Original Research Article

Clustering Balanced Mental Health Data: A Comparative Study of Clustering Techniques

Abstract

Effective clustering of mental health data can provide significant insights into patterns and relationships that are critical for understanding mental health conditions. This study investigated various clustering techniques applied to balanced mental health data to avoid biases associated with an imbalanced data. Clustering of the balanced mental health data was done with respect to the area of Residence feature. Firstly, Random undersampling technique was incorporated to the imbalanced data set as a balancing technique so as to improve model performance. After balancing the data, two clustering techniques were applied to the balanced data. The two techniques were namely: K-means and Divisive techniques. In order to select which of the two clustering techniques is ideal, two test statistics namely Internal Validation and Stability Validation were applied. Results showed that K-means clustering technique indicated lower values as compared to Divisive clustering technique hence has a better performance.

Keywords: Mental Health; Residence; Random Undersampling; K-Means clustering technique; Divisive Clustering technique; Stability Validation; Internal Validation.

1 Introduction

Mental disorders are described as conditions that transforms an individuals behavioral,emotional and thought factor. Statistics according to (6) estimates that 1 in 3 women and 1 in 5 men will experience

major depression in their lives. In addition to that the depression levels may also vary due to a person area of residence. A Study was done on the prevalence of depression among older adults in India dealing with multiple health conditions living in both Urban and Rural areas. The results indicated that depression among they that lived in the rural areas was 9.48% higher than in urban areas.(13)

The integration of both machine learning algorithms and statistical techniques in mental health analysis has proven to be needful and meaningful in the field of psychiatry. This is because the analysis goes an extra mile in giving a better understanding to the scope of psychiatric problems. Statistical analysis of mental health data also helps in improving the quality of life of a person suffering from mental health disorders.(8)

1.1 Data Balancing

This study aims at incorporating various machine learning algorithm in balancing and clustering mental health data. This will be achieved by first and foremost applying a balancing technique to the imbalanced data set. Working with an imbalanced data set has its own share of challenges since it may lead to: model selection biasness towards the majority class, mis-representation of the data, reduction of model performance and also incorrect conclusion. In order to deal with this, Random undersampling technique is incorporated to the imbalanced data set, (1). This in turn: improves model performance, reduces over fitting of the majority class and thus increases the accuracy in identifying the tendencies in the data set.

1.2 Data Clustering

The balanced data set is then clustered based on the observation area of residence. Clustering of mental health data has its significant reasons such as: Identifying subgroups, identifications of patients who are at high risk, stigma reduction and it may also help better the research by enhancing the output or results accuracy. The study integrates two data clustering techniques to the data set namely: Partitional and Hierarchical clustering. K-means clustering technique represents partitional clustering while Divisive clustering technique is used for Hierarchical clustering (2). In order to do a comparison as to which of the two clustering techniques is most preferable two performance test metrics namely Internal and Stability Validation are applied.(7).

2 Literature Review

(11), proposed a psychological management system based on K-means clustering analysis method. The study used college students psychological data comprising of 1000 observations. The data was reutilised by applying the concept of data mining.The aim of the study was to analyse different college students mental health state characteristics. After clustering the applied algorithm divided the data into 3 categories composed of 20.6% ,31.9% and 47.1% of the students respectively.This results differed from those of the data test in the ideal state.

To determine whether there is a relationship between the score of the stroop test and the stress test, (14) applied K-means clustering technique to a data collected through a web application. The data composed of 105 records from the stroop and stress test. Analysis indicated that the most significant value of cohesion and separation in the clusters was achieved when $K = 3$. Findings showed that the relationship between the stroop and stress test does not have a fixed correlation.

For the improvement of both the scientific and targeted work of vocational college students,(15), incorporated clusters analysis to a data composed of 1,300 observations. The study applied K-means clustering algorithm so as to find different students groups and their characteristics in the huge data. The results showed that the students groups were based on their characteristics were in three categories.

(5) Used K means clustering algorithm to study the recognition algorithm of students' mental health problems so as to improve recognitions and management effect of the students. The results showed that the recognition algorithm of students' mental health problems based on K-means clustering had a certain effect in the mental health management of the college students.

In investigating the application and effectiveness assessment of K-means clustering in college students' mental health education. An analysis carried out by (12) with the aim of identifying distinct student profiles based on their similarities in their mental health profiles, demographic attributes and academic performance. K-means algorithms was integrated to a dataset comprising of 500 college students. Results highlighted the need to implement support services and resources to address the diverse needs of students and promote positive mental health outcomes.

(4) Statistically analyzed data comprising of 37 nurses working in closed and open inpatient psychiatric wards in a French University Hospital. The findings gave a categorical structure with four themes and 10 sub-classes,excluding the bias of confirmed expectations associated with a non-statistical analysis. The model displayed interest in an interdisciplinary approach encompassing occupational medicine and social psychology.

For the identification of distinct student anxiety profiles to develop targeted interventions,(10) applied K-means clustering algorithm. During the analysis $K = 3$ was selected as the optimal number of categories resulting in 3 clusters. Findings showed that students with anxiety systems could be categorized into distinct profiles that were open to varying strategies for management and targeted interventions.

(9) Aimed on determining groups in the population that were similar in terms of their mental resources. They applied two types of clustering techniques namely, Hierarchical and K-means and identified three clusters from the randomly selected sampled data. The clusters were namely, high-resource group,medium mental resource and low-resource group with a distribution of 48.3%,39.6% and 12.1% respectively. Findings indicated there is need for mental resources for mental health.

In order to understand heterogeneity among patients with major depressive disorder and acute suicidal ideation or behavior.(16) used Hierarchical algorithm to cluster patients on characteristics of care measured Pre-index. Fifteen patient characteristics measured during Pre-index period and during the index event were chosen for the clustering process. Analysis done on a data composed of 38,876 observations grouped them in three clusters comprising of 16,025, 5,640 and 17,211 patients respectively. These results was dependent on the patients Pre-index exposure to mental health care.

Based on the various studies done on clustering of mental data,limited research has be done on clustering a balanced data. The aim of this study is to cluster a balanced mental health data using both K-means and Hierarchical technique and then determining the best technique of the two.

3 Methods

Quantitative research design was applied in this study for the achievement of the objective. A sampled data comprising of 10,000 observations and 12 features was used for the analysis. The data was sampled from a generated data that had 1,734,982 observations. The features included Gender, Age, Marital status, Family members, Residence, Occupation, Medical test, Diagnosis, Cause, Treatment and payment.

3.1 Experimental Setup.

1. Data Balancing.
   The data was balanced based on the treatment feature in the data set. Here the sampled data was split in two parts the training and test sets at a percentage of 80% and 20% respectively. Random Undersampling method was then integrated to the imbalanced data set for balancing. The method involves balancing class distribution by randomly eliminating of the larger class examples so as to achieve a balanced set. (1),(3)

2. Data Clustering.
   This study applied data clustering based on Residence feature whose observations have been distributed in two classes Urban and Rural. In regards to their simplicity partitional and hierarchical clustering are the most commonly used algorithms. Partitional clustering aims at enhancing a particular objective function and repeatedly improving the quality of the partitions in order to discover the cluster present in the data. This clustering methods are also known as prototype-based clustering algorithm since they require certain user parameters to choose the prototype points that represent each cluster.

   Hierarchical Clustering Algorithms approach the problem of clustering by developing a binary tree-based data structure called the dendogram. This type of clustering can be implemented in two different ways namely, Divisive (top-bottom) and Agglomerative (Bottom-up). This study Applies two types of clustering techniques namely, K-means clustering technique to represent partitional clustering and Divisive clustering technique as a Hierarchical clustering algorithm. Below are the steps that are followed while implementing the fore-mentioned techniques: (2)

   (a) K-Means Clustering Technique.
   - Select K points as initial centroids.
   - Repeat the first step.
   - Form K clusters by assigning each point its closest centroid.
   - Re-evaluate the centroid of each cluster.
   - Repeat until convergence criterion is met.

   (b) Divisive Clustering Technique.
   - Start with the root node consisting all the data points .
   - Repeat the first step.
   - Split parent node into two parts D_1 and D_2 using bisecting k means to maximize wards distance W(D_1, D_2).
   - Construct the Dendogram. Among the current ,choose the cluster with the highest squared error.
   - Repeat until singleton leaves are obtained.

3.2 Performance Metrics

1. Data balancing.
   In order to test the quality and performance of random undersampling balancing technique the following test measures were implemented.(1),(3).

   - Confusion Matrix
     This measure evaluates the functionality accuracy of a given model. The output values comprises of actual negative denoted as $T^{(-)}$, actual positive denoted as $T^{(+)}$, erroneous negative denoted as $F^{(-)}$ and erroneous positive denoted as $F^{(+)}$. $T^{(-)}$, $T^{(+)}$, $F^{(+)}$ and $F^{(-)}$ indicates that the prediction were, correctly positive, correctly negative, incorrectly positive and incorrectly negative respectively. The figure 1 gives an illustration of the confusion matrix.
   - Accuracy
     Accuracy of the model can be computed using the following formula.

     $$Accuracy = \frac{(T^{(-)} + T^{(+)})}{(T^{(-)} + F^{(+)} + F^{(-)} + T^{(+)})} \tag{3.1}$$

     When working with an imbalanced data accuracy can be a little bit misguiding hence the need to apply other measures that are also calculated via the confusion matrix output.
   - Precision
     Precision measure begs to answer the question ,that for all positive predictions how many are actually positive?
     It is calculated based on the formula given below:

     $$Precision = \frac{T^{(+)}}{T^{(+)} + F^{(+)}} \tag{3.2}$$

   - Recall
     This confusion matrix measures is meant to answer the question,for all positive classes how many were accurately predicted?
     It is calculated based on the formula given below

     $$Recall = \frac{T^{(+)}}{T^{(+)} + F^{(-)}} \tag{3.3}$$

   - F Score
     This is a measure that is a combination of both the positive predictive value(Precision, Pre) and the models sensitivity (Recall, Rec) measures.Values greater than 0.7 indicates a good model. Below is a formula for calculating the F score(FS)

     $$FS = 2 \times \left(\frac{Pre \times Rec}{Pre + Rec} \right) \tag{3.4}$$

2. Data clustering.
   In order to test the performance of the two techniques two test metrics are applied namely, Internal validation and Stability Validation.(7).

   (a) Internal Validation.

   i. Connectivity
   Connectivity measures the connectedness which defines proportions to which observations are allocated in the similar groups as their nearest neighbors in that data space. It takes values $0 \Rightarrow \infty$, with a smaller value regarded as ideal.
   To illustrate this, Let;

- M: signifies the complete proportions of observations in a dataset.
- n: indicates complete number of columns that have numerical values.
- $B_{p(q)}$ represent the q^{th} nearest neighbor of observations p.
- Z be the clustering partition of the M occurrences into n dissociate clusters.Where $Z = \{y_1, ..., y_n\}$
- R be the constant value stating the number of nearest neighbors to integrate.

Therefore,

$$Conn(Z) = \sum_{p}^{M} \sum_{q}^{R} L_p \times B_{p(q)} \tag{3.5}$$

where;

$$L_p \times B_{p(q)} = 0$$

if $p = q$
and $\frac{1}{q}$ if otherwise

ii. Silhouette Score
Silhouette score estimates the degree of reliability in the grouping allocation of specific objects. It takes values ranging from $-1 \Rightarrow 1$ and should be maximized. Values > 0.5 indicates that the specific objects are well matched to its own clusters, values ranging between $0.25 \Rightarrow 0.5$ indicates that the objects are fairly matched and those < 0.5 indicates that they are wrongly matched.
It is given as,

$$Sil(p) = \frac{B_p - L_p}{max(B_p, L_p)} \tag{3.6}$$

where,

- L_p represents the mean distance linking obeject x to other objects within a similar cluster
- B_p represents the distance linking object p to occurrences in the nearest neighboring clusters

iii. Dunn Index
This Index indicates the proportion of the least distance linking any two group centroids to the greatest intra-cluster distance. The higher the Dunn index the better defined the clusters are. It takes values between $0 \Rightarrow \infty$ and should be maximized.
Dunn index is given as,

$$DunnIndex = \frac{R(e, f)}{R(p, q)} \tag{3.7}$$

where,

- $R(e, f)$ signifies the distance linking clusters e and f.
- $R(p, q)$ indicates distance linking objects p and q.

(b) Stability Validation.
These measures involves doing a comparison of the results from grouping an entire data set to that of grouping by extracting individual columns at a time . The respective measures are APN, AD, ADM, and the FOM.

i. Average Proportion of Non-overlap(APN)
It measures the mean ratio of objects not allocated similar groups by categorizing with respect to the the entire data set and grouping by extracting a single column. APN takes values $0 \Rightarrow 1$ and a smaller value corresponds with highly compatible group.

ii. Average Distance(AD)
This estimate evaluates the mean distance linking objects allocated to a similar category by grouping with regards to an entire data set and grouping by extracting a single column. APN has the values $0 \Rightarrow \infty$ and a smaller value is preferred.

iii. Average Distance between Means (ADM)
ADM measure calculates the mean distance connecting cluster centroids for objects assigned to a similar cluster by grouping with regards to the whole data set and grouping with respect to the data extracted from a single column. It has the values $0 \Rightarrow \infty$ a smaller value is considered ideal.

iv. Figure Of Merit(FOM)
FOM gives an estimate of the average error by a prognosis in regards to the cluster means. The mean intra-cluster variance of the objects in the extracted column, where grouping is based on the remaining samples are measured. It takes values $0 \Rightarrow \infty$, a lower value is preferable.

4 Results and Discussion

4.1 Results

1. Data Balancing

- Treatment Training and Test Table
Figure [2] and [3] represent the two tables respectively.

		PREDICTED	
		NEGATIVE	POSITIVE
ACTUAL	NEGATIVE	$\mathbf{T}^{(-)}$	$\mathbf{F}^{(+)}$
	POSITIVE	$\mathbf{F}^{(-)}$	$\mathbf{T}^{(+)}$

Figure 1: Confusion Matrix

TREATMENT	PSYCHOTHERAPY	MEDICATION
NUMBER OF OBSERVATIONS	4,790	3,210

Figure 2: Train Sampled Treatment Dataset

- Confusion matrix
  Confusion Matrices for both the imbalanced and Random Undersampled balanced data are represented as [4.1] and [4.2] respectively.

$$\begin{bmatrix} & Reference & \\ Prediction & 0 & 1 \\ 0 & 646 & 377 \\ 1 & 552 & 425 \end{bmatrix} \tag{4.1}$$

$$\begin{bmatrix} & Reference & \\ Prediction & 0 & 1 \\ 0 & 1198 & 0 \\ 1 & 0 & 802 \end{bmatrix} \tag{4.2}$$

- Performance Metrics Results.
  Figure [4] displays the values for the various data balancing test metrics output before and after balancing of the data set.

TREATMENT	PSYCHOTHERAPY	MEDICATION
NUMBER OF OBSERVATIONS	1,198	802

Figure 3: Test Sampled Treatment Dataset

2. Residence Distribution Table
   Figure [5] gives a description of the Residence feature distribution in the sampled data set.

DATA	ACCURACY	RECALL	PRECISION	F SCORE
IMBALANCED	0.5355	0.5392	0.6315	0.5817
RANDOM UNDERSAMPLED	1	1	1	1

Figure 4: Balanced Data Test Statistics Table

3. Balanced Data Clustering Output

   (a) K-Means Clustering Output.

   - Optimal Cluster Point Plot
     Figure [6] represents a plot of the Optimal cluster point for the Balanced data set based on K-Means Clustering technique.

RESIDENCE	URBAN	RURAL
NUMBER OF OBSERVATIONS	6,017	3,983

Figure 5: Residence Distribution Table

- Cluster Plot
  The plot represented in figure [7] gives the diagrammatic display for the distribution of the observations within the three clusters.

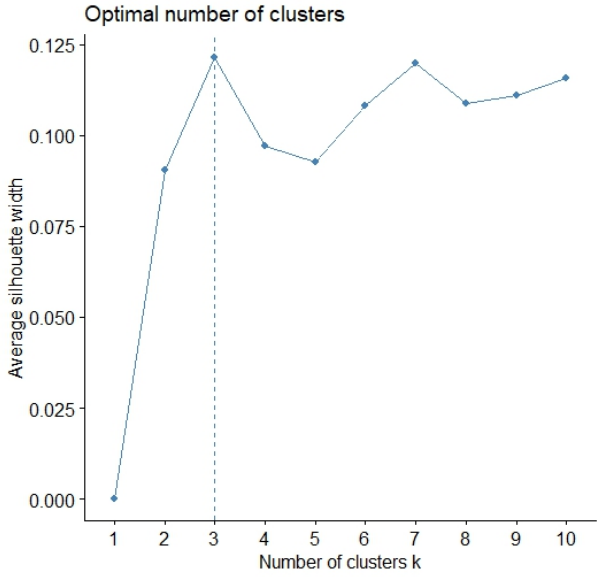

Figure 6: Optimal Balanced Kmeans Cluster Point Plot

- Cluster Table
  Figure [8] represents the values of observations in the three clusters. These observations were drawn from the two area of residence namely Urban and Rural indicated as cluster 1 and 2 respectively.

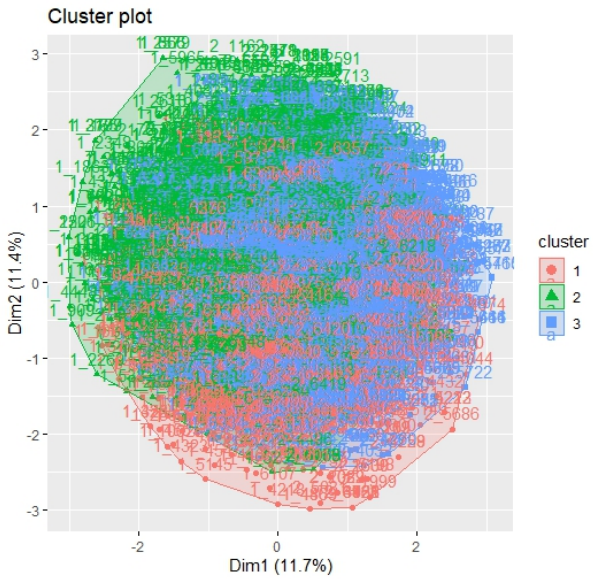

Figure 7: Balanced Kmeans Cluster Plot

(b) Divisive Clustering Output

- Optimal Cluster Point Plot
  Figure [9] represents a plot of the Optimal cluster point for the Balanced data set based on Divisive Clustering technique.

K-MEANS CLUSTERS	URBAN (1)	RURAL (2)
1	624	993
2	667	985
3	1256	1895

Figure 8: Balanced Kmeans Cluster Table

- Cluster Plot
  The plot represented in figure [10] gives the diagrammatic display for the distribution of the observations within the three clusters.

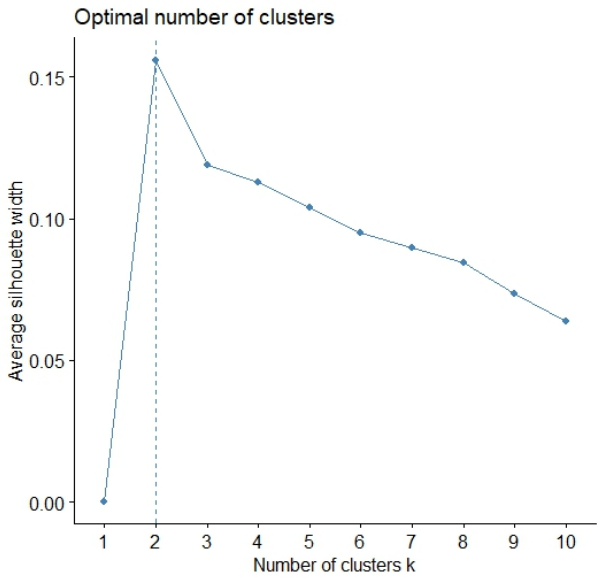

Figure 9: Optimal Balanced Divisive Cluster Point Plot

- Cluster Table
  Figure [11] represents the values of observations in the two clusters. These observations were drawn from the two area of residence namely Urban and Rural indicated as cluster 1 and 2 respectively.

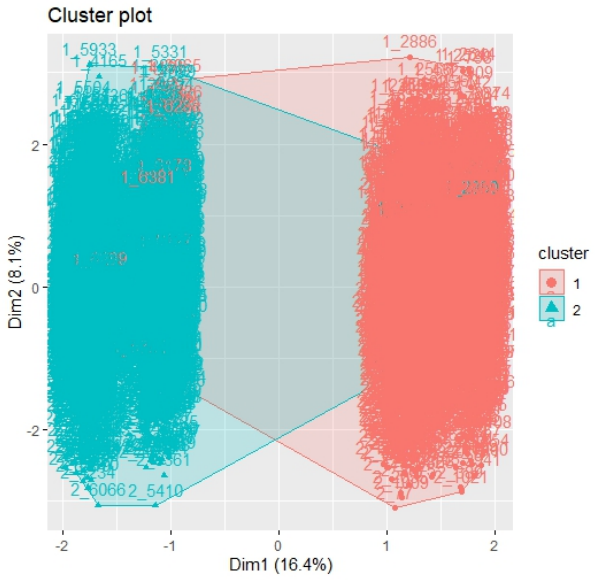

Figure 10: Balanced Divisive Cluster Plot

- Dendogram
  Figure [12] shows the diagrammatic display of the dendogram.

DIVISIVE CLUSTERS	URBAN (1)	RURAL (2)
1	1298	1940
2	1249	1933

Figure 11: Balanced Divisive Cluster Table

(c) Performance Metrics Results

- Internal Validation
  Figure [13] displays the values for the various Internal Validation test metrics output.

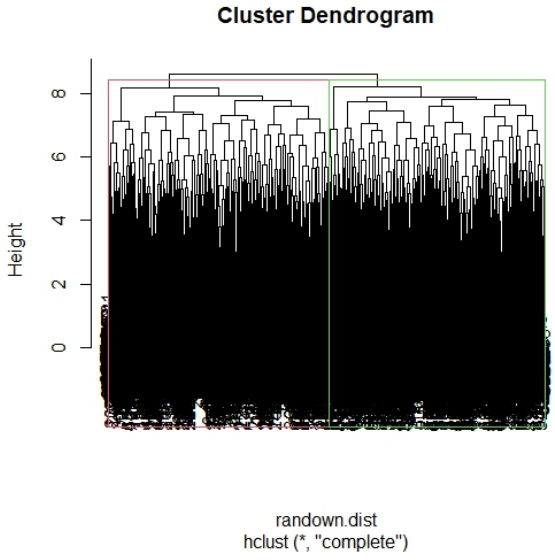

Figure 12: Balanced Divisive Dendogram

- Stability Validation
  Figure [14] displays the values for the various Stability Validation test metrics output.

BALANCED DATA	CONNECTIVITY	DUNN	SILHOUTTE
KMEANS	0.00	0.43	0.25
DIVISIVE	0.00	0.48	0.29

Figure 13: Combined Balanced IVT

4.2 Discussion

1. Data Balancing
   The sampled data with 10,000 observations was split into two part that is the training set and test set at a percentage of 80% and 20% respectively. This gave a distribution of 8,000 and 2,000 observations respectively. Balancing of the data executed based on the treatment feature and its distribution in both the training set and test set are as indicated in figure [2] and figure [3] respectively. The distribution for the training set was given as 4790 for the majority class and 3210 for the minority class. For the test set the distribution was 1198 for the majority class and 802 for the minority class.

   Figure [1],indicate the distribution of the observation before balancing of the data in regards to the reference ,prediction (ref,pred). The observations that were correctly placed are (0,0) with 646 observations and (1,1) with 425 observations. Those that were incorrectly placed were (0,1) and (1,0) with 552 and 377 observations respectively. According to figure 2,those that were correctly placed were (0,0) and (1,1) were 1198 and 802 respectively while none were incorrectly distributed.

   According to the performance metrics output table [4], there are notable changes in the accuracy, recall, precision and F score values from 0.5355, 0.5392, 0.6315 and 0.5817 respectively to 1. This indicates an ideal performance of the model.

2. Data Clustering.
   Figure [5] indicated that the number of observations in urban areas denoted as (1) are 6,017 while those in rural areas denoted as (2) are 3983. After applying silhouette method in order to define the optimal number of clusters, plot [6] and [9] indicated that the ideal number of clusters are 3 and 2 respectively. This suggested that data divided into 3 and 2 respectively are well defined and distinct.

   Based on the two clustering techniques, diagrams [7] and [10] show the visual distribution of the clusters respectively. Though some of the observations were well placed there are those that are wrongly placed hence the overlap. In order to better capture the distributions tables [8] and [11] shows the number observations that were placed in their actual clusters. According to table [8] 624 observations in the urban area and 985 in the rural area were correctly placed while the rest were incorrectly distributed to the other clusters. Table [11] indicates that 1298 observations in the urban area and 1933 in the rural area were correctly placed while the rest were incorrectly distributed to the other clusters.

   Comparing the two clustering techniques, based on the Internal Validation Test table [13] K-means technique is considered a better clustering technique as compared to Divisive technique since its connectivity value, Dunn Index value and silhouette score are lower than that of

Divisive technique. Same applies to the Stability Validation test table [14] where all the values are lower than those of Divisive technique.

5 Conclusion

Applying Random undersampling balancing technique improved model performance and hence increased the accuracy of model. This is notable in table [4] where the model accuracy, recall, precision and F score take an ideal value of 1.

After clustering of the balanced data,K-means has shown to be the most ideal clustering technique with its performance metrics values being lower to those of the Divisive clustering. This is based on both the Internal validation table [13] and the Stability validation table [14].

In order to enhance more knowledge, research work can be done with respect to other machine learning balancing and clustering techniques.

References

[1] Fernandez,A.,Garcia,S.,Galar,M.,Patri,R.C.,Krawczyk,B. and Herrera,F. Learning from Imbalanced data sets.*Springer Nature Switzerland AG*,ISBN 978-3-319-98074-4,(2018).

[2] Aggrawal.C and Reddy.C. Data Clustering:Algorithms and Applications.*Taylor and Francis Group,LLC*,ISBN 13:978-1-4665-5822-9,(2014).

[3] Chege.W.L,Waititu.H.W,Nyakundi.C.O. Comparative Analysis of Data Balancing Techniques in Mental Health Data:Application to Treatment Modalities.*International Journal of Statistics and Application*,(2024).

[4] Cougot.b,Fleury-Bahi.G,Gauvin.J,Armant.A,Durando.P,Dini.G, Gillet.N,Moret.L and Tripodi.D. Exploring perceptions of the work environment among psychiatric nursing staff in France:A qualitative study using hierarchical clustering methods .*International Journal of Environment Research and Public Health* ,(2019).

[5] Chen.Y. and Li.J. Computer Aided Recognition Algorithm for students mental health problems using K means clustering.*Computer Aided and Applications*,(2024).

[6] Dattani.S,Rodes-Guiaro.L,Richie.H and Roser.M. Mental Health.*Our World in Data*,(2023).

[7] Guy.B.,Vasyl.p.,Susmita.D. and Somnath.D.clValid,and R package for cluster validation.*journal of statistical software* , 2008 .

[8] Healy.S.A,Fantaneanu.T.A. and Whitting.S.The importance of mental health in improving quality of life in transition-aged patients with epilepsy.*ELSEVIER* , 2020 .

[9] Khachatryan.K. ,Otten.D.,Beutel.M.E.,Speerforck.S.,Riedel-Heller.S.G,Ulke.C and Brahler.E .Mental resources,mental health and sociodemography:A cluster in a large German city.*BMC Public Health* , 2023 .

[10] Liu.F.,Yang.D.,Liu.Y.,Zhang.Q.,Chen.S,Li.W.,Ren.J., Tian.X and Wang.X.Use of Latent Profile Analysis and K means clustering to identify student anxiety profiles.*BMC Psychiatry* , 2022 .

[11] Liu.Y.Analysis and Prediction of college students' mental Health based on K means clustering Algorithm.*Applied Mathematics and Nonlinear Sciences*, 2021 .

[12] Ouyang.X.Application and Effectiveness Assessment of Big Data analysis algorithm in college students' mental health education.*Journal of Electrical Systems* , 2024 .

[13] Saha.A,Mandal.B,Muhammad.T and Ali.W.Decomposing the rural-urban differences in depression among multimorbid older patients in India:Evidence from a cross sectional study .*BMC Psychiatry* , 2024 .

[14] Sarango.A.A.J,Patino.A,Acosta-Uriguen.M,Sanchez.J.G.F,Cedillo.P and Orellana.M.Analysis of Psychological test data by using K means method.*Science and Technology Publications.Lda* , 2022 .

[15] Sun.Y.Application of cluster analysis in the management of college students' mental health.*EIMSS* , 2023 .

[16] Zhdanava.M,Voelker.J,Pilen.D,Cornwall.T,Morrison.L, Vernnette-Laforme.M,Lefebvre.P,Nash.A.I,Joshi.K and Neslusan.C .Cluster analysis of care pathways in adults with major depressive disorder with acute suicidal ideation or behavior in the USA *PhamacoEconomics* , 2021 .

BALANCED DATA	APN	AD	ADM	FOM
KMEANS	0.12	2.52	0.41	0.9972
DIVISIVE	0.14	2.40	0.42	0.9999

Figure 14: Combined Balanced SVT