Factors affecting student success using EDM: A case study of students in Alentejo region of Portugal

JIAYI, ZHU* YUXIN, MENG**

* Department of Human Development, Teachers Colleage, jz3642@tc.columbia.edu
** Department of Human Development, Teachers Colleage, ym2969@tc.columbia.edu

Abstract
This article uses four educational data mining methods on a dataset consisting of 677 secondary school students from Portugal. We first use liner regression to predict students' Portuguese grades, but found the result is not optimal. We then switch to using random forest classification method to predict the same object and found positive result on the accuracy (86%). To confirm the effectiveness of random forest classification, we applied the same method to predict students' math grades and compared the results of the two models. Specifically, we compared the differences in the main features that contributed to these two models. After discovering the students' academic performance, we investigated other attributes of the students' biographical information using relationship mining (ARM).An interesting result is that students whose family provides educational support, have access to the Internet, and have attended kindergarten are most likely to have intentions to go to college. Finally, we used PCA clustering to identify four clusters among students and found that parents' occupation and guardian status differ among the clusters.

Keywords
EDM, Machine Learning, Classification, Relationship Mining(ARM), Clustering, PCA

1. INTRODUCTION

There has been a lot of research conducted in terms of educational data mining, published in many prestigious journals(Romero & Ventura, 2012) . Even though data mining for educational purposes is still an emerging field, it has received many applications. This research is aiming at using four different popular data mining methods to investigate insightful information from a real-world educational data set. The methods use include linear regression, classification, association rule mining and clustering. We will use regression and classification for prediction; rule mining and clustering for structural discovery.

Firstly, linear regression and classification are used for building a model to predict students' academic performance. There has been many research using one or the mix of these two methods and received different outcomes. For instance (Elbadrawy et al., 2016) used personalized multiregression and matrix factorization approaches to accurately forecast students grades in the future class as well as in-class tests. Sravani and Bala (2020) used linear regression and concluded that students' past academic performances as well as other background features. There has been other reseach focusing on using

classifications. For example, Yang and Li (2018) used Back Propagation Neural Network based on classification to estimate students' performance and evaluate the attributors of it based on students' previous knowledge. As the research went further, there have been some researchers compare multiple classification methods in prediction students academic grades. For example. Dorina Kabakchieva (2012) compared the performance of Neural Network model, Decision Tree model, and the k-NN model on student data and conclude that Neural Network model has the best performance and followed by Decision Tree model.

Another approach that was used to explore the data in this research is relationship mining, also called association rule mining(ARM). Specifically, the Apriori Algorism was used for the present analysis. This algorithm of association rule mining was traditionally used for marketing analysis purpose, for example, supermarkets can use the result of the most frequently purchased pairs to arrange the product distribution in the supermarket (Ünvan, 2020). However, this algorsim have applied to many educational data analysis as well. For example, Yang and Hu (2011) used it to find associations among the courses students often choose and make useful suggestions on course selection arrangement based on the results.

Last but not least, we also use clustering to find certain patterns among students.Clustering is a useful datamining technique for a variety of educational topic. Basically, identifying students' clusters in helpful for giving different education to different students accordingly. Salazar et al. (2005) successfully creat 4 different student clusters for different studying styles. Similar research was done by Wook et al. (2009) to analyze student's annotation habit to cluster students and teach them in the way that aligned to their studying habit using K-means clustering. Another interesting research by Ibrahim, and Rusli (2007) also used K-means clustering to investigate how to teach computer skills to students from urban and rural area.

Therefore, our research questions are as follows:
1. Is it possible to predict student performance in Portuguese (mother tongue) Class? What are the factors that affect student achievement?
2. Do the same analysis of students' math final test performance, does it have the same attribution as Portuguese performance?
3. Can we find any interesting associations between other factors other than student's academic performance?
4. What are the group characteristics of students with high and low grades respectively?

2. RESEARCH METHODOLOGY

2.1. Sample

This data was collected during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal. A total of 677 students' data were collected for the survey, including 395 students taking the Portuguese language course and 649 students taking the math course. Most of students come from the school of Gabriel Pereira (423 students of Portuguese class and 349 students of math class); Fewer students come from the school of Mousinho da Silveira (226 students of Portuguese class and 46

students of math class). There are more female students than male students, 56% in Portuguese class and 52% in math class. The age range is board, which is from 15 to 22, and the majority of them are around 15-18 years old.

2.2. Data Source

The data were obtained in a survey of students' math and Portuguese language courses in secondary school, which is available on Kaggle (Student Alcohol Consumption, n.d.). It contains a lot of interesting biographical attributes of students, including social, gender, family, and study information.

Table 1 shows the dependent variable information of the data and their abbreviations. There are 30 dependent variables, including ordinal variables, categorical variables, and numeric variables. Also, there are 3 independent variables, which are G1, G2 and G3 (first period grade, second period grade and final grade). The distribution of the variables can be seen from Table2 and 3 in the Appendix.

Table 1: Content of variables

Variables	Meanings
school	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	student's sex (binary: 'F' - female or 'M' - male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal: 'mother', 'father' or 'other')
traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)

G3	final grade (numeric: from 0 to 20, output target)

Figure 4*: Correlation map of Portuguese dataset*

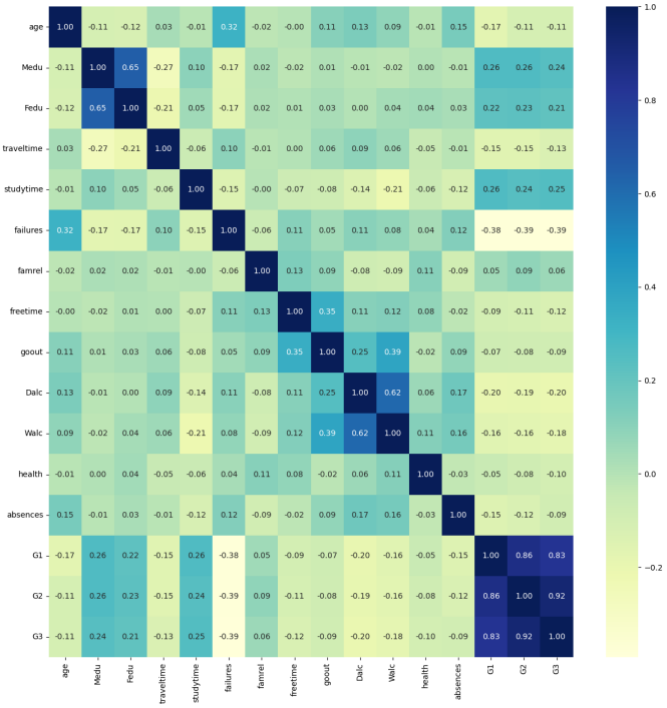
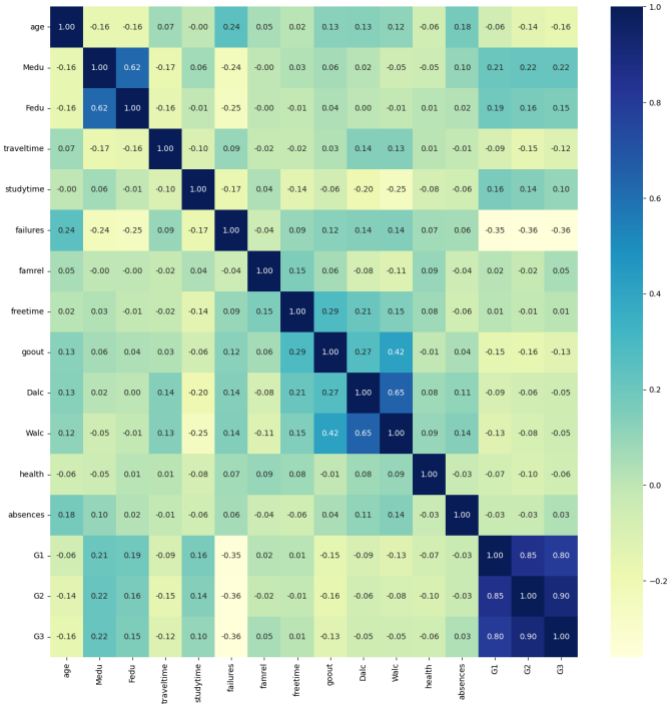


Figure 5*: Correlation map of Math dataset*



Note. The correlation of each pair of attributes are shown in metrics figures, with the number in each cell representing the correlation coefficient score of each pair of variables. The darker colors represent stronger correlation value while the lighter colors represent smaller value.

Figure 4 and 5 shows the correlation map of all variables according to different subjects (Portuguese and Math). As it can be seen from both correlation maps that exam scores evaluated from different periods are highly correlated with each other. It reveals that, generally, students takes almost same grade at each exams.

Among other features, mother and father's education have strongest positive correlation with a student's performance. Past failures has the strongest negative correlation with student's performance. Absences from class and travel time between a student's home and school have the strongest positive correlation with alcohol consumption. Study time has the strongest negative correlation with alcohol consumption.

There are no distinct different correlation relationships between the two datasets.

2.3. Educational Data Mining Methods

The methods we used include regression, classification, relationship mining, and clustering.

Classification and regression are two crucial goals in data mining (DM). Both involve supervised learning, where a model is trained on a dataset consisting of examples indexed by *k* belonging to the set *{1, ..., N}*. Each example maps an input vector (x_{k1}, ..., x_{kl}) to a corresponding target y_k. The primary distinction lies in the representation of the output, with classification having a discrete output and regression having a continuous one. In this case, we need to predict student performance (the Grade) to discover the impact of different factors on student performance.

Clustering and ARM are the unsupervised learning part. ARM can be used to uncover associations and dependencies between different elements in educational data. This method helps reveal intricate connections, enhancing insights into student interactions, and collaboration patterns, and supports network dynamics. Relationship mining complements clustering by providing a holistic view of the educational ecosystem. In EDM, clustering models identify student groups with similar learning patterns, aiding personalized education, resource allocation, and targeted interventions for improved learning outcomes. We want to recognize student groups in this dataset, observe the characteristics of each group, and find if is there any difference in performance level in each group.

Another method we used is random forest classification. Random forest is a powerful and widely used data mining technique, it minimizes the chance of overfitting and generates more accurate outcomes by the aggregation of predictions from multiple trees. Here we use this method for its ability to handle multiple and complex relationships since we are exploring potential attributes to the final grade in the raw data. Also, it is useful if the indicators have no obvious linear relationships.

In order to find out the associations among other features in the dataset, a relationship mining method is used. This method aims at finding possible correlations between various entities. Investigating the association among features in the dataset enables us to understand the relationships between these components and uncover patterns that are hidden in the chaos.

2.4. Data Analysis

Before beginning machine learning, we standardize numeric numbers and convert yes/no into 1/0. standardization and binarization enhance modeling performance by transforming data into a consistent scale, promoting better convergence and interpretation in models. In order to eliminate the effects of multicollinearity, we drop one duplicated column for each category of variable. For example, after the binary, the school column converts into the school_GP and the school_MS column, and we leave only the school_MS column.

Before classification, we also convert the target variable G3 (the final grade scale from 0-20) to binary. 1-10 points become 0, and 10-20 points become 1.

The relationship mining part aims to find relationships among school support, family support, extra paid classes, extracurricular activities, nursery school attendance, aspirations for higher education, internet access, and romantic involvement. First of all, a mapping for replacements is defined, assigning numerical values ('yes': 1, 'no': 0) to binary categorical data for further analysis. Subsequently, the Apriori algorithm is applied from the `mlxtend` library to identify frequent item sets based on a minimum

support threshold of 0.5, indicating the frequency of appearance of certain co-occurrence. A minimum confidence threshold of 0.7 and a lift threshold greater than 1 were then applied, indicating a stronger association than random chance. The result of the association analysis is shown in figure 1.

Figure 6: The association rule mining result

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(famsup)	(higher)	0.612658	0.949367	0.592405	0.966942	1.018512	0.010768	1.531646	0.046925
1	(famsup)	(internet)	0.612658	0.832911	0.529114	0.863636	1.036889	0.018824	1.225316	0.091847
2	(nursery)	(higher)	0.794937	0.949367	0.759494	0.955414	1.006369	0.004807	1.135624	0.030864
3	(higher)	(nursery)	0.949367	0.794937	0.759494	0.800000	1.006369	0.004807	1.025316	0.125000
4	(nursery)	(internet)	0.794937	0.832911	0.663291	0.834395	1.001781	0.001179	1.008958	0.008670
5	(internet)	(nursery)	0.832911	0.794937	0.663291	0.796353	1.001781	0.001179	1.006953	0.010641
6	(internet)	(higher)	0.832911	0.949367	0.792405	0.951368	1.002107	0.001666	1.041139	0.012586
7	(higher)	(internet)	0.949367	0.832911	0.792405	0.834667	1.002107	0.001666	1.010617	0.041534
8	(famsup, internet)	(higher)	0.529114	0.949367	0.513924	0.971292	1.023094	0.011601	1.763713	0.047937
9	(famsup, higher)	(internet)	0.592405	0.832911	0.513924	0.867521	1.041553	0.020503	1.261249	0.097880
10	(famsup)	(internet, higher)	0.612658	0.792405	0.513924	0.838843	1.058604	0.028451	1.288153	0.142922
11	(nursery, internet)	(higher)	0.663291	0.949367	0.630380	0.950382	1.001069	0.000673	1.020448	0.003171
12	(nursery, higher)	(internet)	0.759494	0.832911	0.630380	0.830000	0.996505	−0.002211	0.982874	−0.014375
13	(internet, higher)	(nursery)	0.792405	0.794937	0.630380	0.795527	1.000743	0.000468	1.002888	0.003575
14	(nursery)	(internet, higher)	0.794937	0.792405	0.630380	0.792994	1.000743	0.000468	1.002843	0.003619
15	(internet)	(nursery, higher)	0.832911	0.759494	0.630380	0.756839	0.996505	−0.002211	0.989082	−0.020561

Note. The strength of association is filtered by lift value, organizing in a decreasing order in the figure.

3. RESULTS OR FINDINGS

3.1 Results of Research Question #1 What are the factors that affect student achievement?

An initial linear regression analysis was performed, revealing a mean squared error of 8.7 and an R^2 score of 0.21. These results suggest a significant deficiency in model performance. Consequently, attempts to enhance the model by adjusting variables seem impractical, considering both their performance and the correlation map. As an alternative, converting the dependent variable (Y values) into classes is being contemplated to transition to classification models. The attained testing accuracy of approximately 86% is shown to be satisfactory, indicating favorable results in the classification process.

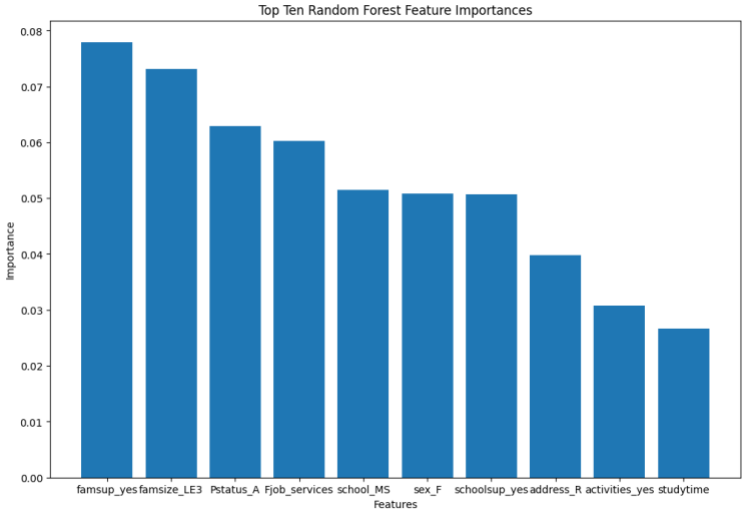
Figure 7: The result of the classification of Portuguese final performance

	precision	recall	Fl-score	Support
0	0.00	0.00	0.00	26
1	0.87	0.99	0.93	169
accuracy			0.86	195
macro avg	0.43	0.50	0.46	195
weighted avg	0.75	0.86	0.80	195
Testing Accuracy	0.8615			
Confusion Matrix		[[0 26] [1 168]]		

Note. The classification efficiency is shown in terms of testing accuracy, precision, recall, F1 score, support, and confusion matrix.

The top four feature important contribute to the classification result is famsup, famsize_LE3, Pstatu, Fjob_service.

Figure 8: Top 10 important features of classification on Portuguese final performance



Note. The top 10 most important contributing features to the model and their corresponding scores of importance are shown, in descending order according to importance.

3.2 Results of Research Question #2 Which factors have the greatest influence on a student's Math final grade?

Confirming the effectiveness of the random forest model, we did a replica analysis to student math score, the result is shown in figure 8. Overall the result is also acceptable, with about 61.34% accuracy.

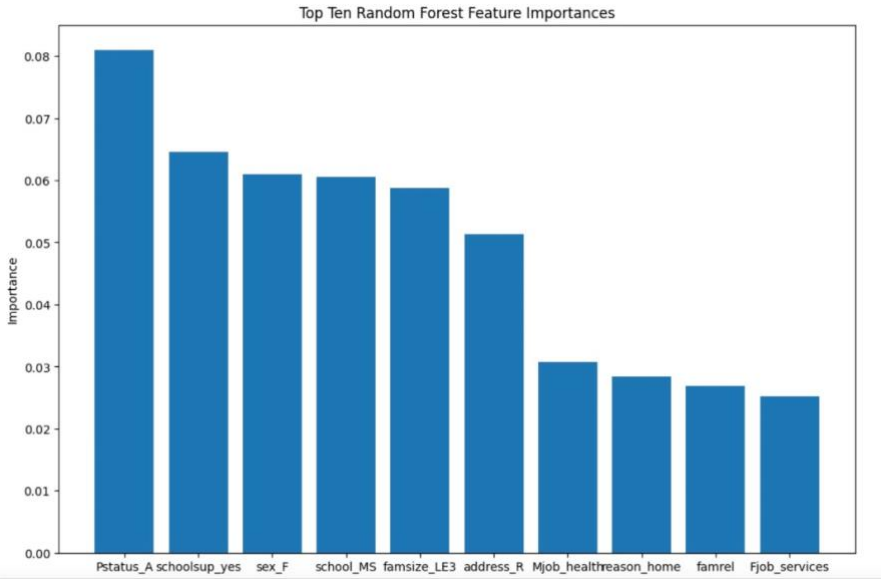
Figure 9: The result of the classification of Math final performance

	precision	recall	Fl-score	Support
0	0.5	0.04	0.08	46
1	0.62	0.97	0.76	73
accuracy			0.61	119
macro avg	0.56	0.51	0.42	119
weighted avg	0.57	0.61	0.49	119
Testing Accuracy	0.6134			
Confusion Matrix		[[2 44] [2 71]]		

Note. The classification efficiency is shown in terms of testing accuracy, precision, recall, F1 score, support and confusion matrix.

The top 10 features that make the most contribution to the model are also sorted out, displayed in Figure 2. The top four feature importance contribute to the classification result is Pstatus(parent's cohabitation status), schoolsup_yes(with extra education support), sex_F(female student), school_MS(the school MS).

Figure 10: Top 10 important features of classification on Portuguese final performance



Note. The top 10 most important contributing features to the model and their cores-

3.3 Comparison of Results of question 1 & question 2

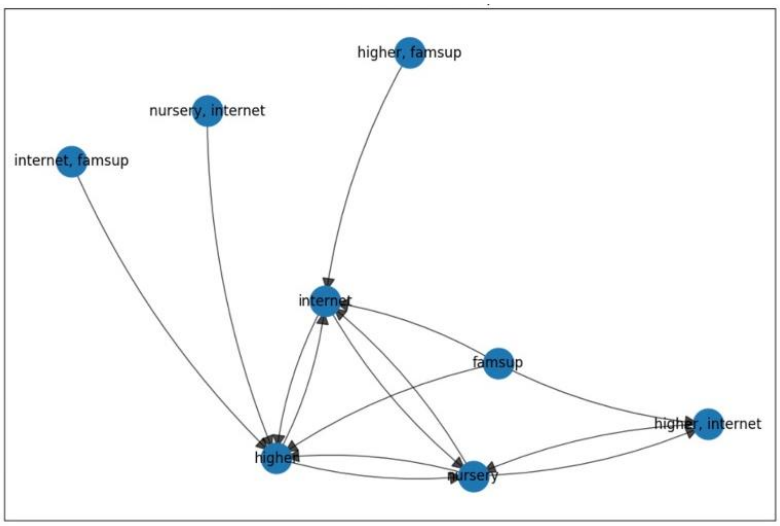
The top four features important contribute to the classification result of dataset 1 are famsup, famsize_LE3, Pstatu_A, Fjob_service. The top five features important contribute to the classification result of dataset 1 are Pstatus, schoolsup_yes, sex_F, school_MS, and famsize_LE3.

Students' performance in Portuguese is mainly influenced by family factors, such as family support, family size, parental relationship, etc. Students' math performance is also affected by school factors, such as the school they attend and school support. We can draw the rough conclusion that children from small, supportive, close-knit families have a better perception of language and words. Good math results are more dependent on the school, such as the quality of teaching, the degree of emphasis and other factors.

3.4 Results of Research Question #3

The results of the relationship mining of research question 3 are shown in Figure 11. Overall, we can see that the characteristics that stand out from the others are higher education intention, kindergarten attendance, Internet access, and family support. Although there could be many interpretations of the figure, we have chosen to interpret the two most important ones. The first and most obvious one is the intention to pursue higher education, since it has the most arrows pointing to it in the figure. Therefore, it can be concluded that students who have access to the Internet at home, who have attended kindergarten, and whose families provide educational support are most likely to want to pursue higher education. The second obvious conclusion is that Internet access is also favored by students who want to pursue higher education, who attended kindergarten and who had family educational support.

Figure 11: The association rules network graph.



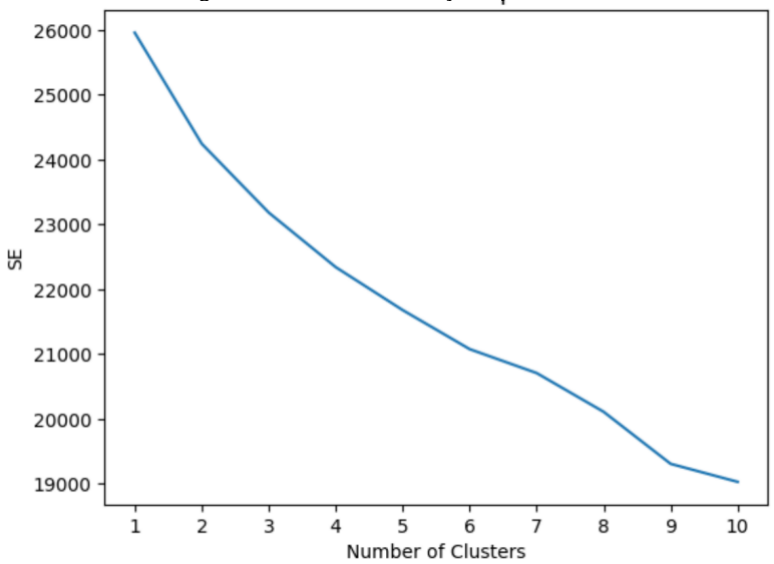
Note. The nodes represent each feature, and the edges with arrows represent each association by pointing from antecedents to consequents.

3.5 Results of Research Question #4 What are the group characteristics of students with high and low grades respectively?

The first step we want to do is doing the PCA. The number of features is about 40. We use PCA dimensionality reduction to transform the dataset into a lower-dimensional one. After adjustment, we chose the number of principal components is 25, which can retain about 90% information.

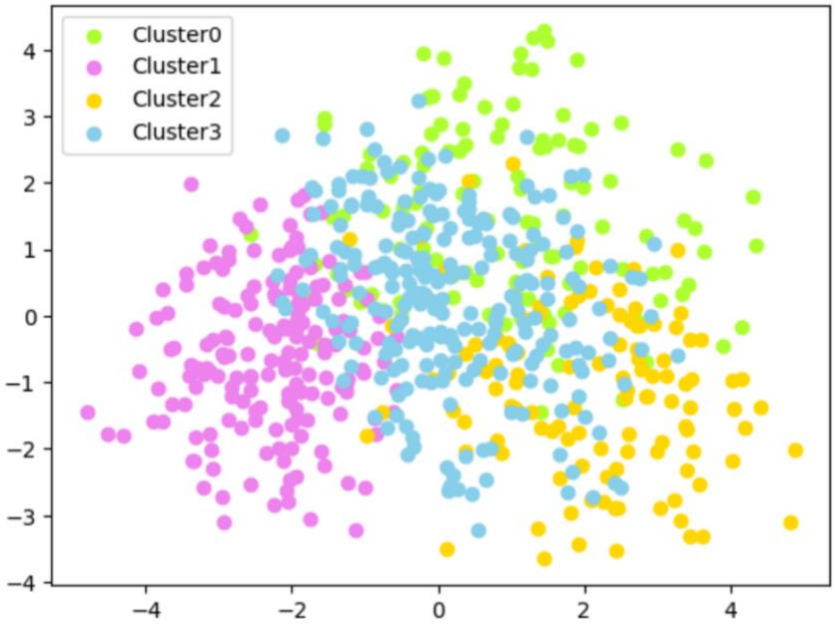
Then we should decide the optimal k for clustering. From the graph, we can see the optimal k can be 2 or 4. We chose 4.

Figure X: Elbow method for optimal k



We can see from the picture that the students are very cleanly divided into four groups. So the next step is to figure out what is the characteristics of each cluster. So we use the KNN method to figure out what features contribute to the final clustering result, making the cluster number y, all the variables as x.

Figure X: Visualization of 4 clusters



The top four features that contribute to clustering are fathers being teachers, fathers working in health-related, studying in school MS, and being guarded by moms. Then we visualize these features in each cluster. These are the results. And I also visualize the mean final score of each cluster. There is little difference in the scores of the students in each group. Students in Cluster 1 have the relatively highest scores, most of whose fathers work in health-related areas and study in school MS. Students in Cluster 2 have the relatively lowest mean score, most of whose fathers are teachers and guarded by moms.

We speculate that the reason is that fathers, as teachers, may be more strict in their children's studies, and children feel higher pressure. But at the same time, most of them are guarded by the mother, and they are closer to the mother, resulting in the father's lower academic requirements for the child. Maybe teachers and schools should pay more attention to these children’s mental health and pressure levels.

Figure X: Number of fathers being teachers of each cluster

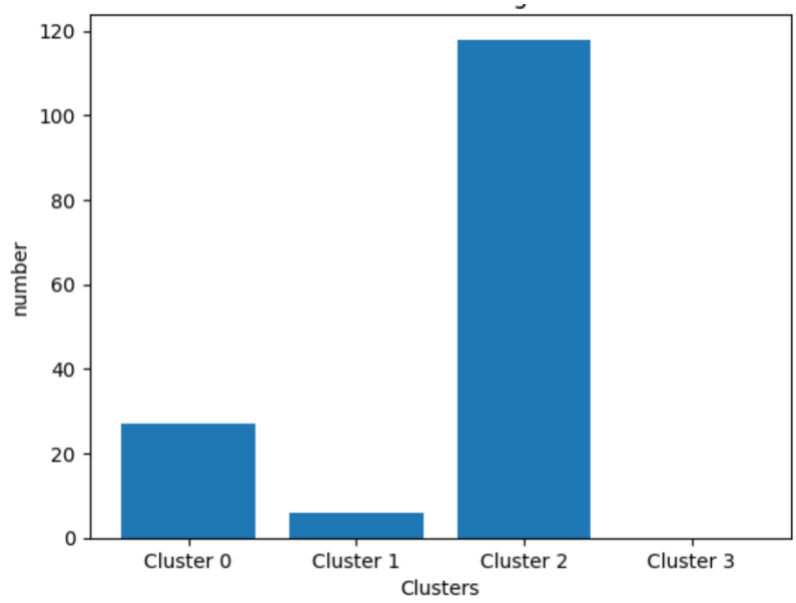


Figure X: Number of fathers working in health departments

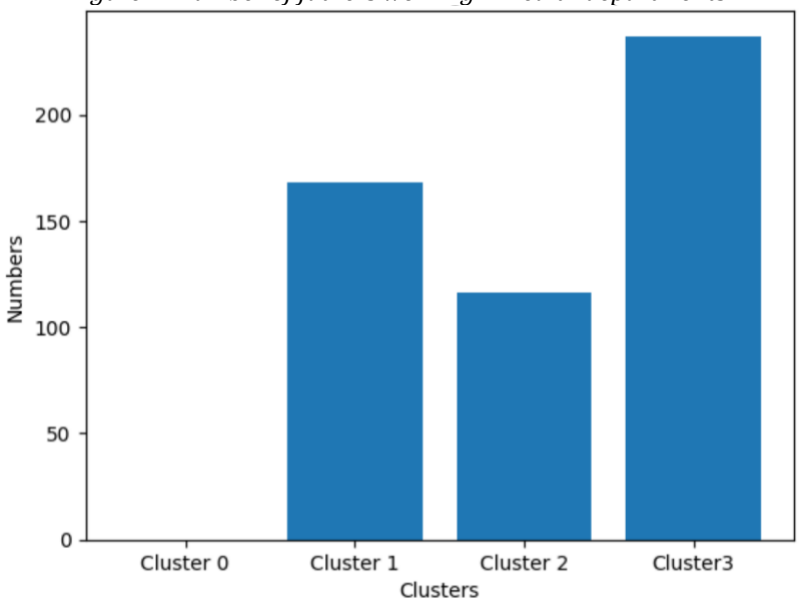


Figure X: Number of students in School Mousinho da Silveira

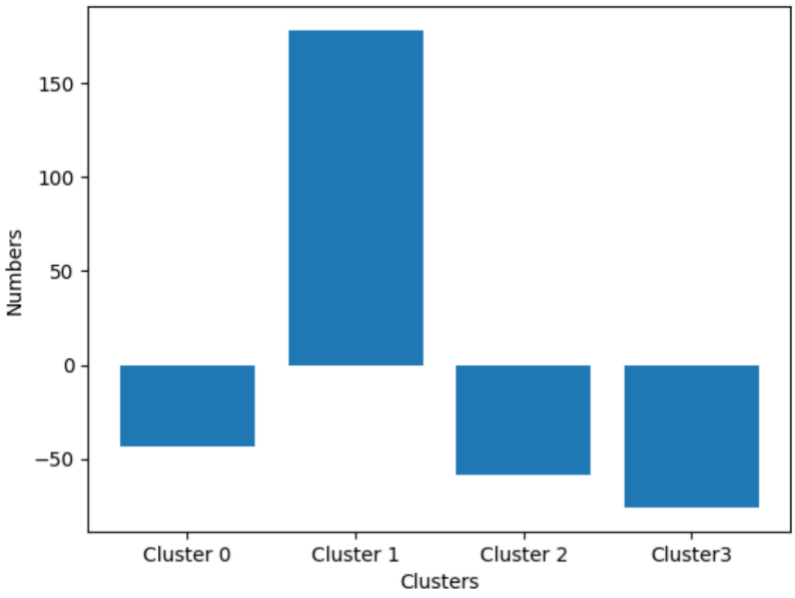


Figure X: number of students guarded by mom

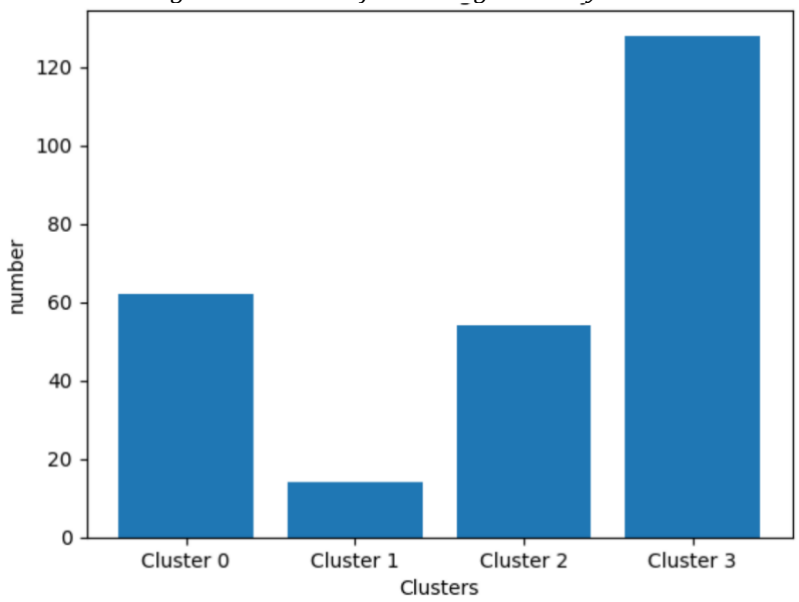
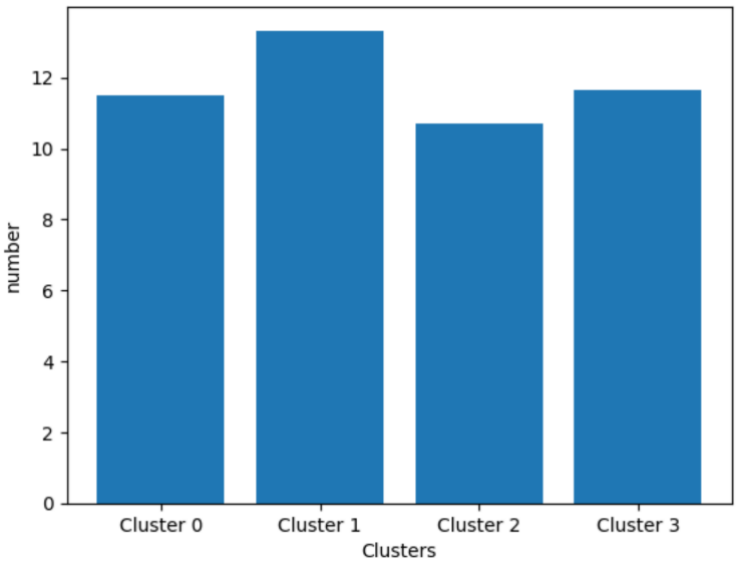


Figure X: mean final score of each cluster



DISCUSSIONS AND CONCLUSIONS

To summarize the main findings for research questions 1 and 2, we found that overall family characteristics are important for both Portuguese and math final grades, with extra school support being more important for math grades. The first aspects of our findings are consistent with previous research on the effect of family on student achievement. For example, both family members and whether parents live together contribute to our model, this result is consistent with Christenson et al. (1992) literature review that home affective environment, discipline and parental involvement are the key factors that family influences on student achievement. There has been more specific research that confirms the positive effect of parental involvement on high school students (Epstein, J. L., 2008; Fehrmann et al. ,1987). However, another important finding is that additional school support helps students improve their final math scores. Future research could be done to further confirm this finding, more specifically, why does extra school support help math improvement instead of language? We speculate that students may receive personalized and detailed explanations from this extra support, which helps them better understand math concepts; on the contrary, language learning requires more repetition and practice instead of concept understanding. However, more research is needed to prove this assumption. Or there could be more discussion if different subjects need different kinds of learning strategies and support, because there might be different methods to learn different disciplines well.

The association mining method found two important results regarding students' willingness to pursue higher education and Internet access. We found that students who have been to kindergarten, have access to the Internet, and have family support are the ones who want to go to college. Note that the data we used was collected from 2005 to 2006, at that time internet access in Portugal was still very low, only 35.2% at the end of 2006, according to OECD (2017). Thus, at that time, the Internet was considered an advanced technology. The reason why families with internet access were more likely to have children willing to pursue higher education could be that students were able to find more information about universities, including resources and courses online. Or it could be that the Internet itself acts as a

learning resource that helps students gather learning materials from a wider resource, which increases students' motivation to learn. Future research could be done on the relationship between information gathering and learning motivation. In addition, students who have attended nursery school or pre-school in their childhood are also those who want to go to university. The reason could be the positive effect of early childhood cognitive development can continue to children's teenager. For example, there was a research shows that higher quality early childhood care predicts higher cognitive achievement at age 15 Vandell et al. (2010). And Campbell et al. (2002) found in their research that, children in the preschool treatment group had significantly higher intellectual and academic achievement in adulthood, had significantly more years of schooling, and were more likely to attend a four-year college than those in the preschool control group.

In terms of the clustering result, we found that all four clusters have similar test score results, but their fathers' occupation and whether their mothers are guardians are different. This result further confirms the influence of family on students. Most interestingly, we found that whether fathers are teachers varies greatly across clusters. This is a counterintuitive result because it is usually assumed that children whose parents are teachers will have higher academic achievement. However, our clustering result shows that this is not the case, because whether the father is a teacher or not produces similar results in the students' grades. This result also contradicts with some previous research which found positive influence of parents' formal occupation like teachers to students' high school performances (Usaini& Abubakar, 2015; Omolade & Salomi, 2011; Odoh et al, 2017). This further concludes that the reasons for students' academic performance are complicated. Future research could be done in terms of whether parents with formal occupations such as teachers influence other aspects of students' academic performance other than test scores.

APPENDIX

Source code:
https://colab.research.google.com/drive/1yhdmN7E_Z_zgb12wsIQi4-1RQ7vvE5aW?usp=sharing
https://colab.research.google.com/drive/1WfyyfqdJ5tE6i-y6mIheVPRIAoU9cCvK?usp=sharing

Table1： Distribution of variables (Math)

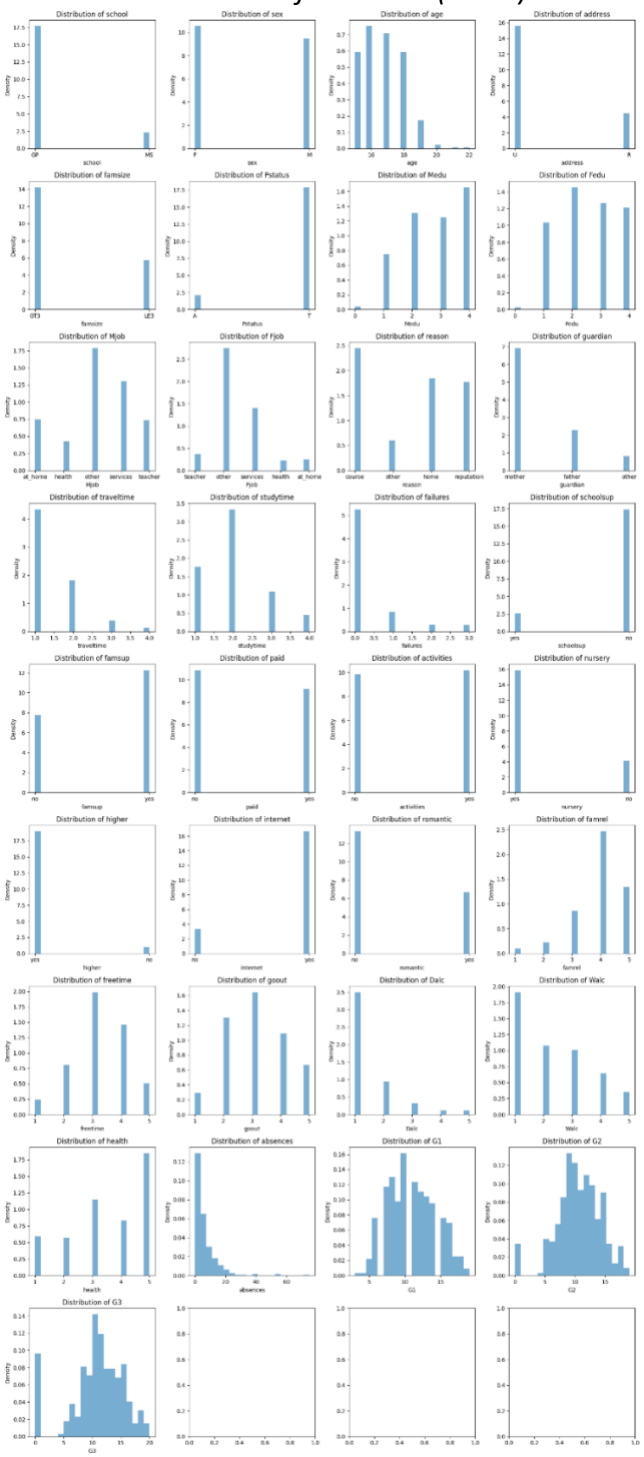
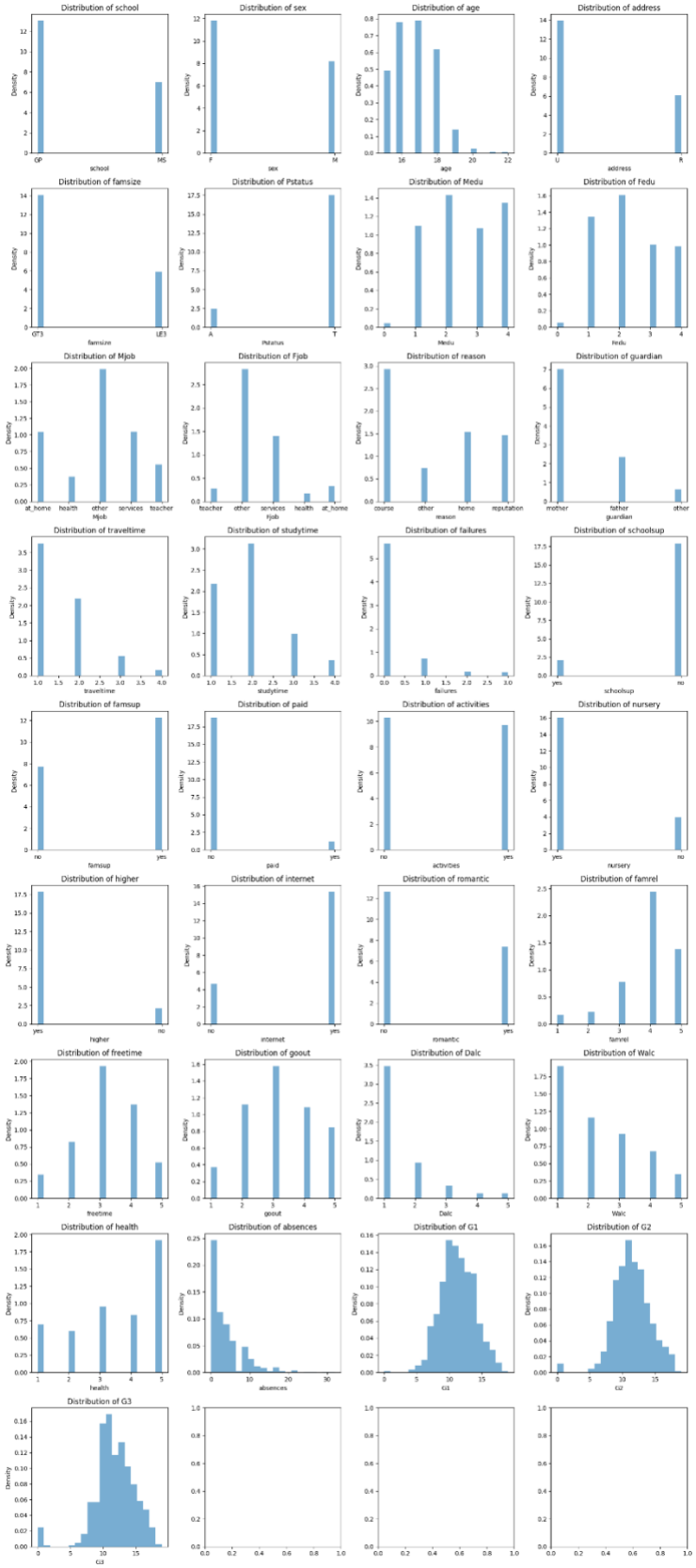


Table2：Distribution of variables (Portuguese)



REFERENCES

- B. Sravani and M. M. Bala, "Prediction of Student Performance Using Linear Regression," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154067.
- Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early Childhood Education: Young Adult Outcomes From the Abecedarian Project. Applied Developmental Science, 6(1), 42–57. https://doi.org/10.1207/s1532480xads0601_05
- Christenson, S. L., Rounds, T., & Gorney, D. (1992). Family factors and student achievement: An avenue to increase students' success. School Psychology Quarterly, 7(3), 178–206. https://doi.org/10.1037/h0088259
- Dorina Kabakchieva. (2012). Student Performance Prediction by Using Data Mining Classification Algorithms.
- Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting Student Performance Using Personalized Analytics. Computer, 49(4), 61–69. https://doi.org/10.1109/mc.2016.119
- Epstein, J. L. (2008). Improving family and community involvement in secondary schools. The Education Digest, 73(6), 9.
- Fehrmann, P. G., Keith, T. Z., & Reimers, T. M. (1987). Home Influence on School Learning: Direct and Indirect Effects of Parental Involvement on High School Grades. The Journal of Educational Research, 80(6), 330–337. https://doi.org/10.1080/00220671.1987.10885778
- Ibrahim, Z., & Rusli, D. (2007). Predicting students' academic performance: comparing artificial neural network, decision tree and liear regression. In 21st Annual SAS Malaysia Forum, 5th September.
- Odoh, L. C., Ugwuanyi, U. B., Odigbo, B. E., & Chukwuani, N. V. (2017). Influence of parental occupation and level of education on academic performance of accounting students in Nigeria. Research on Humanities and Social Sciences, 7(10), 21-27.
- OECD. (2017). Information and communication technology (ICT) - Internet access - OECD Data. TheOECD. https://data.oecd.org/ict/internet-access.htm
- Omolade, A. O. K. A. O., & Salomi, O. M. (2011). RELATIVE EFFECTS OF PARENTS'OCCUPATION, QUALIFICATION AND ACADEMIC MOTIVATION OF WARDS ON STUDENTS'ACHIEVEMENT IN SENIOR SECONDARY SCHOOL MATHEMATICS IN OGUN STATE, NIGERIA. Office Of Research And Development, Ekiti State University, Ado-Ekiti, Ekiti State, Nigeria, 14.
- Romero, C., & Ventura, S. (2012). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12–27. https://doi.org/10.1002/widm.1075
- Salazar, A., J. Gosálbez, Bosch, I., & Vergara, L. (2005). A case study of knowledge discovery on academic achievement, student desertion and student retention. https://doi.org/10.1109/itre.2004.1393665
- Student Alcohol Consumption. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/uciml/student-alcohol-consumption/data
- Ünvan, Y. A. (2020). Market basket analysis with association rules. Communications in Statistics - Theory and Methods, 50(7), 1–14. https://doi.org/10.1080/03610926.2020.1716255

- Usaini, M. I., & Abubakar, N. B. (2015). The impact of parents' occupation on academic performance of secondary school students in Kuala Terengganu. Multilingual Academic Journal of Education and Social Sciences, 3(1), 112-120.
- Vandell, D. L., Belsky, J., Burchinal, M., Steinberg, L., & Vandergrift, N. (2010). Do Effects of Early Child Care Extend to Age 15 Years? Results From the NICHD Study of Early Child Care and Youth Development. Child Development, 81(3), 737–756. https://doi.org/10.1111/j.1467-8624.2010.01431.x
- Wook, M., Yahaya, Y. H., Wahab, N., Isa, M. R. M., Awang, N. F., & Seong, H. Y. (2009, December 1). Predicting NDUM Student's Academic Performance Using Data Mining Techniques. IEEE Xplore. https://doi.org/10.1109/ICCEE.2009.168
- Yang, F., & Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. Computers & Education, 123, 97–108. https://doi.org/10.1016/j.compedu.2018.04.006
- Yang, Q., & Hu, Y. (2011). Application of Improved Apriori Algorithm on Educational Information. https://doi.org/10.1109/icgec.2011.82