Analysis of performance of State based on Fiscal Parameters using K-means Clustering

[1]Kiran Lavavanshi,[2]Priti Rana,[3] Rutuja Kurhekar
[1,2,3] Department of Computer Applications, Shri Ramdeobaba College of Engineering and Management, Nagpur-440013, Maharashtra, India
[1]lavavanshikm@rknec.edu,[2]ranapc@rknec.edu,[3]kurhekarrs@rknec.edu

***Abstract*—Recently, RBI has published "Handbook of Statistics on Indian States" that includes state wise performance on various parameters like total population, literacy rate, birth rate, pattern of land-use, forest cover, Average inflation (CPI),Gross fiscal deficit and so on. Considering dataset of Average Inflation(CPI)-General and Gross fiscal Deficit for year 2021-22,we analyze state performance and tried to classify states according to their performance into different clusters. Clustering is considered as an unsupervised learning technique based on observations. In this paper, we presented how to obtain the Euclidean distance between two or more clusters, how to determine the new centroid using data mining partitioning approach termed as the K-means algorithm and connection of nodes to obtain the result in KNIME tools.**

I. INTRODUCTION

Data analysts, economists and statisticians in early 90s referred to data mining using the terms "data dredging" or "data fishing" which involves the practical analysis of data without an a-priori hypothesis. Most data mining processes are significant and their objectives are basically to obtain vital information that can be easily understood in large data sets. Data mining involves the searching of large information of data or records to discover patterns and utilize these patterns in predicting events in the future. Data mining provides easy access to certain set of methods and tools that is applicable to data processing for discovering hidden patterns. We could describe it as a collection of pure driven data algorithms to get meaningful patterns from raw data. Data mining application to state's performance is known as Educational Data Mining (EDM); which is an important research area aimed at predicting meaningful information obtained from the educational database to enhance the performances and improve the possibilities for a better growth.

EDM utilizes multiple techniques for clustering and classification such as neural network, k- nearest neighbor, naïve Bayes, and support vector machines, decision techniques based on classifying data sets-means, k-medoids and linear regression analysis. Descriptive and predictive analysis can be implemented by rule association mining; classification; and clustering which are the most common techniques When set of data objects are partitioned into various subsets (clusters), then we consider the descriptive analysis known as clustering.

Clustering is one of the important tasks for data analysis exploration which aims to find data structures which have intrinsic state by modifying the data objects into similar groups and the representation of data in classes, for this reason it is called unsupervised classification or learning performed by observation. The main goal of clustering analysis is to group both similar and different objects in the same clusters and different clusters respectively. In clustering, objects in a cluster are identical to one another yet dissimilar to object in other clusters. The semantic of the classes is not known beforehand in clustering techniques. Some typical applications of clustering technique in data mining are: most educational sectors use this technique to group result of students with average, good, excellent performances in various clusters respectively for ease in analyzing the description in future; In biology, clustering technique is used to obtain the taxonomies of both plants and animals to derive the genes with similar function; The marketing sector, clustering helps to discover distinct classes of customers and develop targeted marketing programs; Clustering is applied in insurance firms to identify groups of motor insurance policy holders with a high average claim cost; It is important for identifying areas with similar land use in an earth observation database; Most meteorological sectors use clustering in earth-quake studies by grouping observed earth quake epicenter along continent faults also to understand the earth climate to find similar and different patterns of atmosphere and ocean; it can also be applied to fraud detection, instruction detection and banking sector utilizes clustering. In this work, the partitioning approach to clustering with the k-means method was selected to be utilized on the state wise dataset obtained from website of RBI. Partitioning approach creates different partitions on data and evaluates these partitions using some criteria, for example minimizing the sum of square errors . The data sets were analyzed by using the KNIME software application as an analytical tool. It is an open source software which makes understanding set of data, designing certain data scientific workflows and reusable components, creating an open continuous integration for new development in data science more accessible to everyone.

II. REQUIREMENTS OF CLUSTERING TECHNIQUES.

Certain requirements need to be considered when performing clustering techniques on data sets.
Some of these requirements are given:

- High dimensionality and Scalability: - involves clustering all data into scalable dimensions instead of only samples.
- Ability to deal with attributes of different types such as numerical, binary, categorical, ordinal, linked and mixture of these previously defined types.
- Usability and Interpretability.
- Discovery of clusters with arbitrary shape.
- Ability to handle noisy data with convenience. Constraint based clustering: ability of user to define inputs on constraints through the use of domain knowledge to determine input parameters.

III. CLUSTERING TECHNIQUES APPROACHES

There are several approaches to clustering techniques but for the purpose of this work we selected the partitioning approach with the k-means clustering algorithms for the analysis state's performance. The four main approaches of basic clustering with their various methods are shown in the figure 1.

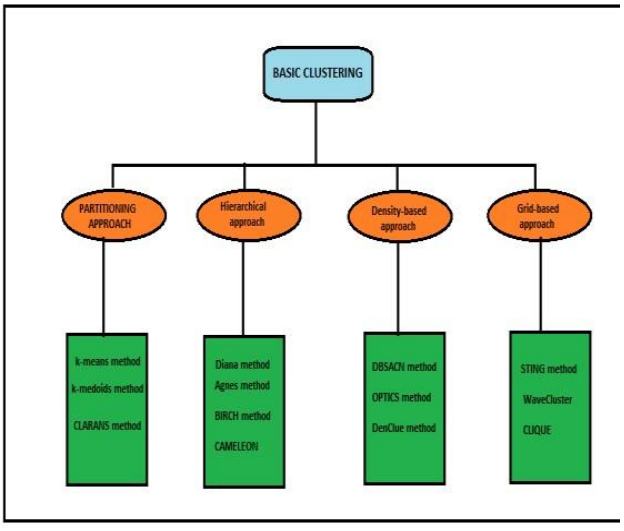


Fig.1. Basic Clustering Approach and their Methods.

The comparative definition of the four main approaches and their respective features is shown the table I:

TABLE I. THE FEATURES OF THE APPROACH TO CLUSTERING TECHNIQUE.

APPROACH.	**FEATURES OF APPROACH TO CLUSTERING TECHINIQUES.**
PARTITIONING	- Find mutually exclusive clusters of spherical shape. - Effective small to medium size data sets. - May use mean or medoid to represent the center of clusters. - Distance – based.
HEIRARCHICAL	- Clustering is a hierarchical decomposition, this means it is multiple levels. - May incorporate other techniques such as micro clustering or considering object linkages.
DENSITY-BASED	- May filter out outliers. -Can find arbitrarily shaped- clusters. - Clusters are dense regions of objects in space that are separated by low-density regions. - Cluster density
GRID-BASED	- Use a multiresolution grid data structure. - Fast processing time.

PARTITIONING CLUSTERING APPROACHES

Partitioning technique helps to improve iteration techniques by mining objects from one graph plot to another. The main objective of partition clustering algorithm is to divide the data points into *K* partitions. Each partition is responsible for reflecting one cluster. Partitioning a database S of n objects into a set k clusters, such that the sum of squared distances are minimized. Given *k*, find a partition of *k* clusters that optimizes the chosen partitioning criterion. *K*-means defines each cluster representation by the center of the cluster and *K*-Medoids defines each cluster representation by one of the objects in the cluster. Mathematically, the error sum of squared Euclidean distances between each observation and its group means using partitioning approach is shown in equation 1:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2 \qquad \text{Equation. 1}$$

where *k*, *C*, (*p* – *mi*) and *mi* represent the number of clusters, the set of objects in a cluster, the distance between *p* and *mi* and the center point of the *i-th* cluster.

K-Means Clustering Algorithm.

K-means is one of the most commonly used in clustering algorithm and future learning. It is method of clustering observations into a specific number of disjoint clusters. The aim of the algorithm is to minimize the measurement between the centroid of the clusters and a given observation by iteratively appending the observation to and clusters when the lowest distance is achieved. K-means performance is determined by initialization and appropriate distance measure. The stages of change of the cluster centers and reassign points are iteratively repeated until, until the border of clusters and location of centroids no longer changes, i.e. at each iteration, every cluster will get the same data point. The secondary goal of K-Means clustering is to reduce the complexities in the data. The illustration of how the algorithm works on a synthetic dataset is in Figure 2:

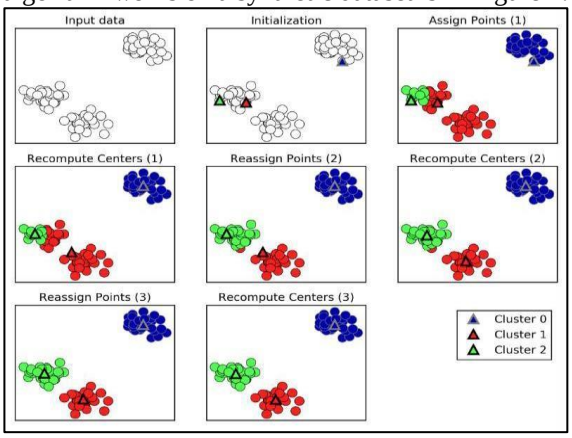


Fig.2. the algorithm working principle on a synthetic set of data.

FLOW CHART OF K-MEANS CLUSTERING.

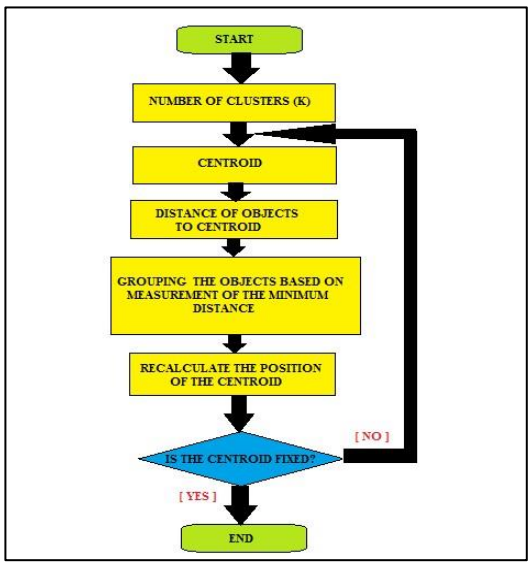


Fig.3. Flow Chart of K-Means Clustering.

Pro and Cons of K-Means Clustering Algorithm.

PRO:

- Ease in the implementation
- Relatively effective and efficient: 0(*tkn*), where the *n* represents the number of data objects; *k* represents the numbers of clusters; *t* represents the numbers of iterations implementation. In the normal state *k* and *t* << *n.*

CONS:

- Applicable in only situations when the mean is defined.
- Not suitable to discover clusters with nonconvex shapes and attributes.
- Highly sensible to noisy data and Outliers.

STEPS FOR K-MEANS ALGORITHM PROCESSING

Randomly choose k objects from database S as the Initial cluster centers; repeat
for each object **do**
Compute distances from the object to cluster centers;
Assign the object to the cluster with the Nearest cluster center, end **for**
each cluster **do**
Calculate the mean value of the objects; end
until no (or minimum) change;

IV. IMPLEMENTATION:

Analysis of state-wise capital expenditure and revenue deficit:

State-wise Fiscal data is our first training data on which k-means clustering has been applied. The dataset consist of State/Union Territory, and their corresponding fiscal deficit, capital expenditure and revenue deficit which is depicted below:

	A	B	C	D	E
1	State/Union Territo	fiscal deficit	capital expenditure	revenue deficit	
2	Andhra Pradesh	37030	47583	5000	
3	Arunachal Pradesh	579	6903	-5747	
4	Assam	15028	23151	-4574	
5	Bihar	22511	41231	-9196	
6	Chhattisgarh	17461	18623	3702	
7	Goa	5875	6915	-59	
8	Gujarat	30783	56568	-1209	
9	Haryana	34004	23003	29194	
10	Himachal Pradesh	7789	9701	1463	
11	Jammu & Kashmir	7200	39502	-28338	
12	Jharkhand	10211	15522	-952	
13	Karnataka	59240	58302	15134	
14	Kerala	30698	27162	16910	
15	Madhya Pradesh	50438	57447	8294	
16	Maharashtra	66641	104829	10225	
17	Manipur	3977	5854	-1550	
18	Meghalaya	1570	3367	-1275	
19	Mizoram	811	1833	-579	
20	Nagaland	1400	3122	-962	
21	Odisha	20465	50433	-6033	
22	Punjab	24240	33022	8622	
23	Rajasthan	47653	42167	23750	
24	Sikkim	1725	2198	-352	
25	Tamil Nadu	100624	67846	58693	
26	Telangana	45510	61343	-6744	
27	Tripura	3680	2651	1717	
28	Uttar Pradesh	90130	144540	-23210	

Fig.4. State-wise Fiscal Data (Training Dataset).

Fiscal Deficit:

The difference between total revenue and total expenditure of the government is termed as fiscal deficit. It is an indication of the total borrowings needed by the government. While calculating the total revenue, borrowings are not included.

The KNIME data analytics tool and Spyder is used for implementation of this work. The dataset with extension **".csv"** was imported into the file reader which reads file and build a workflow which produced the obtained results. The workflow model built for this analysis is shown in figure.5:

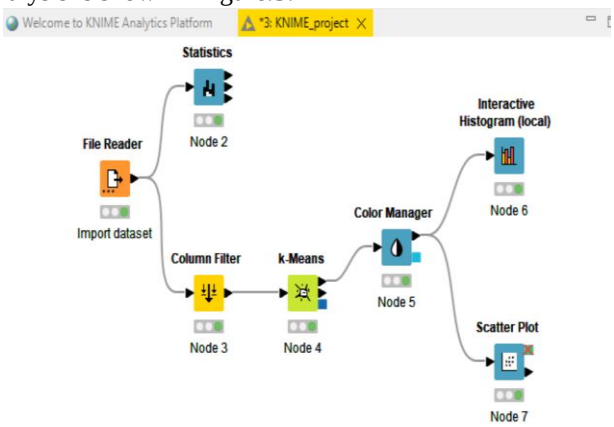
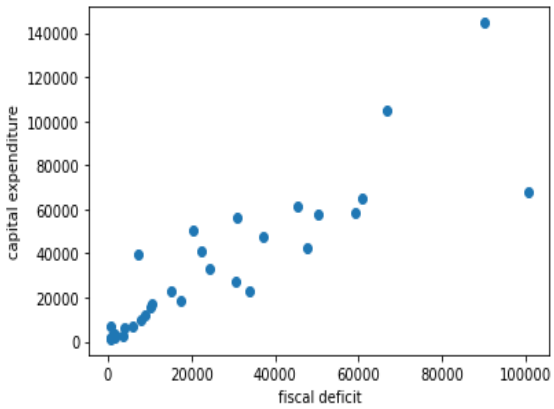


Fig.5. Workflow Model of the K-Means Clustering Analysis.

The workflow consists of the file reader which is the component that read the dataset as the training data sample; then a statistics was generated that gives the statistical description of dataset .A column filter has been applied to focus only on 2021-22 column. Then K-means clustering algorithm has been applied with color Manager node, it creates 3 clusters depicted using 3 colors via an Interactive Histogram and a Scatter plot. Depicted in fig 6,7 and 8.

Initial Scatter plot:



By using Elbow Method, we have tried to calculate the optimize value of k.

Elbow Method:

It is the simplest and commonly used iterative type unsupervised learning algorithm. In this, we randomly initialize the K number of centroids in the data and iterates these centroids until no change happens to the position of the centroid. Let's go through the steps involved in K means clustering for a better understanding.

1) Select the number of clusters for the dataset (K)

2) Select K number of centroids

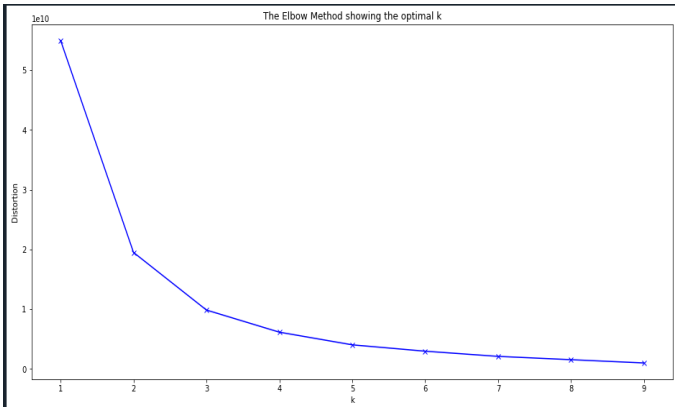
3) By calculating the Euclidean distance or Manhattan distance assign the points to the nearest centroid, thus creating K groups

4) Now find the original centroid in each group

5) Again reassign the whole data point based on this new centroid, then repeat step 4 until the position of the centroid doesn't change.

Finding the optimal number of clusters is an important part of this algorithm. A commonly used method for finding optimal K value is Elbow Method.

For this we have continuously iterate for k=1 to 10 and apply k-Means algorithm. Also plotted a graph for different k values and their corresponding inertia value i.e. distortions.



By observing the plotted graph, after applying Elbow method we have taken optimize value of k =4.

VI. RESULT AND DISCUSSIONS

From the analysis conducted, the representation of the result can categorized into three classes which is shown in the table II:

TABLE II. State Result

S/N	Class	No. of States
1.	Cluster-0	14
2.	Cluster-1	6
3.	Cluster-2	2
4.	Cluster-3	9

The interactive table which identified the number of clusters was generated from the grouped state average inflation rate and it classified the 3 clusters

The interactive table for dataset after processing are shown in figure. 6:

	State/Union Territory	fiscal deficit	...	revenue deficit	cluster_group
0	Andhra Pradesh	37030	...	5000	3
1	Arunachal Pradesh	579	...	-5747	0
2	Assam	15028	...	-4574	3
3	Bihar	22511	...	-9196	3
4	Chhattisgarh	17461	...	3702	0
5	Goa	5875	...	-59	0
6	Gujarat	30783	...	-1209	3
7	Haryana	34004	...	29194	3
8	Himachal Pradesh	7789	...	1463	0
9	Jammu & Kashmir	7200	...	-28338	3
10	Jharkhand	10211	...	-952	0
11	Karnataka	59240	...	15134	1
12	Kerala	30698	...	16910	3
13	Madhya Pradesh	50438	...	8294	1
14	Maharashtra	66641	...	10225	2
15	Manipur	3977	...	-1550	0
16	Meghalaya	1570	...	-1275	0
17	Mizoram	811	...	-579	0
18	Nagaland	1400	...	-962	0
19	Odisha	20465	...	-6033	3
20	Punjab	24240	...	8622	3
21	Rajasthan	47653	...	23750	1
22	Sikkim	1725	...	-352	0
23	Tamil Nadu	100624	...	58693	1
24	Telangana	45510	...	-6744	1
25	Tripura	3680	...	1717	0
26	Uttar Pradesh	90130	...	-23210	2
27	Uttarakhand	8985	...	-115	0
28	West Bengal	60864	...	26755	1
29	NCT Delhi	10665	...	-1271	0
30	Puducherry	800	...	484	0

Fig.6. Interactive table generated by KNIME.

From the interactive table generated, we obtained a histogram for the three classes of clusters (Cluster_0; Cluster_1; Cluster_2, Cluster_3) based on the analysis of average inflation rate. The histogram plot is shown in figure. 7:

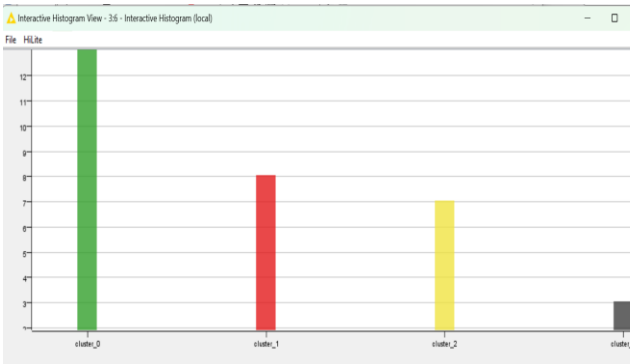
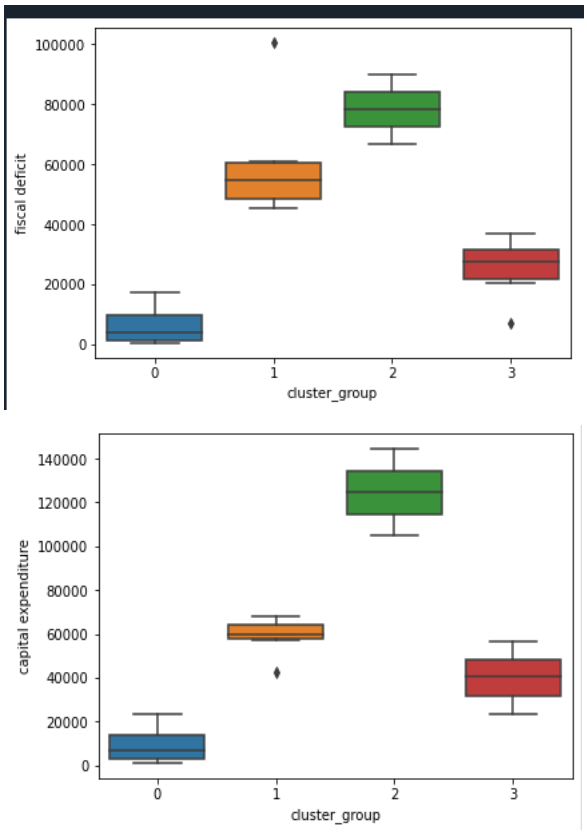


Fig.7. Interactive Histogram for four clusters.

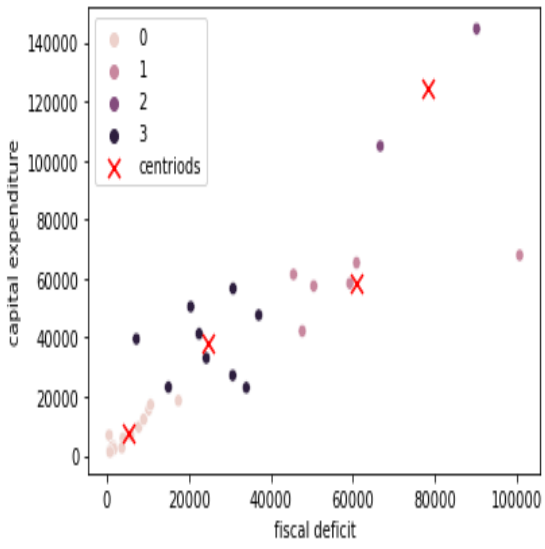
Visualizing Clusters using Box plot:



As we have two dimensions- Fiscal deficit and Capital Expenditure, the possible correct combustion of clusters are

1) High value of Fiscal Deficit and High value of expenditure
2) High value of fiscal deficit and low value of expenditure
3) Low value of Fiscal deficit and high value of expenditure
4) Low value of Fiscal deficit and low value of expenditure

Final Clusters Formation:



CONCLUSION

In this paper, we implemented a qualitative methodology to determine the values of the k-means algorithms from the four clusters obtained. The research work also provided the pro, cons, the algorithm steps and flow chart for the K-Means clustering process. The results of state were analyzed by utilizing the K-means node of the workflow model connection in KNIME tool. The K-means clustering algorithm can serve a better benchmark to observe the performance of state with Fiscal Deficit.

Classification of states on the basis of clusters

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
Arunachal Pradesh	Karnataka	Maharashtra	Andhra Pradesh
Chhattisgarh	Madhya Pradesh	Utter Pradesh	Assam
Goa	Rajasthan		Bihar
Himachal Pradesh	Tamil Nadu		Gujarat
Jharkhand	Telangana		Haryana
Manipur	West Bengal		Jammu and Kashmir
Meghalaya			Kerala
Mizoram			Odisha
Nagaland			Punjab
Sikkim			
Tripura			
Uttarakhand			
NCT Delhi			
Puducherry			

Therefore, Cluster 0 means those states having High value of Fiscal Deficit and High value of expenditure. These states need to work on their Fiscal deficit, which means reduce overall expenditures and increase the total revenue of the states. So, these states need not to urge for loan from central government for further expenditures.

ACKNOWLEDGMENT

The authors appreciate teachers who guided us and provided the necessary tools that were required for the implementation of this research.

REFERENCES

[1] Wikipedia, https://en.m.wikipedia.org/wiki/K-means_clustering

[2] Data Mining-Concepts and Techniques: Jiawei Han, Micheline Kamber Morgan Kaufmann Publishers, Third Edition

[3] Mining of Massive Datasets: Ananad Rajaraman, Jeff Ullman, Jure Leskovec.

[4] Advances In Knowledge Discovery And Data Mining,: Usama M.Fayyad, Gregory Piatetsky-Shapiro, Padhrai Smyth and Ramamsamy Uthurusamy,The M.I.T Press,1996

[5] The DataWarehouse Life Cycle Toolkit: Ralph Kimball, John Wiley & Sons Inc.,1998

[6] HANDBOOK OF STATISTICS ON INDIAN STATES https://m.rbi.org.in//Scripts/AnnualPublications.aspx?head=Handbook%20of%20Statistics%20on%20Indian%20States

[7] Open for KNIME software innovation, [Electronic-Resources] URL: https://www.knime.com/knime-software/