Global Scientific JOURNALS

GSJ: Volume 7, Issue 6, June 2019, Online: ISSN 2320-9186
www.globalscientificjournal.com

DATA MINING USING HIERARCHICAL CLUSTERING TECHNIQUES ON THE POSITION OF EMPLOYEES IN AN INFORMATION TECHNOLOGY FIRM.

Alkadhwi Ali Hussein Oleiwi*[1], Adelaja Oluwaseun Adebayo [1], Ali Alkattan Hussein[2]

Department of System programming,
South Ural State University (National Research University),
76, Lenina Prospect, 454080,
Chelyabinsk, Russia.

Corresponding authors email addresses: alialqady.kirkk@gmail.com, adeleoluwaseun553@gmail.com, hussein_199227@hotmail.com

ABSTRACT

The purpose of this paper is to explain hierarchical clustering, the divisive and agglomerative hierarchical clustering methods. It mainly focus on the concept of the divisive hierarchical processes also known as the top-down approach by generating a workflow model, dendrograms, clustered data table which grouped the clusters based the chosen attribute, and display the distance between each cluster with the aid of a data mining tool called KNIME. The DIANA hierarchical approach used data samples of the list of employees in an Information Technology firm to obtain clusters from the position column in the data sample table. In this work, we also implemented statistical means by generating barchart that shows the ages of the chosen employee sets plotted against the positions which are the Researcher, Programmer and TeamLead.

Keywords: *DIANA, Hierarchical clustering, Divisive, agglomerative, position,Dendrograms,KNIME.*

1. Introduction

Data mining can also be defined as the collection of pure data driven algorithms to obtain meaningful patterns from the raw data which will be helpful in future predictions [1]. In data mining, hidden predictive information are extracted from large databases which makes it is a powerful technology with great potential that helps most companies focus on the most important information in their data warehouses. The prediction of future trends, behaviors, allowing business to create proactive and knowledge-driven decisions can be implemented mostly by data mining tools [2]. The main goal of the clustering techniques used in data mining is to group both identical and distinct objects in the same and different clusters respectively. There are various algorithms utilized in performing clustering, and the criteria of deciding a particular algorithm is dependent mainly on three factors which are the size of the data sets, data dimensionality and the time complexity [3]. Clustering are done mainly by two methods: Hierarchical and Partitioning method, in the scope of data mining hierarchical method groups sets of data objects into a tree of cluster. The hierarchical clustering method which is our main motive in this work can be further classified into divisive and agglomerative hierarchical clustering [4]. Hierarchical approach creates a decomposition of data sets (or objects) in multiple levels of hierarchies using some criterion. Typical methods which functions based on hierarchical approach includes: Diana, Agnes, ROCK, CURE, BIRCH and CAMELEON.

1.1. Hierarchical Clustering Analysis

The divisive and agglomerative clustering does not require the number of clusters ***k*** as an input, but needs a way to compute distance between the clusters. As depicted in the figure 1, the top-down approach from root to leaf is known as the hierarchical divisive clustering (DIANA) while vice-versa is refered to as the bottom-up approach or agglomerative nesting clustering process (AGNES). Dendrograms with different levels and similarity in scale formed as a result of splitting the higher level clusters are typical illustrations of DIANA clustering process. The agglomerative clustering usually starts with a single data point and it merges two or more clusters in a recursive manner. The divisive hierarchical clustering is the reverse of the agglomerative because it starts with a big cluster and then divides this cluster into smaller clusters in a recursive manner [5].

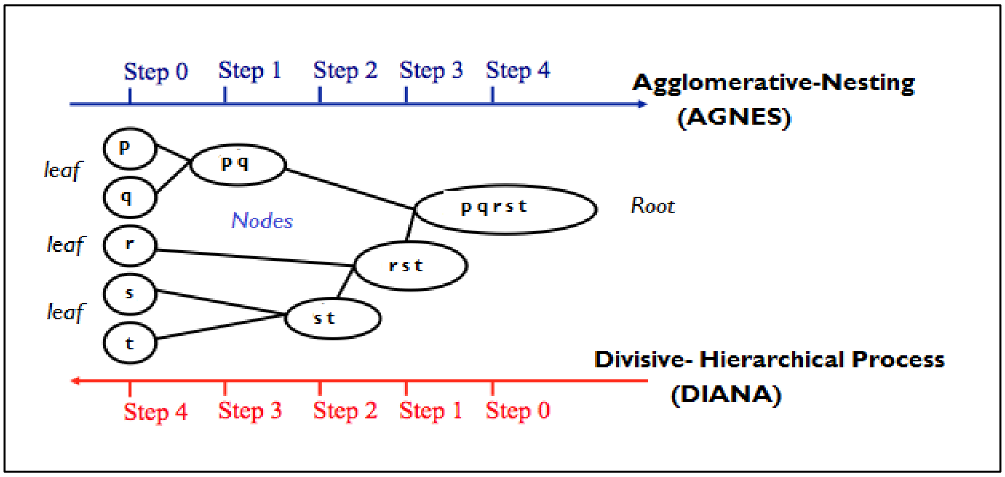


Figure 1: Hierarchical Clustering Techniques

1.2. Dendrogram.

Dendrogram is a diagram that represents the hierarchical relationship between object. From the hierarchical clustering technique in figure 1, the dendrograms are obtained in the following splitting process pattern 1-5 for the levels:

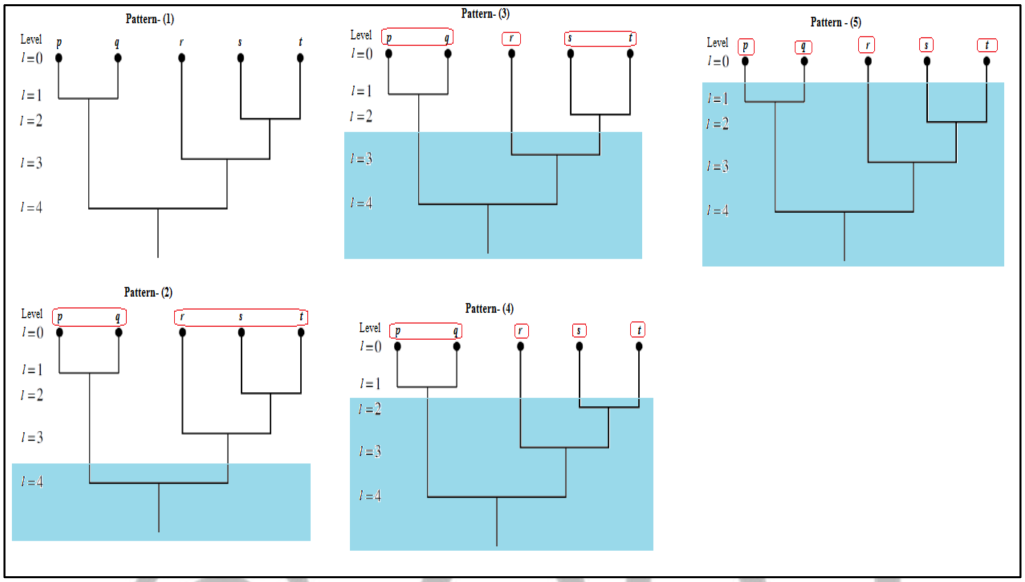


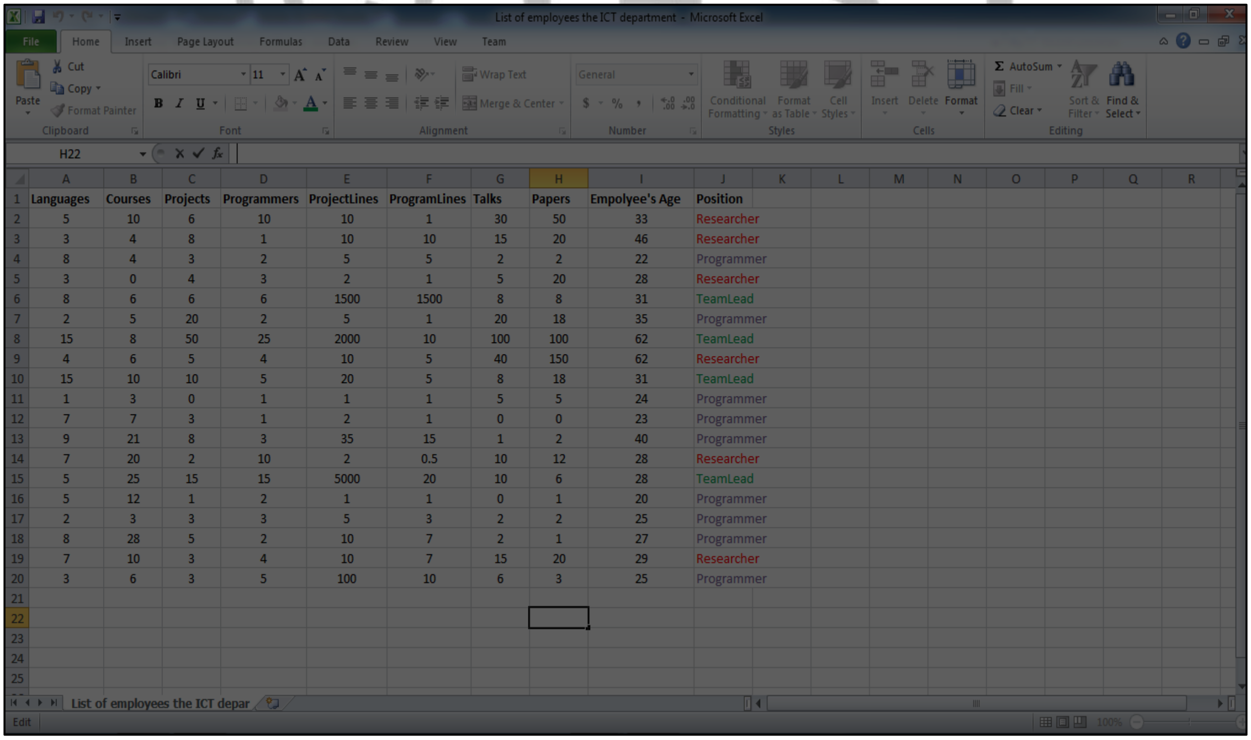
Figure. 2: Dendrogram showing the Hierarchical Splitting Relationship

2.0. Related Work

Odilia Yim and Kylee.T.Ramdeen gives an overview of the hierarchical clustering analysis using the SPSS statistical software to analyse. They also focused on this statistical technique where group are sequentially created by the systematic merging of similar clusters together, as dictated by the linkage measures and the distance. He also comparison these linkage measures (single linkage, complete linkage and average linkage) and applied to psychological data to obtain results [6]. Tian Zhang et al carried out research work that focused on the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) method which is an agglomerative clustering technique and verified that was ideal for large databases. BIRCH was the earliest and first clustering algorithm which was capable of handling noise effectively in a database. It could also generate good cluster with single scan of data and improve the quality. The time/space efficiency, data input order sensitivity and the quality of clusters for BIRCH was evaluated through several experiment conducted by the authors [7]. Vera M.B explained the notion of clustering and used the agglomerative hierarchical algorithm as data mining tool in the capital market to analyse the trade on the Bulgarian Stock exchange with the aim of identifying similar temporal behavior of the traded stock. The author also showed the dendrogram of the clusters of stocks using the average linkage measures [2]. Anna.S et al proposed the hierarchical probabilistic clustering methods used for both supervised and unsupervised learning in data mining application. The probablistics clustering used by these authors was based on the Generalizable Gaussian Mixture Model. The hierarchical scheme proposed by the authors is agglomerative and L_2 distance based metric [8]. Sudipto.G et al presented in their research work, the CURE hierarchical algorithm which identifies clusters with non-spherical shapes and wide variances in size. The CURE combines random sampling and partitioning to handle large databases. CURE represented each cluster by a certain fixed number of points that are generated by selecting well scattered point from the clusters and then shrinking these points towards the cluster's center by specified fraction [9]. Fathi.H.S et al compare different agglomerative algorithms based on the evaluation of clusters quality produced by different hierarchical agglomerative clustering using different criterion functions for the problem of clustering medical documents. The experimental results of their work showed that the agglomerative algorithm that uses *I1* as it criterion function for choosing which clusters to merge produced better clusters quality than the other criterion functions in term of entropy and purity as external measures [10].

3.0. Implementation

The information technology firm consists of ten (10) positions but for the implementation of this experiment we chose to work with only three (3) sets of position for the employees which are the **"Researcher", "Programmer" and "TeamLead"**. The input data named the "List of employess the ICT department" was imported to the KNIME tool for the analysis and also for ease in the building of the workflow diagram. In the experiment conducted, we used 7 nodes repository which includes (3 row splitters each for the the position researcher, programmer and teamlead respectively; 3 hierarchical clustering node for the respective positions also and a single file-reader which directly reads the csv file imported to the KNIME platform for the analysis pupose). Figure 3 shows the **.csv** file in Microsoft excel format.


List of employees the ICT department - Microsoft Excel

H22

	A	B	C	D	E	F	G	H	I	J
1	Languages	Courses	Projects	Programmers	ProjectLines	ProgramLines	Talks	Papers	Empolyee's Age	Position
2	5	10	6	10	10	1	30	50	33	Researcher
3	3	4	8	1	10	10	15	20	46	Researcher
4	8	4	3	2	5	5	2	2	22	Programmer
5	3	0	4	3	2	1	5	20	28	Researcher
6	8	6	6	6	1500	1500	8	8	31	TeamLead
7	2	5	20	2	5	1	20	18	35	Programmer
8	15	8	50	25	2000	10	100	100	62	TeamLead
9	4	6	5	4	10	5	40	150	62	Researcher
10	15	10	10	5	20	5	8	18	31	TeamLead
11	1	3	0	1	1	1	5	5	24	Programmer
12	7	7	3	1	2	1	0	0	23	Programmer
13	9	21	8	3	35	15	1	2	40	Programmer
14	7	20	2	10	2	0.5	10	12	28	Researcher
15	5	25	15	15	5000	20	10	6	28	TeamLead
16	5	12	1	2	1	1	0	1	20	Programmer
17	2	3	3	3	5	3	2	2	25	Programmer
18	8	28	5	2	10	7	2	1	27	Programmer
19	7	10	3	4	10	7	15	20	29	Researcher
20	3	6	3	5	100	10	6	3	25	Programmer

List of employees the ICT depar

Figure.3: The .CSV file of the input data.

4.0. Results.

From the workflow built to obtain the result, the hierarchical clustering node starts with all data points in one huge cluster and the most dissimilar data points are divided into sub-clusters until each cluster conisists of exactly one data point. The work flow we built, using KNIME is shown in figure 4:

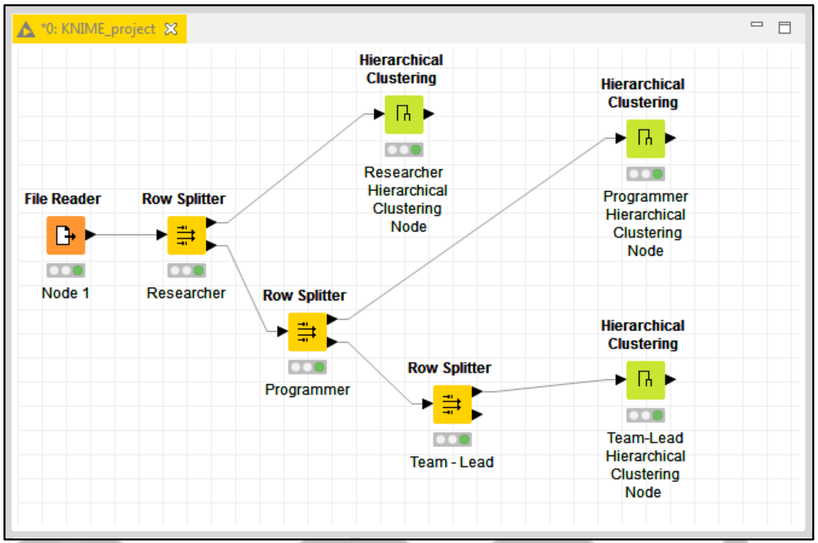
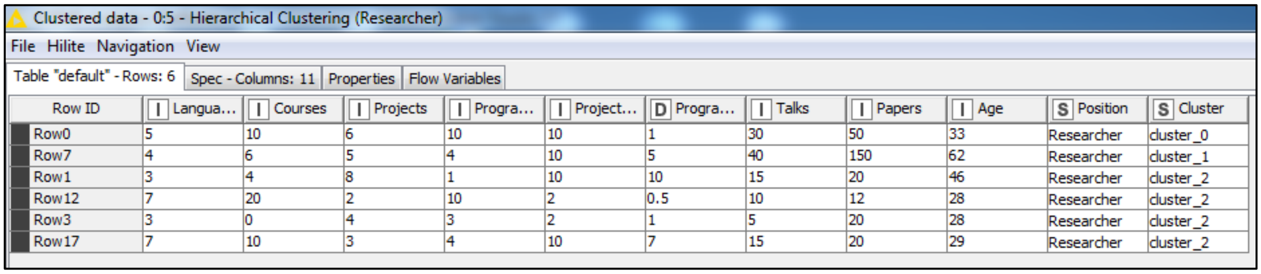


Figure. 4: Hierarchical Clustering WorkFlow Model.

From the result we obtained, the clustered data table for the researcher, the programmer and the Team-Lead positions were generated as shown:



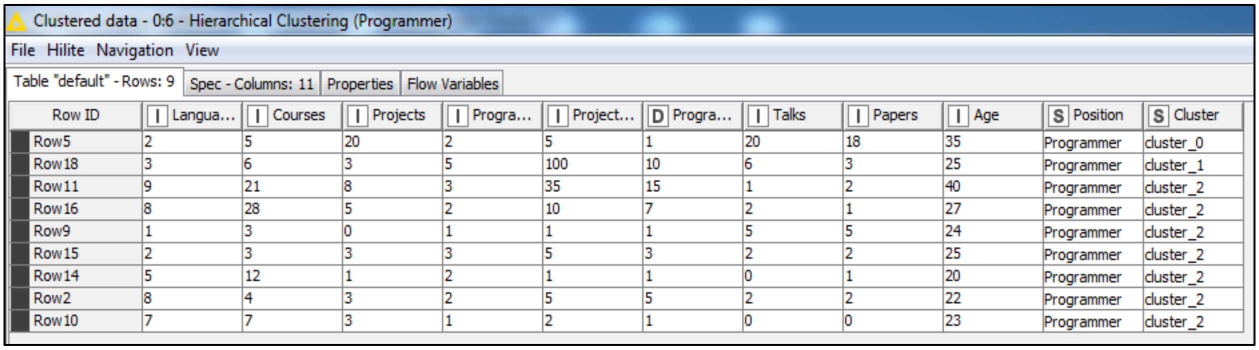
Clustered data - 0:5 - Hierarchical Clustering (Researcher)

File Hilite Navigation View

Table "default" - Rows: 6 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	I Langua…	I Courses	I Projects	I Progra…	I Project…	D Progra…	I Talks	I Papers	I Age	S Position	S Cluster
Row0	5	10	6	10	10	1	30	50	33	Researcher	cluster_0
Row7	4	6	5	4	10	5	40	150	62	Researcher	cluster_1
Row1	3	4	8	1	10	10	15	20	46	Researcher	cluster_2
Row12	7	20	2	10	2	0.5	10	12	28	Researcher	cluster_2
Row3	3	0	4	3	2	1	5	20	28	Researcher	cluster_2
Row17	7	10	3	4	10	7	15	20	29	Researcher	cluster_2

Figure. 5: Clustered data Table for the Researcher.



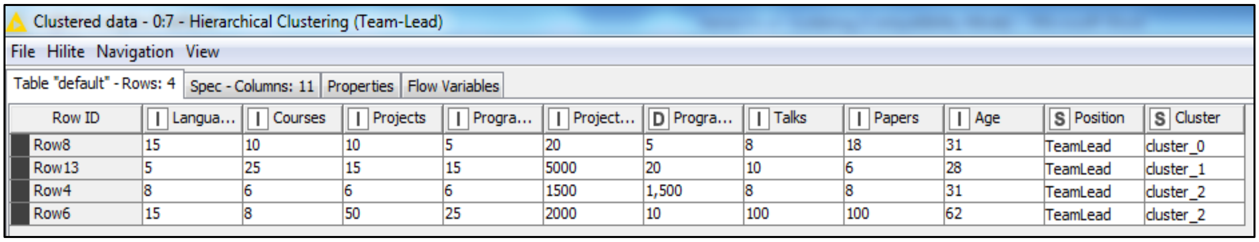
Clustered data - 0:6 - Hierarchical Clustering (Programmer)

File Hilite Navigation View

Table "default" - Rows: 9 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	I Langua…	I Courses	I Projects	I Progra…	I Project…	D Progra…	I Talks	I Papers	I Age	S Position	S Cluster
Row5	2	5	20	2	5	1	20	18	35	Programmer	cluster_0
Row18	3	6	3	5	100	10	6	3	25	Programmer	cluster_1
Row11	9	21	8	3	35	15	1	2	40	Programmer	cluster_2
Row16	8	28	5	2	10	7	2	1	27	Programmer	cluster_2
Row9	1	3	0	1	1	1	5	5	24	Programmer	cluster_2
Row15	2	3	3	3	5	3	2	2	25	Programmer	cluster_2
Row14	5	12	1	2	1	1	0	1	20	Programmer	cluster_2
Row2	8	4	3	2	5	5	2	2	22	Programmer	cluster_2
Row10	7	7	3	1	2	1	0	0	23	Programmer	cluster_2

Figure. 6: Clustered data Table for the Programmer.


Clustered data - 0:7 - Hierarchical Clustering (Team-Lead)
File Hilite Navigation View
Table "default" - Rows: 4
Spec - Columns: 11
Properties
Flow Variables

Row ID	I Langua...	I Courses	I Projects	I Progra...	I Project...	D Progra...	I Talks	I Papers	I Age	S Position	S Cluster
Row8	15	10	10	5	20	5	8	18	31	TeamLead	cluster_0
Row13	5	25	15	15	5000	20	10	6	28	TeamLead	cluster_1
Row4	8	6	6	6	1500	1,500	8	8	31	TeamLead	cluster_2
Row6	15	8	50	25	2000	10	100	100	62	TeamLead	cluster_2

Figure. 7: Clustered data Table for the Team-Lead.

The dendrograms and distances obtained from the clusters splitting for the researcher, programmer and teamlead positions after analyzing the clusters and the row ID for each position are shown:

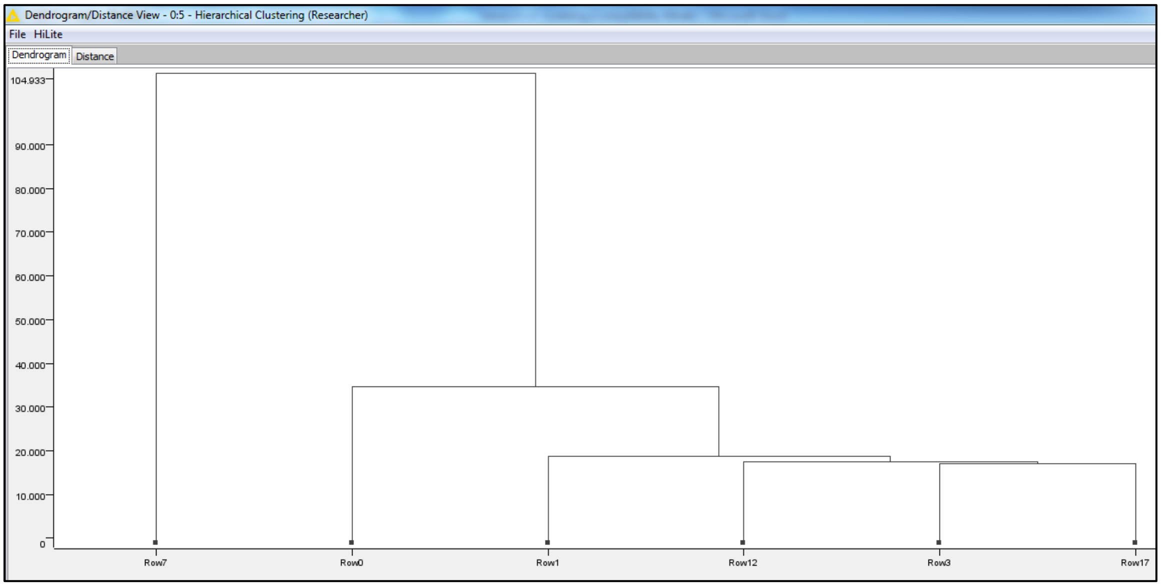


Figure. 8: Dendrogram Of Clusters for Researcher Position.

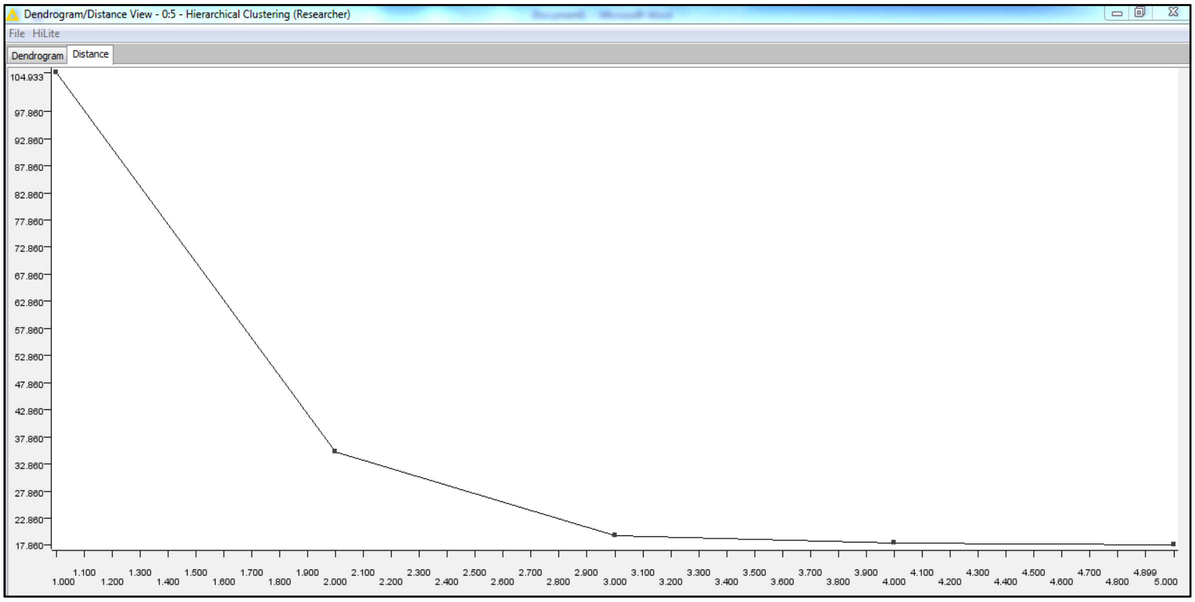


Figure. 8a. Distance between the Clusters (Researcher)

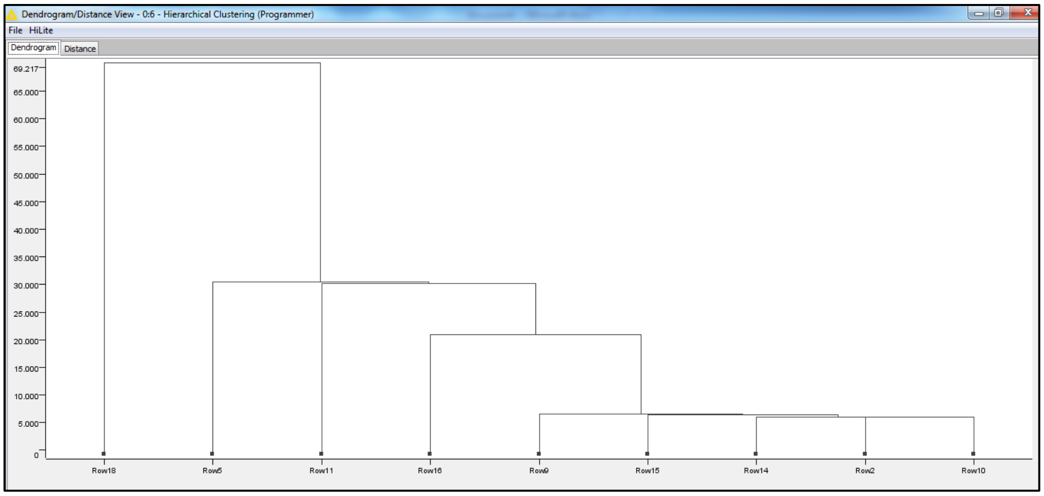


Figure 9: Dendrogram Of Clusters for Programmer Position.

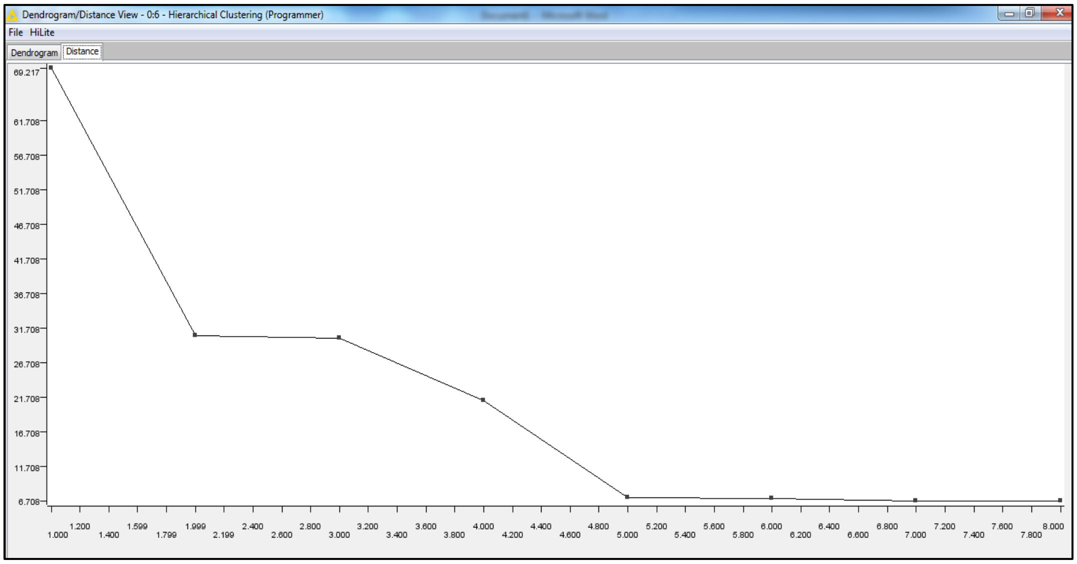


Figure 9a: Distance between the Clusters (Programmer).

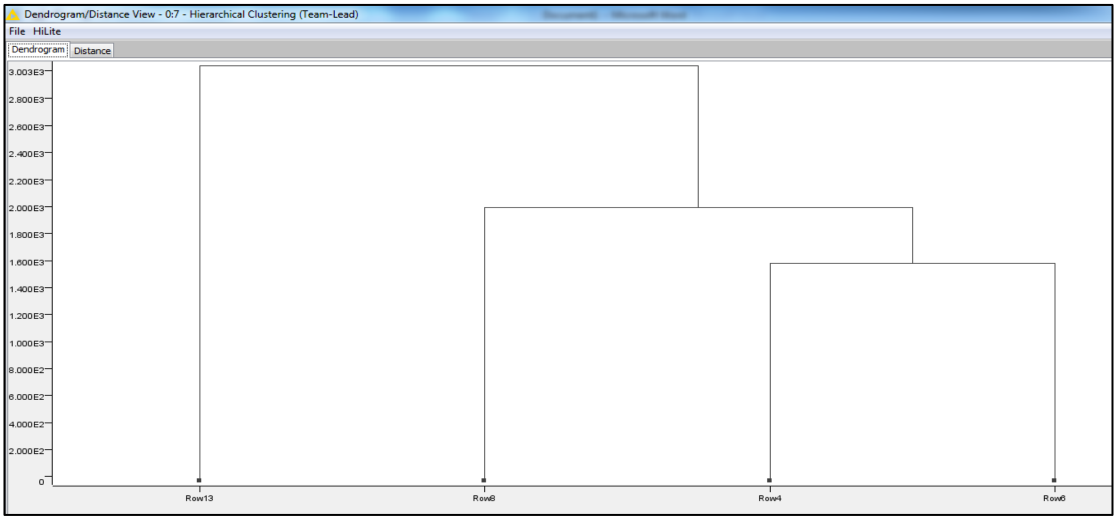


Figure 10: Dendrogram Of Clusters for TeamLead Position.

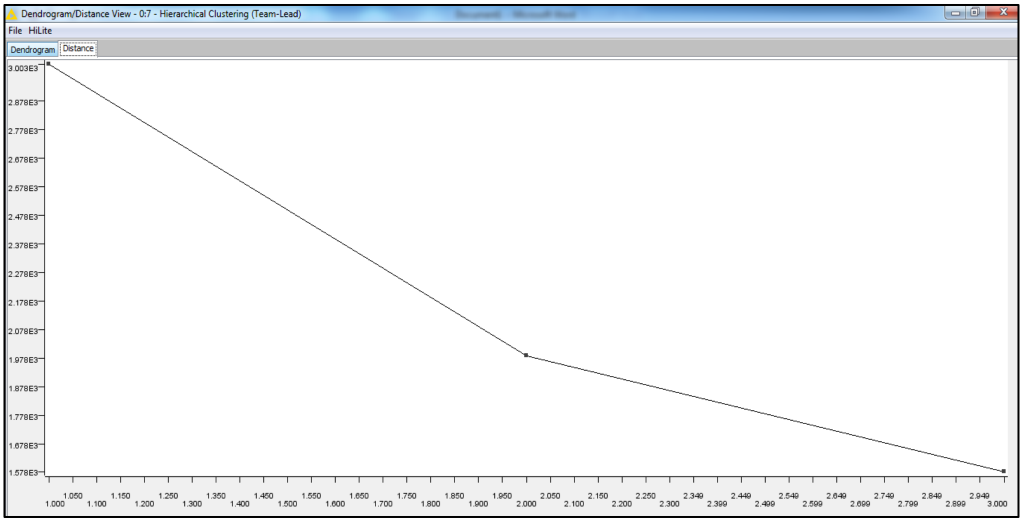
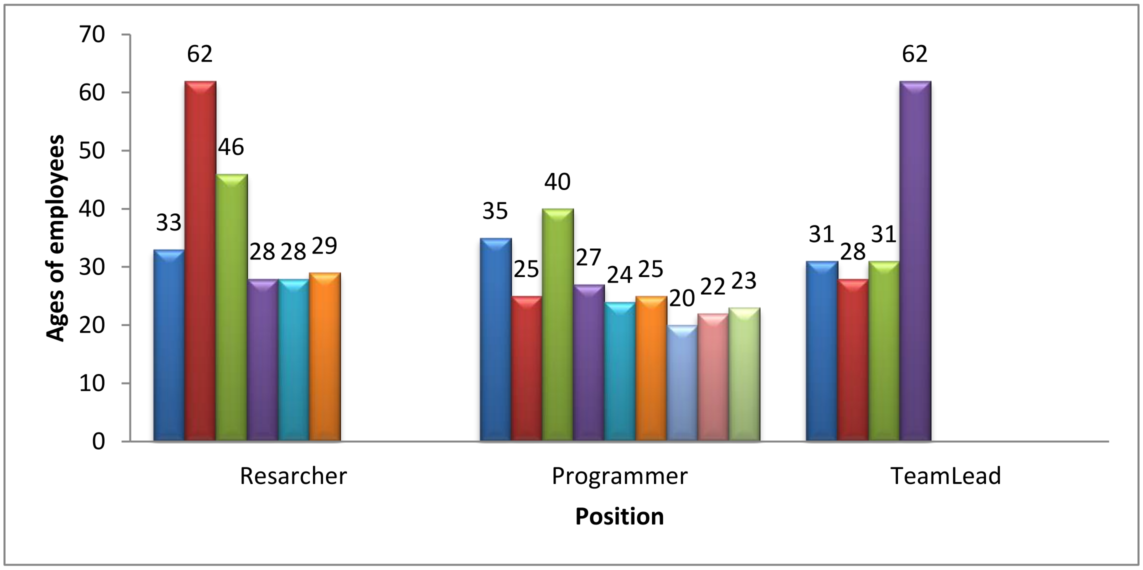


Figure. 10a: Distance between the Clusters (TeamLead).

Statistical means through barchart was used to analyse the ages the employees and the position from the clustered data table was also implemented as shown:



Conclusion

Data mining software allows mostly data analysts and researchers to easily analyze large databases to solve decision problems related to information technology. In this article, it is shown how the divisive hierarchical clustering method will help the information technology firm to regulate and monitor the employee set with the obtained results analysis. The Divise Analysis (DIANA) hierarchical clustering appoarch used is good at identifying large clusters and producing the output in the form of dendrograms.

References:

1. Adelaja O.Adebayo and Mani S. Chaubey, "Data mining classification techniques on the analysis of student's performance". *Global Scientific Journals (GSJ) ISSN 2320-9186*, vol. 7, issue 4, April 2019, pp 79-95.
2. Vera M.B, "Using the agglomerative method of hierarchical clustering as a data mining tool in capital market". *International Journal Information Theories & Applications.* vol. 15, 2008, pp. 382-386.
3. Shuhie A, Parul.P and Seema M, "Hierarchical Clustering- An efficient technique of Data mining for handling voluminous data". *International Journal of Computer Applications (0975 – 8887),* vol 129 – No.13, November2015 pp 31-36.
4. Yogita. R and Dr. Harish.R, "A study of hierarchical clustering algorithm". *International Journal of Information and Computation Technology. ISSN 0974-2239 © International Research Publications House*. Vol. 3, No. 11 (2013), pp. 1225-1232.
5. Chris.D and Xiaofeng. He, "Cluster Merging and Splitting in Hierarchical Clustering Algorithms", 2002.
6. Odilie.Y and Kylee. T.R, "Hierarchical Clustering Analysis: comparison of three linkage measures and application to psychological data", *The Quantitative Methods for Psychology (TQMP)*, vol.11, no 1, 2015, pp 8-21.
7. Tian. Z, Raghu. R, and Miron.L, "BIRCH: an efficient data clustering method for large databases", *International Conference on Management of Data, In Proc. of 1996 ACM-SIGMOD Montreal, Quebec,* 1996, pp 103-114.
8. Szymkowiak.A, Jan.L, Lars.K. H, "Hierarchical Clustering for Datamining", pp. 1-5.
9. Sudipto.G, Rajeev.R and Kyuseok.S, "CURE: An Efficient Clustering Algorithm for large databases", 1998, pp 73-84.
10. Fathi.H.S, Omer.I.E and Rafa.E.A, "Comparison of hierarchical agglomerative algorithms for clustering medical documents", *International Journal of Software Engineering & Applications (IJSEA),* vol.3, No.3, May 2012, pp1-15.