance on track is essential. Over the last decade, the level of Portuguese education has increased. However, the statistics keep Portugal at the end of Europe due to high student failure and dropped out rates.

Secondary education level is one of education that every 13 to 16 years old child must enter (OCDE, 2020). Secondary school is the time every child has a proper education, to be trained in their skill, teaching good and right personalities, preparing them for universities, work, or responsibilities in the future. (Joubish et al., 2010) quoted that "secondary education is important as it not only provides skills and knowledge, but it provides training to sharpen instincts, increase values, and implement right attitudes and habits.

The abundance of data at secondary school can be used optimally according to needs and can be processed into useful information so that we can know the relationship between data attributes in which it can be analyzed and is expected to have an output in the form of student performance. To process data can be done by utilizing data mining in solving the problems. Data mining is a method of finding valuable information from several data that is carried out by utilizing other information such as statistics, mathematics, pattern recognition. In data mining, the analysis and interpretation of student academic performance are regarded as suitable analysis, evaluation, and assessment tools(Dey et al., 2017). By that, the alternative from this application is to analyze the raw data and we can get interesting answers to decide later.

The education rating in Portugal in 2005 was not classified as good(OECD, 2006), it can be seen from the data analyzed that only a few students were listed as pass. This could be caused by several factors which is student performance at school (e.g. absence, activities and their grades), meaning that if students got lot of absences number, it will affect their grades because they didn't attend school and also for some students maybe they are attending school activities (e.g. organization, extracurricular) these kind of activities can affect their grades, because they are busying in activities rather than their lesson, which means they can't manage well their timing. Not only by mentioning the negative relation towards their grades, but there are some variables that considered as supporting their grades too (e.g. Mother education (Medu) and Father education (Fedu)) (Gobena, 2018), these two variables can affect the child's value due to the achievements achieved by the parents, if they provide an education and information based on knowledge then the child can apply it well.

The purpose of this journal is to present some predictive results that can be categorized as important for the secondary student in Portugal. The main point is to predict the cause of the increase or decrease in student grades, types of lessons that makes students interested at, and of course in this journal it will also predict the factors that affect student grades. From the results of this

analysis, we will know what the causes are and give some recommendations for the purpose of creating successful students in the future.

Each student of course has a different type of score based on their demographic and how their performance in school. By using these both of attributes we can get the binary pass or fail for each student. The best accuracy of percentage is by using Naive Bayes. Student's results not only got impacted by demographic sides, but their past school grades also affected their result. We hope by using these 3 kinds of algorithms (i.e., Generalized Linear Model, Random Forest, Naive Bayes) The goal of these projects is to identify what makes education successful and fail and to aim for any new prediction.

LITERATURE REVIEW

Education Introduction

Education as stated above is very important. Each student who receives education will have to do an exam and this is called an examination that has been accepted as a legal measurement for student performance. Examination procedures are different in each country, but it will have the same output which is the score. Those who get high scores will indicate that this student studies well and lower scores will indicate that they don't study hard enough or have poor learning.

Secondary School

Secondary school is a school that occurs between elementary school and college. In Portugal, secondary school is known as *escolassecundarias* and is a must for 15 to 18 years old students. Secondary education lasts for 3 years, and it includes grade 10, grade 11, and grade 12. In secondary school, they provide students with training, general, college-preparation courses, and many more.

Secondary Education in Portugal

1. Education System

Secondary education in Portugal is a must and lasts for 3 years. Children in this phase who are fifteen years old must require a graduation certificate from basic education. Before they enter a secondary class, they must decide whether to choose a general class or vocational training. General classes include science and technology, social, social economics, and language literature. Meanwhile, for vocational training more focusing on administrative, electronic, multimedia, marketing, and sports.

2. Assessment and Grading

Students in secondary level are being qualified by internal examiners throughout their performance and some exams. At the end year of each academic, external exams are being set. Secondary level is being checked with a scheme of 0 - 20, students must get 10 for passing their exams, and they will continue to the next year. For those who did not pass their exams they must

retake again, once they pass it, the students are allowed to continue senior high level.

3. Portugal Education Challenges

It was reported that Portugal has the highest school abandonment rate. Based on the data by(Cheng, 2017), in 2006 the school-leaving rate was 40% and in secondary education, the dropout rate was one-fourth.In 2003, the performance of mathematics and Portuguese language increased but in 2012, the performance was still lower than the average of European countries. The challenge is not only that, but it is also reported that the lack of these core subjects was caused by providing fundamental knowledge. In addition, 62% of adults did not have proper secondary education and only 38% of adults obtained it. This can be categorized as lower than average as the good rate must be 75% (OCDE, 2020).

Secondary Education Impact

Secondary education has lots of impact on student's lives. Starting from impact to individual development to country's economic growth.

1. Individual Development

As a student, we always heard that having education is a gift and we should use it wisely. A person's success and failure is usually determined by their education although we shouldn't have that mindset, but this is a fact. A person who is educated and receives proper education starting from primary, secondary, until tertiary education will have a higher chance to work in a better place, have more opportunities in their lives, better health, and better lives. Therefore, we can see that education plays a big role in a person's individual development.

2. Economic Growth

Believe it or not, if a country has lots of educated people, it will have a significant impact on a country's economic growth. There are empirical results that show educated workers have a positive impact on economic growth and there is an increase from 1% to 1.56% in output and workers' productivity. Therefore, a well-educated individual who completes their secondary education with critical thinking and high skill will be very highly demanded as they can produce greater value for the country's economic growth, increase the efficiency, and promote advanced technology.

Data Mining (DM)

Data mining is a technique used to find anomalies, patterns, and correlation within large or small data sets to make a prediction. It is a difficult process as it contains searching, extracting, and analyzing different types of text, figures, etc. The term of data mining was by misnomer author that the goal is to extract text and pattern. There is also a difference between data mining and data analysis which is from the difficulty, models which are data analysis tend to use mathematical models while data mining using hypotheses.

Machine Learning Algorithm

Machine Learning Algorithm is an AI system or method used to predict and conduct tasks by data given. Machine learning is very important for those who want to learn and find out the underlying patterns inside data. Machine learning usually has 2 processes which are classification and regression. The most common machine learning algorithms are linear regression, logistic regression, SVM, Naive Bayes, Decision Tree, and Random Forest.

Generalized Linear Model (GLM)

Generalized Linear Model (GLM) is one of the RapidMiner's operators that used to operate the data where the distribution of the response variable is a distribution that belongs to the exponential groups. For example, of the exponential group distribution are Poisson distribution and the Binominal. This model describes the structure of the predictor variables, while the linking function specifically describes the relationship between the regression model and the expected value of the response variable.

Random Forest

Random Forest is a RapidMiner operator that is used for classification and regression, and it generates a random forest model. Random Forest ensembles a certain number of random trees and specified by the number of trees parameters. These trees are created to provide nodes for splitting rules. For example, the rule for classification is to separate value to different classes while regression rules are to separate value to reduce error made by estimation. The input for Random Forest is a data set and the output is model example set, and weights. Random Forest also contains lots of parameters but usually used is number_of_trees, number_of _features, and the type_of_trees(Louppe, 2014).

Naive Bayes

Naive Bayes is a RapidMiner operator that is used in classification models. Naive Bayes is usually used in Prescriptive Analytic as it can build a good model even if the data set is small, simple to use, and recommended. The input of Naive Bayes is a data set and the output is model and example set. The parameter for naive Bayes is only 1 which is laplace_correction.

DATA AND METHODOLOGY

DATA

From several journal, researchers found out that, the factor such as past grades, absences, mother's and father's education, age, and sex has the effect to secondary school students performance but the major factor are all the past grades of the students(Mine et al., 2001). Past grades such as G1, G2, and G3 will measure students' performance, if it is low then possibly this student performance is not good and vice versa.

As we are exploring secondary school performance in Portugal for Mathematical subjects, we will use the data which is retrieved from UCI open resource dataset. There are 395 examples, 0 special attributes (no missing

value), and 33 variables. All their characteristics of the data will be explained in Table 1.

Table 1. Secondary School Performance Data

No	Attribute	Description	Role	Data Type
1	School	Student's school	Regular Attribute	Binary (GP or MS)
2	Sex	Student's gender	Regular Attribute	Binary (F or M)
3	Age	Student's age	Regular Attribute	Numeric (15 to 22)
4	Address	Student's home address	Regular Attribute	Binary (U or R)
5	Famsize	Student's family size	Regular Attribute	Binary (≤ 3 or > 3)
6	Pstatus	Student's parent cohabitation status	Regular Attribute	Binary (T or A)
7	Medu	Student's mother education	Regular Attribute	Numeric (0 to 4)
8	Fedu	Student's father education	Regular Attribute	Numeric (0 to 4)
9	Mjob	Student's mother job	Regular Attribute	Nominal
10	Fjob	Student's father job	Regular Attribute	Nominal
11	Reason	Student's reason to choose this school	Regular Attribute	Nominal (home, reputation, course, other)
12	Guardian	Student's guardian	Regular Attribute	Nominal (father, mother, other)
13	Traveltime	Student's home to school range time	Regular Attribute	Numeric (1– <15 min, 2– 15 to 30 min, 3– 30 min to 1 hour or 4 – > 1 hour)
14	Studytime	Student's weekly study time	Regular Attribute	Numeric (1–< 2 hours, 2–2

				to 5 hours, 3–5 to 10 hours or 4– >10 hours)
15	Failures	Student's number of past class failures	Regular Attribute	Numeric (n if 1 ≤ n < 3, else 4)
16	Schoolsup	Student's extra educational support	Regular Attribute	Binary (yes or no)
17	Famsup	Student's educational support from family	Regular Attribute	Binary (yes or no)
18	Paid	Student's extra paid classes within the course subject	Regular Attribute	Binary (yes or no)
19	Activities	Student's extracurricular activities	Regular Attribute	Binary (yes or no)
20	Nursery	Attend on nursery school	Regular Attribute	Binary (yes or no)
21	Higher	Student's willingness to take higher education	Regular Attribute	Binary (yes or no)
22	Internet	Student's internet access at home	Regular Attribute	Binary (yes or no)
23	Romantic	Student's romantic relationship	Regular Attribute	Binary (yes or no)
24	Famrel	Student's family relationship quality	Regular Attribute	Numeric (1,very bad to 5, excellent)
25	Freetime	Student's free time after school	Regular Attribute	Numeric (1, very low to 5,very high)
26	Goout	Student's hanging out with friends	Regular Attribute	Numeric (1, very low to 5,very high)
27	Dalc	Student's workday alcohol consumption	Regular Attribute	Numeric (1, very low to 5,very high)
28	Walc	Student's weekend alcohol consumption	Regular Attribute	Numeric (1, very low to 5,very high)
29	Health	Student's health status	Regular Attribute	Numeric (1, very bad to 5,very good)

30	Absences	Student's number of absence	Regular Attribute	Numeric (0 to 93)
31	G1	First period grade	Regular Attribute	Numeric (0 to 20)
32	G2	Second period grade	Regular Attribute	Numeric (0 to 20)
33	G3	Third period grade	Label Attribute	Numeric (0 to 20)

Source: UCI Machine Learning Repository

1. Categorical Data (Qualitative Data) and Numerical Data (Quantitative Data)

Categorical data or qualitative data is data that is stored in groups or categories or usually aid with names or labels. Although categorical data is qualitative, we can still calculate in numerical value, but the result won't have any value or no meaning as we can't calculate from it. From Table 1, we take out 16 variables that are Qualitative Data.

Table 2. Qualitative Data Summary

Row No.	School	Sex	Address	Famsize	Pstatus	Mjob	Fjob	Reason	Guardian	Famsup	Paid	Activities	Nursery	Higher	Internet	Romantic
1	GP	F	U	GT3	A	at_home	teacher	course	mother	no	no	no	yes	yes	no	no
2	GP	F	U	GT3	T	at_home	other	course	father	yes	no	no	no	yes	yes	no
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
394	MS	M	R	LE3	T	services	other	course	mother	no	no	no	no	yes	yes	no
395	MS	M	U	LE3	T	other	at_home	course	father	no	no	no	yes	yes	yes	no

2. Quantitative data is data that can be used for measurement and very important for statistical analysis. From Table 1, we take out 17 variables that are Quantitative Data.

Table 3. Quantitative Data Summary

Row No.	Age	Medu	Fedu	Traveltime	Studytime	Failures	Schoolup	Famrel	Freetime	Goout	Dalc	Walc	Health	Absence	G1	G2	G3
1	18	4	4	2	2	0	yes	4	3	4	1	1	3	6	5	6	6
2	17	1	1	1	2	0	no	5	3	3	1	1	3	4	5	5	6
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
394	18	3	2	3	1	0	no	3	2	1	3	4	5	0	11	12	10
395	19	1	1	1	1	0	no	4	4	3	3	3	5	5	8	9	9

This data has 395 examples with 0 special attributes,16 regular attributes, and 1 label attribute which is G3. We found out that there is no missing value in this data set, and we directly proceed to the next step. Next step is to set G3 which is term 3 grades as a dependent attribute to examine a student's performance in a Mathematical subject. Student's grades are like a roller coaster, sometimes it goes up, sometimes it goes down. Because of that, teachers must be more extra cautious to check their grades. The process of checking is not hard, but teachers also can slip out and this causes the prediction to be changed. Therefore, a machine learning algorithm such as RapidMiner is very useful to make prediction, prescription, and description analytic of a data.

3. Data Sample

From the table above, we can see lots of factors that can affect a student's performance in a Mathematics subject. The data inside the table has been collected starting from 2005 until 2006 from two public schools in Portugal. In Portugal, the scoring scale is using points which are 0 the lowest and 20 the highest. The result of their past score and data discussion will be discussed in Chapter 4.

METHODOLOGY

Analytic Step

Analyzing steps that we do are searching data set from UCI website, making sure the data is good and usable, transferring the data to RapidMiner to be processed, making descriptive analytic by looking from the data set and using RapidMiner, making predictive analytic using RapidMiner by using GLM and Random Forestt, and making prescriptive analytic using RapidMiner by using Naive Bayes.

Descriptive Analytics

Descriptive analytic, is a simple information gathering and the result is also simple. Descriptive analytics is the most common analytics we can find and see. In business analytics, this analytics is to get the general information from the data set and usually doesn't need any model from RapidMiner. From the student-mat data set, there are 33 variables needed to be analyzed and there are 8 label attributes to be analyzed.

Descriptive statistics usually show us means, median, minimum, maximum, and standard deviation. It also helps us to show the differences between two or more samples, compare samples, and detect any sample that may influence a researcher's decision and conclusion (Thompson, 2009).

In addition, here are several formulas that helped to complete the descriptive analytics, which is by using:

- Mean

Mean is a method that generally uses it in statistics. This method is usually used to find a student's average results in a class. Mean can be calculated by adding up all the total data and then dividing by the amount of data.

$\bar{x} = \sum x / n$

$\sum x$ = the sum of the data

n = number of data

- Median

To get the median result, data must be arranged from smallest to largest value. Also depends on the data, if it is grouping data, we must use this formula:

$Me = Q2 = Tb + [1/2n - fk / fi] p$

Tb = lower edge of the median class
n = sum of all frequencies
fk = the number of frequencies before the median class
fi = median class frequency
p = length of class interval

- Mode

In statistics, the mode is the value that occurs the most. The mode is often calculated with the mean and median. To find mode, is just seeing the data which has the greatest frequency. But in grouping data we can use this formula:

Mo = L + (d1 / (d1 + d2)) i
Mo : the mode result
L : lower edge of mode class
d1 : the frequency of the mode class minus the frequency of the previous class
d2 : the frequency of the mode class minus the frequency of the class after it
i : number of class

Analysis using GLM

GLM is a regression model development for response variables that are not normally distributed. There are three components in this GLM, the first one, random components including the Y response variable from distribution exponential group. Second, systematic component and lastly, link function that relates the random component to the systematic component(Sastri & Setiadi, 2018)

Table 4. GLM's Formula

Distribution	Connector	Connector Function	Inverse Function	Variety
Normal	Identity	Identity	Identity	1
Binomial	Logit	ln[(μ/1 - μ)]	e^n(1+e^n)	μ(1 - μ) / n
Poisson	Log	Ln μ	e^n	μ
Gamma	Inverse	1/μ	1/n	μ^2

One example of the application of GLM is logistic regression which has the assumption that: binomial distribution. In logistic regression, the response variable is a binary variable or dichotomy, which means that it has two values, that is in success / failure. The response variable is Y = 1 if successful and Y = 0 for others. The logistic general form regression probability model is formulated as follows:

$$\pi j = \frac{\exp(\beta 0 + \beta 1X1 + \ldots + \beta kXkj)}{1 + \exp(\beta 0 + \beta 1X1 + \ldots + \beta kXkj)}$$

Figure 1. GLM Forms

This formula the probability Y = 1 at the j-th observation, X1 and Xk is the explanatory variable, β0 is intercept, and βk is the regression coefficient for the corresponding explanatory variable. The πj function is a non-linear function that is difficult to interpret. To interpret the operator easier is to transform the form using the logit function. The result of the transformation is as follows:

$$\ln\left(\frac{\pi j}{1-\pi j}\right) = (\beta 0 + \beta 1 X 1 j + \ldots + \beta p X p j)$$

Figure 2. GLM Form

Analysis using Random Forest

Random Forest is another popular and common machine learning algorithm that is usually used to predict data sets. Random forest is also called as random decision forest and it is used for classification and regression tasks. Important note is whether random forest and random decision forest sound the same, but their output is different. Random forest output is usually the majority result of decision trees.

Steps in Random Forest is quite many, it also needs trees. First step is of course to collect our data set, preprocessing, training, testing, and we are done with the step.

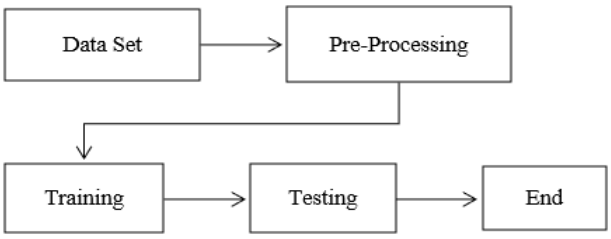


Figure 3. Random Forest Step

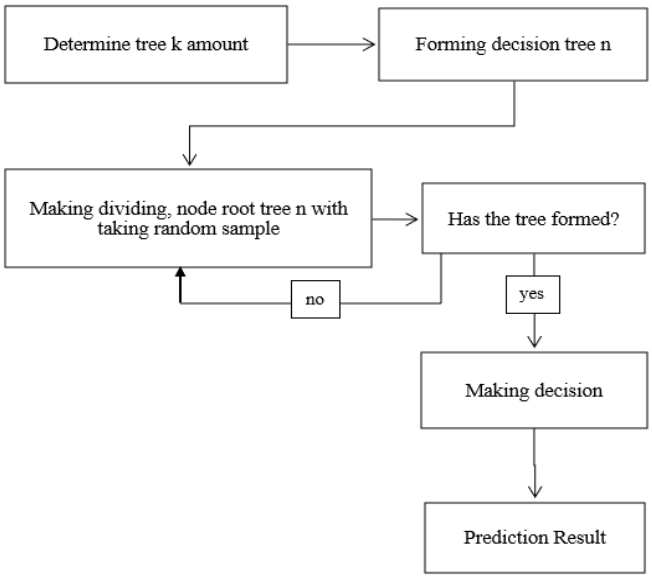


Figure 4. Random Forest Training Step

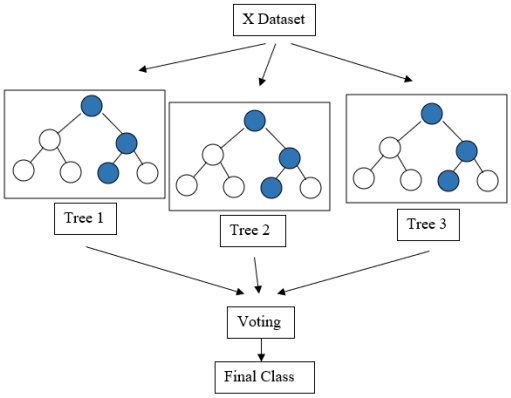


Figure 5. Random Forest

Usually, the result of Random Forest will be Confusion Matrix, which looks like this:

Table 5. Confusion Matrix

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

This matrix shows us that random forest algorithms do classification rightly by seeing the accuracy level and positive level.

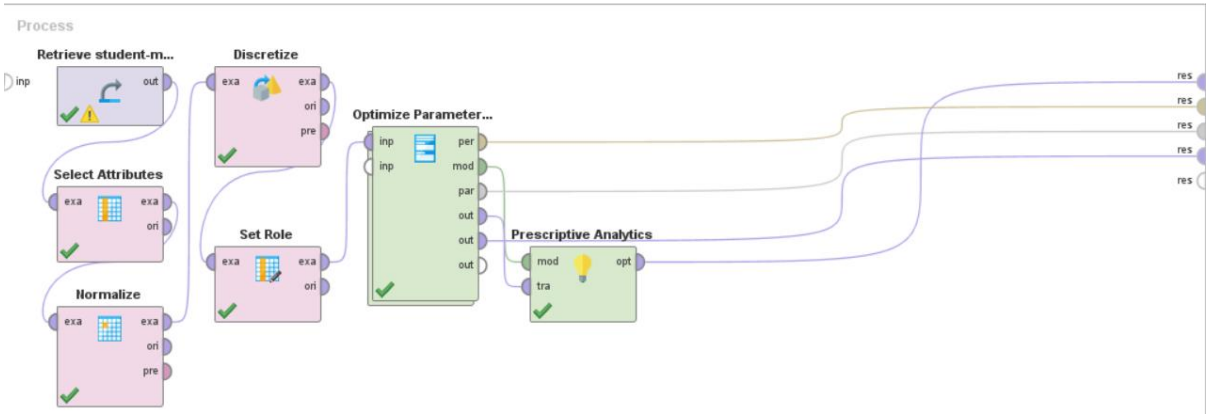
Analysis using Naive Bayes

Naive Bayes is one of the most popular algorithms used as the result is always good and works much better. Naive Bayes is founded by Thomas Bayes. The advantages of this algorithm are a lot, such as training small data and handling big data, easy, simple, time efficiency, and many more. This is the basic Bayes formula:

$$P(Q|x) = \frac{P(X|Q)P(Q)}{p(x)}$$

Figure 6. Naive Bayes Formula

Models



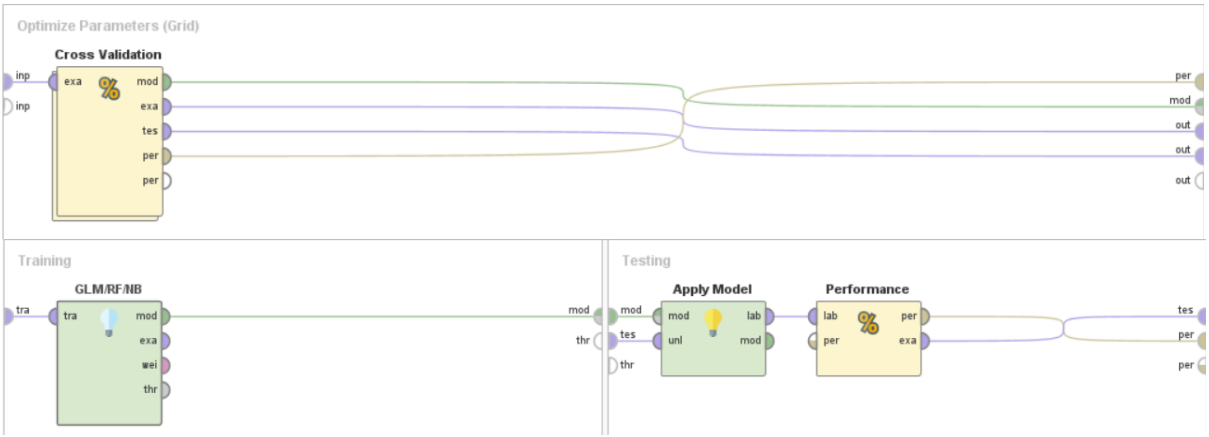


Figure 7. Data Mining Models

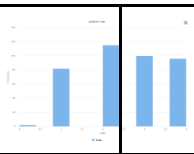
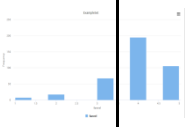
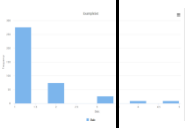
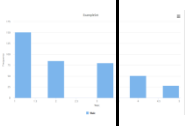
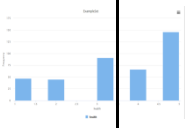
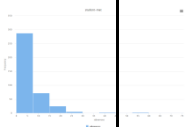
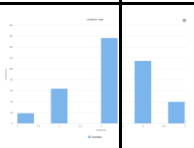
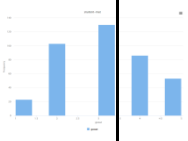
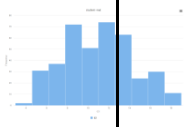
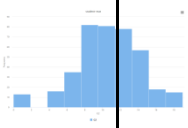
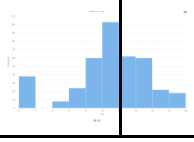
RESULTS AND DISCUSSIONS

Descriptive Analytic Result

As stated above, descriptive analytic represents means, median, minimum, maximum, and standard deviation. These are some data that we took out from 33 variables that we think are important and the factors that can influence a student's performance.

Table 6. Descriptive Analytic Result

Attribute	Data Type	Missing Value	Visualization	in	ax	Average	Deviation
Age	Integer	0		5	2	16.696	1.276
Traveltime	Integer	0				1.448	0.698
Studytime	Integer	0				2.035	0.839
Failure	Integer	0				0.334	0.744
Medu	Integer	0				2.749	1.095

Fedu	Integer	0				2.522	1.088
Famrel	Integer	0				3.944	0.897
Dalc	Integer	0				1.481	0.891
Walc	Integer	0				2.291	1.288
Health	Integer	0				3.554	1.390
Absences	Integer	0			5	5.709	8.003
Freetime	Integer	0				3.235	0.999
Goout	Integer	0				3.109	1.113
G1	Integer	0			9	10.909	3.319
G2	Integer	0			9	10.714	3.762
G3	Integer	0			0	10.415	4.581

By seeing those results numbers of the average can be considered big as it is close to max data. Average can affect standard deviation; a big average can indicate a small standard deviation. Meaning that a low standard deviation is the data can be considered as a good data because not much outlier inside the data and vice versa also each variables have different standard deviation because it can be affected by the outliers.

Starting from the age of each student, itcontain 15 years old until 22 years old. Student's age is not important as age can't become the benchmark of good performance. All students have their own uniqueness despite their age, so we can consider that this attribute is good, have low standard deviation and large average that close to the max. Travel time and study time surprisingly can effect ones performance. As a student who goes to school by transportation such as car or bus, they can study inside it, therefore if the travel time is high, probably this student performance will high as they have more time to revise. Same goes with study time, higher the study time, their performance of course will be good. Both attributes have good data. As for failure, the number of students who fail is only few.

A child inherited their intelligence is from their parents especially their mother. Therefore, mother and father education are important. Parents who have high education will have intelligence child. Seeing from the table, we can conclude that parents who have high education (4), their child's performance is also high. Family relationship is often ignored while relationship with parents can also affect student's performance. Student who has good relationship with their parents tend to have healthy body and stable mental health. One's mental health is important, if students too tired and depressed, their performance will likely drop. From our analysis, we conclude students averagely have a good relationship with their parents. Dalc and Walc means their alcohol consumption during workday and weekend. It might be weird for student to drink, but the legal age to drink is 18 and law make an exception for minors if they consume it in a restaurant. From the analysis, the alcohol consumed during workday is less than during weekend. This can also affect student performance as during school time, they probably stop drinking to have their brain work perfectly without any distraction from alcohol side effect.

Another attributes we analysis is health. A healthy body will produce healthy mind and we can do everything. Based on the analysis, we discovered that there are a lot of healthy students. It is also proven that healthy students have high performance. Next is about student absences in schools is considered an important problem in the management of students at school because this is very closed related to student performance and by seeing the average and standard deviation value is quite high. Going out and free time can be considered as factor to effect performance, but by seeing the result, the average and standard deviation is not high. Finally, past grades such as G1, G2, and G3, we found that their performance is not bad.

Predictive Result

1. Generalized Linear Model (GLM)

Table 7. GLM Confusion Matrix

	True Fail	True Sufficient	True Good	True Excellent	True Satisfactory	True Very Good	class precision
Pred. Fail	125	55	0	0	1	0	69.06%
Pred. Sufficient	5	40	54	15	55	20	21.05%
Pred. Good	0	6	5	2	4	2	30.00%
Pred. Excellent	0	0	0	0	0	0	0.00%
Pred. Satisfactory	0	2	0	0	2	0	50.00%
Pred. Very good	0	0	0	0	0	0	0.00%
class recall	95.15%	38.83%	10.00%	0.00%	3.23%	0.00%	

This is the result of predictive result using GLM. 395 dataset, 32 regular attributes, and 8 special attributes were analyzed in RapidMiner and provides us with 43.80% of accuracy. It also shows us that no students passed the test, the number of students who fail is larger than those who pass.

2. Random Forest

Table 8. Random Forest Confusion Matrix

	True Fail	True Sufficient	True Good	True Excellent	True Satisfactory	True Very Good	class precision
Pred. Fail	133	48	0	0	1	0	69.75%
Pred. Sufficient	17	23	12	5	17	1	30.67%
Pred. Good	0	17	30	9	22	15	32.26%
Pred. Excellent	0	1	1	1	0	0	33.33%
Pred. Satisfactory	0	13	15	1	21	6	37.50%
Pred. Very Good	0	1	2	2	1	0	0.00%
class recall	86.92%	22.33%	50.00%	5.56%	33.87%	0.00%	

This is the result of predictive result using Random Forest. 395 dataset, 32 regular attributes, and 8 special attributes were analyzed in RapidMiner and provides us with 47.61% of accuracy. It also shows us that no students passed the test, the number of students who fail is larger than those who pass.

3. Naive Bayes

Table 9. Naive Bayes Confusion Matrix

	True Fail	True Sufficient	True Good	True Excellent	True Satisfactory	True Very Good	class precision
Pred. Fail	110	42	0	0	1	0	71.90%
Pred. Sufficient	20	25	4	1	13	0	39.68%
Pred. Good	0	9	29	6	18	11	39.73%
Pred. Excellent	0	2	8	3	0	5	16.67%
Pred. Satisfactory	0	22	13	4	25	6	35.71%
Pred. Very Good	0	3	6	4	5	0	0.00%
class recall	84.62%	24.27%	48.33%	16.67%	40.32%	0.00%	

This is the result of predictive result using Naive Bayes. 395 dataset, 32 regular attributes, and 8 special attributes were analyzed in RapidMiner and provides us with 48.59% of accuracy. As stated above, Naive Bayes produce a very good accuracy so we can trust the accuracy level. It also shows us that no students passed the test, the number of students who fail is larger than those who pass.

Prescriptive Result

1. Generalized Linear Model (GLM)

Table 10. GLM Prescriptive Analytic

	row no	prediction (G3)	confidence (fail)	confidence (sufficient)	confidence (good)	confidence (excellent)	confidence (satisfactory)	confidence (very good)
	1	sufficient	0.286	0.344	0.132	0.06	0.104	0.073

reason	pstatus	walc	schoolup	G1	G2	fedu	traveltime	studytime	school
home	A	fail	yes	excellent	fail	fail	fail	excellent	GP

romantic	fjob	guardian	famrel	absences	failures	address	freetime	famsize	famsup	sex
no	other	mother	fail	excellent	fail	R	excellent	LE3	no	M

health	medu	dalc	mjob	nursery	activities	paid	goout	age	internet	higher
fail	excellent	fail	services	no	yes	yes	fail	fail	yes	yes

Generalized Linear Model (GLM) provides us with this result. Predicted score for G3 is sufficient, and if seeing the table, it is true because confidence level of sufficient is 34.4%. This student almost fail, by looking from their past performance, they only get one excellent score. If we see from the table, student who have 0 travel time and have excellent study time also can fail in the examination. Somehow, this student parents especially father must look up their child performance and adding more educational support to their child. Although this student does not have any failures, but still best for they to lessen their activities and spend more time to study.

2. Random Forest

Table 11. Random Forest Prescriptive Analytic

row no	prediction (G3)	confidence (fail)	confidence (sufficient)	confidence (good)	confidence (excellent)	confidence (satisfactory)	confidence (very good)
1	excellent	0.021	0.04	0.133	0.602	0.045	0.159

reason	pstatus	walc	schoolup	G1	G2	fedu	traveltime	studytime	school
course	A	fail	no	excellent	excellent	excellent	fail	fail	GP

romantic	fjob	guardian	famrel	absences	failures	address	freetime	famsize	famsup	sex
no	teacher	father	excellent	fail	fail	U	fail	GT3	yes	M

health	medu	dalc	mjob	nursery	activities	paid	goout	age	internet	higher
excellent	excellent	fail	services	yes	yes	no	fail	fail	yes	yes

Random Forest provides us with this result. Predicted score for G3 is excellent and if seeing the table, it is true because confidence level of excellent is 60.2%. To achieve this score, we can see from the table what are needed. First is of course to keep maintain G1 and G2 score, be healthy, no absence, have good relationship with family, and getting educational support from parents. These students also not have much study time and travel time, but he can maintain his score very well. We believe other attributes also effect their score.

3. Naive Bayes

Table 12. Naive Bayes Prescriptive Analytic

row no	prediction (G3)	confidence (fail)	confidence (sufficient)	confidence (good)	confidence (excellent)	confidence (satisfactory)	confidence (very good)
1	excellent	0	0.002	0.003	0.987	0.007	0

reason	pstatus	walc	schoolup	G1	G2	fedu	traveltime	studytime	school
home	A	fail	no	excellent	excellent	excellent	fail	excellent	GP

romantic	fjob	guardian	famrel	absences	failures	address	freetime	famsize	famsup	sex
no	teacher	other	excellent	excellent	excellent	U	excellent	LE3	no	M

health	medu	dalc	mjob	nursery	activities	paid	goout	age	internet	higher
fail	excellent	fail	services	yes	yes	no	excellent	fail	yes	yes

Naive Bayes provides us with this result. This student has predicted score for G3 is excellent, and if seeing the table, it is true because confidence level of excellent is 98.7%. Another founding for this student to achieve excellent score is they must maintain their score from G1 to G3. Although some attributes are 0 such as health, other attributes such as mother and father education, internet, study time, etc. can effect this student score to be excellent. We believe this student study a lot from their extracellular activities and during free time as their free time is very high.

Machine Learning Algorithm Comparison

Table 13. Algorithm Comparison

Operators	Accuracy
Generalized Linear Model	43.80%
Random Forest	47.61%
Naïve Bayes	48.59%

The table above, show us that each machine learning result will have different accuracy level. The highest accuracy level is obtained by using Naive Bayes and lowest is GLM. Low accuracy doesn't mean the result is wrong, but it just not precise.

CONCLUSION AND RECOMMENDATION

The study to analyze Portuguese's students' performance by using data mining and to predict their performance. This dataset has 33 variables and 1 dependent or label variable involved. The algorithm that used in this study is Generalized Linear Model (GLM), Random Forest, and Naive Bayes and algorithm that have the highest accuracy is Naive Bayes which is 48.59%.

Education is important, knowing our future generation performance in Mathematics subject is low will of course make them not confident. As a core subject, Mathematics is very useful and knowing that Portuguese student performance is not really good, they might want to change the way of the study, teaching, or the source. Lots of factor starting from internal and external can affect their performance, therefore this study is made to see what their exact problems is.

Analyzing students' performance is very important as it can tell us how students' performance is. We are hoping that this topic will keep used to help student study and to give them advice on how to get higher grades. Gathering more data from school may help to achieve more complete dataset too so analyzing will be easier to do.

REFERENCES

Cheng, L. (2017). Exploring the Factors that Affect Secondary school's Mathematical and Portuguese Performance in Portugal. *Masters Dissertation Technological University Dublin*. https://doi.org/10.21427/D7P33K

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *15th European Concurrent Engineering Conference 2008, ECEC 2008 - 5th Future Business Technology Conference, FUBUTEC 2008*, *2003*(2000), 5–12.

Dey, N., Ashour, A. S., & Nguyen, G. N. (2017). Deep learning for multimedia content analysis. *Mining Multimedia Documents*, *1*(4), 193–203. https://doi.org/10.1201/b21638

Farooq Joubish, M., Memon, G., & Ashraf Khurram, M. (2010). Impact of Parental Socio-Economic Status on Students' Educational Achievements at Secondary Schools of District Malir, Karachi. *Middle-East Journal of Scientific Research*, *6*(6), 678–687.

Gobena, G. A. (2018). Family Socio-economic Status Effect on Students' Academic Achievement at College of Education and Behavioral Sciences, Haramaya University, Eastern Ethiopia. *Journal of Teacher Education and Educators*, *7*(3), 207–222.

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. *July*. http://arxiv.org/abs/1407.7502

Mine, Y., Hiraishi, H., & Mizoguchi, F. (2001). Collaboration of networked home electronics using multi-agent technology. *Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS*, *5*(9), 2648–2652. https://doi.org/10.1109/nafips.2001.943641

OCDE. (2020). Education Policy Outlook: Portugal. *OECD Journals*, 31.

https://www.oecd.org/education/policy-outlook/country-profile-France-2020.pdf

OECD. (2006). Tertiary Education in Portugal Background Report. In *Higher Education*. http://dx.doi.org/10.1787/104853273381 and

Sastri, R., & Setiadi, Y. (2018). *Laporan Penelitian Dosen Stis Generalized Linear Mixed Model Untuk Data Kematian Bayi Di Indonesia Generalized Linear Mixed Model*. 1–17.

Thompson, C. B. (2009). Descriptive Data Analysis. *Air Medical Journal*, *28*(2), 56–59. https://doi.org/10.1016/j.amj.2008.12.001