

# STATISTICAL METHODS IN ARTIFICIAL INTELLIGENCE COURSE PROJECT

## Understanding of Internal Clustering Validation Measures

**BY TEAM 62:**

CHIRAG JAIN SHANTILAL(2021122004)

K PAVAN KUMAR(2021122006)

SHUBHAM PRIYADARSHAN(2020102027)

YELLAPRAGADA SRI KRISHNA SARAT CHANDRA(2021122002)

---

### INTRODUCTION:

Clustering is perhaps one of the most important tasks when it comes to machine-learning, it helps us in understanding the natural grouping in the dataset, finding which is a huge task in itself.

The reason for saying this is because there is no single criteria that can be said to reflect a good clustering.

However, it is necessary to validate how good our clustering actually is, in order to give some authority to our output.

In general, we can divide cluster validation into two classes, they are:

- i) Internal clustering validation: This does not need any external information other than the data itself, which makes this a more generalised application in most of the cases.
- ii) External clustering validation” In this, we make use of external data in order to validate our clustering . However, this kind of data might not always be available. To use this, one must know the true cluster number beforehand, which makes them applicable in only a few cases.

## GOAL OF THE PAPER:

The main goal of the document is to present a detailed study of different clustering measures using some conventional aspects of clustering in general. We focus on 11 internal clustering validation measures and five aspects of clustering to be precise.

## BASICS OF INTERNAL CLUSTER VALIDATION MEASURES:

In general, Internal cluster validation has two main components, they are:

- i) Compactness: It tells us how tightly the cluster is packed. Different measures compute this in different ways. One can compute this using the variance of the points within the cluster or can be computed with the help of intra cluster distance.
- ii) Separation: This tells us how well the clusters are separated from other clusters. This can also be realised as the inter cluster measure.

In general, we can determine the optimal cluster number by the general procedure of:

Step 1: Initialize a list of clustering algorithms which will be applied to the data set. Step 2: For each clustering algorithm, use different combinations of parameters to get different clustering results. Step 3: Compute the corresponding internal validation index of each partition obtained in step 2. Step 4: Choose the best partition and the optimal cluster number according to the criteria.

Most indices consider both of the evaluation criteria (compactness and separation) in the way of ratio or summation, such as  $DB$ ,  $XB$ , and  $S Dbw$ . On the other hand, some indices only consider one aspect, such as  $RMSSTD$ ,  $RS$ , and  $\Gamma$

## DIFFERENT INDICES USED FOR THE MEASURE:

### **Davies-Bouldin Index**

In this metric, for a cluster maximum similarity with other clusters is computed. The average of similarity values of all clusters corresponds to Davies-Bouldin Index.

They capture this by estimating the cluster distribution and the distance between clusters. More nearer to the clusters and more similar the cluster distribution, more the similarity value.

Since, similarity is considered, minimum over all  $k$ 's corresponds to optimal value.

### **Xie-Beni Index**

The metric takes the form  $\frac{compactness}{separation}$ . Compactness is calculated as average intra cluster distance from centre. Separateness is calculated as maximum inter cluster distance.

Optimal value of Xie-Beni Index has to be low as compactness should be minimum and separation should be maximum.

### **SD Validity Index**

Metric considers ratio of average standard deviation of all the clusters after training to standard deviation of total data. The ratio has to be minimum as good clustering should result in less average standard deviation. This captures compactness.

The separation value is captured using distance between the cluster centres.

### **Root-mean-Square std dev**

Metric only measures compactness by calculating mean square of distances of sample points to their respective cluster centres.

### **R squared**

Metric captures compactness by calculating sum of squares of distances of sample points to their respective centres. As  $k$  increases, sum of squares decreases and elbow point is taken as optimal value.

### **Modified Hubert T(Tow) Statistic**

Metric captures separation by considering distance between cluster centres for each sample pair. As the number of clusters increases, inter cluster distance decreases so  $K$  corresponding to elbow point is optimal.

### **Calinski – Harabasz Index**

Metric considers the form  $\frac{separation}{compactness}$ . Optimal  $K$  is given by maximum index value.

### **I Index**

Metric is of the form  $\frac{separation}{compactness}$ . Separation captures minimum inter cluster distance and compactness captures sum of distances of sample points to respective clusters. Optimal  $K$  is given by maximum  $I$  index value.

### **Dunn's Index:**

Metric is of form *separation/compactness*. Separation is given by minimum distance between two sample points belongs to different clusters. Compactness is given by maximum diameter among all the clusters.

### **Silhouette Index:**

Index computes Silhouette Value for all the sample points in all the clusters. Silhouette value measures similarity of the sample point to its cluster and how distinct from the other clusters. Average value is considered as Silhouette Index.

### **S\_Dbw validity Index:**

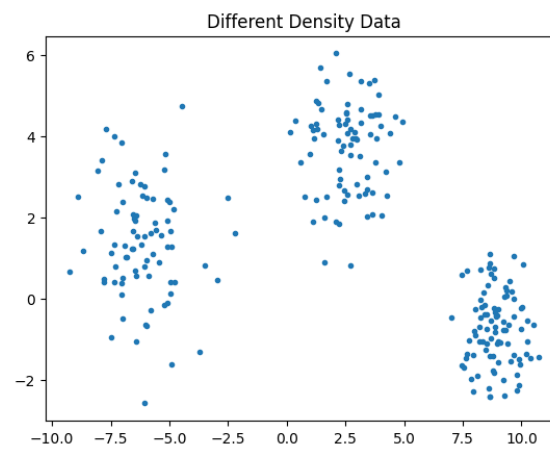
Metric computes scattering value by computing ratio of variance of sample points belongs to a particular cluster to total data variance. The index considers density of clusters to calculate separation. K corresponding to Lowest S\_Dbw value is optimal value.

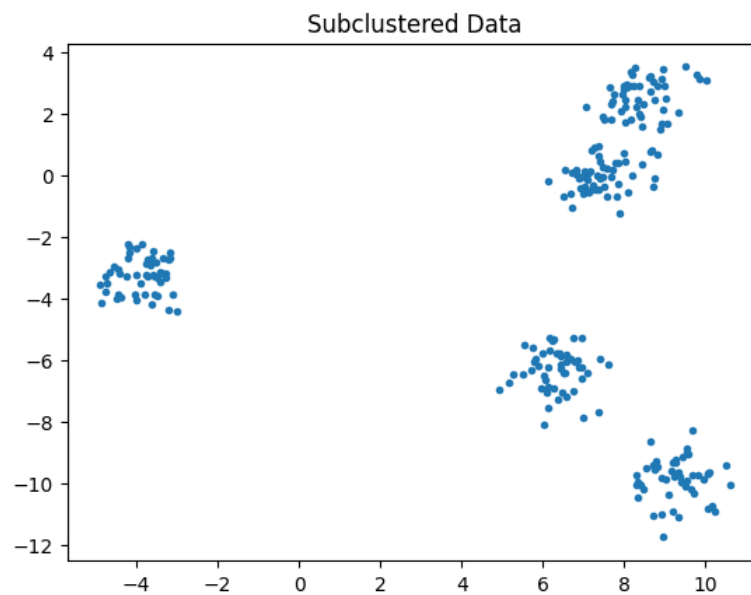
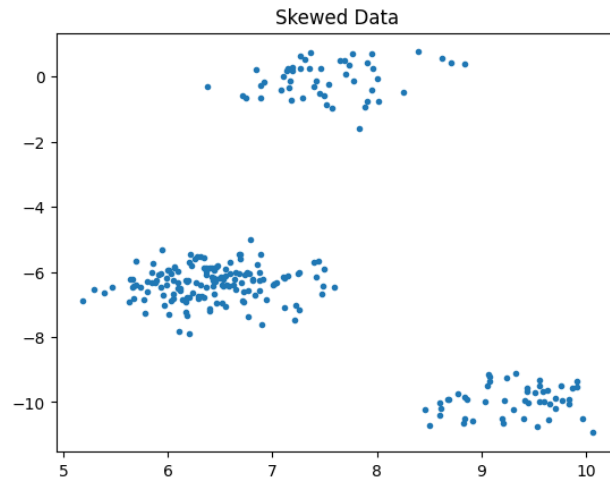
## WHAT WE IMPLEMENTED:

Our input data (synthetic):



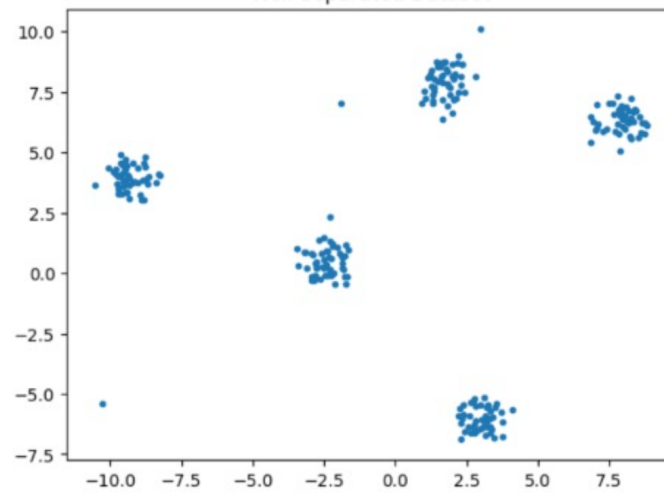
Number of clusters: 2 S\_Dbw: 2.041826970531833  
Number of clusters: 3 S\_Dbw: 0.6613393448914942  
Number of clusters: 4 S\_Dbw: 0.17127587255620919  
Number of clusters: 5 S\_Dbw: 0.08924097873839398  
Number of clusters: 6 S\_Dbw: 0.1798547698635274  
Number of clusters: 7 S\_Dbw: 0.2753976763412074  
Number of clusters: 8 S\_Dbw: 0.26051731684561996  
Number of clusters: 9 S\_Dbw: 0.26894815779311887



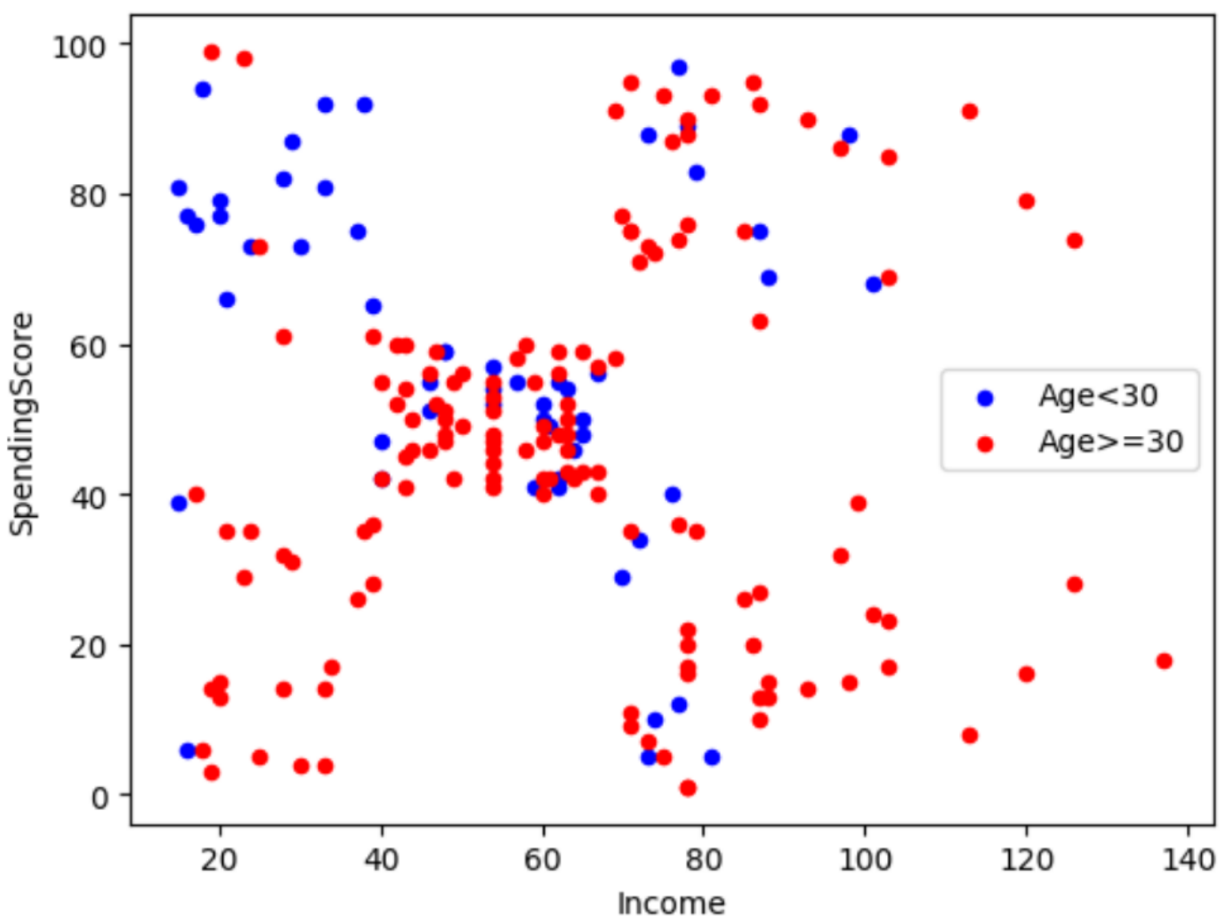


```
Number of clusters: 2 S_Dbw: 1.0643224147555246
Number of clusters: 3 S_Dbw: 0.8496724913507592
Number of clusters: 4 S_Dbw: 0.3788862164259825
Number of clusters: 5 S_Dbw: 0.3592774903104059
Number of clusters: 6 S_Dbw: 0.40883601811613
Number of clusters: 7 S_Dbw: 0.43804885090742407
Number of clusters: 8 S_Dbw: 0.39846676938302394
Number of clusters: 9 S_Dbw: 0.4218640101928597
```

On data with noise



The dataset of Mall\_Customers.csv (Real life data)



## CONCLUSION:

SDBW IS MOST OPTIMAL

WHEN IRIS DATASET (NOISE, SKEWED, WELL SEPERATED) IS APPLIED, SDBW FAILED.

SDBW WORKED WELL FOR ALL THE SYNTHETIC DATA.

REST ALL FAILED FOR ATLEAST ONE OF VARIATION IN DATA.

## WHAT MORE CAN BE DONE?

INVESTIGATE FURTHER ON WHY SDBW FAILED ON THE REAL WORLD DATA SET.

## CONTRIBUTIONS:

JAIN CHIRAG SHANTILAL(2021122004): Validation measures, data generation and Analysis on synthetic data set.

SHUBHAM PRIYADARSHAN(2020102027): Analysis of real life data.

K PAVAN KUMAR(2021122006): Validation measures and analysis of Synthetic data set.

YELLAPRAGADA SRI KRISHNA SARAT CHANDRA(2021122002): Analysed synthetically generated data.