Mining association rules in national testing results

Chandler Calderon, chandlercalderon@lewisu.edu

I. Introduction

Standardize testing is a common practice in the educational context, students take these test throughout primary school, secondary school and post-secondary education [1]. What makes standardize tests different from regular tests is that they are designed to be standard in their administration, scoring and interpretation; making them a more objective measure than non-standardize tests [1]. Some of these tests are evaluative, their function is to measure student's knowledge in a subject, others are diagnostic, meaning they are used as a tool to detect student's difficulties and their progress throughout time [1]. Some standardize tests are design to massive populations, like is the case for Dominican Republic's "Pruebas Nacionales" which 147,493 students took last year [2]. The goal of "Pruebas Nacionales" being to measure achievement and keep track of established educational indicators [2].

Data Mining is a suited technology to make sense of all this data, with Data Mining we can find patterns in the data that otherwise would not be possible to find [1]. These patterns can provide us additional insights that can aid the analysis of the data [1]. In this study we are going to use the same dataset as last report to add to our last analysis and paint a more complete picture of this data. The dataset contains "Pruebas Nacionales" testing scores sorted by school and other features, this time we are going to use different features and Associate Rule Mining, a Data Mining algorithm used to find rules regarding frequent item sets [3].

In the last report [4] we used three different clustering methods to create and label cluster groups, three clusters were found and labeled as:

- First cluster (C1): big schools with average scores.
- Second cluster (C2): average size schools with high scores.
- Third cluster (C3): small schools with average scores.

With this information it was inferred that the size of the school was related to the test scores. Most of the high scoring schools were small to average sized, in average scoring schools the size appeared to matter and very low scoring schools were almost exclusively small [4].

To try to respond with more certainty if there is a relationship between scores and the size, Associate Rule mining was used to try and find association rules regarding size and school modality through frequent sets.

The same dataset was used, although the data extracted from the dataset was different. Because the data is not the same, this study can't answer the posed questions in the last report of the same data, but it can help answer those questions in a broader manner because more instances were used. The Association Rule Mining was done with "Orange" Data Mining software for Windows, the data was cleaned with Python's programming language Pandas library.

II. Data

For the Association Rule Mining a dataset provided by the Dominican Republic's ministry of education (MINERD) was used, it was downloaded through the government's open source data repository (see [2] for more information). This dataset is 2016-2018 "Pruebas Nacionales" national testing results sorted by schools, each instance is a school [2]. The dataset contains 30694 instances and 17 features, out of the 17 features, 7 are of the categorical type and 10 are of the numerical type [2].

A. Categorical Features

Categorical features: Year of test, Summoning number, Regional school number, School district number, Modality, School code and School name. For the Association Rule Mining only the Modality was used, the Modality refers to the school type of school and it contains 5 modalities: Adult education, General, Technical, Basic, Art. Only the General, Technical and Adult education were used.

B. Numerical Features

Numerical features: Not promoted, Promoted, students, Male students, Female students, Number of students, Natural science mean score, Social sciences mean score, Math subject mean score and Spanish mean score. Only the total number of students was used.

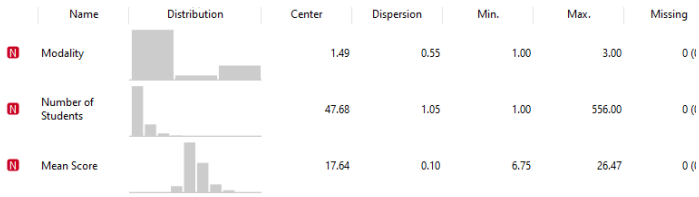
C. Data Cleaning

The dataset was cleaned and manipulated with Python Pandas library to only present the data that used in the Association Rule Mining. All categorical data except Modality were dropped from the dataset. A new feature called "Mean score" was created from the mean of all subject score's features: Math mean score, Spanish mean score, Social sciences mean score, Natural Sciences mean score. Subject average test scores are normalized, scores interval is 0-30 and they represent the

actual test scores. The resulting dataset contains 9,440 instances and 3 features.

D. Descriptive Statistics

The general descriptive statistics of the dataset features were examined. The mean Number of students was 47.68 with a minimum of 1 and a maximum of 556. The Mean Score was 17.64 with a minimum of 6.75 and a maximum of 26.47, following a normal distribution skewed to the left as seen in Fig. 1.



	Name	Distribution	Center	Dispersion	Min.	Max.	Missing
N	Modality		1.49	0.55	1.00	3.00	0 ((
N	Number of Students		47.68	1.05	1.00	556.00	0 ((
N	Mean Score		17.64	0.10	6.75	26.47	0 ((

Fig. 1. Descriptive statistics of data.

III. Methodology

For the Association Rule Mining, "Orange" Data Mining program "Frequent Itemsets" and "Association Rules" widgets were used with their default algorithm.

A. Association Rule Mining

Association Rule Mining can be defined as the process of extracting frequent sets of items from data and presenting them as a collection of if-then rules [3].

The container of the set of items is called a "basket" and the implication of each association rule is how likely an item is to appear in a basket, which is called the "confidence" of the rule [3]. The confidence tells how likely is an item to appear in a specified number of baskets [3].

The confidence can tell us how likely an item appears but it can't tell us if there is a true relationship, meaning the item we are interested affect other items [3]. To help determine that relationship, we can use the "lift" measure of an association rule [5]. The lift is the confidence of a rule divided by the expected confidence [5]. The expected confidence refers to what is expected if the items were occurring independently [5]. A lift value of 1 means that the items are independent to each other, a value greater than 1 means that the items occur more often than if they were independent to each other, meaning a relationship might exist between items [5].

IV. Results

In this section the results of the Association Rule Mining done by the Orange default algorithm are presented.

Since the dataset follows close to a normal distribution, the frequent itemsets represent previously discussed descriptive statistics. For example, the general modality had 72.34 support which represents the frequency compared to the other two modalities, 72.34% percent of instances belong to modality 1, the general school modality. The remaining 2 features were evenly distributed in three by the discretization of the numerical values. The number of students was discretized in less than 20 students, between 20 and 50 students and more than 50 students. The mean score was discretized in less than 16.6812, between 16.6812 and 18.1163 and more than 18.1163 as seen in Fig 2.

Itemsets	Support	%
Modality=< 2	6829	72.34
Number of Students=< 20	3187	33.76
Number of Students=20 - 50	3113	32.98
Number of Students=≥ 50	3140	33.26
Mean Score=< 16.6812	3143	33.29
Mean Score=16.6812 - 18.1163	3147	33.34
Mean Score=≥ 18.1163	3150	33.37

Fig 2. Frequent itemsets.

Regarding the association rules found in the data. Taking in consideration that school modalities 2 and 3 instances were disproportionally less than modality 1. Rules with a minimum support of 1% and 60% minimum confidence were selected and ordered by lift value as seen in Fig. 3. Of all the rules the following were the most interesting:

- Modality 3 and more than 50 students with a mean score lower than 16.6812, the lift value being 2.070, meaning that this rule was twice more likely to appear than random with a confidence of 69%.
- Modality 2 and more than 50 students with a mean score more than 18.1163, the lift value being 1.878 with a confidence of 63%.
- Modality 3 with a mean score lower than 16.6812, the lift value being 1.874 with a confidence of 62%.
- Modality 2 with a mean score higher than 18.1163, the lift value being 1.874 with a confidence of 62%.

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.039	0.689	0.056	5.919	2.070	0.020	Modality=≥ 2, Number of Students=≥ 50	→	Mean Score=< 16.6812
0.025	0.627	0.040	8.400	1.878	0.012	Modality=2 - 2, Number of Students=≥ 50	→	Mean Score=≥ 18.1163
0.131	0.624	0.209	1.591	1.874	0.061	Modality=≥ 2	→	Mean Score=< 16.6812
0.043	0.623	0.069	4.821	1.870	0.020	Modality=≥ 2, Number of Students=20 - 50	→	Mean Score=< 16.6812
0.042	0.623	0.067	4.953	1.866	0.019	Modality=2 - 2	→	Mean Score=≥ 18.1163
0.098	0.852	0.115	6.277	1.178	0.015	Number of Students=20 - 50, Mean Score=≥ 18.1163	→	Modality=< 2
0.111	0.829	0.134	5.398	1.146	0.014	Number of Students=< 20, Mean Score=≥ 18.1163	→	Modality=< 2
0.092	0.804	0.114	6.358	1.112	0.009	Number of Students=20 - 50, Mean Score=16.6812 - 18.1163	→	Modality=< 2
0.266	0.798	0.334	2.168	1.104	0.025	Mean Score=≥ 18.1163	→	Modality=< 2
0.257	0.772	0.333	2.170	1.067	0.016	Mean Score=16.6812 - 18.1163	→	Modality=< 2
0.083	0.765	0.108	6.682	1.058	0.005	Number of Students=< 20, Mean Score=16.6812 - 18.1163	→	Modality=< 2
0.247	0.750	0.330	2.194	1.036	0.009	Number of Students=20 - 50	→	Modality=< 2
0.083	0.744	0.111	6.498	1.029	0.002	Number of Students=≥ 50, Mean Score=16.6812 - 18.1163	→	Modality=< 2
0.237	0.711	0.333	2.175	0.983	-0.004	Number of Students=≥ 50	→	Modality=< 2
0.240	0.709	0.338	2.143	0.981	-0.005	Number of Students=< 20	→	Modality=< 2
0.097	0.707	0.137	5.286	0.977	-0.002	Number of Students=≥ 50, Mean Score=< 16.6812	→	Modality=< 2
0.057	0.676	0.084	8.568	0.935	-0.004	Number of Students=≥ 50, Mean Score=≥ 18.1163	→	Modality=< 2
0.200	0.600	0.333	2.173	0.829	-0.041	Mean Score=< 16.6812	→	Modality=< 2

Fig 3. Association rules, minimum support 1% with minimum confidence 60%.

The other rules were not selected because they didn't add any new information to the analysis of the data.

V. Conclusion

Taking in consideration the Association Rule Mining algorithm results, descriptive statistics, visualizations and previous report on the 2018 school national testing results. Is concluded that school modalities 2 and 3 have unique

characteristics that may explain the differentiation in the previous report clusters as seen in Fig. 4.

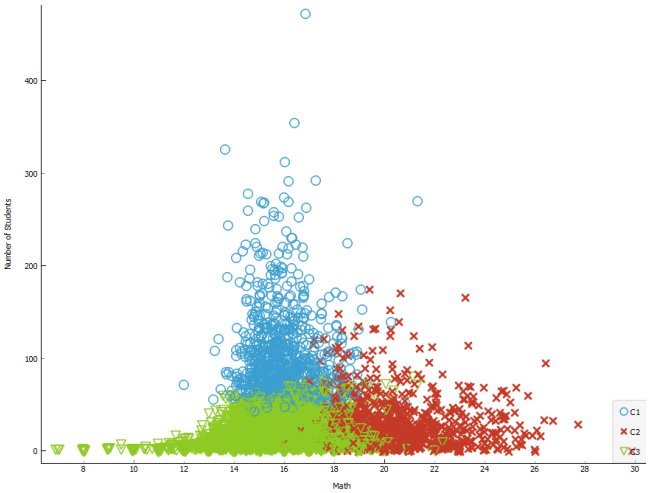


Fig. 4. Scatterplot representing Number of Students feature (y-axis) and math score feature (x-axis). Blue, red, green represent C1, C2, C3 clusters respectively. Math score was highly correlated to the other subjects, meaning it might serve as a proxy to the mean of all subject scores.

The rules with the highest lift value, meaning they were more likely to occur than the other rules, consisted of rules regarding modality 2 and 3, even though modality 2 and 3 represented a small minority of the instances.

The modality 3 with a low mean score rule and the modality 3 with a high student count rule might explain cluster 2 peak. The modality 2 with a high mean score rule and the modality 2 with high student count rule might explain the sparse high count and high mean score schools. Meaning that modality 2 schools being the exception, schools with a high number of students tend to perform worse than average to low number of student's schools.

Modality 2 schools are technical schools, schools were students graduate with a technical degree in a specific area. Technical schools have a high number of students because of the nature of technical teaching resource management. Modality 3 schools represent adult learning schools, schools designed to students that didn't finish school in time. These population of students contains a high number of previously not promoted students which might explain the lower score average.

The conclusion of the previous report regarding the number of students per school and score performance was "Small to average sized schools ranged between low and high scores. After around 100 students per school the rate of higher than average scores are drastically reduced" [4]. With the Association Rule Mining analysis, we can add that the exception are technical schools that have a high student count which tend to score high and adult learning schools which have a high count of students and tend to score low.

REFERENCES

[1] R. Rabbany and O. R. Zaiane, "Mining Large Scale Data from National Achievement Tests," in ASSESS Dat. Min. Edu. Ass. Fee. Wor. with KDD, NYC, New York, USA, Aug. 24, 2014.

[2] MINERD, "Estadisticos de Pruebas Nacionales, 2016-2018," Nov. 30, 2018. [Online]. Available: https://datos.gob.do/dataset/pruebas-nacionales

[3] J. Leskovec, A. Rajaraman and J. Ullman, "Mining of Massive Datasets," 2nd ed. Cambridge, U.K.: Cam. Pres., 2014.

[4] C. Calderon, "Clustering Dominican Republic's secondary school national testing results", 2019.

[5] S. Bagui, J. Just and S. Bagui, "Deriving strong association mining rules using a dependency criterion, the lift measure", Int. Jou. Dat. Ana. Tec. Stra., vol. 1, Mar. 2009.