CLOUD MODEL FOR BEHAVIOURAL ANALYSIS

[1]Vishal Singh, [2]Vimal Yadav, [3]Shripad Raaj Jha

Department of Applied Mathematics
Delhi Technological University
New Delhi, India.

***Abstract*—Analyzing Human behaviour has become essential for almost every organization for it to succeed, as it helps in tracking the people's efficiency and factors associated with said efficiency. Behavioral Analytics utilizes a combination of artificial intelligence and big data analytics on consumer behavioural data and identifies patterns, trends, anomalies, outliers. Many Individuals have been training as practitioners and using this technology to address its various applications which are quite diverse such as education, verbal behaviour. During recent years, we went through a rough phase due to COVID-19 which affected our lives as well as our education. Most of the educational classes took place online. We want to use behaviour analysis to do a study of our peers' behaviour towards online classes during our lockdown. We, in this project, are trying to apply various machine learning techniques to find which are more suitable for monitoring/ helping students of our college.**

***Index Terms*—Clustering, K-means, Student Metrics.**

I. INTRODUCTION

Unexpected incidents have occurred in recent years, requiring us to go into lockdown for months. Everyone has been affected by Covid-19 or SARS-COV2. Many people have lost loved ones as a result of this sickness.

For nearly two years, students from all around the world had to attend online classes. Everyone did not have similar resources to deal with the difficulties of online classrooms while also scoring well on tests. Many institutions took various steps to make sure that their expectations were met. Some offered supplementary online classes, while others organised student groups.

With all the uncertainty all around we wanted to analyse the impact of the same on students' performance. Were they successful in getting students to learn about all of the subjects? We were also keen to see Post Covid's performance.

This project aims to answer our questions as well as generate a reliable performance monitoring system.

II. MOTIVATION

Many people's mental health has suffered as a result of the pandemic; common problems include increased stress and loneliness.

There are many Socio-economic problems too. Not every student has equal access to facilities like fast internet or quiet study areas. It would be crucial to examine the impact that socioeconomic circumstances have had on students' capacity to succeed in online courses. What has been the impact of COVID on the students of our time? We are trying to find the answer to some of these questions by creating our model to detect student behaviour, Cause of their distress, be it financial or medical.

Once we have a clear understanding of the current state of the art, the next step is to understand the tools and technologies used for our specific problem. We will assess the suitability of different tools and technologies for our project, and identify the ones that are most appropriate for our needs.

Based on our research and analysis, we will then develop a machine learning model that incorporates multiple techniques we have learned during our review. The model will be designed to analyze human behaviour and cluster it into meaningful patterns. We will also ensure that the model is scalable and can handle large amounts of data, as this is crucial for real-world applications.

III. LITERATURE REVIEW

Clustering is a strong data mining approach that is employed in a variety of applications such as classification, and pattern recognition. There are various clustering algorithms, each with its own distinct approach to grouping things based on their attributes. These algorithms differ in complexity, scalability, and capacity to handle various forms of data. k-means, hierarchical clustering, density-based clustering, and fuzzy clustering are some of the most common clustering methods. These algorithms come with their own pros and cons, which is why it is crucial to select the best algorithm for a certain application. Since it is a powerful tool for identifying patterns and connections in data, clustering is a crucial technique for data analysis and knowledge discovery.

Partition-based clustering is an iterative procedure that divides data points into clusters by arranging them amongst clusters. All data points are initially regarded as a single cluster, which is then separated into smaller clusters. K-Means, K-Medoids, and K-Modes are some popular partitioning techniques [3]. These methods optimise clustering by iteratively adjusting the positions of cluster centroids.

Hierarchical clustering algorithms, on the other hand, adopt a different approach. Hierarchical clustering can be divided into two approaches: agglomerative and divisive [1]. Each observation begins in its own cluster in the agglomerative process, and pairs of these clusters are joined as we advance up the hierarchy. This method includes merging the two nearest clusters iteratively until only one cluster remains. The results are often shown as dendrograms, which show the clusters' hierarchical structure.

There are various advantages of using agglomerative hierarchical clustering over other clustering algorithms. One of the primary benefits is that it can generate a hierarchical structure or ordering of the objects, which can be useful for data display and interpretation. The hierarchical structure of the data can aid in understanding the links between clusters and finding patterns or

trends. Furthermore, it does not require a predetermined number of clusters, which can be advantageous when the optimal number of clusters is uncertain [6].

Furthermore, agglomerative hierarchical clustering can handle mixed-dataset datasets and can be employed with a number of similarity measures. However, this method can be computationally expensive, particularly for large datasets, and the quality of the clustering result can be affected by the similarity measure and linking mechanism used.

Divisive hierarchical clustering, on the other hand, begins with all observations in a single cluster and then splits the clusters into smaller clusters recursively. The divisive technique is breaking the largest cluster into smaller ones iteratively until all clusters contain just one observation. Dendrograms can also be used to display the hierarchical structure of divisive hierarchical clustering clusters [2].

IV. METHODOLOGY

To gather the data needed for our analysis, we conducted a survey using Google Forms. The survey was distributed to students in our college. The dataset we obtained can be divided into three parts: demographic data, prompts, and metrics.

The demographic data includes information about the students' environment and specific characteristics, such as their age, gender, and area of housing, parents' job.

The third part of the dataset consists of metrics, which allow us to track changes in student behaviour. The two metrics we focused on were attendance and marks scored by the students after the prompt. By tracking changes in these metrics, we can assess the impact of the prompts on student behaviour.

V. ANALYSIS

We started by cleaning up the data and removing any extraneous columns.

The correlation matrix for the data frame df is calculated with the code correlation = df.corr(). The generated correlation variable is a square matrix with rows and columns representing the df variables and each entry representing the correlation coefficient between two variables. Because each variable is completely associated with itself, the diagonal elements of the matrix are always 1. The correlation matrix is an effective tool for investigating the correlations between variables in a dataset.

We generate a heatmap plot of the correlation matrix. The x labels and y labels inputs give the labels for the x- and y-axes, which are obtained from the corr matrix's column names.

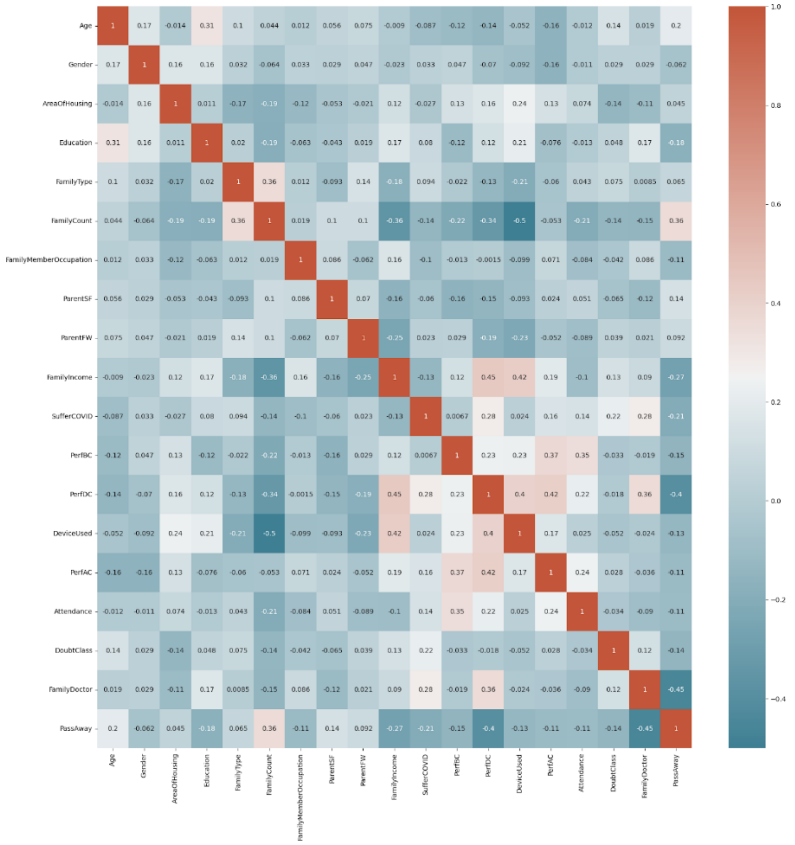

Fig. 1. Heatmap for correlation matrix

Now for the clustering

K-means clustering uses distance for unsupervised clustering that groups data points that are close to each other into a set number of clusters/groups.

The elbow method is a graphical representation of the process of determining the most appropriate value of 'K' in our Analysis. It calculates the WCSS (stands for Within-Cluster Sum of Squares), which is the sum of the square distances between cluster points and the cluster centroid.

The graph derived from the elbow method indicates WCSS values (on the y-axis) correlating to various K values (on the x-axis).

In the derived graph we check for an elbow shape and pick the value of K at the point where curve is formed. This is known as the Elbow point.

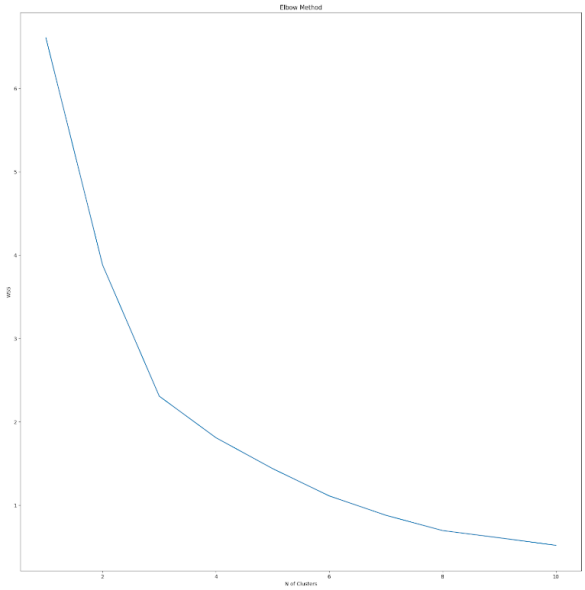

Fig. 2 elbow method for finding k (x: No. of Clusters; y: within cluster sum of squares)

From the above graph we know our k=3

We generate a K-Means clustering object with three clusters and the 'k-means++' initialization method. K-Means is an unsupervised learning technique that attempts to divide the collected data into K groups, where K is the number of cluster. The 'k-means++' initialization approach is a methodology for selecting the initial centroids for the K-Means algorithm in a way that can improve the algorithm's convergence speed and lessen the likelihood of getting trapped in local optima. k-means.fit(X) applies the K-Means model to the data in feature matrix X. The algorithm's goal is to reduce the sum of squared distances between the data points and their corresponding centroids within the cluster.

We add cluster labels to each data point in the feature matrix X after fitting the kmeans model to the data. The k_means_labels variable that results is an array of numbers, with each integer representing the cluster assignment for the corresponding data point in X.

We assign the K-Means algorithm's cluster labels to the variable pounds.The code kmean_df=dataset duplicates the original dataset and saves it in a new variable called kmean_df.

Finally, We add a new column to kmean_df called 'cluster_no_km', which holds the K-Means algorithm's cluster labels.

	metric	0	1	2	Overall Dataset
column					
Age	mean	22.195652	21.831325	22.333333	22.000000
Gender	mean	0.391304	0.349398	0.466667	0.375000
AreaOfHousing	mean	0.847826	0.795181	0.733333	0.805556
Education	mean	1.326087	1.361446	1.200000	1.333333
FamilyType	mean	0.347826	0.216867	0.400000	0.277778
FamilyCount	mean	4.543478	4.228916	6.533333	4.569444
FamilyMemberOccupation	mean	0.543478	0.590361	0.733333	0.590278
ParentSF	mean	0.956522	0.843373	1.000000	0.895833
ParentFW	mean	0.652174	0.710843	0.733333	0.694444
FamilyIncome	mean	1.695652	1.927711	1.733333	1.833333
SufferCOVID	mean	0.847826	0.927711	0.666667	0.875000
PerfBC	mean	80.260870	75.626506	67.733333	76.284722
PerfDC	mean	82.434783	86.168675	70.600000	83.354167
DeviceUsed	mean	0.891304	0.939759	0.800000	0.909722
PerfAC	mean	76.565217	74.397590	66.133333	74.229167
Attendance	mean	89.065217	68.987952	38.733333	72.250000
DoubtClass	mean	0.847826	0.963855	1.000000	0.930556
FamilyDoctor	mean	0.804348	0.963855	0.800000	0.895833
PassAway	mean	0.195652	0.048193	0.466667	0.138889

Fig. 3. Summary of summary statistics for the dataset and each cluster using K-Means

Hierarchical Clustering

We use 4 different types of Hierarchical Clustering techniques:

- Complete
- Average
- Single
- Ward

The linkage mechanism selected can have a substantial impact on the resulting clustering. Single linkage results in long, stringy clusters, whereas complete linkage results in more compact clusters. Average linkage provides a balance between the two, and Ward linkage is especially effective for minimising variance within each cluster.

Here are the results for the WARD linkage:

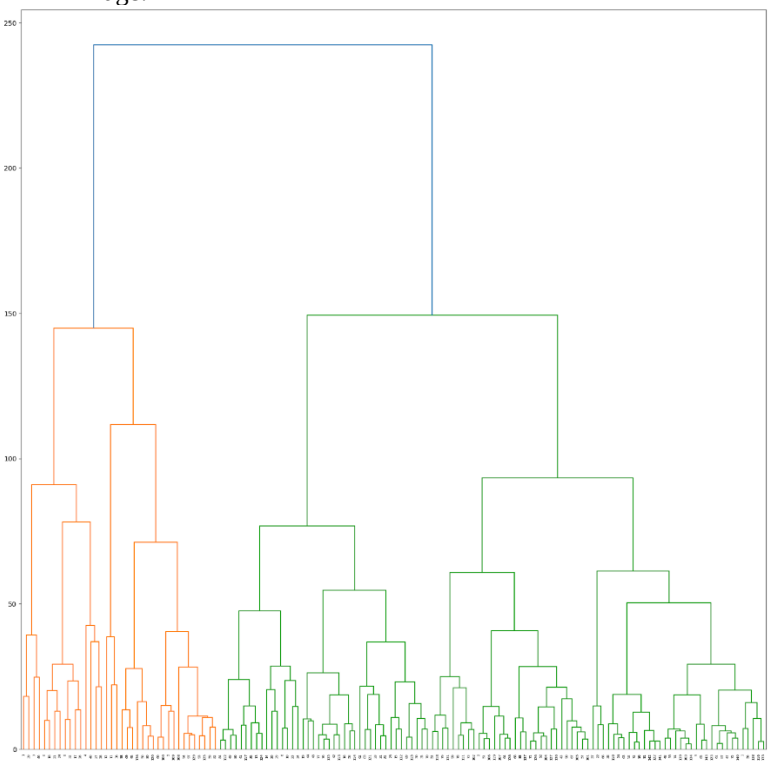

Fig. 4. Dendrogram of ward linkage method

column	metric	1	2	3	4	5	Overall Dataset
Age	mean	21.857143	21.909091	22.027778	22.428571	22.400000	22.000000
Gender	mean	0.285714	0.409091	0.388889	0.571429	0.400000	0.375000
AreaOfHousing	mean	0.809524	0.840909	0.805556	1.000000	0.600000	0.805556
Education	mean	1.309524	1.250000	1.555556	1.428571	1.066667	1.333333
FamilyType	mean	0.238095	0.318182	0.222222	0.285714	0.400000	0.277778
FamilyCount	mean	4.142857	4.590909	3.972222	5.714286	6.600000	4.569444
FamilyMemberOccupation	mean	0.595238	0.590909	0.527778	0.857143	0.600000	0.590278
ParentSF	mean	0.809524	0.954545	0.861111	1.000000	1.000000	0.895833
ParentFW	mean	0.714286	0.659091	0.611111	0.714286	0.933333	0.694444
FamilyIncome	mean	2.000000	1.772727	1.916667	2.428571	1.066667	1.833333
SufferCOVID	mean	0.952381	0.840909	1.000000	0.428571	0.666667	0.875000
PerfBC	mean	84.309524	80.704545	66.861111	76.714286	63.266667	76.284722
PerfDC	mean	87.690476	84.500000	88.138889	83.000000	56.533333	83.354167
DeviceUsed	mean	0.976190	0.909091	1.000000	1.000000	0.466667	0.909722
PerfAC	mean	82.809524	77.704545	65.166667	71.428571	63.066667	74.229167
Attendance	mean	70.738095	88.454545	67.777778	32.285714	58.333333	72.250000
DoubtClass	mean	0.976190	0.840909	0.972222	0.714286	1.066667	0.930556
FamilyDoctor	mean	1.000000	0.795455	1.000000	1.000000	0.600000	0.895833
PassAway	mean	0.023810	0.181818	0.027778	0.285714	0.533333	0.138889

Fig. 5 Summary of summary statistics for the dataset and each cluster using K-Means

VI. CONCLUSION

A Student's performance during COVID-19 is directly correlated with Area of Housing, Total Family Income, and their performance before COVID, and inversely proportional to Family Size meaning the more the members the worse the performance.
The correlation between gender and performance during covid shows that this does not affect the performance of the student. This is because most of our sample data comes from affluent households where females get similar opportunities as their male counterparts.

- K-means clustering shows that Cluster 0 had similar performance before and after COVID-19, but their performance dropped after COVID-19. Their attendance during COVID-19 was good compared to the other clusters. Cluster 1 and 2 both had students who performed well during COVID-19, possibly due to using the internet for help. However, Cluster 2 consisted of generally poor-performing students with the worst attendance out of all clusters.
- From Hierarchical clustering we found that Cluster 5 had the worst performance throughout the entire period and had the most family members, with many living in a joint family setup. Cluster 3 had suspiciously high scores during COVID-19, indicating possible cheating, and their performance dropped after COVID-19. Cluster 1 and 2 had good performers with decent attendance during COVID-19. Cluster 1, which performed the best, had more male students, possibly due to females having to devote more time to household activities.

Through this study we have been able to segregate all the students into various groups or clusters on the basis of their demographics and performance. This helps us understand the factors that might be affecting their performance.
The study has helped us a lot in understanding the complexities associated with a student's behaviour and performance in academics. The project has helped us in getting insights on ways that can improve the quality of education being imparted to the students.

REFERENCES:

1. Fionn Murtagh and Pedro Contreras, Methods of Hierarchical Clustering, Science Foundation Ireland, May 3,2011.
2. Shreya Tripathi, Aditya Bhardwaj and Poovammal, Approaches to Clustering in Customer Segmentation, IJET, Vol-7 No. 3.12 802-807, 2018
3. J.A Hartigan and M. A. W, A K-Means Clustering Algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1, 1979
4. Camila Maione, Donald R. Nelson and Rommel Melgaço Barbosa, Research on social data by means of cluster analysis, Applied Computing and Informatics (2018), doi: https://doi.org/10.1016/j.aci.2018.02.003
5. Santi Setyaningsih, Using Cluster Analysis Study to Examine the Successful Performance Entrepreneur in Indonesia, Procedia Economics and Finance Vol 4 286 – 298, 2012
6. K.Sasirekha and P.Baby, Agglomerative Hierarchical Clustering Algorithm- A Review, International Journal of Scientific and Research Publications, Volume 3, Issue 3, March 2013
7. Zhongying Yang and Xiaolong Su, Customer Behavior Clustering Using SVM, Physics Procedia Vol 33 1489 – 1496, 2012
8. Nor Bahiah Ahmada, Umi Farhana Aliasa, Nadirah Mohamada and Norazah Yusof, Principal Component Analysis and Self-Organizing Map Clustering for Student Browsing Behaviour Analysis, Procedia Computer Science Vol-163 550–559, 2019
9. R. C. T. Lee , CLUSTERING ANALYSIS AND ITS APPLICATIONS, Advances in Information Systems Science, 169–292. doi:10.1007/978-1-4613-9883-7_4
10. Kristina P. Sinaga and Miin-Shen Yang, Unsupervised K-Means Clustering Algorithm, IEEE Access, 1–1. doi:10.1109/access.2020.2988796