**Project Definition:**

I used a dataset of cereals along with different percentage of protein, calorie, fat, Sodium etc. in this project which was given to us. I used hierarchical clustering to discover knowledge from the given dataset.

**Literature Survey:**

The dataset is given below:

<http://www.cs.umd.edu/hcil/hce/examples/cereal/cereal-updated.txt>

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

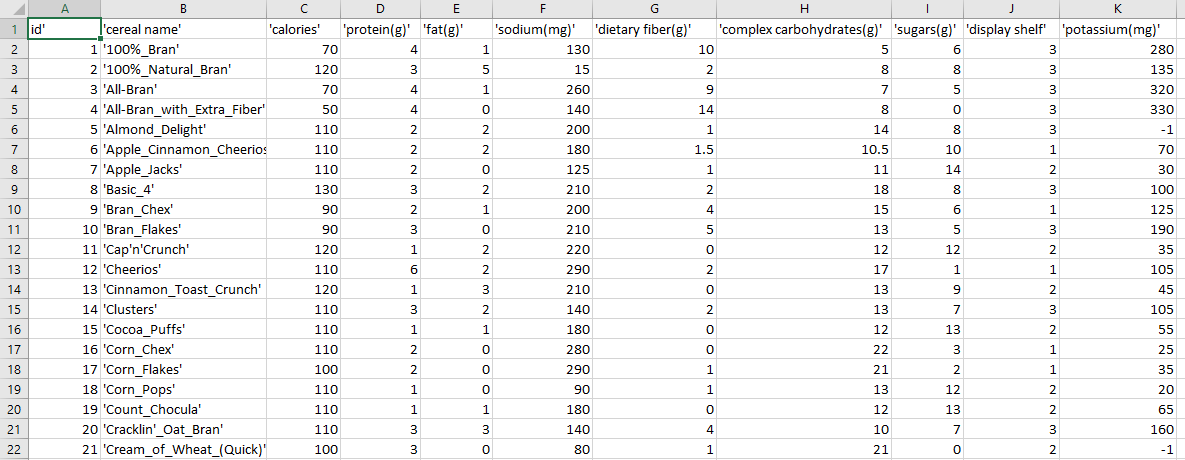
11. Shelf: display shelf (1, 2, or 3, counting from the floor)

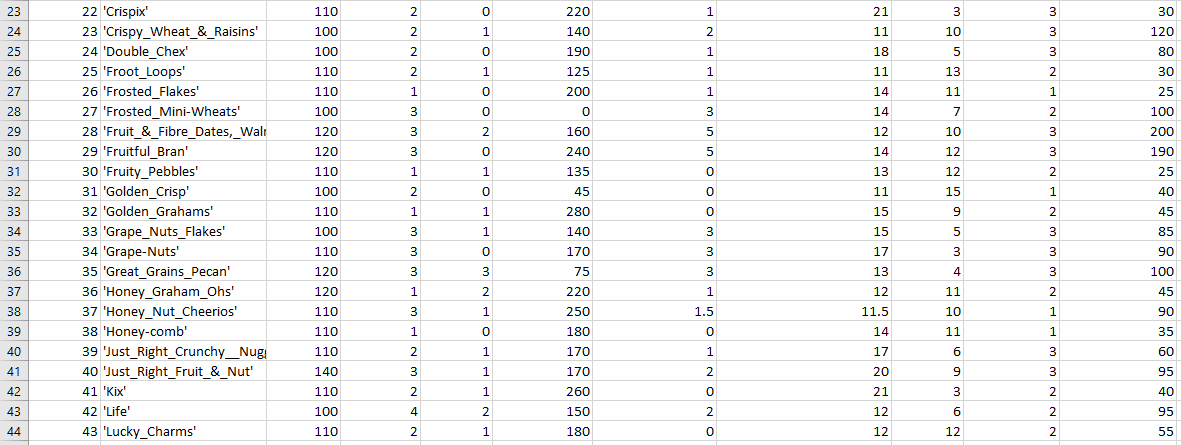
12. Rating: a rating of the cereals (calculated by Consumer Reports)

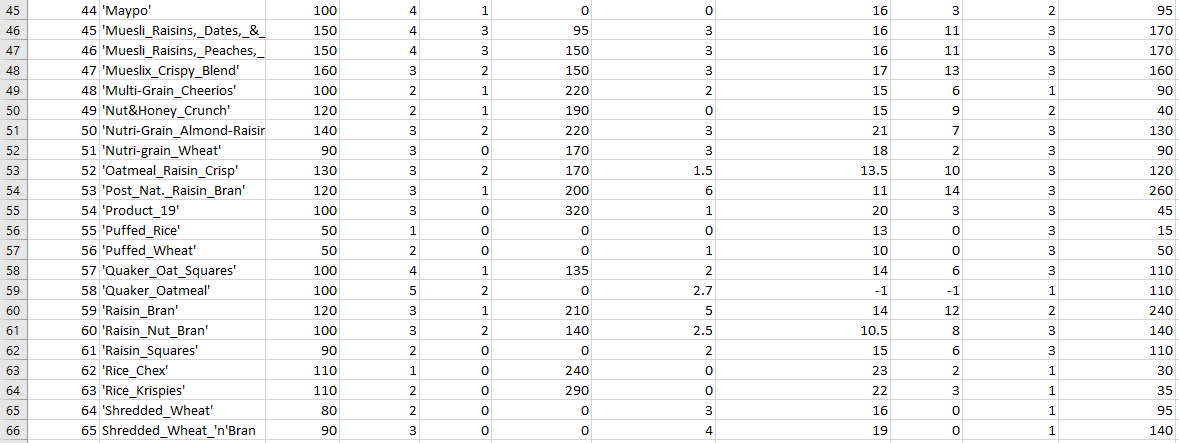
**Methods:**

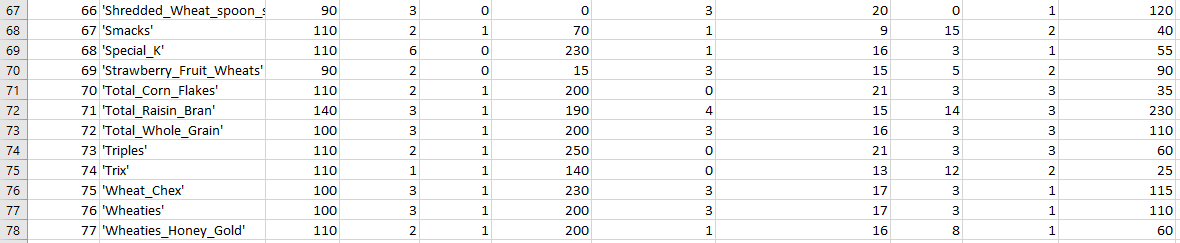
The dataset has been clustered by the hierarchical clustering technique. The cluster tree has been cut in several places. Then similarities between instances of individual clusters and dissimilarities between instances of different clusters have been analyzed.

**Dataset(without vitamin):**









**Hierarchical cluster tree with cutting point:**

=== Run information ===

­­­­­­­­­­

Scheme:weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

Relation: cerealsWithoutVitamin

Instances: 77

Attributes: 10

id

calories

protein(g)

fat(g)

sodium(mg)

dietaryfiber(g)

complexcarbohydrates(g)

sugars(g)

displayshelf

potassium(mg)

Test mode:evaluate on training data

=== Model and evaluation on 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):0.04835):0.04479,(24:0.24792,39:0.24792):0.01877):0.03473):0.01636,23:0.31778):0.01982,28:0.33761):0.00004):0.02293,35:0.36057):0.02023,54:0.3808):0.01647,((29:0.3668,(53:0.30072,71:0.30072):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,(48:0.207,77:0.207):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):0.09441,17:0.13787):0.11011,62:0.24798):0.22706):0.04806):0.01632,59:0.53943):0.00315,(((27:0.26217,69:0.26217):0.18008,44:0.44226):0.07239,61:0.51464):0.02793):0.00624,42:0.54881):0.00309,31:0.5519):0.00512,21:0.55703):0.0063):0.03183,(55:0.26816,56:0.26816):0.32701):0.00491,(64:0.28591,(65:0.10242,66:0.10242):0.1835):0.31415):0.06013):0.02044,(12:0.49034,68:0.49034):0.19029)

Time taken to build model (full training data) : 0.04 seconds

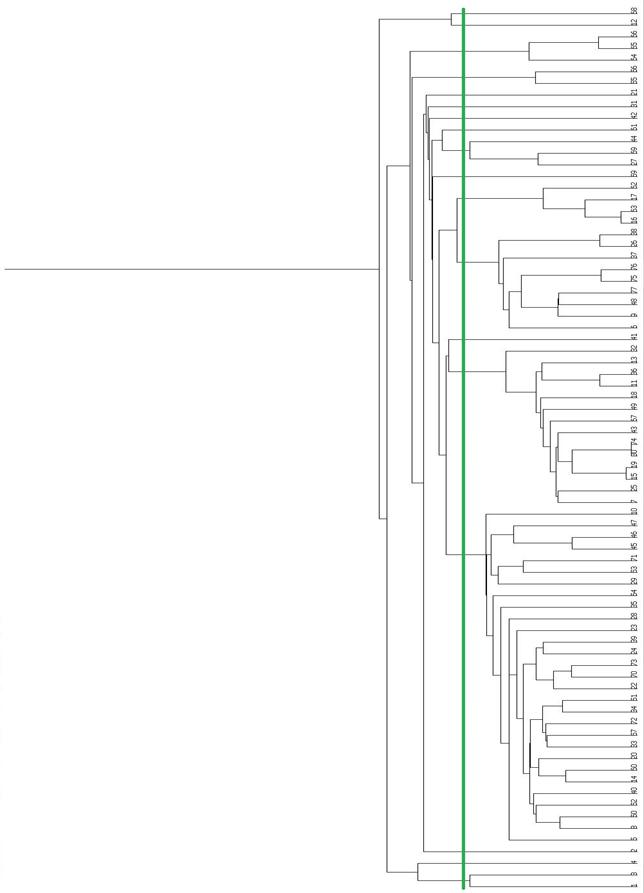
=== Model and evaluation on training set ===

Clustered Instances

0 76 ( 99%)

1 1 ( 1%)

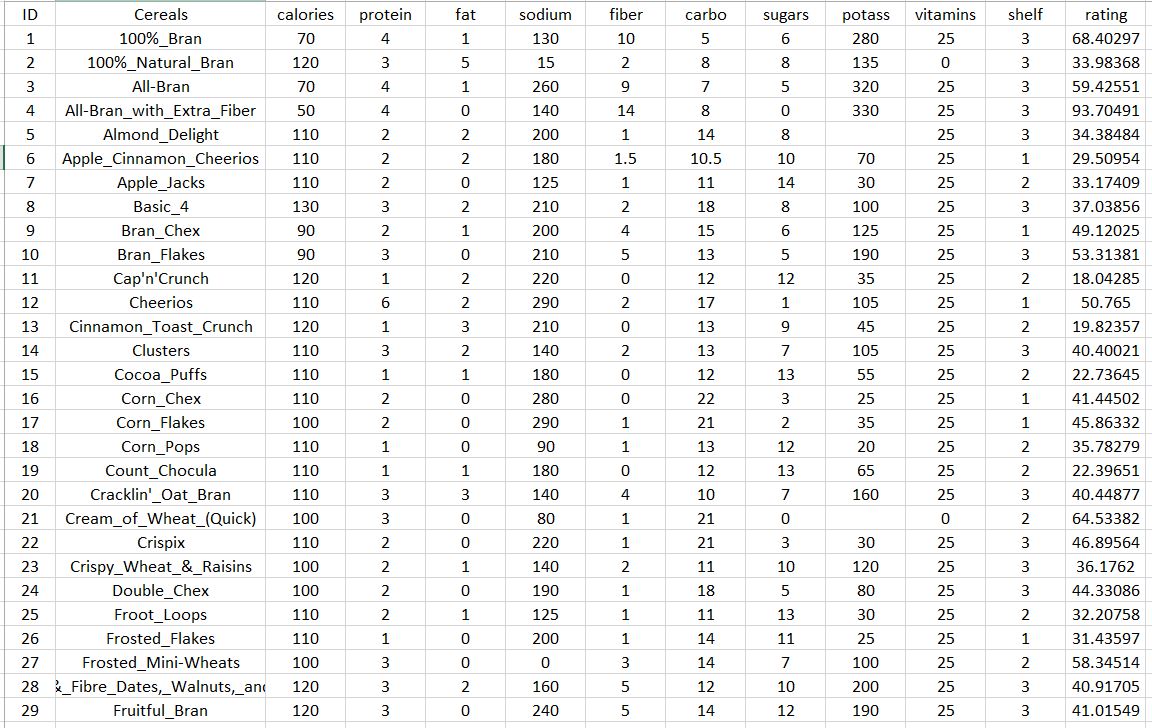
**Visualization:**

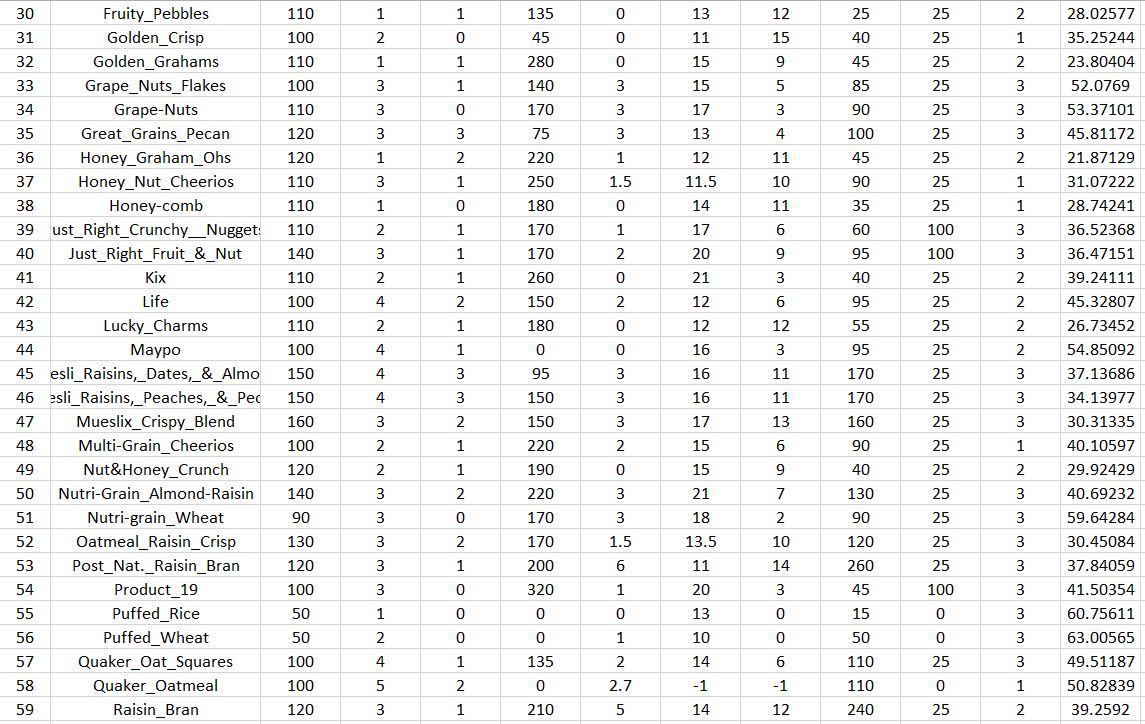
****

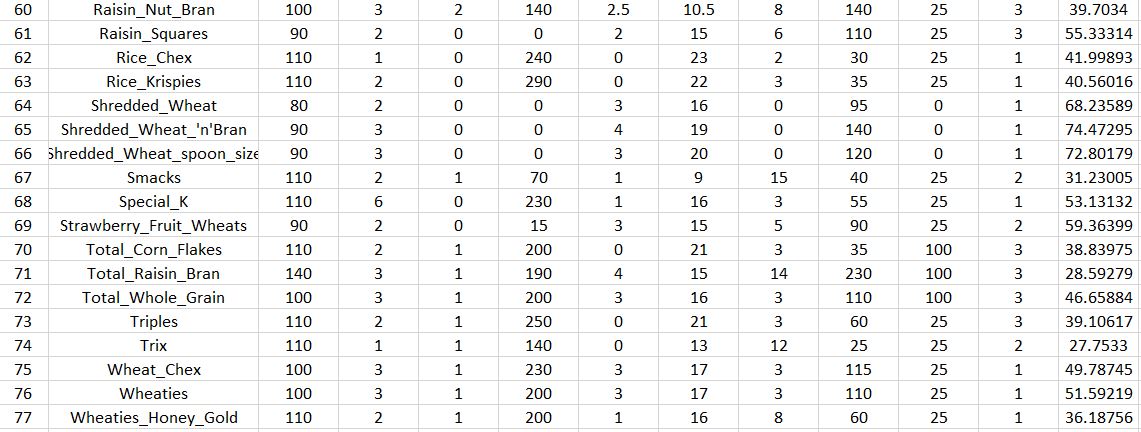
**Analysis:**

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster no. | ‘id’ no | Number of ID | Analysis Result |
| 1 | 1,3 | 2 | High potass, high sodium, medium cal |
| 2 | 4 | 1 | High sodium, zero sugar |
| 3 | 2 | 1 | High cal |
| 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 | Low fat |
| 5 | 7,25,15,19,30,74,43,67,49,18,11,36,13,32 | 14 | Low fat high sodium |
| 6 | 41 | 1 | High cal ,high sod |
| 7 | 6,9,48,77,75,76,37,26,38 | 9 | High sodium,low protein,fat |
| 8 | 16,63,17,62 | 4 | Low fat, zero fiber,suger |
| 9 | 59 | 1 | zero sodium |
| 10 | 27,69,44 | 3 | zero sodium, zero fiber, low protein,fat |
| 11 | 61 | 1 | High cal, high potass |
| 12 | 42 | 1 | High sodium, high cal |
| 13 | 31 | 1 | High cal,high sod |
| 14 | 21 | 1 | High cal, high sodium, high potass |
| 15 | 55,56 | 2 | zero fat, zero sodium, zero sugar |
| 16 | 64,65,66 | 3 | zero sugar , zero sodium, Low fat |
| 17 | 12 | 1 | zero fiber |
| 18 | 68 | 1 | Low fiber and protein shelf |

**Dataset(with vitamin):**







**Hierarchical cluster tree with cutting point:**

=== Run information ===

Scheme: weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

Relation: relation

Instances: 77

Attributes: 12

name

calories

protein

fat

sodium

fiber

carbo

sugars

potass

vitamins

shelf

rating

Test mode: evaluate on training data

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

Cluster 0

(((((68.402973:1.09985,59.425505:1.09985):0.10114,(((33.983679:1.18502,((((((((((34.384843:1.09306,((((((((((37.038562:1.02197,40.69232:1.02197):0.01674,30.450843:1.03871):0.00917,((40.400208:1.01763,39.7034:1.01763):0.01573,40.448772:1.03336):0.01452):0.00243,36.176196:1.05031):0.00012,((52.076897:1.02851,49.511874:1.02851):0.01078,(53.371007:1.02253,59.642837:1.02253):0.01677):0.01113):0.00082,((46.895644:1.03551,44.330856:1.03551):0.00021,39.106174:1.03571):0.01554):0.00422,40.917047:1.05547):0.00995,45.811716:1.06542):0.01137,(41.015492:1.06699,37.840594:1.06699):0.0098):0.00061,((37.136863:1.01544,34.139765:1.01544):0.0377,30.313351:1.05314):0.02427):0.01566):0.00411,53.313813:1.09716):0.02294,39.241114:1.1201):0.00879,(((((((33.174094:1.0218,32.207582:1.0218):0.00349,(((22.736446:1.00047,22.396513:1.00047):0.01633,(28.025765:1.00013,27.753301:1.00013):0.01667):0.00628,26.734515:1.02308):0.00221):0.00096,31.230054:1.02625):0.00518,29.924285:1.03143):0.00553,((18.042851:1.00622,21.871292:1.00622):0.02537,19.823573:1.03159):0.00538):0.00016,35.782791:1.03712):0.02138,23.804043:1.0585):0.07039):0.00269,((((29.509541:1.05926,((49.120253:1.02857,(40.105965:1.02251,36.187559:1.02251):0.00606):0.01957,(49.787445:1.00478,51.592193:1.00478):0.04335):0.01112):0.00341,31.072217:1.06266):0.00374,(31.435973:1.00558,28.742414:1.00558):0.06083):0.0419,(((41.445019:1.00101,40.560159:1.00101):0.01088,45.863324:1.01189):0.01843,41.998933:1.03032):0.078):0.02328):0.00478,39.259197:1.13637):0.00216,(((58.345141:1.03388,59.363993:1.03388):0.06052,54.850917:1.09441):0.03151,55.333142:1.12592):0.0126):0.00352,45.328074:1.14204):0.00339,35.252444:1.14543):0.0332,64.533816:1.17862):0.0064):0.00955,(60.756112:1.03576,63.005645:1.03576):0.15882):0.0039,(68.235885:1.04182,(74.472949:1.00547,72.801787:1.00547):0.03635):0.15665):0.00252):0.00174,93.704912:1.20273):0.00773,(50.764999:1.11419,53.131324:1.11419):0.09627):0.05153,(((((36.523683:1.0409,38.839746:1.0409):0.03387,46.658844:1.07476):0.01218,36.471512:1.08694):0.02566,41.50354:1.1126):0.04609,28.592785:1.15869):0.1033)

Time taken to build model (full training data) : 0.02 seconds

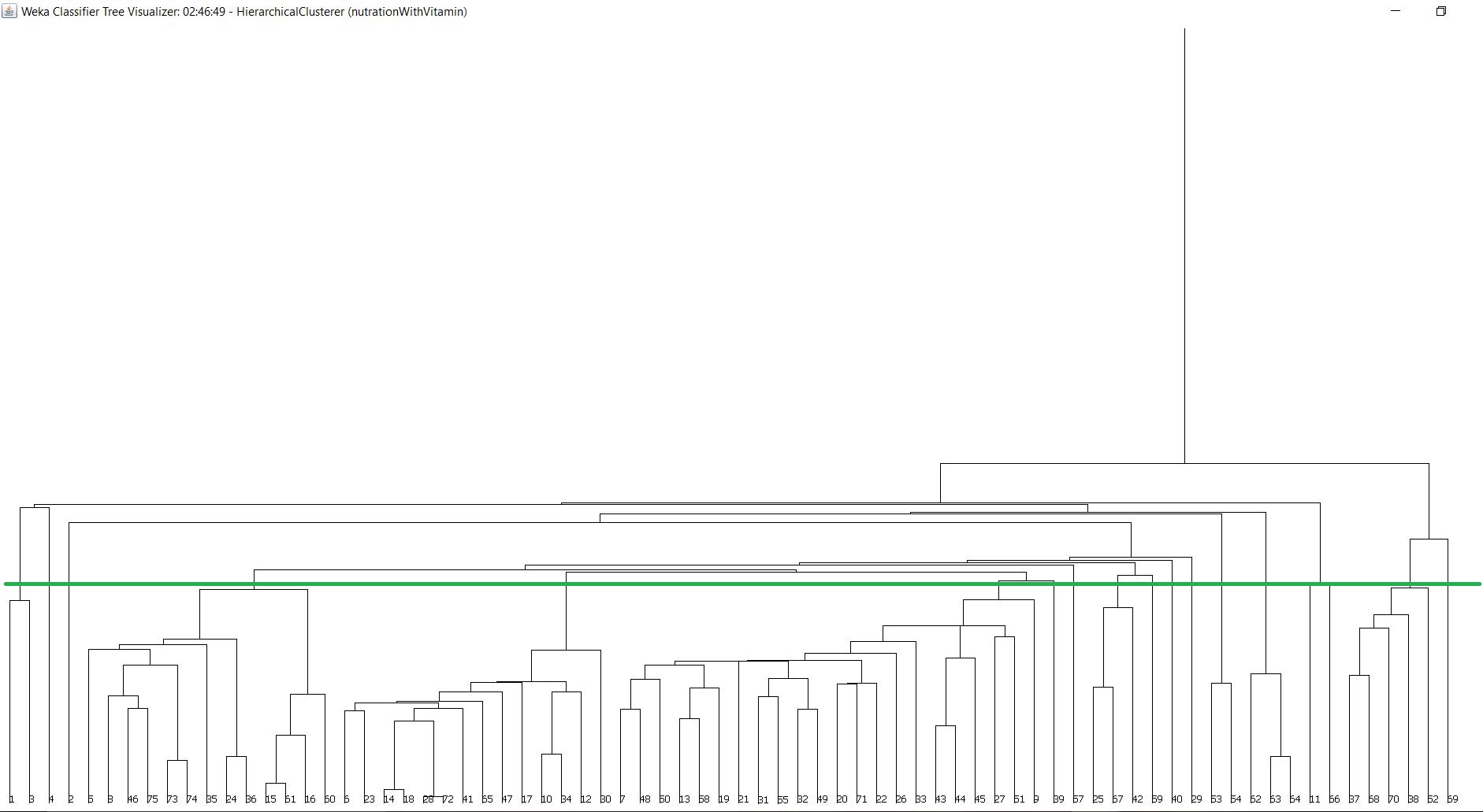
=== Model and evaluation on training set ===

Clustered Instances

0 76 ( 99%)

1 1 ( 1%)

**Visualization:**



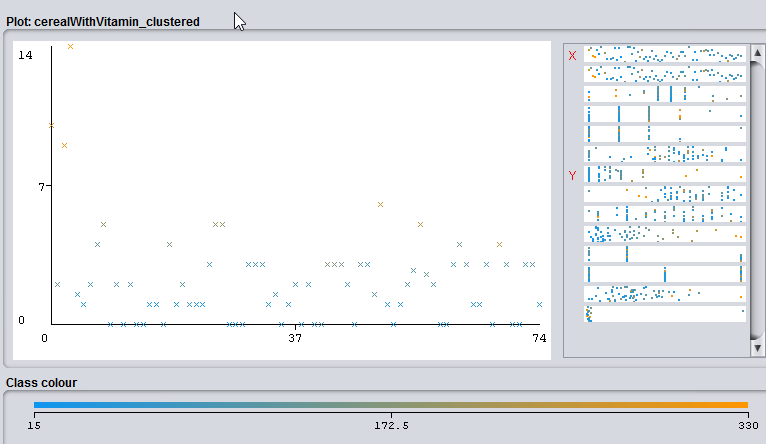
**Analysis:**

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster No | ID no | Number of ID | Analysis Result |
| 1 | 1,3 | 2 | High potass,low fat |
| 2 | 4 | 1 | zero suger,high potass,high rating |
| 3 | 2 | 1 | zero vitamin |
| 4 | 5,8,46,75,73,74,35,24,36,15,61,16,60 | 9 | Low fat,high sodium |
| 5 | 6,23,14,18,28,72,41,65,47,17,10,34,12,30 | 18 | Average rating,,high sodium&potass |
| 6 | 7,48,50,13,58,19,21,31,55,32,49,20,71,22,26,33,43,44,45,27,51,9 | 22 | Low fat, high cal&sodium |
| 7 | 39 | 1 | High vitamin,low fat |
| 8 | 57 | 1 | High cal,average rating |
| 9 | 25,67,42 | 3 | Low sodim,low vitamin high rating |
| 10 | 59 | 1 | zero sodium |
| 11 | 40 | 1 | High vitamin |
| 12 | 29 | 1 | zero fat,high potass |
| 13 | 53,54 | 2 | High cal,low fat,high sodium |
| 14 | 62,63,64 | 3 | zero fat, low fiber, average rating |
| 15 | 11,66 | 2 | High potass,high sdium |
| 16 | 37,68,70,38,52 | 5 | Low fat high sodium |
| 17 | 69 | 1 | zero fat and high cal |

**Questions:**

**Is a strong correlation between dietary fiber and potassium?**

**Ans:**

****

There is a correlation between dietary fiber and potassium. When potassium increases or decreases the dietary fiber also increases or decreases.

**Are groups of cereals from which we can choose according to our preferences?**

From the analysis we can see that the clusters can be used according to our preferences.

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



Vitamin remains constant for different values of ratings. Rating increases but vitamin remains same.