**STUDENT PERFORMANCE CLUSTERING**

**PROJECT REPORT**

**INSTRUCTOR: Dr. Xiaofei Zhang**

**Naga Mani Yalaka, U00869534,** [**nyalaka@memphis.edu**](mailto:nyalaka@memphis.edu)

**Department of Computer Science, University of Memphis, Memphis, TN**

Dataset - Student Performance Dataset (CSV dataset)

**ABSTRACT**

The performance of the students is concentrated in this project. Clustering of student performance is done in this study using the K-Means, DBSCAN, and Agglomerative techniques. This study aims to determine whether grouping student performance by grades or scores may be used to identify performance. This will make it easier to determine which students, based on the input and output criteria, have received good marks and which ones have received low marks. Metric parameters, such as recall, accuracy, and precision, can help in determining the capabilities and performance of the K-Means algorithm.

**INITIAL APPROACH:**

Examining each student’s marks manually is very hard because of the student’s strength and also it requires more time for the teachers to divide the students into sections. Using data mining clustering techniques, it will be very easy to cluster the students into different sections and this will be used to predict the data of the students whether they will pass or fail the next exam.

**1 Introduction: The Dataset**

The Kaggle dataset is used in this project, and it has the data of students. It has around 1000 students’ information.

Link: [Student performance Analysis | Kaggle](https://www.kaggle.com/code/vardhansiramdasu/student-performance-analysis/data)

The Student Performance Dataset does have 8 columns of which three are numerical variables and the rest are categorical variables.

The dataset columns include: -

* Gender – Gender of the student (male/female)
* Ethnicity/race – Ethnicity to which the student belongs (group A, B, C, D, E)
* Parental level of education – The education level of parents
* Lunch – Standard of the lunch provided to the student
* Test preparation course – Whether the student took a preparation course or not
* Math score – Score obtained by the student in mathematics
* Reading score - Score obtained by the student in reading
* Writing score - Score obtained by the student in writing

**2 Problem Statement**

2.1 Unsupervised Learning

The unsupervised learning challenge that this research address is the clustering of student performance based on test scores or grades.

**3 Fixing the Issues**

Computer systems can generally accomplish a task without precise instructions through a technique called machine learning. In this section, I detail the issues at hand and how a machine-learning model might be created to solve them.

3.1 Scaled Features

After selecting the features, scaling the features is done.

* StandardScalar: data is scaled to have a mean of 0 and a variance of 1.
* MinMaxScalar: This is helpful when it makes sense to think of the data as a percentage of the maximum value.

3.2 Transformed data

After scaling, the data is transformed before training is commenced. Transforming data involves:-

* Feature Construction
* Handling skewness
* Data Scaling

3.3 Proposed System

I used three clustering algorithms and found the accuracy for all three. The description of various clustering models used in the project is as follows:

**Clustering:** Clustering is a way to group a set of data points in a way that similar data points are grouped together. As a result, clustering algorithms search for patterns of similarity or dissimilarity among data points. There is no label attached to the data points because clustering is an unsupervised learning technique. The program seeks to identify the underlying organization of the data.

3.3.1 Partition-based Clustering (K-Means):

K-Means is an unsupervised learning technique that partitions the data into a predetermined number of clusters. Each cluster has a centroid. Each point is assigned to a cluster with the closest centroid. The number of clusters K must be specified. K-means can be more powerful and applicable after appropriate modifications. K-means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

3.3.2 Density-based Clustering (DBSCAN):

DBSCAN stands for density-based **s**patial clustering of applications with noise. It is able to find arbitrary-shaped clusters and clusters with noise (i.e. outliers). It is an Unsupervised learning Non-linear algorithm. It groups together point that are closely packed together. Does not require specifying the number of clusters beforehand. Performs well with arbitrary shapes clusters. DBSCAN is robust to outliers and able to detect the outliers.

3.3.3 Hierarchical Clustering (Agglomerative):

Agglomerative hierarchical Clustering takes a bottom-up approach. Each item in this approach is first treated as a singleton cluster. Following that, pairs of clusters are sequentially combined into a single large cluster that contains all items. A dendrogram-a tree-based representation of the objects-is the outcome.

**4 Interesting Points to Ponder**

4.1 Challenges

Scaling was a key component of clustering in the student performance clustering challenge. It's challenging to choose which columns to remove, though. I had to consider all the factors, both numerical and categorical, before deciding which column, if deleted, may have a detrimental impact on the data frame. I concluded that it could be fair to exclude some categorical factors in favor of numerical variables.

1. Removed unnecessary attributes. 2. Removed low correlated attributes. 3. Scaled the attributes. 4. Converted the attributes to binary values.

4.2 Friendliness to Users

The model is pretty simple to comprehend and apply when it comes to user-friendliness. One can quickly follow the processes or methods and acquire useful insights from them.

4.3 Efficiency in computation and space

This learning strategy is computationally quite effective, to put it in more technical terms. Models can be trained in under a minute, depending on how the parameters are configured, and training was done over the CSV document. Training is particularly effective in terms of spatial efficiency because it uses less memory.

**5 Deliverables**

5.1 Data extraction and cleaning

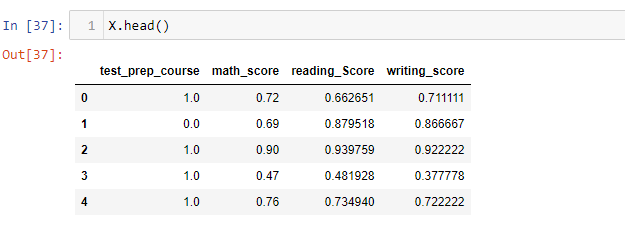
We will clean up the data so that we have useful information to work with. Additionally, feature modeling will be used to provide relevant data that is devoid of null and repeating values.

5.2 Data analysis and modeling

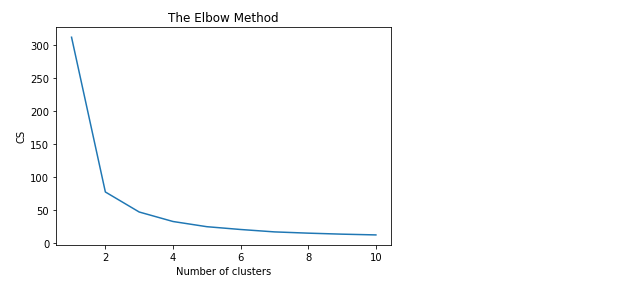
After collecting the useful data, we will put it through our K means machine learning model to assess the model's accuracy. We will cluster the student performance into the appropriate categories based on the results.

1. K\_Means

K-Means Scaled Data frame

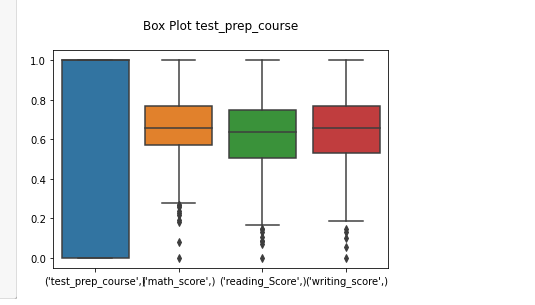


K\_Means Elbow Graph



The optimal k value is that having k=2

Boxplot



Different K-means scores

K=4 has an Accuracy score: 0.55

K=2 has an Accuracy score: 1.00

Different K-Means inertia values

K=4 has an inertia of 32.167604177172976

K=3 has an inertia of 46.60496696149129

K=2 has an inertia of 77.03808486882392

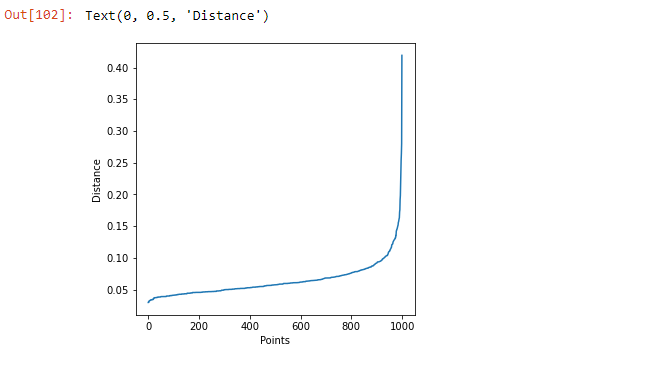
1. DBSCAN METHOD

Dbscan with k=2 with 0 noise points has a v\_measure of 100% while

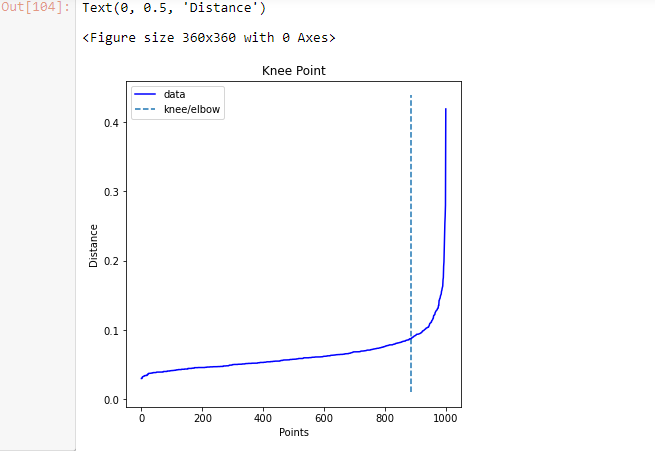
dbscan with k=2 with 16 noise points has a v\_measure of 93%

dbscan with k= 3 with 21 noise points has a V\_measure of 89%

Optimal Epsilon value

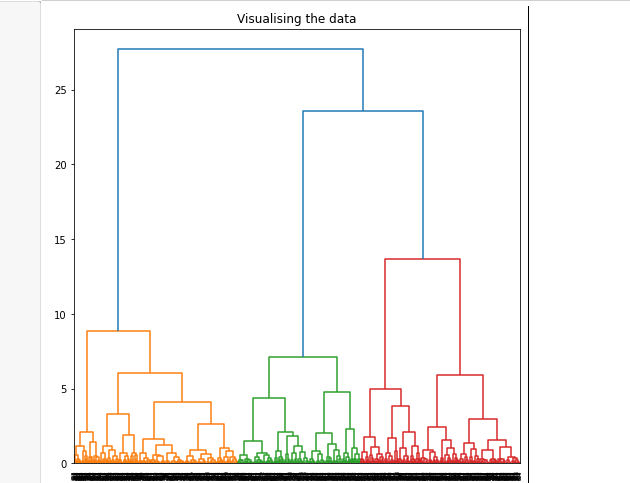


Elbow Point Detection

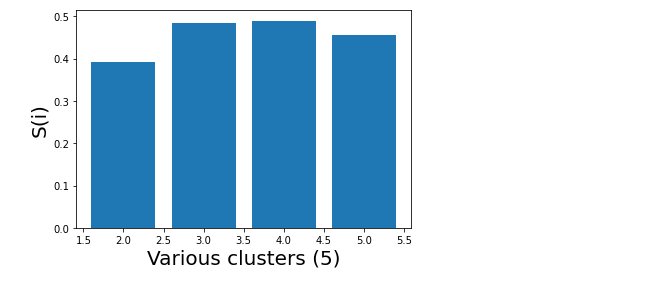


1. AGGLOMERATIVEIVE METHOD

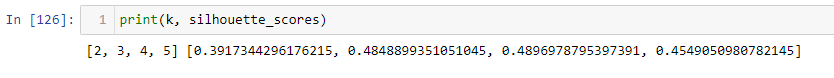
Plotting the Dendrogram



The bar graph compares the different agglomerative model results



Various silhouette scores for the various k values



From the results above, it is evident that when the Agglomerative -Hierarchical Clustering is used, the following scores are produced: - k=2, the score is 39% k=3, the score is 48% k=4, the score is 49% k=5, the score is 45%

**6 Conclusion**

* By clustering the similar behavior students based on their marks we can easily divide sections and encourage students in their studies.
* This predictive analysis can be applied to most of the university student’s records.
* The project mainly concentrates on the development of a system for student performance analysis.
* From the 3 different algorithms used {K-means clustering algorithm, DBSCAN clustering algorithm, and lastly, Agglomerative clustering method which is a type of Hierarchical Clustering Method}, it is evident that K-Means make use of accuracy to determine the preciseness of the algorithm, DBSCAN makes use of V\_measure to determine the preciseness of the algorithm and finally, Agglomerative -Hierarchical clustering method makes use of the silhouette scores to determine the preciseness of the algorithm.
* At last we can conclude that K-Means is the best algorithm to group the data for the student dataset used in this project.

I have uploaded my project to GitHub.

Link to GitHub: [NagaMani2110/Student-Performance-analysis: Data mining project on clustering (github.com)](https://github.com/NagaMani2110/Student-Performance-analysis)

**7 REFERENCES**

* Pasina I, Bayram G, Labib W, Abdelhadi A, Nurunnabi M. Clustering students into groups according to their learning style. MethodsX. 2019 Sep 24;6:2189-2197. doi: 10.1016/j.mex.2019.09.026. PMID: 31667119; PMCID: PMC6812368.
* Cui, J., Nian, M., Zhang, J., Chen, B. (2021). Cluster Analysis of Student Scores Based on Global K-Means Algorithm. In: Liang, Q., Wang, W., Mu, J., Liu, X., Na, Z., Cai, X. (eds) Artificial Intelligence in China. Lecture Notes in Electrical Engineering, vol 653. Springer, Singapore. https://doi.org/10.1007/978-981-15-8599-9\_51
* Shashikant Pradip Borgavakar, Mr. Amit Shrivastava, 2017, Evaluating Student’s Performance using K-Means Clustering, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 06, Issue 05 (May 2017), http://dx.doi.org/10.17577/IJERTV6IS050