Clustering Dominican Republic's secondary school national testing results

Chandler Calderon, chandlercalderon@lewisu.edu

***Abstract*— In this study a dataset containing Dominican Republic's 2018 secondary school national testing scores is used to perform clustering analysis. The dataset contains subject test scores results ordered by schools, 6136 instances and 10 features. Three clustering algorithms were used: Agglomerative Hierarchical Clustering, K-means and DBSCAN. The results of the clustering analysis helped create three clusters (C1, C2, C3) they represent respectively, big schools with average scores, average size schools with high scores and small schools with average scores. From the results of the cluster analysis is inferred that the school student size presents itself as a factor influencing student's test performance.**

I. Introduction

EDUCATIONAL Data Mining (EDM) is an emerging discipline that is concerned with analyzing and studying data from academic databases [1]. Through the use of various data mining methods, unique patterns can be identified which may help with the study, prediction and improvement of student's academic performance [1]. EDM can aid the decision making for the benefit of different stakeholders in the educational system, being an important tool for taking more informed decisions [1]. In this study three clustering algorithms (Agglomerative Hierarchical clustering, K-means, DBSCAN) are used to perform a cluster analysis of an educational dataset.

Every year students all around the world take standardized exams, these exams or tests assess students in nearly identical conditions with the purpose of providing an accurate measure of their knowledge in different school subjects [1]. Some of these tests are diagnostic, the data they provide is used to keep track of student's progress [1]. In some cases, standardized exams are used as a requirement, as in the case of the Dominican Republic's national test "Pruebas Nacionales" [2]. This is a 2-hour standardized test in 4 subjects (Spanish, math, natural sciences and social sciences) imparted nationally to last year secondary school students (equivalent to U.S. high school seniors) [2]. Students have to score an established average of the 4 subjects to graduate from secondary school [2].

In this study a dataset of Dominican Republic's secondary school national exam was used for the clustering analysis [2]. The dataset is provided by the Ministry of Education (MINERD) and contains each school average test results (countrywide) from 2016-2018 and other school features like the number of students who took the test in that school, number of promoted students, number of students who failed, number of female and male students and other categorical school features [2].

The motivation for this analysis is the possibility of finding differentiated groups based on different school features, previously mentioned like tests results and other school features. Founding patterns in this dataset is of value for the future betterment of the DR's educational system [1]. knowing which features are tied to the school's performance can benefit schools and students in the future [1]. Also, the chance of using Educational Data Mining methods in educational datasets is important for the understanding of EDM (EDM is my group's survey paper subject).

For the cluster analysis Orange Data Mining program (Windows version) and Python programming language was used. Before performing the analysis, the data was cleaned and prepared with Python's Pandas library. The dataset contained 16 features, only 10 were used. The dataset contains 2016-2018 national school results, only the results from 2018 were used because the relation between years is not of interest in this particular analysis, another reason for only using 2018 data was to reduce the instances to a more manageable number.

This report is divided in five sections: Introduction, Data, Methodology, Results, Conclusion and Recommendations. In the next section, section II, an overview of the data and details of the dataset used are explained. Then in section III, an overview of the clustering methods used is provided along the steps taken for the clustering analysis. In section IV, the results of the clustering methods and analysis are discussed. Lastly, in section VI, the conclusions and recommendations for future clustering analysis are presented.

II. Data

For this clustering analysis 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 [2]. This year (2019) data is not available as of writing, the most recent data was used. 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, school modality, school code and school name. For the clustering analysis none of these features was used, the only categorical featured used was the cluster number meta-feature created by the Orange program after performing any of the clustering algorithms.

B. *Numerical Features*

Numerical features: number of not-promoted students, number of promoted students, number of male students, number of female students, total number of students, natural sciences subject average test score, social sciences subject average test score, math subject average test score and Spanish subject average test score by school. All numerical features were used for the analysis. Subject average test scores are normalized, scores interval is 0-30 and they represent the actual test scores, to better reflect the real scores in the analysis the scores were not normalized further.

C. *Data Cleaning*

The dataset was cleaned and manipulated with Python Pandas library to only present the data that was to be used in the clustering analysis. Only 2018 as year of the test instances were used, lowering the number of instances from 30,694 to 6,136, this number is lower than a third of the data because 2016 instances include primary school national testing, primary school exams were discontinued after 2016, explaining the instances being lower than expected from what is supposed to be close to a third of the data.

All the categorical features were dropped, for the purposes of the clustering analysis they were not needed.

D. *Descriptive Statistics*

The general descriptive statistics of the dataset features were examined. All subject scores had only 0.54 out of 30 mean difference. This statistic can be explained because subject scores in standardized tests have a high correlation and as seen in Fig. 1 they follow a normal distribution skewed to the left in Spanish, social sciences and natural sciences, meaning there are fewer students with a score higher than the opposite; math subject scores are skewed to the right, students scored the lowest in math, meaning there were more students who scored higher than the average than the opposite. Math subject also had the highest max score (30), next highest max scores were Spanish (27.14), social sciences (27) and natural sciences (26.50), with a 2.86 score difference between math and the second highest max score. These statistics are to be expected for the most part, math is usually the lowest score subject in standardized tests in the western hemisphere, although math having the highest score by 2.86 is not expected [3].

The mean Number of Students (per school) was 31.07, mean Number of Promotions was 26.86, mean Not Promoted was 4.04 as seen in Fig. 1. Indicating that 13% of students were not promoted.

Name	Distribution	Center	Dispersion	Min.	Max.	Missing
Spanish		17.12	0.14	2.00	27.14	12 (0%)
Math		16.58	0.15	7.00	30.00	9 (0%)
Social Studies		16.68	0.13	2.00	27.00	4 (0%)
Natural Sciences		16.69	0.12	7.00	26.50	1 (0%)
Not Promoted		4.04	1.81	0.00	99.00	0 (0%)
Female Students		15.71	1.30	0.00	266.00	0 (0%)
Number of Promotions		26.86	1.25	0.00	469.00	0 (0%)
Number of Students		31.07	1.22	1.00	472.00	0 (0%)
Male Students		15.35	1.18	0.00	206.00	0 (0%)

Fig. 1. Descriptive Statistics of data

III. Methodology

For this clustering analysis three different clustering algorithms were used, Agglomerative Hierarchical clustering, K-means and DBSCAN executed with Orange Data Mining Windows software. Hierarchical Clustering and K-means algorithms were chosen because they are the most common algorithms used in EDM (and per instructions it was recommended to choose these two algorithms) [1]. DBSCAN was chosen because after performing Hierarchical Clustering and K-means the clusters presented a high number of outliers in the data, DBSCAN can be used to detect these outliers [3].

A. *Clustering*

Clustering can be known as the process of grouping a set of objects into classes of similar objects [4]. A cluster is a collection of data or objects that are similar to one another within the same cluster and dissimilar to other clusters [4]. To group together data there needs to exist a similarity measure, this is done by calculating a distance measure (Ex. Euclidean distance) [4].

Clustering is useful in cases where the most common categories within the dataset are not known in advanced, in unlabeled data like the one chosen for this study [4]. Some EDM clustering applications, schools with some similarity can be clustered together or students and their behavior can be clustered for example [1].

There are many clustering algorithms that perform clustering on datasets, each algorithm performs clustering with a different method [4]. The two main types of clustering are hierarchical approaches like Agglomerative Hierarchical Clustering and non-hierarchical approaches like K-means [4]. The main difference between methods is that hierarchical approaches assume that clusters, cluster themselves together, non-hierarchical approaches assume that clusters are separate from each other [4].

B. *Agglomerative Hierarchical Clustering*

A hierarchical clustering algorithm group a data set into various clusters via an agglomerative or a divisive approach based on a dendrogram [4]. Agglomerative clustering begins from all the points, each point representing a single cluster and obtains a hierarchy by successively merging clusters [4]. Divisive Clustering begins from a cluster with all the points and divides itself into many clusters [4].

The Agglomerative Hierarchical Clustering algorithm works with every single point being its own cluster, then as times goes on larger clusters will be constructed by combining smaller clusters until there isn't better combinations of clusters to merge [4]. The combination of clusters is done by a distance Measure, depending in what measure is chosen the clusters will be merging by that rule [4]. For example, in Euclidean distance the average points are represented as the cluster centroid, if there's only one point in one cluster that point value is the centroid, clusters with the lowest distance are merge together until there isn't more combinations [4].

The result of the algorithm can be visualized in what is called a "dendrogram" which is a graph showing all connections in a hierarchical manner, starting (agglomerative) from all the points as clusters then merging to more sparse clusters until reaching a single connection [4]. From the dendrogram the clusters are selected depending on the height of the dendrogram or number of clusters [4].

This method was favored because it provides an accurate clustering measure and is often used in EDM datasets, educational data tends to be of a hierarchical and non-independent nature making this method an ideal candidate for this type of data [1]. For our data we selected the "ward" distance measure which is based on a sum-of-squares criterion, this type of distance measure produces groups that minimize within-group dispersion [5]. Ward distance worked the best in the dataset, creating a clearer separation between clusters in the dendrogram [5].

C. *K-means*

K-means clustering algorithms belong to the point-assignment algorithms and they are the best-known algorithms of this type [6]. K-means principal characteristics is that they assume a Euclidean space and also the number of clusters (k) in advance [6]. Although a k is known in advance, there isn't a straightforward method of selecting the best k number of clusters [6].

The algorithm starts by selecting the number of k clusters and their starting points [6]. Selecting good starting points can improve the efficiency of the algorithm, a rule of thumb is to select starting points that are as far away from each other as possible [6]. This method allows to have a good chance of these points lying in different clusters [6].

The next step is to assign the points other than k to the closest cluster, specifically to the nearest centroid of a cluster, in this case the centroid will be the starting point [6]. After assigning points the centroids of the clusters are fixed, the centroids are recalculated with the new points [6]. This process repeats itself until there isn't any change on the cluster's point membership [6].

K-means was chosen because it provides a more efficient clustering process than Hierarchical Clustering although not always as accurate [6]. It was used as a comparison measure.

D. *DBSCAN*

The DBSCAN clustering algorithm is based on a density method is design to discover clusters for arbitrary shape [3]. The algorithm selects a minimum density level estimation depending on a threshold for the numbers of neighbors within a radius (distance measure), called the "minPts" [3]. Objects with more than minPts neighbors within the radius are considered a "core point" [3]. The goal of DBSCAN is trying to find those areas which satisfy the minimum threshold, points that are density reachable [3].

All neighbors within a radius are considered to be part of the same cluster [3]. Non-core points are called border points, these points are not density reachable from any core point, they are considered noise and do not belong to any cluster [3]. Finding these border points makes the DBSCAN algorithm a good method for finding outliers in the data [3].

This algorithm was particularly picked because the data presented what seemed to be many outlier points. This algorithm can separate the outlier points with more precision [3].

IV. Results

In this section the results of the cluster analysis utilizing the three, previously mentioned, clustering methods are presented with visualizations.

A. *Agglomerative Hierarchical Clustering*

Agglomerative Hierarchical Clustering was performed using a ward distance measure, after examining the clusters in a scatterplot and through descriptive statistics three clusters were selected as the ideal number of clusters, C1, C2 and C3 as seen in Fig. 2.

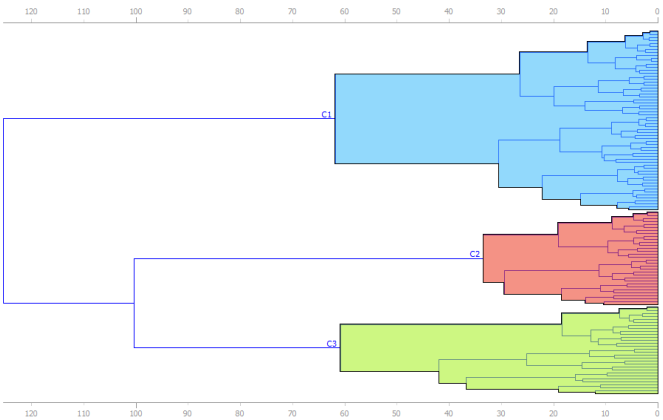

Fig. 2. Dendrogram Agglomerative Hierarchical Clustering, three clusters highlighted. Blue, red, green represent C1, C2, C3 clusters respectively.

In terms of the distribution of clusters, C1 was the smallest cluster containing 834 instances, then C2 followed with 1055 and C3 with the remaining 4247 out of 6132 instances as seen in Fig. 3.

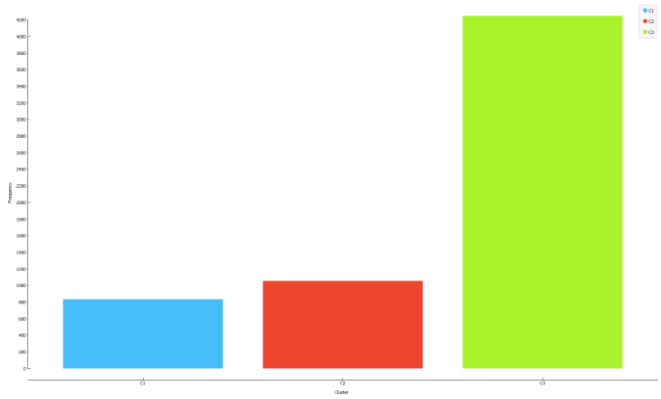

Fig. 3. Instances distribution by clusters. Blue, red, green represent C1, C2, C3 clusters respectively.

The descriptive statistics and visualizing scatterplots of features helped labeling each cluster. Each cluster represents a group of school differentiated mostly by the average number of students per school and any subject score which are correlated (explained in the Data section). Average promoted or not promoted students feature differentiated between groups, which one should expect be correlated with the average subject score because the cause of not being promoted is having a below minimum average score. Average number of students was 102 (C1), 27.42 (C2) and 17.96 (C3). Percentage of not promoted students was 15.39% (C1), 1.39% (C2), 13%(C3).

The three different clusters can be best visualized with a scatterplot between the math score feature and the number of students as seen in Fig. 4. C3 had the lowest number of students and had the lowest scores, although it overlaps with C2 who had the highest scores, meaning this group had the most variation in scores. C1 had the highest number of students and average scores, it also had more extreme outliers than the other two clusters. C2 had the highest scores and average number of students, this group was more disperse in the number of students than C3.

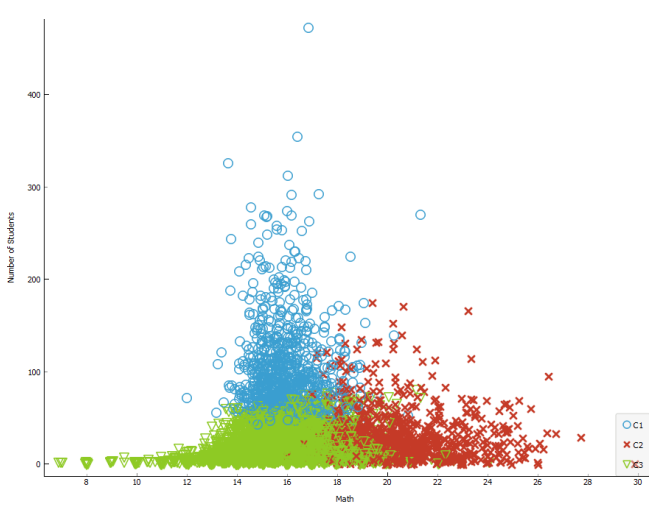

Fig. 4. Scatterplot representing Number of Students feature (y-axis) and math score feature (x-axis). Blue, red, green represent C1, C2, C3 clusters respectively.

With that information cluster labels were defined. C1 is defined by big schools with average scores and a high percentage of not promoted students (15%). C2 is defined by mostly average size schools with high scores and very low percentage of not promoted students (1.39%). C3 is defined mostly of small schools with average scores and a high percentage of not promoted students (13%).

B. K-means

The K-means algorithm was performed with 5 clusters, since Orange Data Mining Software recommended this number with the highest silhouette score. Since K-means uses Euclidean distance measure K-means had the same problem as the Hierarchical Clustering when using the Euclidean measure, a high number of very small clusters representing outlier data. Small clusters representing a few instances was detrimental for the analysis of the data, this result might be caused by the long-distance difference between some of the outlier data. For this reason, smaller k's were selected (2 and 3) and the k-means algorithm was run again. The result of the second and third try was an inaccurate version of the Hierarchical Clustering results.

After examining the k-means results, particularly attributing the inaccuracy to the outlier data, the next algorithm chosen was DBSCAN, which is used for data outlier detection.

C. DBSCAN

The DBSCAN algorithm was performed with Orange Data Mining software. A Euclidean distance metric and the program's recommended core point neighbors (4) and the neighborhood distance (1.51) based on the dataset as seen in Fig. 6.

After running the DBSCAN the result was examined in a data table, the algorithm created one cluster and identified the outlier data. A scatterplot with the same features as Fig. 5 was used to visualized the results as seen in Fig. 6.

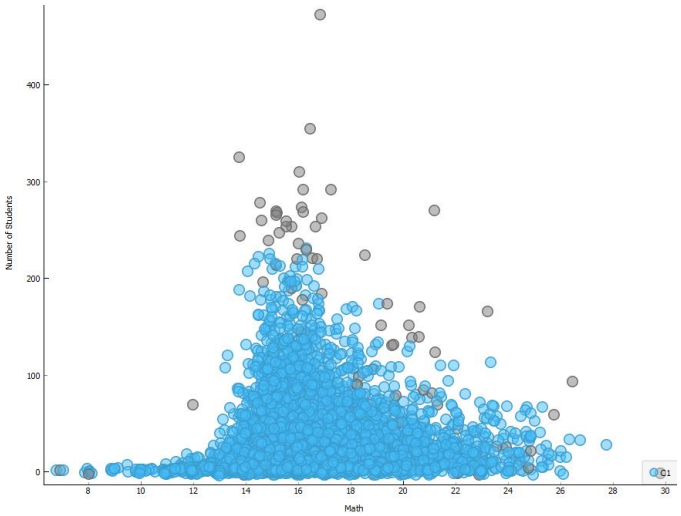

Fig. 5. Scatterplot DBSCAN cluster (blue) and outlier data points (black).

The algorithm differentiates the outlier data contained in the dataset as expected. Although the algorithm differentiates these abnormal data points, it doesn't indicate if these points are errors in the data or real outliers, for this reason these points were examined with a data table. The majority of the outlier data was attributed to features characteristic not readily apparent in the dataset.

The first group of outliers and the minority was attributed to errors in data entry. The second group also a minority was attributed to schools with extreme data like a high number of students and very high scores. The majority of cases were attributed to the third and fourth group. The third group contained schools of the modality of adult learning, a modality created for adults that didn't graduate secondary school in time. Most of the third group schools actually not be schools but testing centers that agglomerate a high number of adult students for the testing. The fourth group consisted on schools with the names of school professors, is suspected that these might not be schools but professors who for special cases test students themselves, like home school students and other cases. These last two groups of outlier data might explain the dispersion found in C1 from the Hierarchical Clustering.

The DBSCAN algorithm was very useful not only it identified the outlier data, it helped discovering unique dataset characteristics that otherwise would not have been known or taken in consideration.

V. Conclusion and Recommendations

Taking in consideration the previous results and its analysis three clusters were chosen and labeled accordingly.

The first cluster C1 represents big schools with average scores. The second cluster C2 represents average size schools with high scores. The third cluster C3 represents small schools with average scores. With this information we can infer that the size of the school might impact the test scores. Most of the high scoring schools were small to average sized. In Average scoring schools the size did not matter. The very low scoring schools were almost exclusively small.

This insight can be used to examine classroom size as a factor of academic performance. 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.

For future clustering analysis of this dataset is recommended to use a dimensionality reducing technique like Primary component Analysis (PCA) before performing K-means. Also performing outlier detection like a DBSCAN before other clustering to better identify and clean the data of outliers, depending on the goal of the clustering. Also is recommended to use other clustering methods which might result in better clusters. Calculating the exact number of students per school that reduces performance might be useful to more precisely determine the causes of the reduced performance.

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] E. Schubert et al, "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN," ACM Tran. Dat. Sys., vol. 42, art. 19, Jul. 2017.

[4] S. Zhou, Z. Zu and F. Liu, "Method for Determining the Optimal Number of Clusters Based on Agglomerative Hierarchical Clustering," IEEE Tran. Neu. Net. Lea. Sys., 2016.

[5] F. Murtagh and P. Legendre, "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?," Jou. Class., vol. 31, pp. 274-295, 2014.

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