**Project 1**

**Problem:** All the patients suffered heart attacks at some point in the past. Some are still alive and some are not. The survival and still-alive variables, when taken together, indicate whether a patient survived for at least one year following the heart attack.

The problem addressed by past researchers was to **predict from the other variables whether or not the patient will survive at least one year**. The most difficult part of this problem is correctly **predicting that the patient will NOT survive.** (Part of the difficulty seems to be the size of the data set.)

**Attribute Information:**

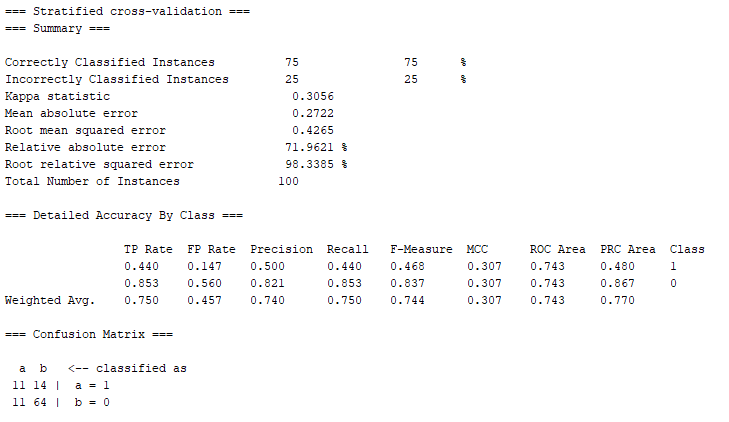
1. **survival** -- the number of months patient survived (has survived, if patient is still alive). Because all the patients had their heart attacks at different times, **it is possible that some patients have survived less than one year (alive-at-1 = 0) but they are still alive. Check the second variable to confirm this. Such patients cannot be used for the prediction task mentioned above*.***
2. **still-alive** -- a binary variable. 0=dead at end of survival period, 1 means still alive
3. **age-at-heart-attack** -- age in years when heart attack occurred
4. **pericardial-effusion** -- binary. Pericardial effusion is fluid around the heart. 0=no fluid, 1=fluid
5. **fractional-shortening** -- a measure of contractility around the heart lower numbers are increasingly abnormal.
6. **epss** -- E-point septal separation, another measure of contractility. Larger numbers are increasingly abnormal.
7. **lvdd** -- left ventricular end-diastolic dimension. This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts.
8. **wall-motion-score** -- a measure of how the segments of the left ventricle are moving
9. **wall-motion-index** -- equals wall-motion-score divided by number of segments seen. Usually 12-13 segments are seen in an echocardiogram. **Use this variable INSTEAD of the wall motion score**.
10. **mult** -- a derivate var **which can be ignored**.
11. **name** -- the name of the patient (I have replaced them with "name").
12. **group** -- meaningless, **ignore it**
13. **alive-at-1** -- Boolean-valued. Derived from the first two attributes. 0 means patient was either dead after 1 year or had been followed for less than 1 year. 1 means patient was alive at 1 year.

**Dataset Preprocessing:**

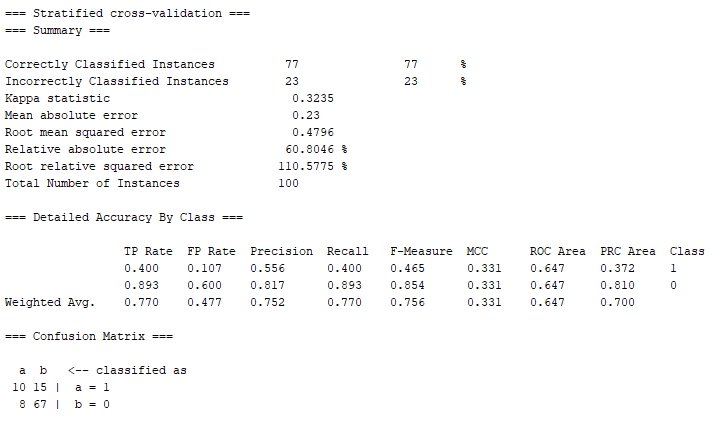
The dataset was prepared by the rules given below.

1. The attributes ‘**mult**’, ‘**name**’, ‘**wall-motion-score’** and ‘**group**’ were **removed** from the dataset as three of them were told to be ignored and the value of ‘name’ is same for all attributes.
2. All the instances having one or more missing value except **alive-at-1** attribute is removed.
3. For **survival <= 12 and still-alive = 1**, missing values of **alive-at-1 filled as 1**. All other missing values of alive-at-1 filled as 0.
4. For **survival > 12 and still-alive = 1 and alive-at-1 = 0**, the instances were removed.
5. Finally, **survival** and **still-alive** attributes were removed.

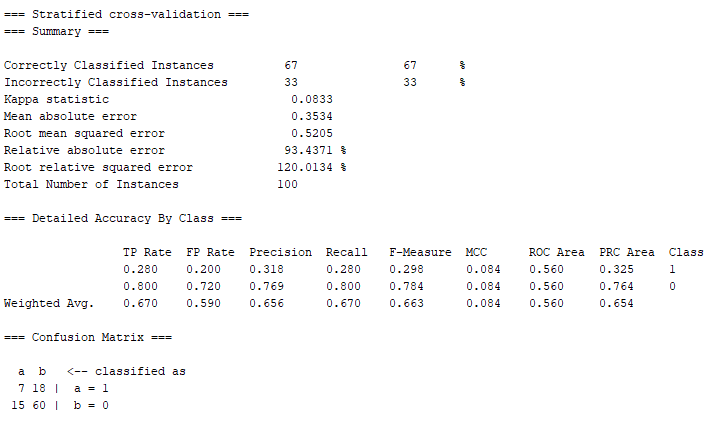
**Naïve Bias Classifier:**



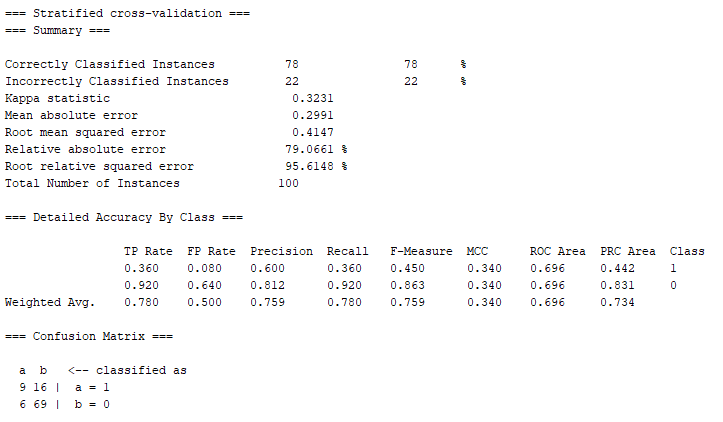
**SGD:**



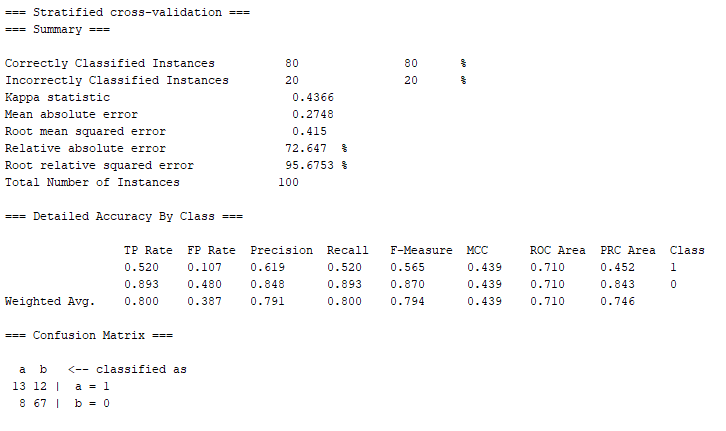
**KStar:**



**JRip:**



**J48:**



ROC Space:

|  |  |  |
| --- | --- | --- |
| Classifier | FPR | TPR |
| NaiveBayes | 0.147 | 0.440 |
| SGD | 0.107 | 0.400 |
| KStar | 0.200 | 0.280 |
| jrip | 0.080 | 0.360 |
| J48 | 0.107 | 0.520 |

**Comment:**

In this study,

TP = Number of instances correctly classified as survived.

FP = Number of instances incorrectly classified as survived.

FN = Number of instances incorrectly classified as not survived.

TN = Number of instances correctly classified as not survived.

Our main goal is to correctly predict the not survive class. So, we need to minimize the False negative.

From ROC space, J48 is giving me the best interest. Because it is incorrectly predicting only 12 instances as false negative which is lowest among all. So, for prediction, I choose J48.

**Project 2**

**With vitamin and rating**

**Project definition:** In this project we used a dataset of cereals along with their different percentage of calories, protein, fat etc. Besides, we did hierarchical clustering for finding out who needs which type of diet or food.

**Literature Survey:**

(With 'vitamin' and 'rating' columns): 77 x 11 here:

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. Potassium: 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)

**Method:** 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.

**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: cereal

Instances: 75

Attributes: 12

cerials

calories

protein

fat

sodium

fiber

carbo

sugars

potass

vitamins

shelf

rating

Test mode: evaluate on training data

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

Cluster 0

((((((100%-Bran:0.45956,All-Bran:0.45956):0.21047,All-Bran\_with\_Extra\_Fiber:0.67003):0.00706,(((((((((Apple\_Cinnamon\_Cheerios:0.34946,((Bran\_Chex:0.24314,(Multi-Grain\_Cheerios:0.21537,Wheaties\_Honey\_Gold:0.21537):0.02777):0.07118,(Wheat\_Chex:0.09803,Wheaties:0.09803):0.2163):0.03513):0.01125,Honey\_Nut\_Cheerios:0.36071):0.0113,(Frosted\_Flakes:0.10622,Honey-comb:0.10622):0.26578):0.11218,(((Corn\_Chex:0.04606,Rice\_Krispies:0.04606):0.10861,Corn\_Flakes:0.15467):0.09347,Rice\_Chex:0.24814):0.23605):0.04578,((((((((Apple\_Jacks:0.20993,Froot\_Loops:0.20993):0.01771,(((Cocoa\_Puffs:0.03206,Count\_Chocula:0.03206):0.15431,(Fruity\_Pebbles:0.01603,Trix:0.01603):0.17034):0.02973,Lucky\_Charms:0.2161):0.01153):0.0032,Smacks:0.23084):0.02228,Nut&Honey\_Crunch:0.25311):0.02191,Corn\_Pops:0.27503):0.00006,((Cap\_'n\_Crunch:0.11214,Honey\_Graham\_Ohs:0.11214):0.14118,Cinnamon\_Toast\_Crunch:0.25332):0.02177):0.07193,Golden\_Grahams:0.34702):0.17681,((((((((((((Basic\_4:0.21278,Nutri-Grain\_Almond-Raisin:0.21278):0.06881,Oatmeal\_Raisin\_Crisp:0.28159):0.03189,((Clusters:0.19164,Raisin\_Nut\_Bran:0.19164):0.06953,Cracklin\_Oat\_Bran:0.26118):0.05231):0.00829,Crispy\_Wheat\_&\_Raisins:0.32177):0.00038,((Grape\_Nuts\_Flakes:0.24174,Quaker\_Oat\_Squares:0.24174):0.0414,(Grape-Nuts:0.21347,Nutri-grain\_Wheat:0.21347):0.06966):0.03902):0.00223,((Crispix:0.27122,Triples:0.27122):0.002,Double\_Chex:0.27322):0.05116):0.01553,Fruit\_&\_Fibre\_Dates,\_Walnuts,\_and\_Oats:0.33991):0.02772,Great\_Grains\_Pecan:0.36763):0.03526,((Muesli\_Raisins,\_Dates,\_&\_Almonds:0.17638,Muesli\_Raisins,\_Peaches,\_&\_Pecans:0.17638):0.15406,Mueslix\_Crispy\_Blend:0.33044):0.07245):0.00072,(Fruitful\_Bran:0.37831,Post\_Nat.\_Raisin\_Bran:0.37831):0.02529):0.05821,Bran\_Flakes:0.46182):0.04317,Kix:0.50499):0.01885):0.00614):0.01013,Raisin\_Bran:0.5401):0.00472,(((Frosted\_Mini-Wheats:0.2627,Strawberry\_Fruit\_Wheats:0.2627):0.18199,Maypo:0.44469):0.07307,Raisin\_Squares:0.51776):0.02706):0.00696,Life:0.55179):0.0068,Golden\_Crisp:0.55859):0.1185):0.00506,(Cheerios:0.49375,Special\_K:0.49375):0.18839):0.3331,(((((Just\_Right\_Crunchy\_\_Nuggets:0.28993,Total\_Corn\_Flakes:0.28993):0.10692,Total\_Whole\_Grain:0.39685):0.03047,Just\_Right\_Fruit\_&\_Nut:0.42733):0.0605,Product\_19:0.48783):0.11208,Total\_Raisin\_Bran:0.59991):0.41534):0.14361,100%\_Natural\_Bran:1.15886)

Cluster 1

((Puffed\_Rice:0.27195,Puffed\_Wheat:0.27195):0.81547,(Quaker\_Oatmeal:1.05563,(Shredded\_Wheat:0.29323,(Shredded\_Wheat\_'n'Bran:0.10657,Shredded\_Wheat\_spoon\_size:0.10657):0.18666):0.7624):0.03178)

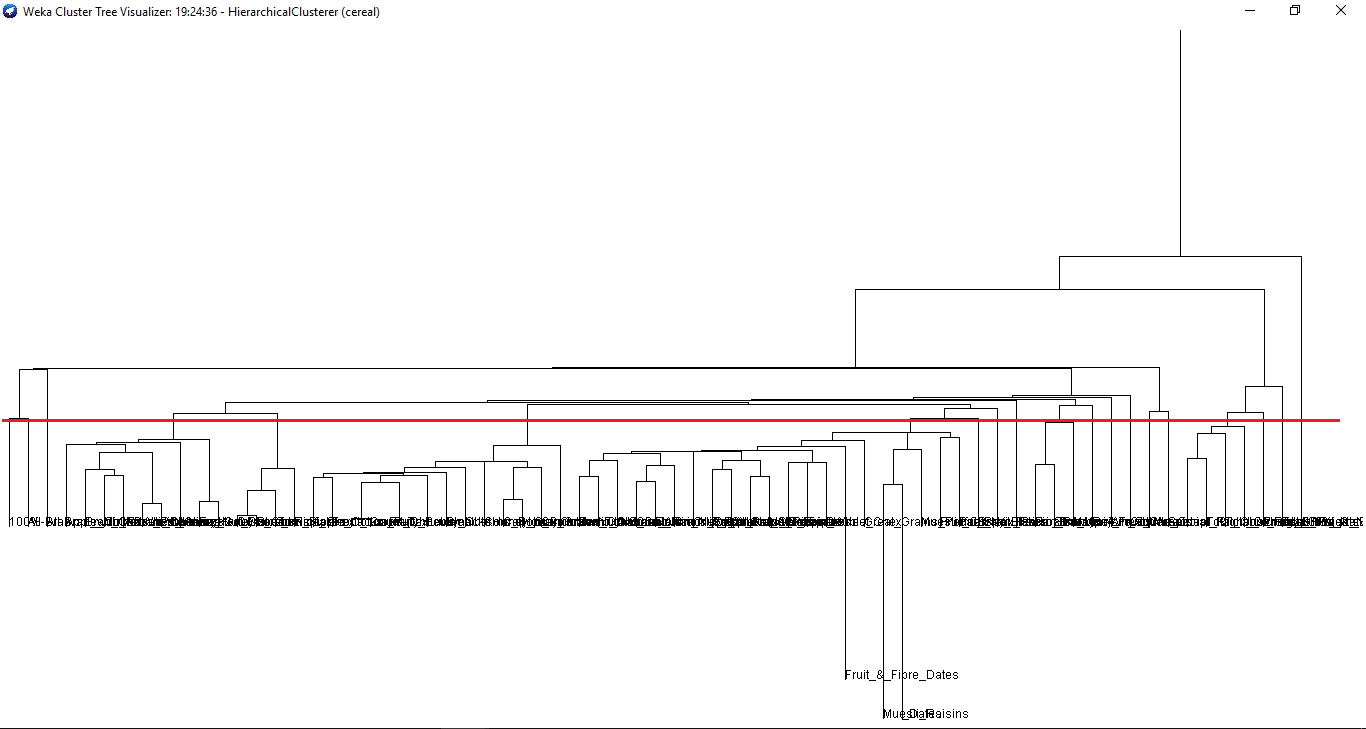
Time taken to build model (full training data) : 0.02 seconds

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

Clustered Instances

0 69 ( 92%)

1 6 ( 8%)

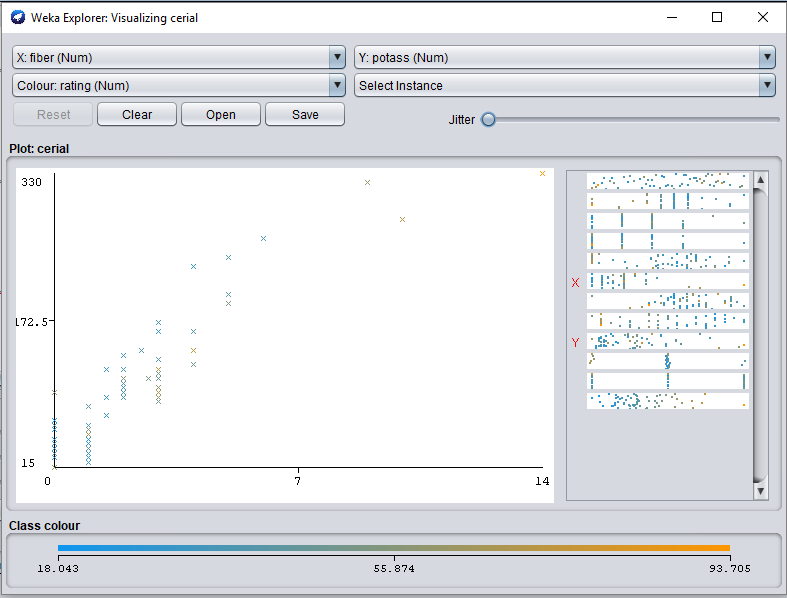


|  |  |  |
| --- | --- | --- |
| Cluster no | Cereal Name | Analysis |
| 1 | 100%\_Bran  All-Bran | High Potassium, low fat, high fiber |
| 2 | All-Bran\_with\_Extra\_fiber | Zero Sugar, high potassium rating fiber |
| 3 | 100%\_Natural\_Bran | No vitamin |
| 4 | Almond\_Delight  Basic\_4 Muesli\_Raisins,\_Peaches,\_&\_Pecans  Wheat\_Chex  Triples  Trix  great\_Grains\_Pecan  Double\_chex  Honey\_Graham\_Ohs | Low fat, high sodium |
| 5 | Cocoa\_Puffs  Raisin\_Squars  Corn\_Chex  Raisin\_Nut\_Bran | Average rating, high sodium and potassium |
| 6 | Apple\_Cinnamon\_Cheerios  Crispy\_Wheat&\_Raisins  Clusters,Corn\_Pops Fruit\_&\_Fibre\_Dates,\_Walnuts,\_and\_Oats  Total\_Whole\_Grain  Kix  Shredded\_wheat\_’n’\_Bran  Mueslix\_Crispy\_Blend  Corn\_Flakes  Bran\_Flakes  Grape-Nuts  Cheerios  Fruity\_Pebbles | Low fat, high calories & sodium |
| 7 | Apple\_Jacks  Multi-Grain\_Cheerios  Nutri-Grain\_Almond Raisin  Cinnamon\_Toast\_Crunch  Quaker\_Oatmeal  Count\_Chocula  Cream\_of\_Wheat\_(Quick)  Golden\_Crisp  Puffed\_Rice  Golden\_Grahams  Nut&Honey\_crunch  Cracklin’\_Oat\_Bran  Total\_Raisin\_Bran  Crispix  Frosted\_Flakes  Grape\_Nuts\_Flakes  Lucky\_Charms  Maypo Muesli\_Raisins,\_Dates,\_&\_Almonds Frosted\_Mini-Wheats  Nutri-GRain\_Wheat  Bran\_Chex | High vitamin, low fat, low fiber |
| 8 | Just\_Right\_Crunchy\_\_Nuggets | High calories, average rating |
| 9 | Quaker\_Oat\_Squares | High calories, average rating |
| 10 | Froot\_Loops  Smack  Life | No sodium |
| 11 | Raisin\_Bran | High vitamin |
| 12 | Just\_Right\_Fruit\_&\_Nut | No fat, high potassium |
| 13 | Fruitful\_Bran | High calories, low fat, high sodium |
| 14 | Post\_Nat.\_Raisin\_Bran  Product\_19 | 0 fat, low fiber, average rating |
| 15 | Rice\_Chex  Rice\_Krispies  Shredded\_wheat | High potassium, high sodium |
| 16 | Cap’n’Crunch | Low fat, high sodium, zero fiber |
| 17 | Shredded\_Wheat\_spoon\_size  Honey\_Nut\_Cheerios  Special\_K  Total\_Corn\_Flakes | No fat, high calories |
| 18 | Honey-com | High potassium, low fat |
| 19 | OatMeal\_Raisin\_Crisp | Zero sugar, high potassium, high rating |
| 20 | Strawberry\_Fruit\_Wheats | No fat |

**Questions:**

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

Answer:



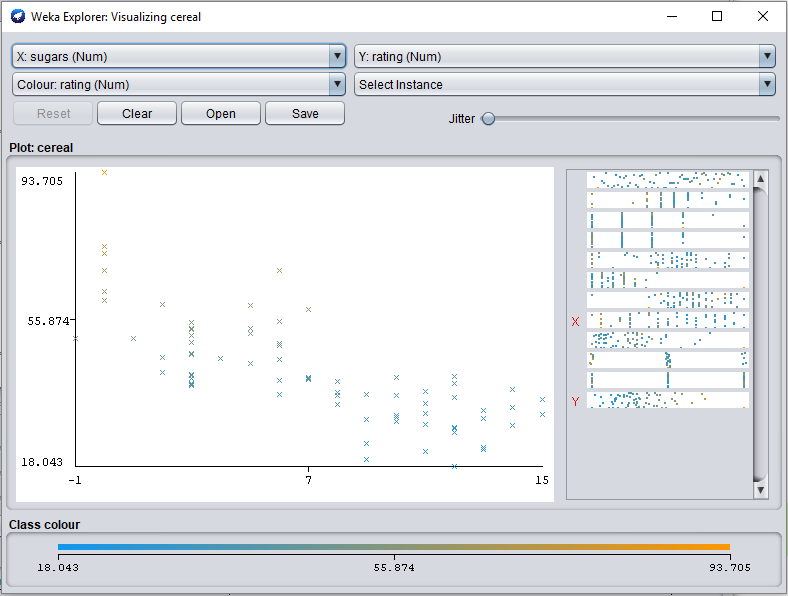
As dietary fiber increases, the amount of potassium increases. So, they have strong correlation.

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

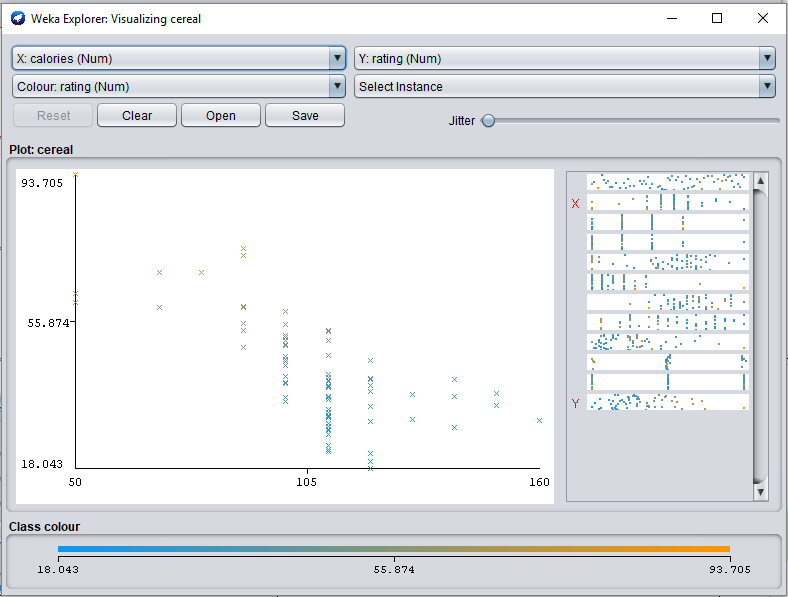
Answer: We can get to choose our preferred cereals from the cluster table given above.

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

Answer:



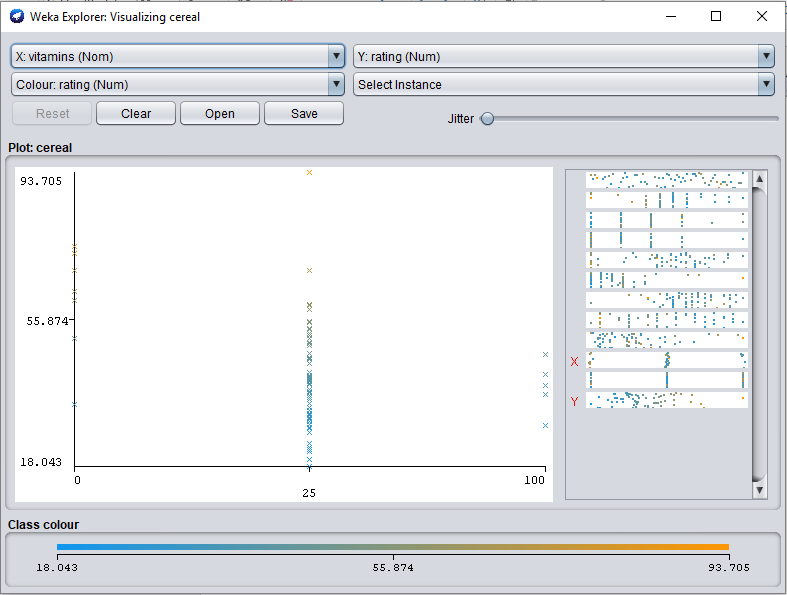
Sugar and customer rating is negatively correlated.



Customer rating and calories are negatively correlated.

1. Try to find other information.

Answer: There is only three types of vitamin. So, it is easy to find cereal based on vitamin.



**Without Vitamin and rating**

**Project definition:** In this project we used a dataset of cereals along with their different percentage of calories, protein, fat etc. Besides, we did hierarchical clustering for finding out who needs which type of diet or food.

**Literature Survey:**

(With 'vitamin' and 'rating' columns): 77 x 10 here:

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. Shelf: display shelf (1, 2, or 3, counting from the floor)
10. Potassium: milligrams of potassium.

**Method:** 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.

**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: cereal

Instances: 77

Attributes: 10

cerials\_name

calories

protein(g)

fat(g)

sodium(mg)

dietary\_fiber(g)

complex\_carbohydrates(g)

sugars(g)

display\_shelf

potassium(mg)

Test mode: evaluate on training data

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

Cluster 0

((((100%\_Bran:0.44226,All-Bran:0.44226):0.1363,All-Bran\_with\_Extra\_Fiber:0.57856):0.08163,(((100%\_Natural\_Bran:0.56333,((((((((((((Almond\_Delight:0.33764,(((((((Basic\_4:0.20516,Nutri-Grain\_Almond-Raisin:0.20516):0.06193,Oatmeal\_Raisin\_Crisp:0.26708):0.00672,Just\_Right\_Fruit\_&\_Nut:0.2738):0.01039,(((Clusters:0.18836,Raisin\_Nut\_Bran:0.18836):0.0719,Cracklin'\_Oat\_Bran:0.26026):0.02185,(((Grape\_Nuts\_Flakes:0.2381,Quaker\_Oat\_Squares:0.2381):0.00319,Total\_Whole\_Grain:0.24129):0.00859,(Grape-Nuts:0.19672,Nutri-grain\_Wheat:0.19672):0.05316):0.03223):0.00208):0.01722,((Crispix:0.22189,(Total\_Corn\_Flakes:0.17355,Triples:0.17355):0.04835):0.04479,(Double\_Chex:0.24792,Just\_Right\_Crunchy\_\_Nuggets:0.24792):0.01877):0.03473):0.01636,Crispy\_Wheat\_&\_Raisins:0.31778):0.01982,Fruit\_&\_Fibre\_Dates,\_Walnuts,\_and\_Oats:0.33761):0.00004):0.02293,Great\_Grains\_Pecan:0.36057):0.02023,Product\_19:0.3808):0.01647,((Fruitful\_Bran:0.3668,(Post\_Nat.\_Raisin\_Bran:0.30072,Total\_Raisin\_Bran:0.30072):0.06607):0.01968,((Muesli\_Raisins,\_Dates,\_&\_Almonds:0.17188,Muesli\_Raisins,\_Peaches,\_&\_Pecans:0.17188):0.15453,Mueslix\_Crispy\_Blend:0.3264):0.06007):0.0108):0.00187,Bran\_Flakes:0.39914):0.10546,((((((((Apple\_Jacks:0.20954,Froot\_Loops:0.20954):0.00491,(((Cocoa\_Puffs:0.03021,Count\_Chocula:0.03021):0.14149,(Fruity\_Pebbles:0.01562,Trix:0.01562):0.15608):0.03784,Lucky\_Charms:0.20954):0.00491):0.01581,Smacks:0.23027):0.01888,Nut&Honey\_Crunch:0.24915):0.00601,Corn\_Pops:0.25516):0.01214,((Cap'n'Crunch:0.0996,Honey\_Graham\_Ohs:0.0996):0.15227,Cinnamon\_Toast\_Crunch:0.25187):0.01542):0.07879,Golden\_Grahams:0.34608):0.15117,Kix:0.49725):0.00735):0.0185,((((Apple\_Cinnamon\_Cheerios:0.33799,((Bran\_Chex:0.20919,(Multi-Grain\_Cheerios:0.207,Wheaties\_Honey\_Gold:0.207):0.00219):0.09722,(Wheat\_Chex:0.09496,Wheaties:0.09496):0.21145):0.03158):0.01512,Honey\_Nut\_Cheerios:0.35311):0.01197,(Frosted\_Flakes:0.0996,Honey-comb:0.0996):0.26547):0.10996,(((Corn\_Chex:0.04347,Rice\_Krispies:0.04347):0.09441,Corn\_Flakes:0.13787):0.11011,Rice\_Chex:0.24798):0.22706):0.04806):0.01632,Raisin\_Bran:0.53943):0.00315,(((Frosted\_Mini-Wheats:0.26217,Strawberry\_Fruit\_Wheats:0.26217):0.18008,Maypo:0.44226):0.07239,Raisin\_Squares:0.51464):0.02793):0.00624,Life:0.54881):0.00309,Golden\_Crisp:0.5519):0.00512,Cream\_of\_Wheat\_(Quick):0.55703):0.0063):0.03183,(Puffed\_Rice:0.26816,Puffed\_Wheat:0.26816):0.32701):0.00491,(Shredded\_Wheat:0.28591,(Shredded\_Wheat\_'n'Bran:0.10242,Shredded\_Wheat\_spoon\_size:0.10242):0.1835):0.31415):0.06013):0.02044,(Cheerios:0.49034,Special\_K:0.49034):0.19029)

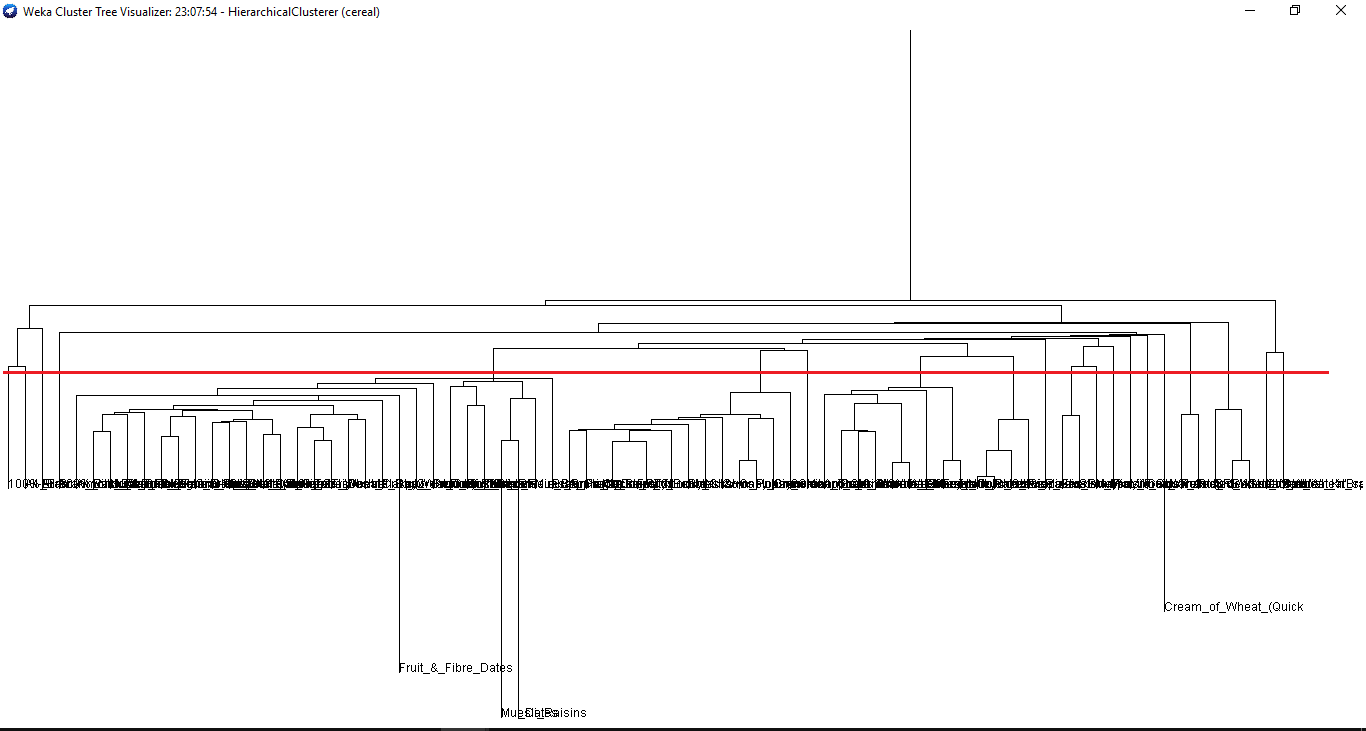
Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0 76 ( 99%)

1 1 ( 1%)

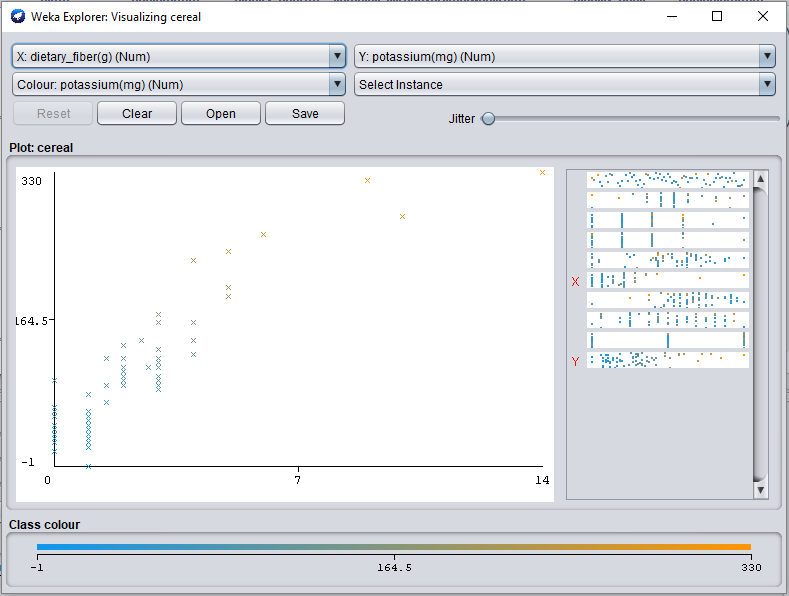


|  |  |  |
| --- | --- | --- |
| Cluster | Cereal | Analysis |
| 1 | 100%\_Bran | High sodium, high potassium, medium calories |
| 2 | All-Bran | High sodium, high potassium, medium calories |
| 3 | All-Bran\_with\_Extra\_Fiber | High sodium |
| 4 | 100%\_Natural\_Bran | High calories |
| 5 | Almond\_Delight  Basic\_4  Bran\_Flakes  Clusters  Cracklin'\_Oat\_Bran  Crispix  Crispy\_Wheat\_&\_Raisins  Double\_Chex  Fruit\_&\_Fibre\_Dates,\_Walnuts,\_and\_Oats  Fruitful\_Bran  Grape\_Nuts\_Flakes  Grape-Nuts  Great\_Grains\_Pecan  Just\_Right\_Crunchy\_\_Nuggets  Just\_Right\_Fruit\_&\_Nut  Muesli\_Raisins,\_Dates,\_&\_Almonds  Muesli\_Raisins,\_Peaches,\_&\_Pecans  Mueslix\_Crispy\_Blend  Nutri-Grain\_Almond-Raisin  Nutri-grain\_Wheat  Oatmeal\_Raisin\_Crisp  Post\_Nat.\_Raisin\_Bran  Product\_19  Quaker\_Oat\_Squares  Raisin\_Nut\_Bran  Total\_Corn\_Flakes  Total\_Raisin\_Bran  Total\_Whole\_Grain  Triples | Low fat |
| 6 | Apple\_Cinnamon\_Cheerios  Cap'n'Crunch  Cinnamon\_Toast\_Crunch  Cocoa\_Puffs  Corn\_Pops  Count\_Chocula  Froot\_Loops  Fruity\_Pebbles  Golden\_Grahams  Honey\_Graham\_Ohs  Lucky\_Charms  Nut&Honey\_Crunch  Smacks  Trix | Low fat, high sodium |
| 7 | Kix | High calories and sodium |
| 8 | Apple\_Cinnamon\_Cheerios  Bran\_Chex  Frosted\_Flakes  Honey\_Nut\_Cheerios  Honey-comb  Multi-Grain\_Cheerios  Wheat\_Chex  Wheaties  Wheaties\_Honey\_Gold | High sodium, low protein and fat |
| 9 | Corn\_Chex  Corn\_Flakes  Rice\_Chex  Rice\_Krispies | Low fat, no fiber and sugar |
| 10 | Raisin\_Bran | No sodium |
| 11 | Frosted\_Mini-Wheats  Strawberry\_Fruit\_Wheats | No fiber, low protein and fat |
| 12 | Maypo | High calories and potassium |
| 13 | Raisin\_Squares | High calories and potassium |
| 14 | Life | High sodium and calories |
| 15 | Golden\_Crisp | High calories and sodium |
| 16 | Cream\_of\_Wheat\_(Quick) | High calories, sodium and potassium |
| 17 | Puffed\_Rice  Puffed\_Wheat | Zero fat |
| 18 | Shredded\_Wheat  Shredded\_Wheat\_'n'Bran  Shredded\_Wheat\_spoon\_size | Zero fat |
| 19 | Cheerios | Zero fiber |
| 20 | Special\_K | Low fiber and protein, same shelf |

**Questions:**

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

Answer:



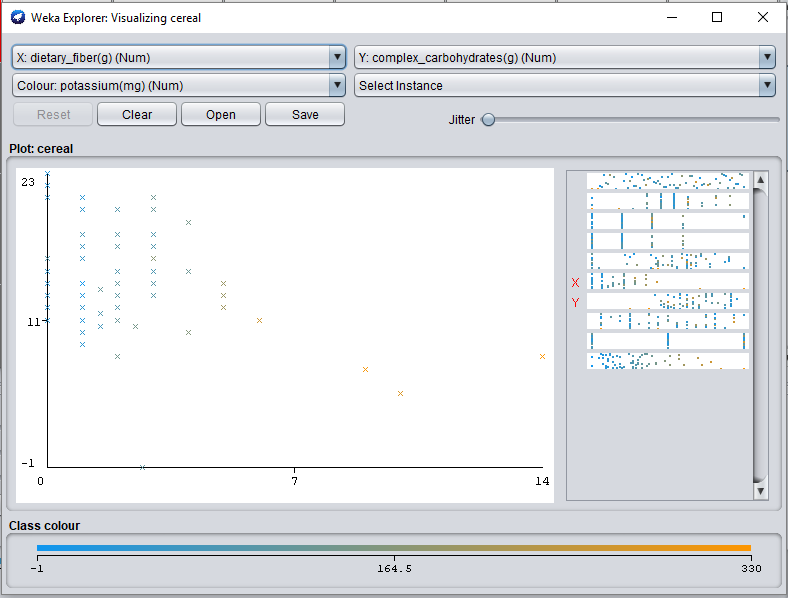
From the diagram we can say that there is a strong correlation between potassium and dietary fiber. Because if one is increasing, another one is increasing too.

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

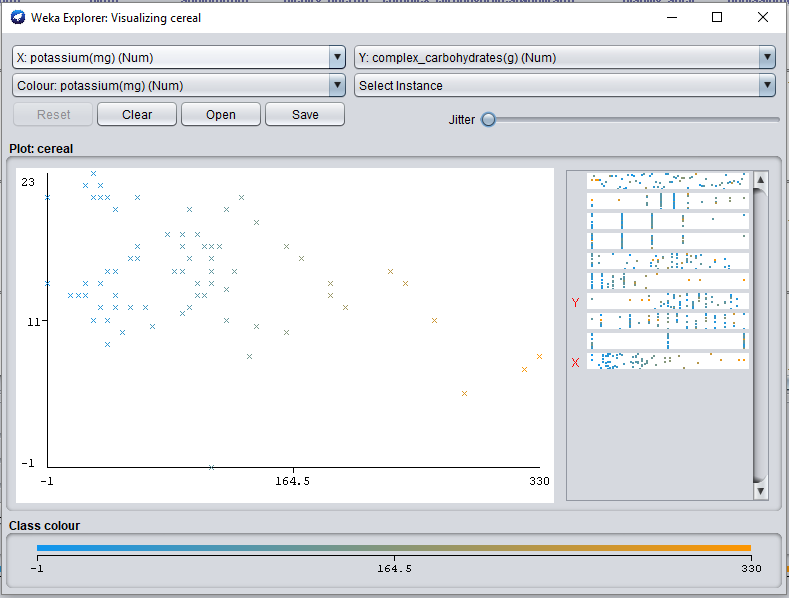
Answer: We can get to choose our preferred cereals from the cluster table given above.

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

Answer:



Complex carbohydrate and dietary fiber is negatively correlated.



Potassium and carbohydrate is negatively correlated.