Exploring Student Academic Performance in Portugal using Machine Learning

Kenechukwu Otito Ajufo
School of Computing
National College of Ireland
Dublin, Ireland
x19190174

I. Introduction

Education is crucial and often thought to be the most important aspect of every human life. It can be described as a means or a series of activities through which information, skills, values, beliefs, and behaviors are acquired. It is often iterative and involves individuals (teachers) teaching other individuals (students) and vice versa.

There are three forms of education: Formal, Informal, and Non-Formal. Education can be characterized as Formal when the individual, that is, the student receives teaching within the four walls of a structured building like a classroom in a school or university, or through virtual means as has become universal in these days of COVID-19 restrictions.

Education is Informal when it occurs at home or by experience, for example, parents teaching their children or wards how to cook.

Finally, it can also be Non-Formal. An instance can be apprenticeship, technical education or learning off the streets.

Education should be fundamental for every person as it gives us the opportunity to understand the environment – the things that surround us, our body and mind, etc. Advancement in education can potentially aid humans in obtaining a better quality of life, enhancing economies, and overall, making the society safer.

The focus of this research is on an aspect of Formal Education. In Formal Education, after lessons are taught, assessments are made to measure the extent to which the students have learned. There are various factors which can influence the outcome of the performance of students in the assessments. These range from age, health status, study time, absence from school and family stability etc.

The application of Data Mining and Machine Learning can help gain insights into how learning in schools occurs and help to form new teaching methods. Educational Data Mining (EDM) is the domain of research that is involved with the processing of data from educational environments using data mining, machine learning and statistics.

Educational Data Mining is a tool for structuring and storing student academic records in a form that is adaptive to evaluation, predicting failure or success, or identifying factors that are responsible for performance using the concept learned from the massive accumulated database.

Paulo Cortez and Alice Silva in [1] utilized Business Intelligence (BI) and Data Mining to gather information to boost the education domain in Portugal. A survey approach was implemented, and the answers collected from the respondents, and their grades made up the data set that was analyzed in the research.

In [2], the authors proposed EDM techniques to predict academic performance at the undergraduate level. The purpose of the study is to arm educators with a guide to using these techniques in predicting and improving student performance.

An approach to predict the examination scores of a student in the United Arab Emirates (UAE) was implemented in [3] while also discussing the origins of EDM. The authors made use of a Linear Regression model to determine the factors which influenced the examination scores the most.

In [4], the authors build a model using open data from edX to predict student performance. They based the student's ability to get the certificate as the metric of the student's performance. They attributed the lack of work done on E-Learning area to the difficulty in extracting data from those websites.

Building a classification model by implementing data mining techniques to predict the possibility of a student being successful in the Licensure Examination for Teachers (LET) in the Philippines is the focal point of the researchers in [5]. They reported that the model built identified student who were likely to fail the LET.

The authors in [6] argued that predicting the performance of any student should not be solely academic excellence, interpersonal skills and behaviours should be considered to get an overview of the real excellence of students. They utilized various classification algorithms to achieve the objectives of their research.

Decision Tree as a Data Mining technique was explored to classify the student's performance in [7], [8]. Both papers also talk about the use of classification algorithms in EDM to predict student performance.

The data set analyzed in this research work was obtained from the University of California Irvine (UCI) Machine Learning Repository. It comprises of 33 attributes and 649 instances. The attributes and instances are answers from respondents to a survey in Portugal and include school related details like student grades and other socioeconomic factors.

II. Research Objectives

The following are the objectives of this research paper are two fold, viz:

- To identify and classify the important factors that determine student performance.
- To predict student performance using machine learning techniques.

The research question is:-

1) What are the factors that are crucial to Students' academic performance in Portugal?

III. Ethical Concerns

Before executing any research, ethical issues should ideally be considered as it can affect the outcome of the research. The ethical concerns considered while conducting research on secondary school children are highlighted below:

A. Explicit Consent

Researchers should make an effort to get explicit consent from the respondents (students) in the survey with no use of coercion or force. Most importantly since the respondents are under the legal age, authorization from their parents or wards is essential because minors cannot give consent.

B. Privacy and Transparency

Researchers should comply with the standards and laws that guide the use of the data gathered ensuring that the respondent's personal information is secure. No information that can potentially lead to the respondent identity should not be released publicly.

C. Norms

Researchers should appropriately recognise and take into consideration norms. They can be cultural or gendered. It is key because norms are sometimes unspoken laws that guide a social group.

D. Data Integrity

While carrying out quantitative or qualitative analysis, the researcher must make sure the sample is an appropriate representation of the population to guarantee data integrity. Not getting an accurate representation of the population can lead to incorrect results, which potentially leads to incorrect action plans being implemented.

Although the authors of the data set do not state their ethical considerations in [1], there is no instance that can reveal any respondent of the survey, hence it is assumed that ethics were considered before the research was conducted.

In this study the data set chosen is in the public domain and it can be safely assumed that all the concerns aforementioned were taken into cognizance before the publication of the data set.

IV. Strategy

The application of Predictive Analytics in the area of education can have an immense impact on the stakeholders involved. The stakeholders in this area are the educational institutions - primary, secondary schools, universities etc., and the students, teachers, guidance counsellors. Some benefits of using Predictive Analytics in the area are:-

A. Institutional Level

At the institutional level the use of predictive analytics can aid educational institutions in making more informed decisions when it comes to finance, risk reduction, implementation of best practices and resource allocation.

B. Educational Administrators

Teachers, Guidance Counsellors can successfully assess the effectiveness of their teaching methods, customize learning styles to fit each individual student, and also identify at risk students to avail them support.

C. Students

It can help understand students' learning styles and behaviours, then harness them to yield improvement in student performance, lessen drop out rates, and increase overall literacy rates in the society.

V. Preliminary Visualizations

This is an initial look at the dataset. Visualizations can capture the relationship between the Final Grade and factors which can influence it. The subsequent figures show the preliminary visualizations done on the dataset.

A. Average Final Grade by Sex

The Pie Chart below shows the average final grade of the students by their sex. The average final grade for females is 9.97 while for the males is 10.91. The Legend shows the Males in dark blue and Females in sky blue.

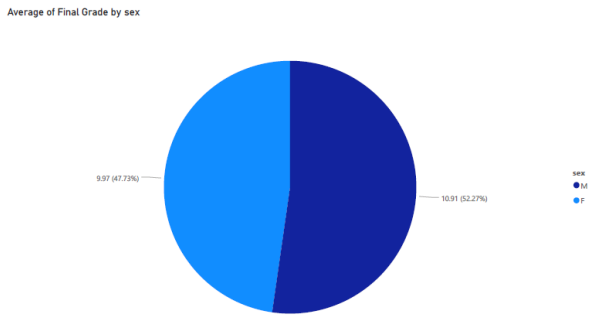

Fig. 1. Average of Final Grade by Sex

B. Average Final Grade by Age and Sex

The Stacked Column Chart shows the average final grade for the students by their sex and grade. The Average Final Grade is on the y axis and the Age is on the x axis. The Legend shows the Males in dark blue and Females in sky blue.

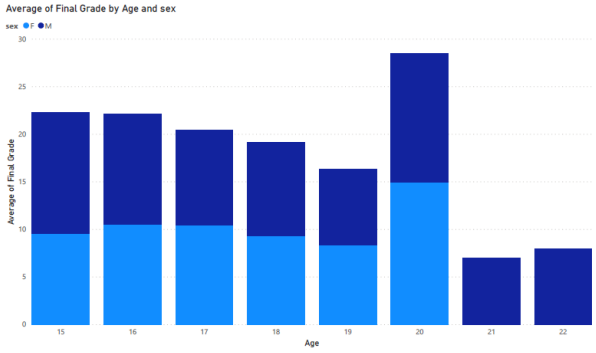

Fig. 2. Average Final Grade by Age and Sex

C. Standard Deviation of the Final Grade by Free time after School

Figure 3 shows an area chart of the standard deviation of the final grades secured by the student, by the amount of free time they have after school. This helps imagine the influence of free time after school and the grades.

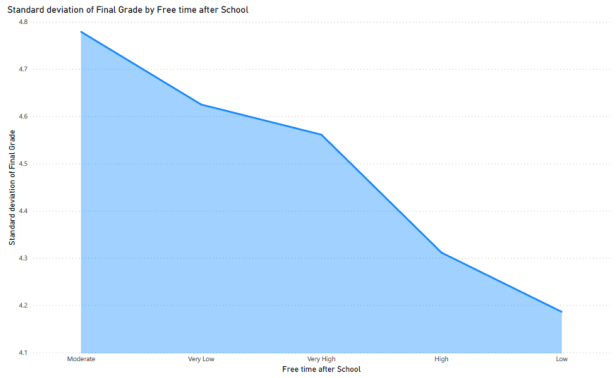

Fig. 3. Standard Deviation of the Final Grade by Free time after School

D. Average Final Grade by Study Time and Extra-curricular Activities

The Clustered Column Chart shows the average final grade of the students by the time the students spent studying and their extra-curricular activities. This can help visualize the effect of studying and extra-curricular activities have on the final grade. The Average Final Grade is on the y axis and the Study Time is on the x axis. The Legend shows the Extra-curricular activities, dark blue for the average final grades for those who had extra-curricular activities and the sky blue for those who did not have.

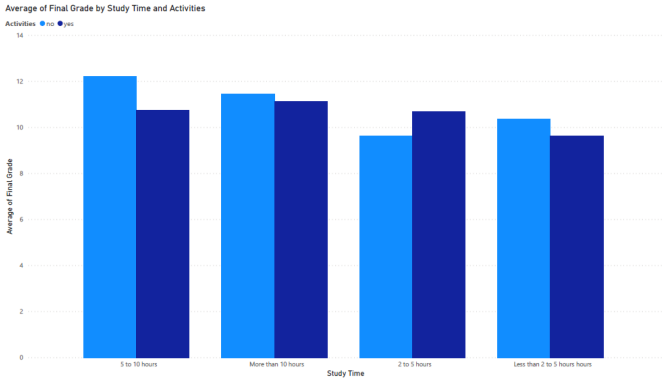

Fig. 4. Average Final Grade by Study Time and Extra-curricular Activities

E. Average Final Grade by Internet Access

The Doughnut chart shows the average final grade of students by the access to the internet. The average final grade of students who had internet access is 10.62, while 9.41 for those who did not have access to the internet. This means on average students who have internet do better.

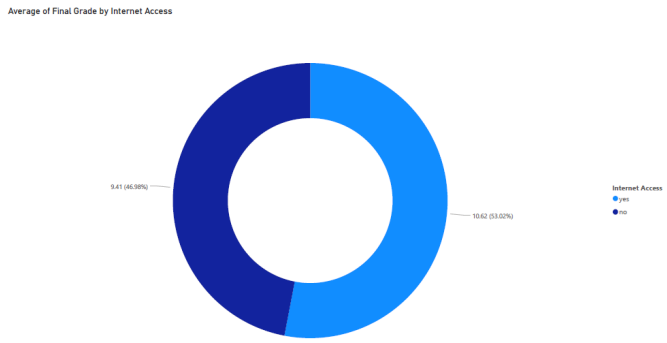

Fig. 5. Average Final Grade by Internet Access

F. Average Final Grade by Study Time and Extra Educational Support

The Clustered Column Chart shows the average final grade by study time and extra educational support from the school. The Study Time is on the y axis and the x axis is the average final grade by each student. This can aid us in understanding whether support from school has an effect on the final grade. The dark blue in the Legend shows the students that got extra support while the sky blue shows the ones that did not.

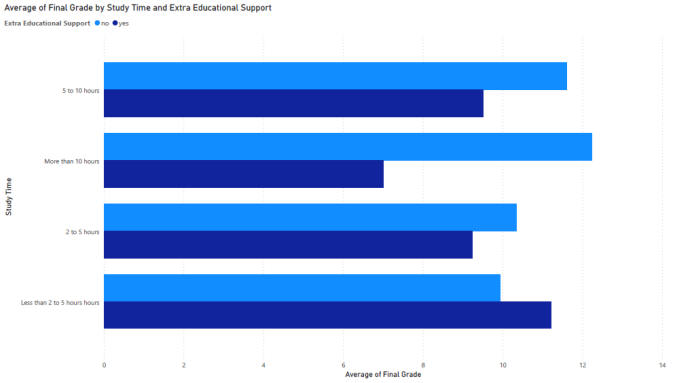

Fig. 6. Average Final Grade by Study Time and Extra Educational Support

G. Average Final Grade by Intention to Get Higher Education

The Pie Chart visualizes the average final grade of the students grouped by their intention to get Higher Education after secondary school. This shows the impact of the intention of get higher education on the final grade. The average grade of the student who intend on getting Higher Education in sky blue is 10.61 while those who do not want in dark blue is 6.80. On Average the student who intend to get Higher Education get better grades.

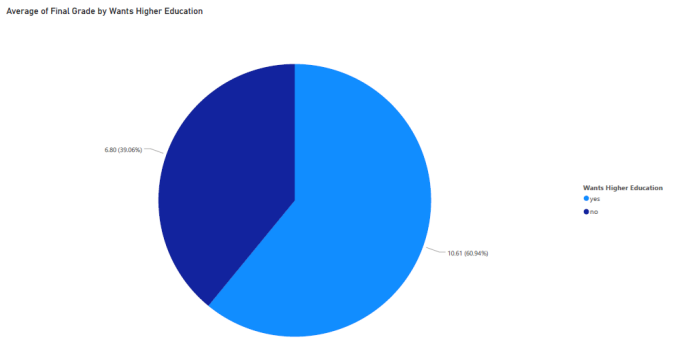

Fig. 7. Average Final Grade by Intention to Get Higher Education

H. Average Final Grade by Quality of Family Relationship and Family Educational Support

The Clustered Column Chart displays the average final grade by the quality of the student's family relationship and the educational support the student's family gives. This reveals the influence family relationship and support on the final grade. The y axis is the average final grade, x axis is the quality of the family support. The legend is the Family Educational Support, dark blue shows the average final grade of students that have family educational support while sky blue shows those who do not.

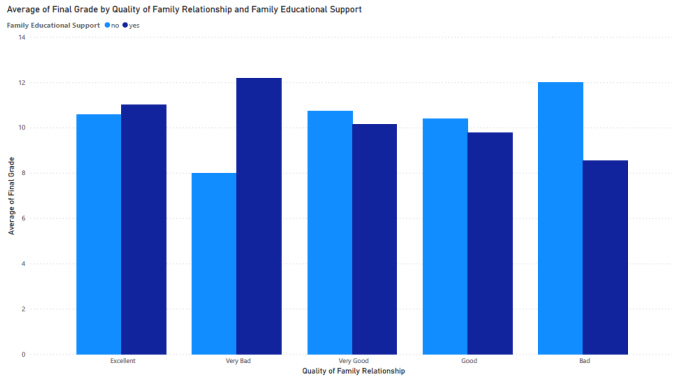

Fig. 8. Average Final Grade by Quality of Family Relationship and Family Educational Support

I. Average Final Grade by Health Status

The pie chart below shows the average final grade of the students grouped by their health status. This gives us a picture of the effect of health on the final grade.

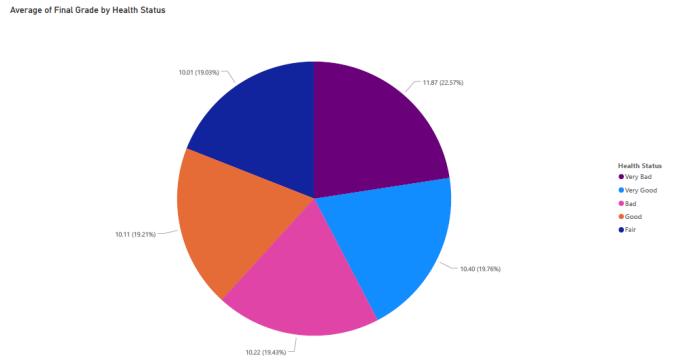

Fig. 9. Average Final Grade by Health Status

J. Average Final Grade by Parent's Cohabitation Status

The doughnut chart displays the average final grade of the students grouped by their parents' cohabitation status. It describes the influence the parents cohabitation status has on the grades of their wards. On average, the students whose parents live together scored 10.32, while the ones who did not scored 11.20.

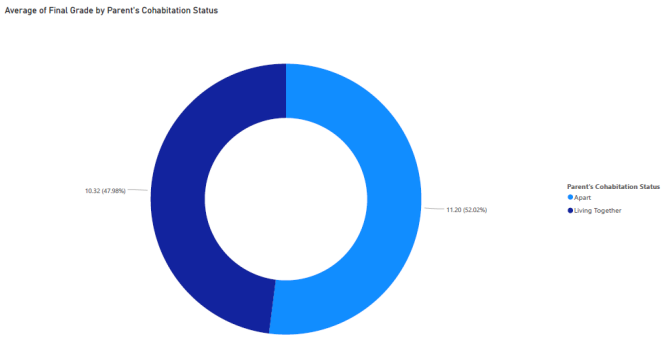

Fig. 10. Average Final Grade by Parent's Cohabitation Status

K. Average Final Grade by Student Residential Address

The pie chart below shows the average final grade by the student residential address to help understand the effect of where the student lives on the final grade. The students who live Rural areas on average scored 9.51, while the students in Urban areas on average scored 10.67

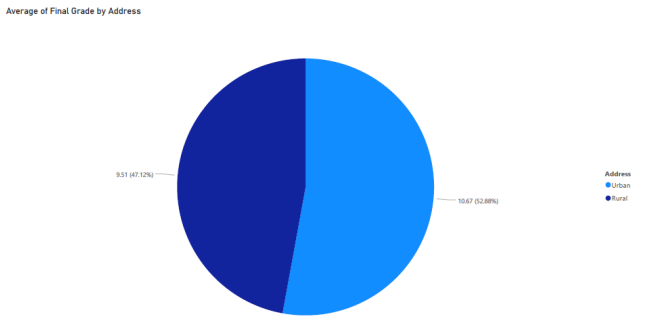

Fig. 11. Average Final Grade by Student Residential Address

L. Average Final Grade by Family Size

The Clustered Column Chart displays the average final grade of student by the size of the family. On average students with family size less than or equal three scored 11, while the students with family sizes greater than 3 scored 10.

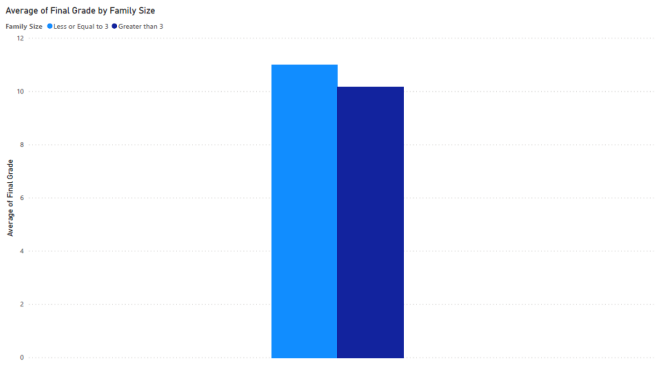

Fig. 12. Average Final Grade by Family Size

M. Average Final Grade by Extra Paid Classes

The doughnut chart shows the average final grade the students earned by the extra paid classes. This gives an idea of the impact of getting extra paid classes have on the final grade. The students that paid for extra classes on average scored 10.92 while the students who did not scored 9.99.

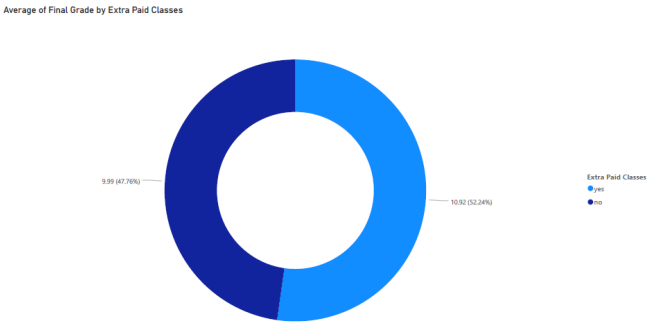

Fig. 13. Average Final Grade by Extra Paid Classes

N. Average Final Grade by Romantic Relationship Status

The pie chart in figure 13 shows the average final grade by their romantic relationship status to identify the influence of the student relationship status have on the final grade. The students who are not in a relationship on average scored 10.84, while the student who had scored 9.58.

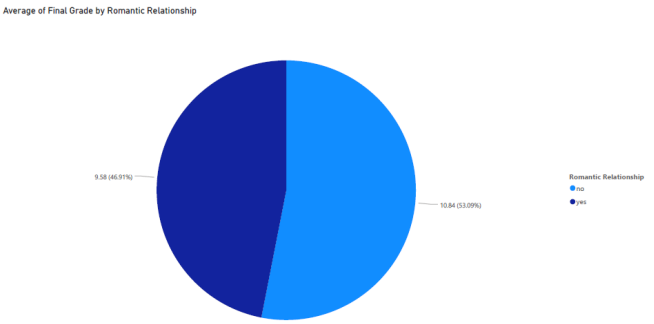

Fig. 14. Average Final Grade by Romantic Relationship Status

O. Average Final Grade by Travel Time to School

The doughnut chart below shows the average final grade of the students by the amount of time spent going to school.

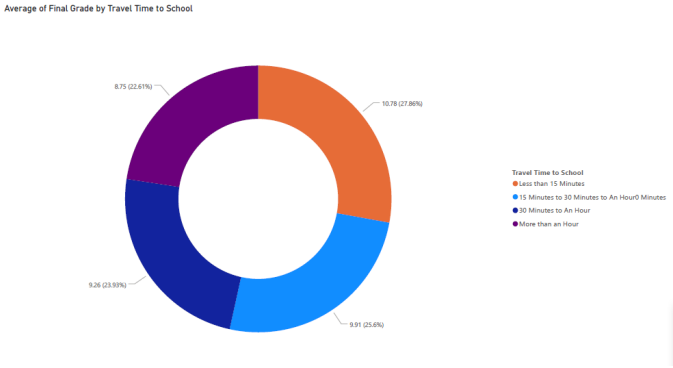

Fig. 15. Average Final Grade by Travel Time to School

VI. Applicable Techniques

The following shows the various Machine Learning techniques in a tabular form which can be utilized to achieve the research objectives.

TABLE I
Applicable Machine Learning Techniques

Classification Algorithms	Regression
Decision Trees	Multiple Linear Regression
Bayesian Networks	Random Forest Regressor
Naive Bayes Classifiers	Support Vector Machines
Random Forest Classifiers	-
Neural Networks	-

References

[1] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[2] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, p. 3, 2020.

[3] A. Anzer, H. A. Tabaza, and J. Ali, "Predicting academic performance of students in uae using data mining techniques," in *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, 2018, pp. 179–183.

[4] C. Ma, B. Yao, F. Ge, Y. Pan, and Y. Guo, "Improving prediction of student performance based on multiple feature selection approaches," in *Proceedings of the 2017 International Conference on E-Education, E-Business and E-Technology*, ser. ICEBT 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 36–41. [Online]. Available: https://doi.org/10.1145/3141151.3141160

[5] R. A. Rustia, M. M. A. Cruz, M. A. P. Burac, and T. D. Palaoag, "Predicting student's board examination performance using classification algorithms," in *Proceedings of the 2018 7th International Conference on Software and Computer Applications*, ser. ICSCA 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 233–237. [Online]. Available: https://doi.org/10.1145/3185089.3185101

[6] E. Chandra and K. Nandhini, "Predicting student performance using classification techniques," in *Proceedings of SPIT-IEEE Colloquium and International Conference, Mumbai, India, p. no83-87*, 2005.

[7] R. Kabra and R. Bichkar, "Performance prediction of engineering students using decision trees," *International Journal of computer applications*, vol. 36, no. 11, pp. 8–12, 2011.

[8] A. B. Raut and M. A. A. Nichat, "Students performance prediction using decision tree," *International Journal of Computational Intelligence Research*, vol. 13, no. 7, pp. 1735–1741, 2017.