AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh

Assignment Title: Project: Unsupervised

Assignment No: 1 Date of Submission: 15/05/2020

Course Title: Data Warehousing and Data Mining

Course Code: Section: A

Semester: Spring 20 19 - 20 Course Teacher: Rahman Mohammod Hafizur

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

Group Name/No.:

No	Name	ID	Program	Signature
1	Sarkar, Christine Monisha	16-31255-1	CSE	

Faculty use only		
FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Project Title: Unsupervised Learning

Introduction: In this project using the breakfast cereal dataset (without vitamin and ratings) unsupervised learning was explored. Unsupervised learning focuses on discovering patterns in data from a dataset and mostly handles unlabeled data. In this dataset the names, percentages of protein, sugar, potassium, fat etc. are given of many breakfast cereals and using hierarchical clustering an analysis and observation was made to group these data into few patterns to address nutrition based concerns for individual people as per their needs.

Background:

A brief description of the dataset [1] used in this project is given here. There are few extra columns mentioned here which are not used for this project (vitamin and rating).

The meaning of each column :

1. 1st column : Name of cereal
2. calories: calories per serving
3. protein: grams of protein
4. fat: grams of fat
5. sodium: milligrams of sodium
6. fiber: grams of dietary fiber
7. carbo: grams of complex carbohydrates
8. sugars: grams of sugars
9. potass: milligrams of potassium
10. vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
11. shelf: display shelf (1, 2, or 3, counting from the floor)
12. rating: a rating of the cereals (calculated by Consumer Reports)

Procedure:

- Firstly the dataset was copied from the source and was saved in an individual file with arff extension.
- The instances of attribute “Cereal name” were decided to be the items to be clustered in this project.
- For reducing computational complexity, instances of “cereal name” attribute were converted into “string” data type from “nominal” data type on arff file before running on weka. (Because using these instances as nominal data type was giving arbitrary weighted values instead of the actual names of the instances after the dendrogram was generated).

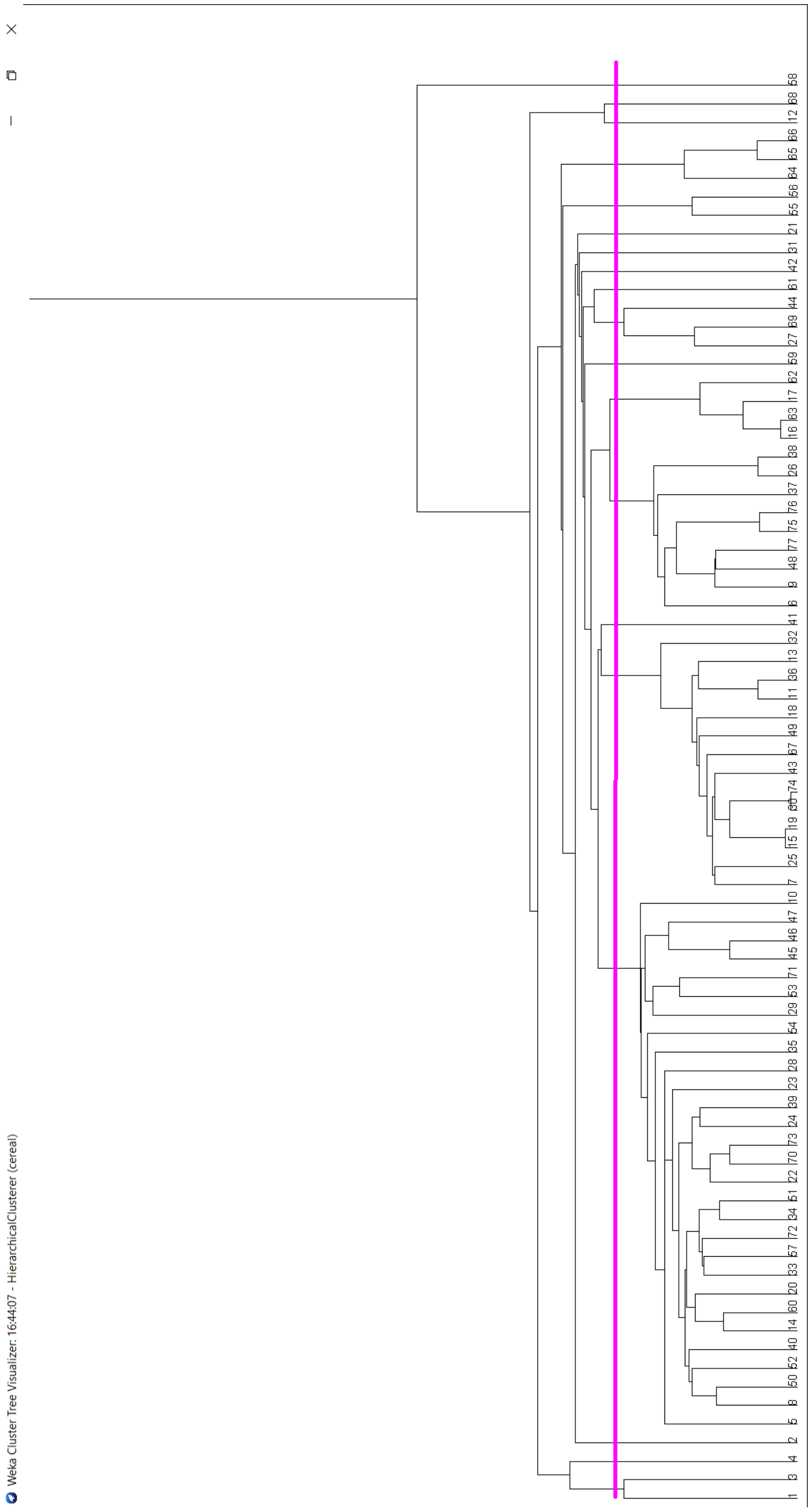
- Also, for a clear view of the final dendrogram, instances of “cereal name” attribute were given a unique ID number each instead of the long names before running on weka on the same arff file. Hence when the final dendrogram was visualized each nodes had a different ID number instead of a long name of cereal.
- There was no missing values in the dataset, hence there was no need to deal with missing values by “discarding instances” or “replace by most frequent/ average value” methods.
- In weka the arff file was opened and in the cluster tab the scheme named “weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"” was chosen.
- The scheme was run and a dendrogram was generated.
- Then the generated tree was cut into a cutting point with a line and cluster analysis was done.
- A pattern for distinguishing similarities between instances of different clusters was then observed and tabulated.

Hierarchical Clustering of the dataset in Weka:

```
=== Run information ===

Scheme:       weka.clusterers.HierarchicalClusterer -N 1 -L SINGLE -P -A
"weka.core.EuclideanDistance -R first-last"
Relation:     cereal
Instances:    77
Attributes:   10
              cereal_name
              calories
              protein(g)
              fat(g)
              sodium(mg)
              dietary_fiber(g)
              complex_carbohydrate(g)
              sugars(g)
              display_shelf
              potassium(mg)
Test mode:    evaluate on training data


=== Clustering model (full training set) ===

Cluster 0
```

```
(((((1:0.44226,3:0.44226):0.1363,4:0.57856):0.08163,(((2:0.56333,((((((((((((5:0.33764,(((((((8:0.
20516,50:0.20516):0.06193,52:0.26708):0.00672,40:0.2738):0.01039,(((14:0.18836,60:0.18836)
:0.0719,20:0.26026):0.02185,(((33:0.2381,57:0.2381):0.00319,72:0.24129):0.00859,(34:0.19672
,51:0.19672):0.05316):0.03223):0.00208):0.01722,((22:0.22189,(70:0.17355,73:0.17355):0.048
35):0.04479,(24:0.24792,39:0.24792):0.01877):0.03473):0.01636,23:0.31778):0.01982,28:0.337
61):0.00004):0.02293,35:0.36057):0.02023,54:0.3808):0.01647,((29:0.3668,(53:0.30072,71:0.30
072):0.06607):0.01968,((45:0.17188,46:0.17188):0.15453,47:0.3264):0.06007):0.0108):0.00187
,10:0.39914):0.10546,((((((((7:0.20954,25:0.20954):0.00491,(((15:0.03021,19:0.03021):0.14149,
(30:0.01562,74:0.01562):0.15608):0.03784,43:0.20954):0.00491):0.01581,67:0.23027):0.01888,
49:0.24915):0.00601,18:0.25516):0.01214,((11:0.0996,36:0.0996):0.15227,13:0.25187):0.01542
):0.07879,32:0.34608):0.15117,41:0.49725):0.00735):0.0185,((((6:0.33799,((9:0.20919,(48:0.20
7,77:0.207):0.00219):0.09722,(75:0.09496,76:0.09496):0.21145):0.03158):0.01512,37:0.35311):
0.01197,(26:0.0996,38:0.0996):0.26547):0.10996,(((16:0.04347,63:0.04347):0.09441,17:0.1378
7):0.11011,62:0.24798):0.22706):0.04806):0.01632,59:0.53943):0.00315,(((27:0.26217,69:0.26
217):0.18008,44:0.44226):0.07239,61:0.51464):0.02793):0.00624,42:0.54881):0.00309,31:0.55
19):0.00512,21:0.55703):0.0063):0.03183,(55:0.26816,56:0.26816):0.32701):0.00491,(64:0.285
91,(65:0.10242,66:0.10242):0.1835):0.31415):0.06013):0.02044,(12:0.49034,68:0.49034):0.190
29):0.28553,58:0.96617)


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      77 (100%)
```

Dendrogram of the cluster tree with cutting point:



Analysis of the clusters:

From the dendrogram a cutting point is taken as such that the total number of clusters here is 19, which is not too much or too little for this dataset. Too many clusters would have caused lengthy analyzing process and too little clusters would have caused faults in precision while taking decisions from the analysis.

The table below shows the 19 clusters and their instances found after cutting the tree and some decisions from matching each of those instances (in respect to each other in a cluster) is given at the right most columns for each of the clusters.

The 19 clusters are color coded in 19 colors to make it easy for detecting them from the Dataset and find out the names and other attribute values for each of the instances.

Cluster No	ID_number	Number of instances	Decision		
			High	Low	Medium
1	1,3	2	Sodium, potassium, shelf	Fat, Carb, calories	Fiber, protein, sugar
2	4	1	Fiber, potassium, shelf	calories, fat(0), sugar(0)	protein, sodium, carb
3	2	1	Fat, Shelf, calories	Sodium, Fiber	protein, carb, sugar, potassium
4	5,8,50,52,40,14,60,20,33,57,72,34,51,22,70, 73,24,39,23,28,35,54,29,53,71,45,46,47,10	29	Self	Fat, Fiber, potassium	Protein, sodium, carb, sugar, calories
5	7,25,15,19,30,74,43,67,49,18,11,36,13,32	14	Sodium, sugar	Protein, fat, fiber, potassium	Carb, calories, self
6	41	1	Sodium, Carb, calories	protein, fat, fiber(0), sugar, potassium	shelf
7	6,9,48,77,75,76,37,26,38	9	Sodium	Protein, fat, fiber, self, potassium	Calories, carb, sugar

8	16,63,17,62	4	Sodium, carb	Protein, fat, fiber, sugar, self, potassium	carb
9	59	1	Sodium, sugar, potassium	fat, fiber	calories, protein, carb, shelf
10	27,69,44	3		fat(0), Sodium(0), fiber, potassium	calories, protein, carb, sugar, shelf
11	61	1	shelf, calories	protein, fat(0), sodium(0), fiber	carb, sugar, potassium
12	42	1	Sodium	Fat, Fiber, potassium	calories, protein, carb, sugar, shelf
13	31	1	sugar, calories	protein, fat(0), sodium, fiber(0), shelf, potassium	carb
14	21	1	carb, calories	fat(0), sodium, fiber, sugar(0), potassium(0)	protein, shelf
15	55,56	2	Fiber	Calories, protein(0), fat(0), sodium (0), carb(0), sugar, potassium	
16	64,65,66	3	Carb	fat(0), sodium(0), fiber, sugar(0), shelf	calories, protein, potassium
17	12	1	Calories, protein, sodium	fat, fiber, sugar, shelf, potassium	carb

18	68	1	Calories, protein, sodium	fat, fiber, sugar, shelf, potassium	carb
19	58	1	protein	fat, sodium(0), fiber, carb(0), sugar(0), shelf, potassium	calories

Dataset:

The dataset is shown below with the ID_number and color coding to determine which instance belong to which cluster.

ID_number	cereal name	calories	protein	fat	sodium	dietary fiber	complex carbohydrates	sugars	display shelf	potassium
1	100%_Bran	70	4	1	130	10	5	6	3	280
2	100%_Natural_Bran	120	3	5	15	2	8	8	3	135
3	All-Bran	70	4	1	260	9	7	5	3	320
4	All-Bran_with_Extra_Fiber	50	4	0	140	14	8	0	3	330
5	Almond_Delight	110	2	2	200	1	14	8	3	-1
6	Apple_Cinnamon_Cheerios	110	2	2	180	1.5	10.5	10	1	70
7	Apple_Jacks	110	2	0	125	1	11	14	2	30
8	Basic_4	130	3	2	210	2	18	8	3	100
9	Bran_Chex	90	2	1	200	4	15	6	1	125
10	Bran_Flakes	90	3	0	210	5	13	5	3	190
11	Cap'n'Crunch	120	1	2	220	0	12	12	2	35
12	Cheerios	110	6	2	290	2	17	1	1	105
13	Cinnamon_Toast_Crunch	120	1	3	210	0	13	9	2	45
14	Clusters	110	3	2	140	2	13	7	3	105
15	Cocoa_Puffs	110	1	1	180	0	12	13	2	55

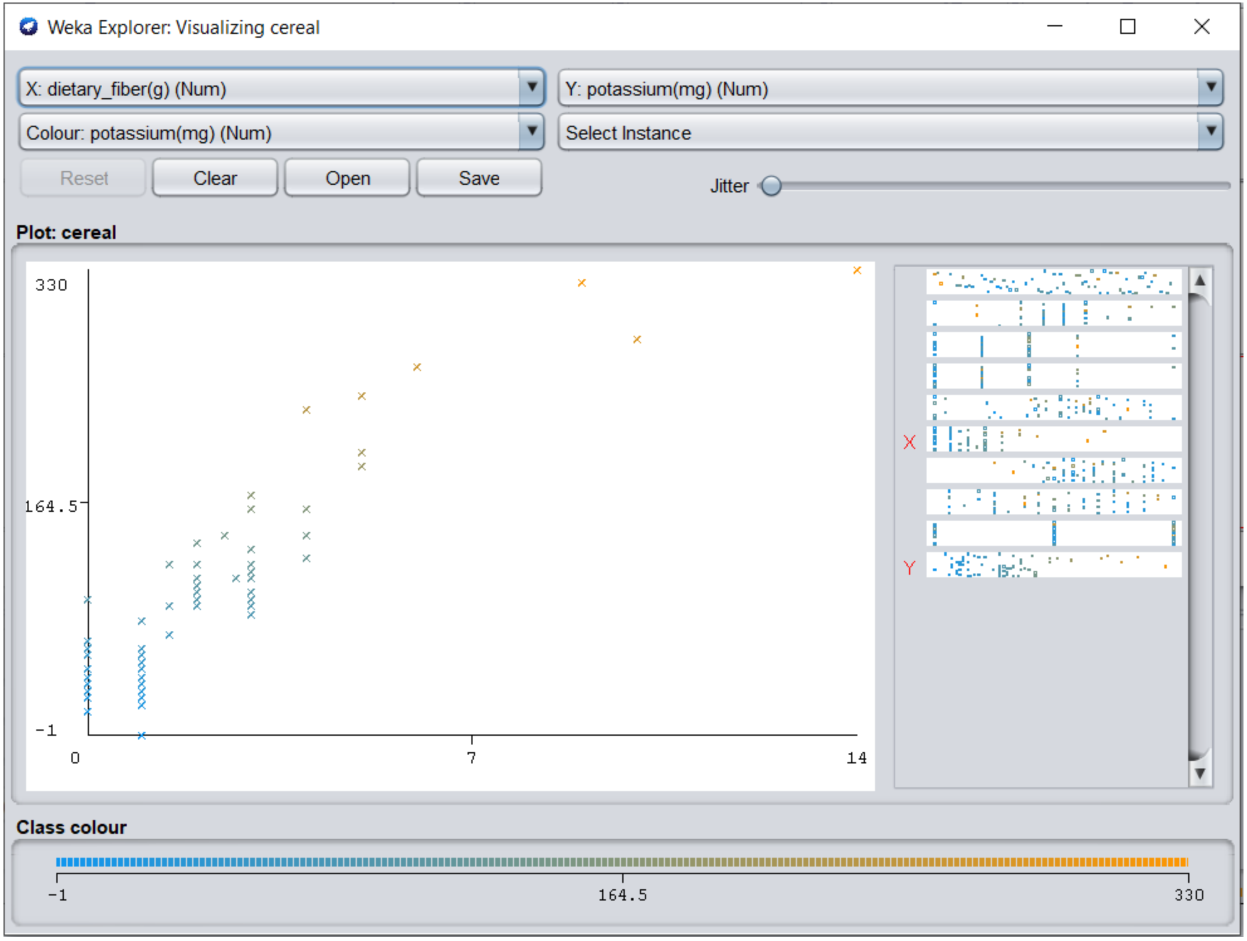
16	Corn_Chex	110	2	0	280	0	22	3	1	25
17	Corn_Flakes	100	2	0	290	1	21	2	1	35
18	Corn_Pops	110	1	0	90	1	13	12	2	20
19	Count_Chocula	110	1	1	180	0	12	13	2	65
20	Cracklin'_Oat_Bran	110	3	3	140	4	10	7	3	160
21	Cream_of_Wheat_(Quick)	100	3	0	80	1	21	0	2	-1
22	Crispix	110	2	0	220	1	21	3	3	30
23	Crispy_Wheat_&_Raisins	100	2	1	140	2	11	10	3	120
24	Double_Chex	100	2	0	190	1	18	5	3	80
25	Froot_Loops	110	2	1	125	1	11	13	2	30
26	Frosted_Flakes	110	1	0	200	1	14	11	1	25
27	Frosted_Mini-Wheats	100	3	0	0	3	14	7	2	100
28	Fruit_&_Fibre_Dates,_Walnuts,_and_Oats	120	3	2	160	5	12	10	3	200
29	Fruitful_Bran	120	3	0	240	5	14	12	3	190
30	Fruity_Pebbles	110	1	1	135	0	13	12	2	25
31	Golden_Crisp	100	2	0	45	0	11	15	1	40
32	Golden_Grahams	110	1	1	280	0	15	9	2	45
33	Grape_Nuts_Flakes	100	3	1	140	3	15	5	3	85
34	Grape-Nuts	110	3	0	170	3	17	3	3	90
35	Great_Grains_Pecan	120	3	3	75	3	13	4	3	100
36	Honey_Graham_Ohs	120	1	2	220	1	12	11	2	45
37	Honey_Nut_Cheerios	110	3	1	250	1.5	11.5	10	1	90
38	Honey-comb	110	1	0	180	0	14	11	1	35
39	Just_Right_Crunchy__Nuggets	110	2	1	170	1	17	6	3	60
40	Just_Right_Fruit_&_Nut	140	3	1	170	2	20	9	3	95
41	Kix	110	2	1	260	0	21	3	2	40
42	Life	100	4	2	150	2	12	6	2	95
43	Lucky_Charms	110	2	1	180	0	12	12	2	55
44	Maypo	100	4	1	0	0	16	3	2	95
45	Muesli_Raisins,_Dates,_&_Almonds	150	4	3	95	3	16	11	3	170
46	Muesli_Raisins,_Peaches,_&_Pecans	150	4	3	150	3	16	11	3	170
47	Mueslix_Crispy_Blend	160	3	2	150	3	17	13	3	160
48	Multi-Grain_Cheerios	100	2	1	220	2	15	6	1	90
49	Nut&Honey_Crunch	120	2	1	190	0	15	9	2	40
50	Nutri-Grain_Almond-Raisin	140	3	2	220	3	21	7	3	130
51	Nutri-grain_Wheat	90	3	0	170	3	18	2	3	90

52	Oatmeal_Raisin_Crisp	130	3	2	170	1.5	13.5	10	3	120
53	Post_Nat._Raisin_Bran	120	3	1	200	6	11	14	3	260
54	Product_19	100	3	0	320	1	20	3	3	45
55	Puffed_Rice	50	1	0	0	0	13	0	3	15
56	Puffed_Wheat	50	2	0	0	1	10	0	3	50
57	Quaker_Oat_Squares	100	4	1	135	2	14	6	3	110
58	Quaker_Oatmeal	100	5	2	0	2.7	-1	-1	1	110
59	Raisin_Bran	120	3	1	210	5	14	12	2	240
60	Raisin_Nut_Bran	100	3	2	140	2.5	10.5	8	3	140
61	Raisin_Squares	90	2	0	0	2	15	6	3	110
62	Rice_Chex	110	1	0	240	0	23	2	1	30
63	Rice_Krispies	110	2	0	290	0	22	3	1	35
64	Shredded_Wheat	80	2	0	0	3	16	0	1	95
65	Shredded_Wheat_'n'Bran	90	3	0	0	4	19	0	1	140
66	Shredded_Wheat_spoon_size	90	3	0	0	3	20	0	1	120
67	Smacks	110	2	1	70	1	9	15	2	40
68	Special_K	110	6	0	230	1	16	3	1	55
69	Strawberry_Fruit_Wheats	90	2	0	15	3	15	5	2	90
70	Total_Corn_Flakes	110	2	1	200	0	21	3	3	35
71	Total_Raisin_Bran	140	3	1	190	4	15	14	3	230
72	Total_Whole_Grain	100	3	1	200	3	16	3	3	110
73	Triples	110	2	1	250	0	21	3	3	60
74	Trix	110	1	1	140	0	13	12	2	25
75	Wheat_Chex	100	3	1	230	3	17	3	1	115
76	Wheaties	100	3	1	200	3	17	3	1	110
77	Wheaties_Honey_Gold	110	2	1	200	1	16	8	1	60

Q/A:

1. Is there a strong correlation between dietary fiber and potassium?

Ans: In the graph below, a strong correlation between dietary fiber and potassium can be seen. X axis represents dietary fiber and Y axis represents potassium. From this graph generated by weka it can be seen that when the value of one axis increases or decreases the other one also increases or decreases similarly.

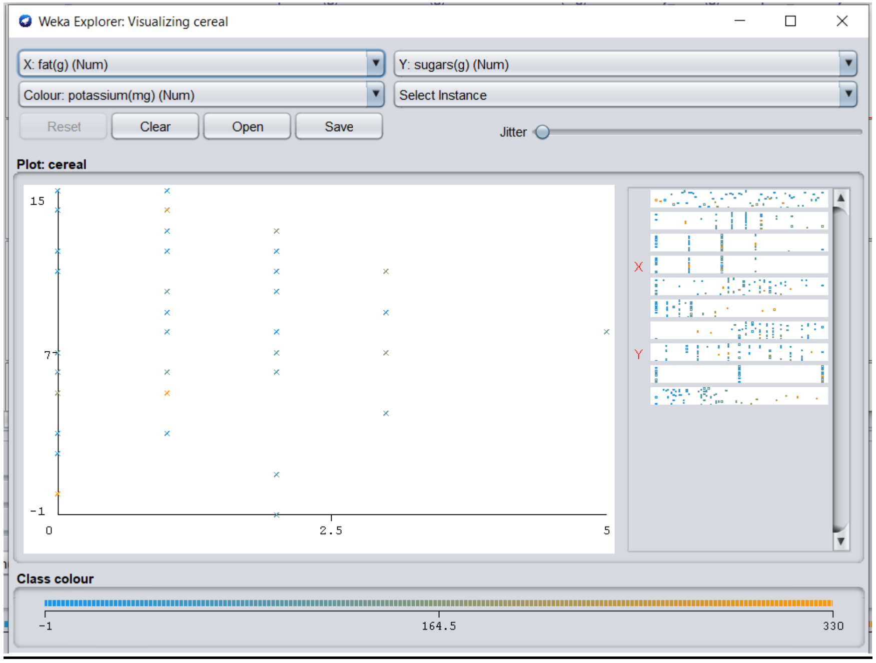


2. Are groups of cereals from which we can choose according to our preferences?
With some common health concerns and their required food habit a table is made to easily analyze the dataset to prescribe food for each group of people.

Health Concerns	**Prescription**		**Suggested Clusters**
	Take	**Avoid**	
High pressure	------	Sodium, **High** fat	10,11,15,16,19
Low pressure	**High** Sodium, fiber and protein	----	1,2,5,6,7,12,17,18
Obesity	**Low** calories	Sugar, fat	2,8,10,14, 15,16
Diabetics	-----	Sugar	2,14,16
child	**High** Sugar and **Moderate** sodium, protein	--------	5,9,13,4,12
Pregnant woman	**High** fiber, protein, calories	-------	2,3,14,17,18,19
Constipation	**High** Fiber, **Less** carbohydrate, sugar	-------	1,2,15
Diarrhea	-----	**High** Fiber, Protein	5,6,7,8,11,12,13

3. See other correlation between the data given in the files.

Fat and Sugar: From the graph, it can be seen that when sugar (y axis) increases (or decreases) the value of Fat(x axis) remains the same. So here the correlation is that fat remains the same for different values of sugar and vice versa.



Weka Explorer: Visualizing cereal

X: fat(g) (Num)

Y: sugars(g) (Num)

Colour: potassium(mg) (Num)

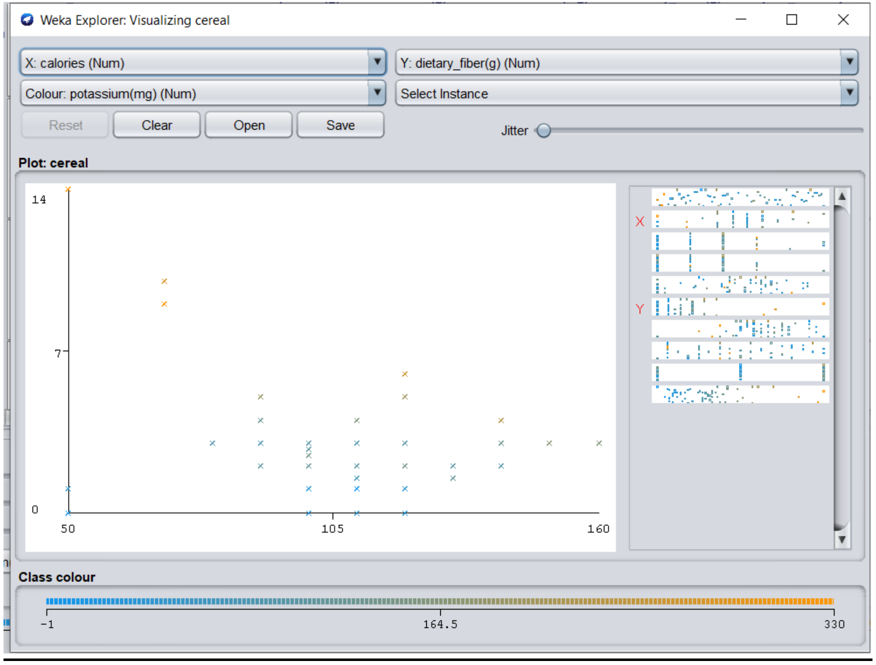
Select Instance

Jitter

Plot: cereal

Class colour

Calories and Dietary Fiber: In this graph the values for dietary fiber and calories are totally mismatched. Here it can be seen that for a same fiber value the value of calories are different multiple times. So no pattern can be found from here hence no correlation can be established.



Weka Explorer: Visualizing cereal

X: calories (Num)

Y: dietary_fiber(g) (Num)

Colour: potassium(mg) (Num)

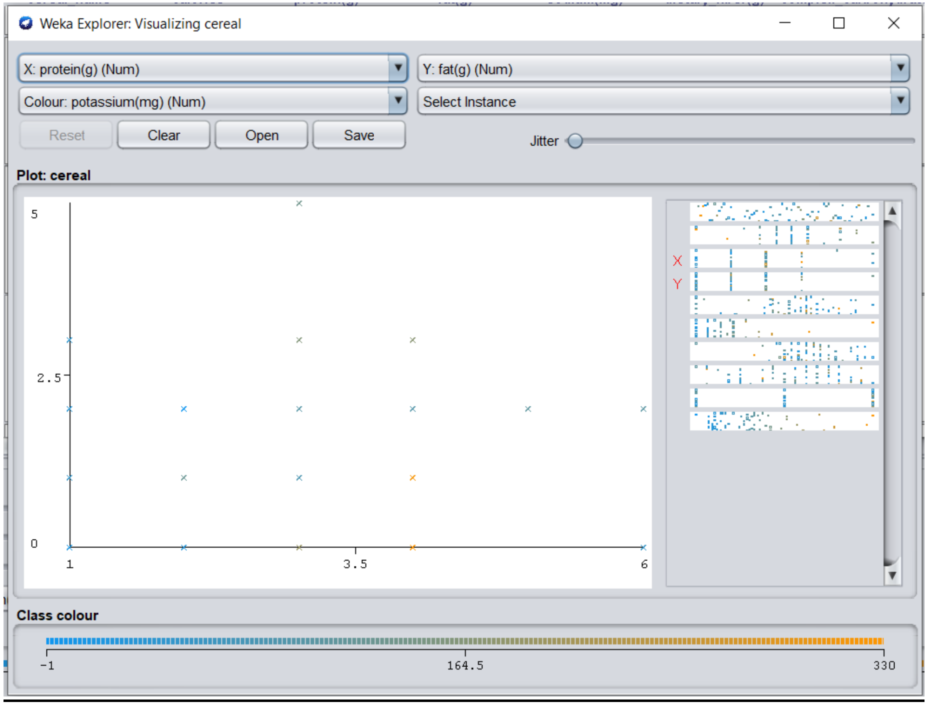
Select Instance

Jitter

Plot: cereal

Class colour

<u>Fat and Protein:</u> Here, when fat increases or decreases the value of protein remains the same and vice versa. So there is a correlation between fat and protein.



Weka Explorer: Visualizing cereal

X: protein(g) (Num)

Y: fat(g) (Num)

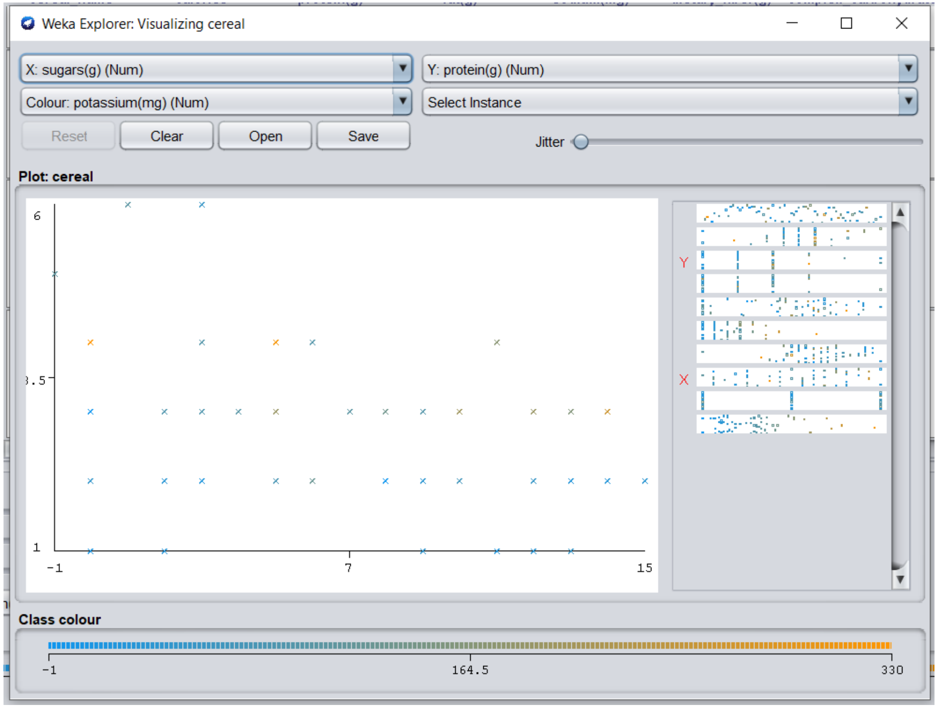
Colour: potassium(mg) (Num)

Select Instance

Plot: cereal

Class colour

<u>Protein and Sugar:</u> Here, a positive correlation between protein and sugar is observed. If sugar increases at certain point the protein remains same.



Weka Explorer: Visualizing cereal

X: sugars(g) (Num)

Y: protein(g) (Num)

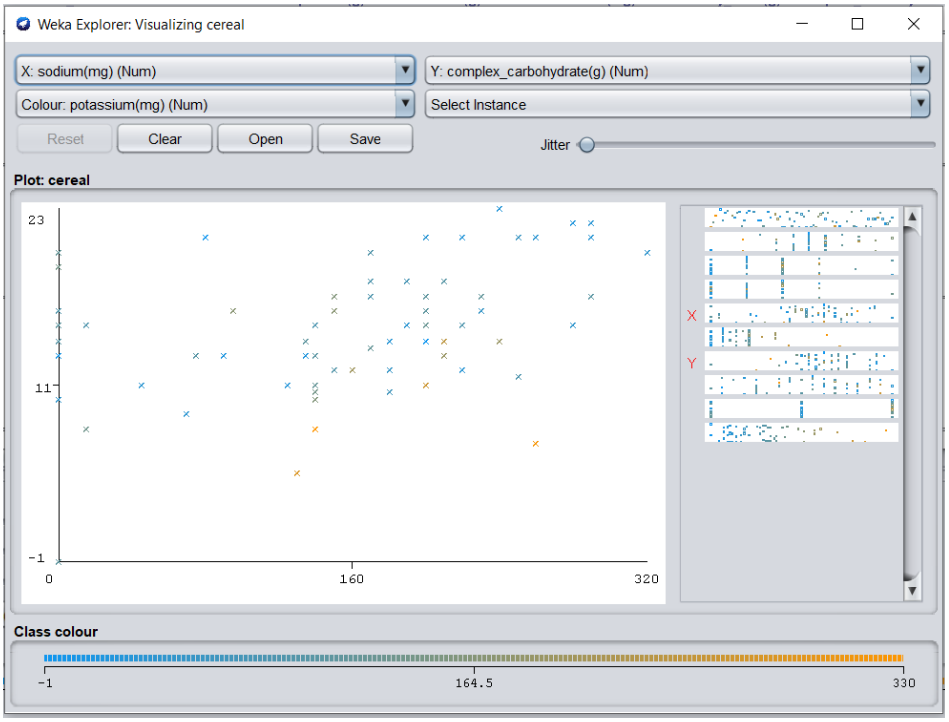
Colour: potassium(mg) (Num)

Select Instance

Plot: cereal

Class colour

<u>Sodium and Complex Carbohydrate:</u> The values here are scattered and are increasing/ decreasing at random. So no relatable relation can be established.



Weka Explorer: Visualizing cereal

X: sodium(mg) (Num)

Y: complex_carbohydrate(g) (Num)

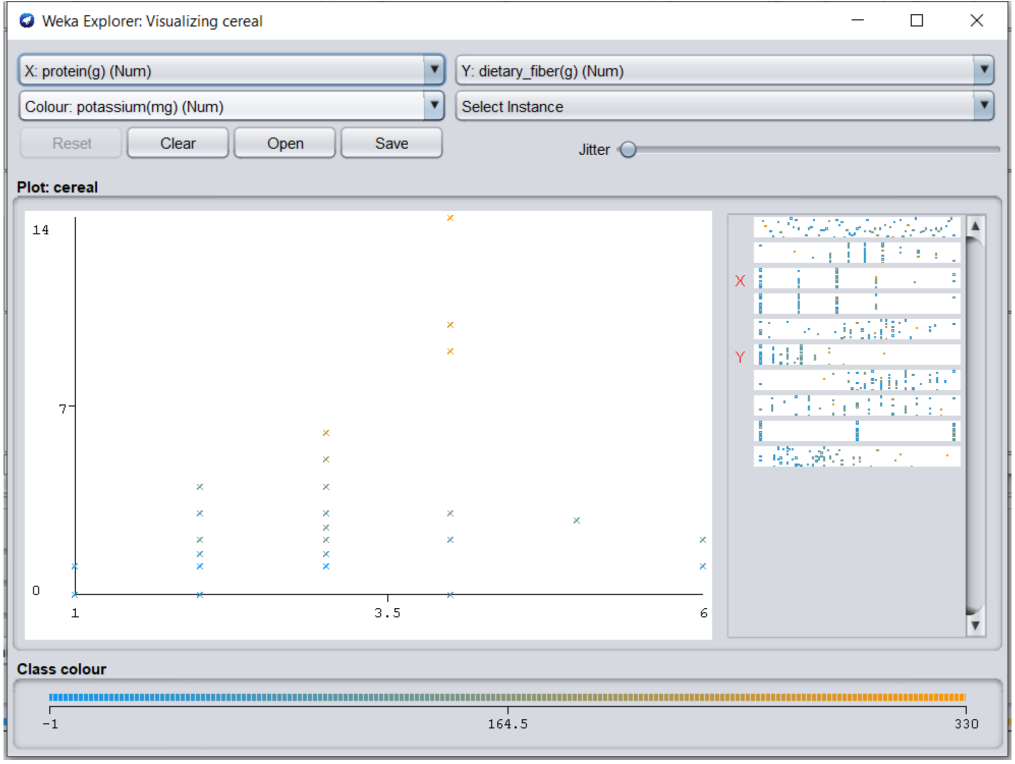
Colour: potassium(mg) (Num)

Select Instance

Plot: cereal

Class colour

<u>Dietary Fiber and Protein:</u> Here, it can be seen that the protein remains the same in certain points while fiber increases. After a ratio protein also increases same.



Weka Explorer: Visualizing cereal

X: protein(g) (Num)

Y: dietary_fiber(g) (Num)

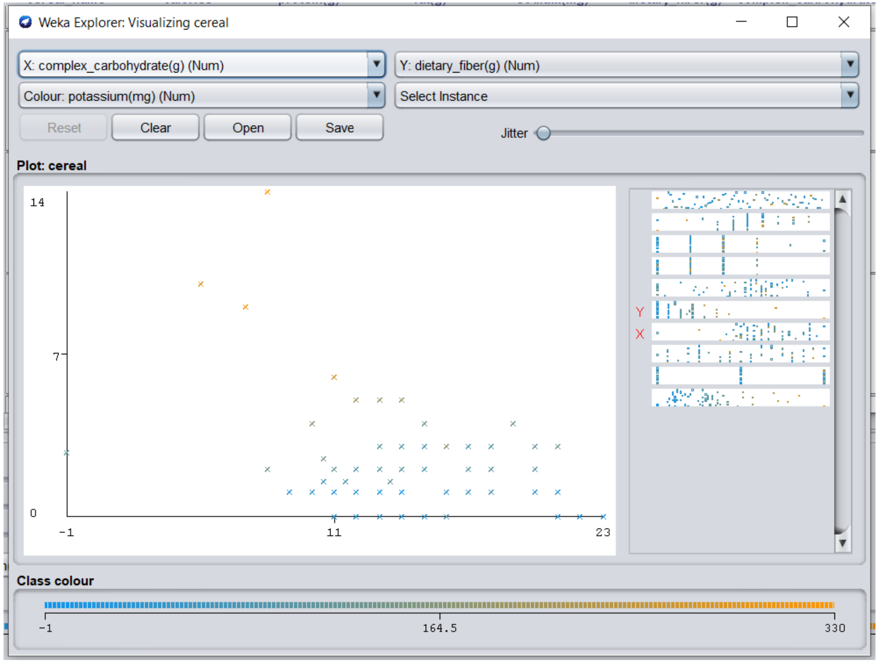
Colour: potassium(mg) (Num)

Select Instance

Plot: cereal

Class colour

Complex Carbohydrate and Dietary fiber: When fiber increases the value for carbohydrate remains constant for certain points.



Weka Explorer: Visualizing cereal
X: complex_carbohydrate(g) (Num)
Y: dietary_fiber(g) (Num)
Colour: potassium(mg) (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: cereal
14
7
0
-1
11
23
Y
X
Class colour
-1
164.5
330

Reference:

[1] http://www.cs.umd.edu/hcil/hce/examples/cereal/cereal.txt