Analysis on Nutrition Facts for Food using Hierarchical Clustering Technique

Myat Thuzar Soe[#1], Laet Laet Lin[#2]
[#1, #2] *Faculty of Computer Science, University of Information Technology*
Yangon, Myanmar

[1]myatthuzarsoe@uit.edu.mm
[2]laetlaetlin@uit.edu.mm

***Abstract* – Clustering is a strategy of unsupervised learning that is a typical procedure for data analysis which is utilized in various fields. It is used for partitioning of information focuses on numerous groups that are from different information objects. A forceful method called hierarchical clustering grants you to manufacture tree structures from information similarities. It is divided into two classes indicated by their clustering measures: Divisive and Agglomerative. In this paper, the hierarchical clustering method is utilized for the actual workflow model, clustered data table, dendrograms, statistic view, and distance between each cluster is displayed by a data mining tool called KNIME. The divisive hierarchical clustering approaches to information tests of the list of foods for acquiring clusters. The single-linkage approach is applied to calculate the distance of two clusters. The result shows calories (kcal) which can be obtained among the nutrition facts from foods for users by using Power BI statistical bar chart.**

Keywords* - *Hierarchical Clustering, Single linkage, Divisive, Dendrograms, Food, KNIME, Power BI

I. INTRODUCTION

In data mining, the objective of the clustering technique is a division of information into groups of comparable objects. A cluster is a group of information objects that each other inside a similar cluster or group but not information objects inside another group or cluster [1]. There are only some algorithms that can perform clustering, the standards of choosing a specific algorithm essentially rely upon three components which are the size of the informational indexes, information dimensionality, and the time complexity [2]. Clustering techniques are primarily separated into two groups: hierarchical and partitioning techniques [3].

The creation of hierarchy includes sorting of information into a lot of groups submitted at various levels by using the hierarchical clustering technique. A binary tree or a dendrogram is used to outline the various clusters. This paper targets the divisive hierarchical clustering technique with a dendrogram [4]. In divisive methodology, information focuses are considered as a single group and are divided into numerous clusters based on specific standards. A portion of the various hierarchical clustering strategies is DIANA, AGNES, LEGCLUT, CHAMELEON, BIRCH, CURE, and ROCK.

II. BACKGROUND THEORY

A. Hierarchical Clustering

A group or information objects in a hierarchical clustering is designed in an order form, i.e. a tree of clusters, and are also called dendrograms [5]. Any ideal number of groups can be obtained by cutting a dendrogram at the best possible level. In hierarchical clustering, pre-determining the quality of groups is not needed, but an approach is needed to register separation between the clusters [6]. There are two types of hierarchical clustering strategies, the divisive (DIANA) and the agglomerative (AGNES). The agglomerative clustering approach is also known as the bottom-up approach, starting with each object that are shaping individual cluster or groups. It is repetitively merged into at least two groups near each other until all groups are merged into single separated groups. The divisive clustering approach is also known as a top-down approach, starting with all the objects are inside one cluster or group. It recursively divides a single cluster into small groups or clusters, until each object is in one cluster. In AGNES and DIANA, the merging/splitting choices are basic, and the cycle of agglomerative and divisive clustering as shown in Fig. 1.

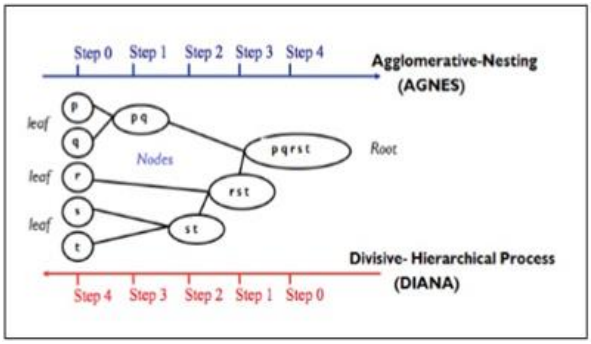


Fig. 1. Hierarchical Clustering Techniques

B. Single Linkage Hierarchical Clustering

Distance between two clusters is characterized as the briefest distance between two focuses in single-linkage hierarchical clustering. For example, the length of the bolt between two nearest focuses is equivalent to the distance between groups "r" and "s" to one side that is generated shown in Fig. 2.

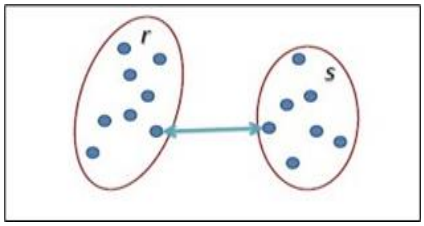


Fig. 2. Single Linkage Hierarchical Clustering

The single-linkage approach is applied to calculate the distance of two clusters r and s which is equal to the minimum of the distance between point xi and xj such that xi belongs to r and xj belongs to s. Mathematically this can be formulated as,

$$L(r,s) = \min(D(xri, xsj)), xi \in r \text{ and } xj \in s \quad (1)$$

where L(r,s) is the distance between clusters r and s, xi is the object in cluster r, xj is an object in cluster s, and then L(xi,xj) denotes the distance between objects belonging to these clusters.

C. Dendrogram

The hierarchical connection between objects is displayed with a dendrogram. Mostly in regular, it is made as an output from hierarchical clustering. The primary usage of a dendrogram is to find the best approach to divide the objects into groups. A dendrogram from below shows the hierarchical clustering steps in Fig. 3

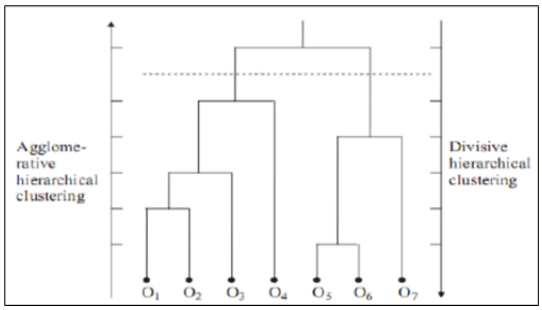


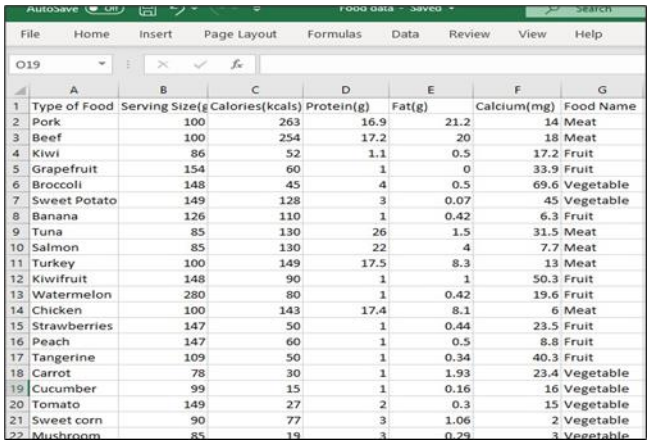
Fig. 3. Dendrogram Showing the Hierarchical Clustering Process

III. RELATED WORK

Studipto Guha et al. [7] suggested another hierarchical clustering algorithm considered as CURE, which is more grounded to exceptions and recognizes clusters having non-circular shapes and wide differences in size. Tian Zhang et al. [8] suggested two strategies, an agglomerative hierarchical clustering strategy called BIRCH (Balanced Iterative Reducing and Clustering utilizing Hierarchical) and ratified that it is appropriated for many information bases. Odilie Yim and Kylee.T.Ramdeen [9] showed a diagram of hierarchical clustering analysis by using the SPSS statistical programming to evaluate. Vera M.B [10] stated the idea of clustering and made the agglomerative hierarchical algorithm as a data mining tool in the capital market to access the exchange on the Bulgarian Stock trade with the point of recognizing comparable temporary conduct of the exchanged stock. The merging and splitting measures in the hierarchical clustering strategy are presented by Chris ding and Xiaofeng [11]. The author made a wide range of clustering research by testing among eight strategies and found that normal linkage is the best technique in divisive clustering and the Min-Max linkage is the best in agglomerative clustering.

IV. IMPLEMENTATION AND EXPERIMENT

There is a huge amount of nutrition facts that can be obtained from foods. A sample data set includes three types of foods. They are Meat, Fruit, and Vegetable. The input data named “List of Food” was imported to the KNIME tool for the analysis and ease in the building of the workflow diagram. The following Fig. 4. shows the .csv file in Microsoft Excel format.



	A	B	C	D	E	F	G
1	Type of Food	Serving Size(g	Calories(kcals)	Protein(g)	Fat(g)	Calcium(mg)	Food Name
2	Pork	100	263	16.9	21.2	14	Meat
3	Beef	100	254	17.2	20	18	Meat
4	Kiwi	86	52	1.1	0.5	17.2	Fruit
5	Grapefruit	154	60	1	0	33.9	Fruit
6	Broccoli	148	45	4	0.5	69.6	Vegetable
7	Sweet Potato	149	128	3	0.07	45	Vegetable
8	Banana	126	110	1	0.42	6.3	Fruit
9	Tuna	85	130	26	1.5	31.5	Meat
10	Salmon	85	130	22	4	7.7	Meat
11	Turkey	100	149	17.5	8.3	13	Meat
12	Kiwifruit	148	90	1	1	50.3	Fruit
13	Watermelon	280	80	1	0.42	19.6	Fruit
14	Chicken	100	143	17.4	8.1	6	Meat
15	Strawberries	147	50	1	0.44	23.5	Fruit
16	Peach	147	60	1	0.5	8.8	Fruit
17	Tangerine	109	50	1	0.34	40.3	Fruit
18	Carrot	78	30	1	1.93	23.4	Vegetable
19	Cucumber	99	15	1	0.16	16	Vegetable
20	Tomato	149	27	2	0.3	15	Vegetable
21	Sweet corn	90	77	3	1.06	2	Vegetable
22	Mushroom	85	19	3	0.29	3	Vegetable

Fig. 4. CSV File of the Food Data

A. Data Collection and Implementation of Workflow Model

As a sample dataset, this study uses food data such as type of food, serving size, calories, protein, fat, calcium are available from “https://www. nutritionvalue.org/” [12]. This dataset contains three categories of food: Meat, Fruit and Vegetable. It incorporates 7 attributes and 300 occurrences. It tends to be saved in Microsoft Excel format. Although numerous different highlights

exist, the examination of divisive hierarchical clustering measure depends on the attributes, for example, Type of Food Name, Serving Size, Calories, Protein, Fat, Calcium, and Food Name. These samples data indexes stacked on the workflow model are represented in Table I.

Table I. Description of Attributes of the Food

Attributes	Types
Type of Food	nominal
Serving Size	numeric
Calories	numeric
Protein	numeric
Fat	numeric
Calcium	numeric
Food Name	nominal

The system flow of this study by using divisive clustering method is shown in Fig. 5.

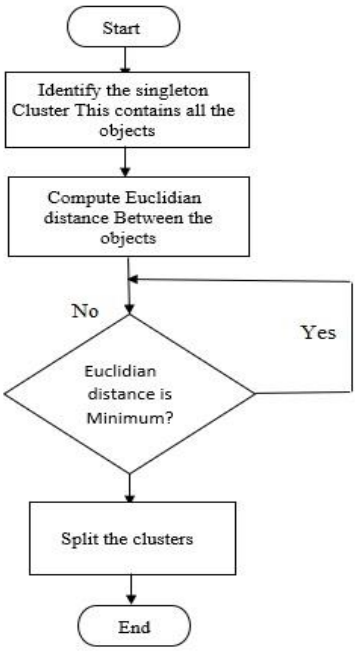


Fig. 5. System Flow of Propose Method

The experiment was conducted by utilizing ten nodes store which incorporates three-row splitter nodes, three hierarchical clustering nodes, three statistics nodes, and these nodes are utilized for the food of Meat, Fruit and Vegetable separately: and one file reader node. In the workflow model, the file reader node which straightforwardly reads the CSV record imported to the KNIME stage for the analysis, the row splitter node eliminates at least one data rows from the input data table as per some filtering standards, the hierarchical clustering node begins with all data focuses in one huge group and the most unique data focuses are partitioned into sub-group until each group comprises of precisely one data point and the statistics node evaluate the statistical values of the clusters. With the KNIME stage, the structure of the workflow model as shown in Fig. 6.

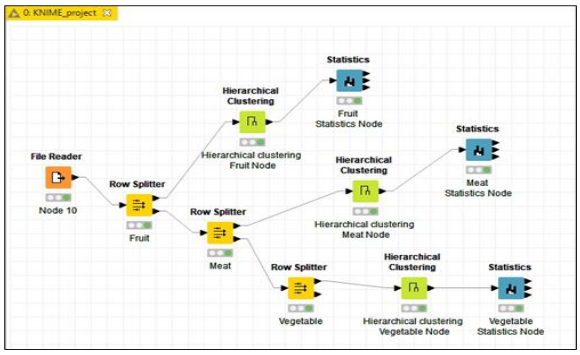


Fig. 6. Hierarchical Clustering Workflow Model

B. Clustered Data Table

The divisive hierarchical clustering begins with all data focuses on one huge cluster and the most unique data focuses are partitioned into sub-clusters until each cluster comprises precisely one data point. The clustered data table for Meat is got from the hierarchical clustering node relying upon the attribute values of Food records. The quantities of two clusters of Meat are gathered in the clustered data tables that are produced as shown in Fig. 7.

Clustered data – 0:6 – Hierarchical Clustering (Meat)

File Edit Hilite Navigation View

Table "default" - Rows: 6 Spec - Columns: 8 Properties Flow Variables

Row ID	Type of...	Serving...	Calorie...	Protein(g)	Fat(g)	Calcium...	Food N...	Cluster
Row0	Pork	100	263	16.9	21.2	14	Meat	cluster_0
Row1	Beef	100	254	17.2	20	18	Meat	cluster_0
Row3	Tuna	85	130	26	1.5	31.5	Meat	cluster_1
Row4	Salmon	85	130	22	4	7.7	Meat	cluster_1
Row2	Chicken	100	143	17.4	8.1	6	Meat	cluster_1
Row5	Turkey	100	149	17.5	8.3	13	Meat	cluster_1

Fig. 7. Clustered Data Table for Meat

The clustered data table for Fruit is acquired from the hierarchical clustering node relying upon the attribute values of Fruit. The quantities of three clusters of Meat are gathered in the clustered data tables that are produced as shown in Fig. 8.

Clustered data – 0:5 – Hierarchical Clustering (Fruit)

File Edit Hilite Navigation View

Table "default" - Rows: 8 Spec - Columns: 8 Properties Flow Variables

Row ID	Type of...	Serving...	Calorie...	Protein(g)	Fat(g)	Calcium...	Food N...	Cluster
Row11	Banana	126	110	1	0.42	6.3	Fruit	cluster_0
Row7	Grapefruit	154	60	1	0	33.9	Fruit	cluster_1
Row12	Peach	147	60	1	0.5	8.8	Fruit	cluster_1
Row6	Kiwi	86	52	1.1	0.5	17.2	Fruit	cluster_1
Row10	Strawberries	147	50	1	0.44	23.5	Fruit	cluster_1
Row13	Tangerine	109	50	1	0.34	40.3	Fruit	cluster_1
Row8	Kiwifruit	148	90	1	1	50.3	Fruit	cluster_2
Row9	Watermelon	280	80	1	0.42	19.6	Fruit	cluster_2

Fig. 8. Clustered Data Table for Fruit

Clustered data – 0:7 – Hierarchical Clustering (Vegetable)

File Edit Hilite Navigation View

Table "default" - Rows: 10 Spec - Columns: 8 Properties Flow Variables

Row ID	Type of...	Serving...	Calorie...	Protein(g)	Fat(g)	Calcium...	Food N...	Cluster
Row17	Sweet Potato	149	128	3	0.07	45	Vegetable	cluster_0
Row19	Sweet corn	90	77	3	1.06	2	Vegetable	cluster_1
Row14	Broccoli	148	45	4	0.5	69.6	Vegetable	cluster_2
Row21	Bean	12	42	3	0.15	13.56	Vegetable	cluster_2
Row20	Mushroom	85	19	3	0.29	3	Vegetable	cluster_2
Row16	Cucumber	99	15	1	0.16	16	Vegetable	cluster_2
Row23	Lettuce	100	17	2	0.3	33	Vegetable	cluster_2
Row15	Carrot	78	30	1	1.93	23.4	Vegetable	cluster_2
Row18	Tomato	149	27	2	0.3	15	Vegetable	cluster_2
Row22	Pumpkin	100	26	1	0.1	21	Vegetable	cluster_2

Fig.9. Clustered Data Table for Vegetable

The clustered data table for Vegetable is acquired from the hierarchical clustering node relying upon the attribute values of Vegetable. The quantities of three groups of Vegetables are gathered in the clustered data tables that are produced as shown in Fig. 9.

The Table II shows the experimental test on the quantity of samples dataset of Food for the hierarchical clustering. The brief of the utilized datasets is introduced in it.

Table II. Description of Datasets of the Food List

Datasets	No. of Data Instances	No. of Clusters
Meat	100	2
Fruit	100	3
Vegetable	100	3

C. Dendrogram of the Clusters

A dendrogram which shows the entire cluster hierarchy. At the lower part of the dendrogram are all data focus and the closet data focuses are associated, where the height of the association shows the separation between them. The dendrogram of Meat shows the connections between comparative arrangements of clustered data present calories. It is gotten from the clusters data table for the Meat after analyzing the clusters with the row ID for Meat as shown in Fig. 10.

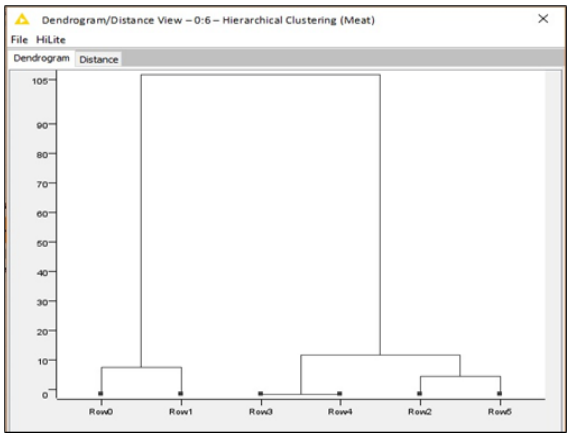


Fig. 10. Dendrogram of Clusters for Meat

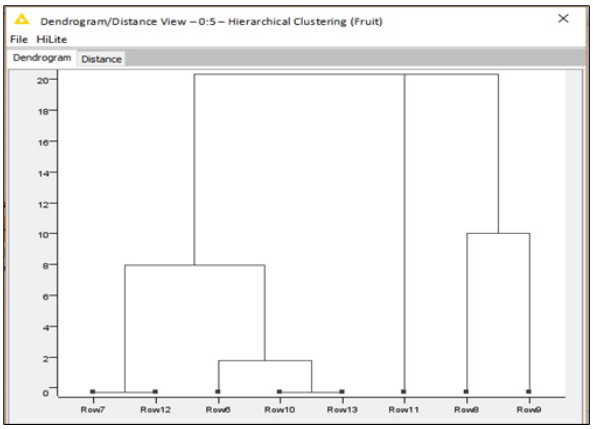


Fig. 11. Dendrogram of Clusters for Fruit

The dendrogram of Fruit shows the connections between similar sets of clustered data present calories. It is gotten from the clusters data table for the Fruit after analyzing the clusters with the row ID for Fruit as shown in Fig. 11.

The dendrogram of Vegetable shows the connections between similar sets of clusters data as indicated by calories. It is gotten from the clusters data table for the vegetable after analyzing the clusters with the row ID for vegetable are shown in Fig. 12

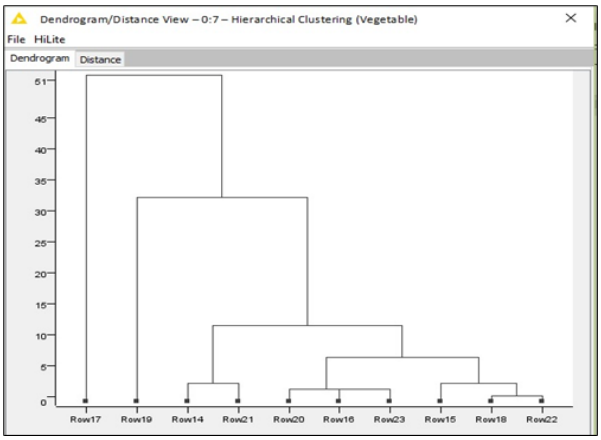


Fig. 12. Dendrogram of Clusters for Vegetable

D. The distance of the Clusters

In the hierarchical clustering problem, single linkage is used for calculating the distance between two clusters is important to merge or divide the clusters. Separations between the groups for each number of clusters are shown by the distance plot. The clusters parting as indicated by the dendrogram of Meat gives the distance plots for Meat. The distance between the clusters of Meat is shown in Fig. 13.

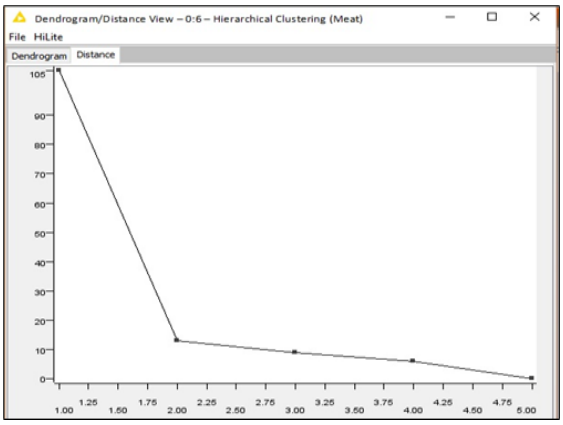


Fig. 13. Distance between the Clusters for Meat

The distance plots for Fruit are acquired from the clusters splitting as indicated by the dendrogram of Fruit. The distance between the clusters of Fruit is shown in Fig. 14.

The distance plots for Vegetable are acquired from the clusters splitting as indicated by the dendrogram of

Vegetable. The distance between the clusters of Vegetable is shown in Fig. 15.

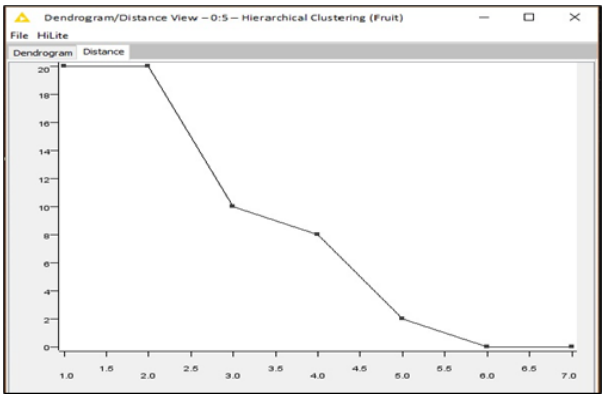


Fig. 14. Distance between the Clusters for Fruit

E. Statistics View

F. The statistics node evaluates statistical values, for example, minimum, maximum, mean, and standard deviation values of the clusters. The statistics shows the minimum calories and maximum calories of the Meat with histogram. The statistics view of the Meat is shown in Fig. 16.

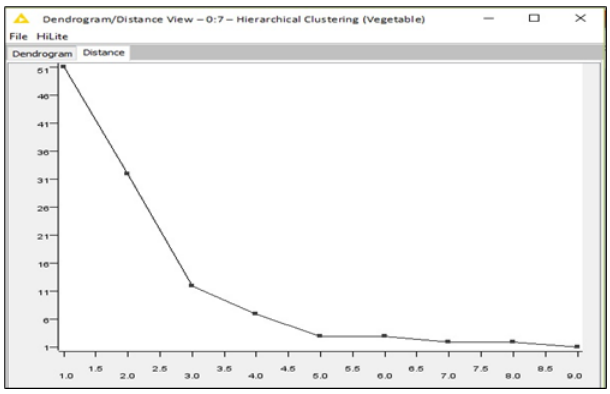


Fig. 15. Distance between the Clusters for Vegetable

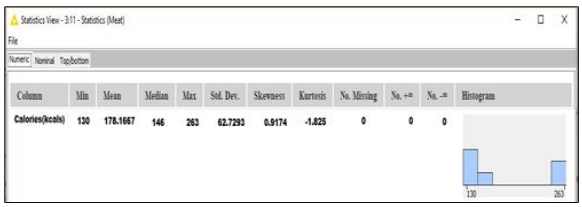


Fig. 16. Statistics View for Meat

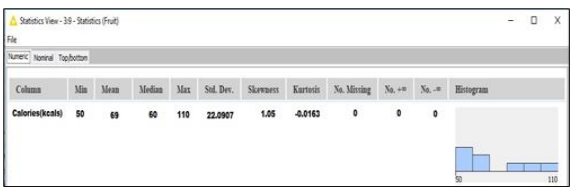


Fig. 17. Statistics View for Fruit

The statistics view shows the minimum calories and maximum calories of the Fruit with histogram. The statistics view of the Fruit is shown in Fig. 17.

The statistics view shows the minimum calories and maximum calories of the Vegetable with histogram. The statistics view of the Vegetable is shown in Fig. 18.

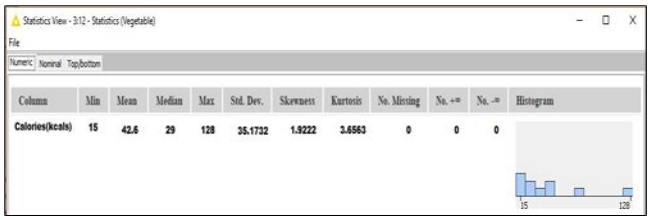


Fig. 18. Statistics View for Vegetable

In Table III, the obtained outcomes are described as far as a minimum, maximum, mean, and standard deviation values of the food as indicated by calories from the statistics view.

Table III. Description of Statistics Data of the Food List

Food	Mini-mum	Maxi-mum	Mean	Std.
Meat	130	263	178.1667	62.7293
Fruit	50	110	69	22.0907
Vegetable	15	128	42.6	35.1732

V. EXPERIMENTAL RESULT

This paper describes the experimental result dependent on the workflow model utilizing a divisive hierarchical clustering (DIANA) approach. A statistical bar chart is utilized to analyze the calories of each food. It shows the various calories of the food from the clustering of the data samples. The statistical bar chart shows that meats can give calories most and fruits in second whilst vegetables are least. The statistical bar chart is represented from the clustered data table as shown in Fig.19.

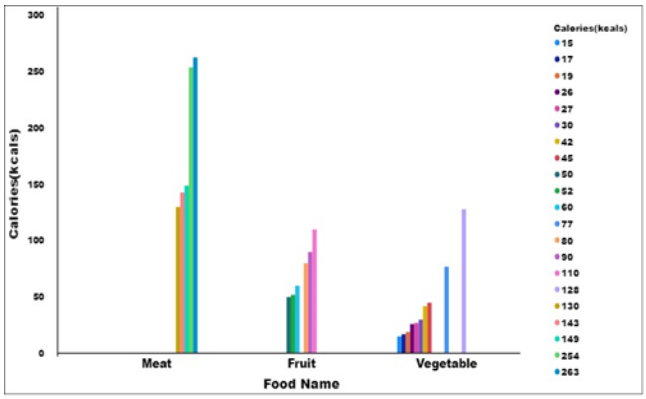


Fig. 19. Statistical Bar Chart

IV. CONCLUSIONS

Although cluster analysis comprises different techniques, using the divisive analysis (DIANA) hierarchical clustering approach is more practicable for distinguishing huge clusters and making the yield as

dendrograms.

In this research, the divisive hierarchical clustering strategy is valuable for users when predicting the number of calories from nutrition facts of foods with the acquired result analysis.

REFERENCES

[1] P. Rai, and S. Singh, "A Survey of Clustering Techniques", *International Journal of Computer Applications (0975-8887)*, vol. 7-No.12, Oct. 2010.

[2] S. Aggarwal, P. Phoghat and S. Maitrey, "Hierarchical Clustering- An Efficient Technique of Data Mining for Handling Voluminous Data", *International Journal of Computer Applications (0975-8887)*, vol. 129- No.13, Nov. 2015, pp. 31-36.

[3] Y. Rani and Dr.H. Rohil, "A Study of Hierarchical Clustering Algorithm", *International Journal of Information and Computation Technology*, ISSN 0974-2239, vol. 3, Number 11(2013), pp. 1225-1232.

[4] P. R. Saraiya and Y. Ganage, "Study of Clustering Techniques in the Data Mining Domain", *International Journal of Computer Science and Mobile Computing*, vol. 7, Issue. 11, Nov. 2018, pg. 31-37.

[5] S. S. Kankal, A. R. Dhakne and Y. R.Tayade, "A Brief Survey on Clustering Algorithm in Data Mining", *International Journal for Scientific Research and Development*, vol.4, Issue 11,2017, ISSN(online):2321-0613.

[6] K. Kameshwaran and K. Malarvizhi, "Survey on Clustering Techniques in Data Mining", *International Journal of Computer Science and Information Technologies*, vol. 5(2), 2014, 2272-2276.

[7] S. Guha, R. Rastogi and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases", *Technical report, Bell Laboratories, Murray Hill*, 1998.

[8] T. Zhang, R. Ramakrishnan and M. Livny, "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.

[9] O. Yim and K. T. .Ramdeen, "Hierarchical Clustering Analysis: Comparison of Three Linkage Measures and Application to Psychological Data", *The Quantitative Methods for Psychology (TQMP)*, vol. 11, No.1, 2014, pp.8-21.

[10] V. M. Boncheva, "Using the Agglomerative Method of Hierarchical Clustering as a Data Mining Tool in a Capital Market*",International Journal Information Theories and Applications*, vol.15, 2008, pp. 382-386.

[11] C. Ding and X. He, "Clustering Merging and Splitting in Hierarchical Clustering Algorithms", IEEE International Conference on, Feb. 2002.

[12] (2020) The nutritionvalue website. [online]. Available: https://www.

nutritionvalue.org/