Form 2 (Task 2)

You may fill in this form in either Dutch or English. This form will be available as a separate file on the yOUlearn website. An assignment will be created on yOUlearn in Week 11 of the course; the form must also be submitted via yOUlearn.

1.	Student information	
a.	Name	Ronald Koeman
b.	Student number	835152731
c.	Date	

2. Question 1. Task identification.
Given the assigned data set, is this an association rule mining problem or a text mining problem? Explain your reasoning. (Max 400 words)

Association Rule Mining:
Association rule mining is a suite of techniques for finding associations between items or attributes of a record in a data warehouse.

Text Mining:
Text mining is an important topic in data mining since a lot of information is nowadays simply published online. Very often companies need to monitor news and categorise this according to a set of predefined categories.

Until recently, IT specialists in the enterprise data world focused on “data mining”, which we can define as the discovery of knowledge from structured data (data contained in structured databases or data warehouses.) Today the majority of available business data is unstructured information; even though it may also contain numbers, dates and facts in structured fields, unstructured information is typically text (articles, website text, blog posts, etc.). The presence of unstructured information makes it more difficult to effectively perform knowledge management activities using traditional business intelligence tools.

The discovery of knowledge sources that contain text or unstructured information is called “text mining”. So, the main difference between data mining and text mining is that in text mining data is unstructured.

Data mining vs text mining approaches

Just as data mining is not just a unique approach or a single technique for discovering knowledge from data, text mining also consists of a broad variety of methods and technologies such as:

● Keyword-based technologies: The input is based on a selection of keywords in text that are filtered as a series of character strings, not words nor “concepts”.

● Statistics technologies: Refers to systems based on machine learning. Statistics technologies leverage a training set of documents used as a model to manage and categorize text.

● Linguistic based technologies: This method may leverage language processing systems. The output of text analysis allows a shallow understanding of the structure of the text, the grammar and logic employed. (For a better understanding of how this works, this post on text mining and NLP is helpful.)

All these approaches have a common feature: they are all concerned with processing text in an

approximate way since they are not capable to understand them.

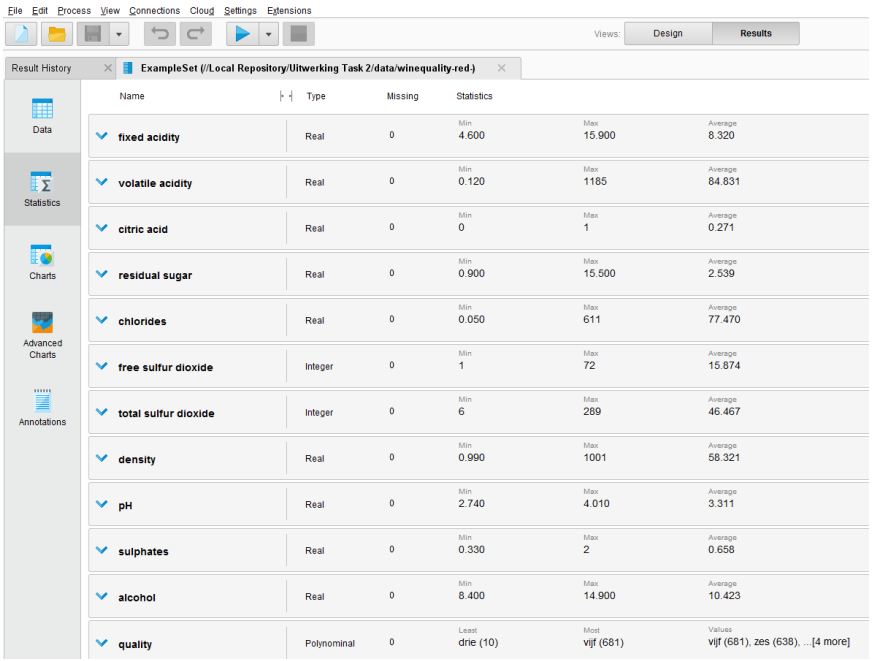
Unlike these technologies, a cognitive technology such as Cogito is designed to understand and analyze text not by guessing at the meaning of words, but by relying on a deep semantic analysis and a rich knowledge graph to ensure a precise, complete and more effective understanding of text as a person would.

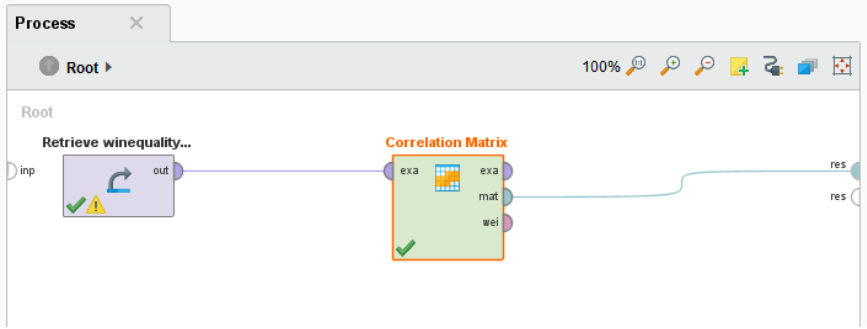
So this is a associating rule mining problem.

3. Question 2. Analyse the dataset potential.
Does the dataset you have been assigned have a specific class associated to the instances of the dataset? Can this be considered a classification problem, or would clustering be a better choice? Explain your reasoning. (400 words max.)

So there are 12 features. Target variable, in the machine learning context is the variable that is or should be the output. We can find a group kind of wine, called quality. I also change the attribute quality into string. Rapid Miner sees that as a polynomial.

Feature Selection: From many features to a few that are useful. Not all features are created equal.Those attributes that are irrelevant to the problem need to be removed. There will be some features that will be more important than others to the model accuracy. There will also be features that will be redundant in the context of other features. Feature selection addresses these problems by automatically selecting a subset that are most useful to the problem. Feature selection algorithms may use a scoring method to rank and choose features, such as correlation or other feature importance methods.





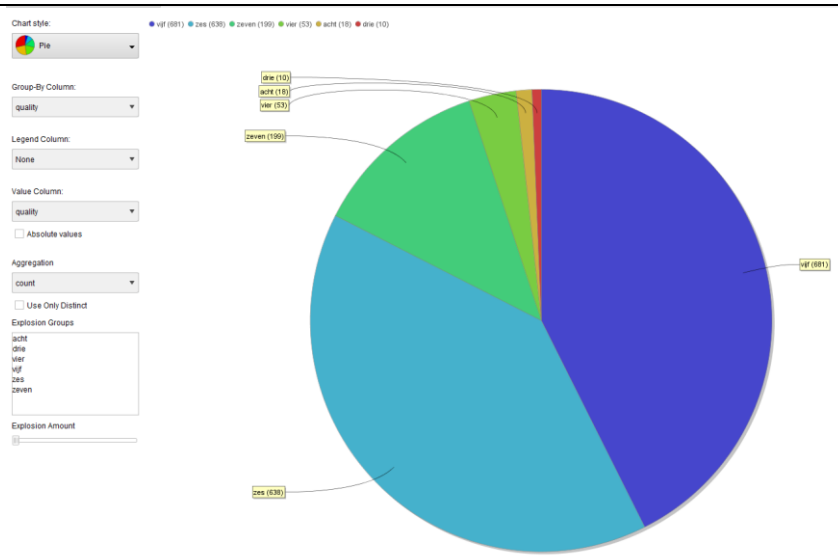
Correlation Matrix (Correlation Matrix) ExampleSet (/Local Repository/Uitwerking Task 2/data/winequality-red)

Attributes	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dio...	total sulfur dio...	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.066	0.672	0.115	0.085	-0.154	-0.113	0.055	-0.683	0.183	-0.062	0.067
volatile acidity	-0.066	1	-0.203	0.037	0.042	-0.061	-0.010	-0.001	0.056	-0.083	-0.079	-0.022
citric acid	0.672	-0.203	1	0.144	0.153	-0.061	0.036	0.052	-0.542	0.313	0.110	0.110
residual sugar	0.115	0.037	0.144	1	0.049	0.187	0.203	-0.007	-0.086	0.006	0.042	0.031
chlorides	0.085	0.042	0.153	0.049	1	-0.033	-0.009	0.025	-0.189	0.224	-0.173	-0.074
free sulfur dioxi...	-0.154	-0.061	-0.061	0.187	-0.033	1	0.667	-0.004	0.071	0.052	-0.070	-0.119
total sulfur dioxi...	-0.113	-0.010	0.036	0.203	-0.009	0.667	1	0.026	-0.066	0.043	-0.206	-0.242
density	0.055	-0.001	0.052	-0.007	0.025	-0.004	0.026	1	-0.011	0.018	-0.032	-0.017
pH	-0.683	0.056	-0.542	-0.086	-0.189	0.071	-0.066	-0.011	1	-0.197	0.206	0.038
sulphates	0.183	-0.083	0.313	0.006	0.224	0.052	0.043	0.018	-0.197	1	0.094	0.151
alcohol	-0.062	-0.079	0.110	0.042	-0.173	-0.070	-0.206	-0.032	0.206	0.094	1	0.393
quality	0.067	-0.022	0.110	0.031	-0.074	-0.119	-0.242	-0.017	0.038	0.151	0.393	1

Conclusion: There are no relations from 0,8 or greather, so in my opinion there is no evident dependency between the features

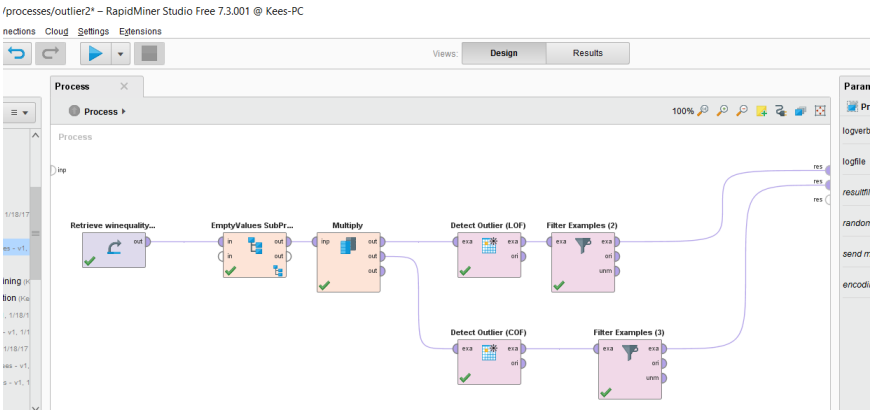
Supervised learning is the Data mining task of inferring a function from labeled training data.The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a “reasonable” way… In this case we have a label called class, so this is a supervised data mining task.

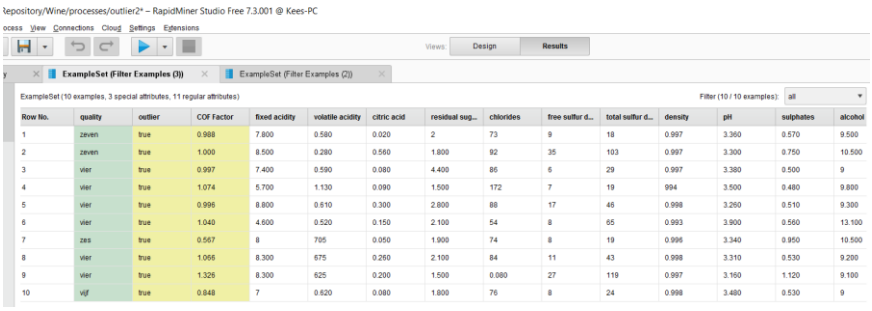
There are no missing values in the data:
- Look at the statistics of the data in RapidMiner



The classes are in very unbalance.

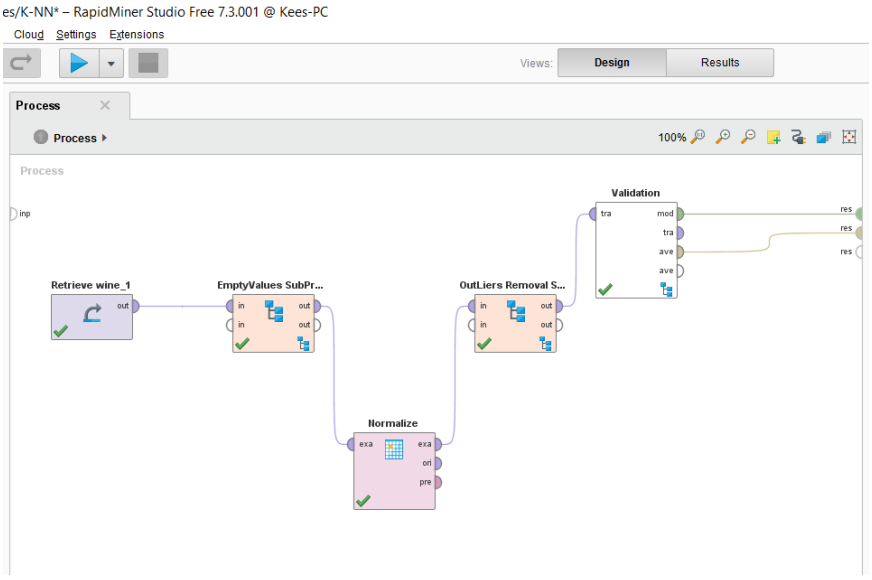
Process en beschrijving toevoegen

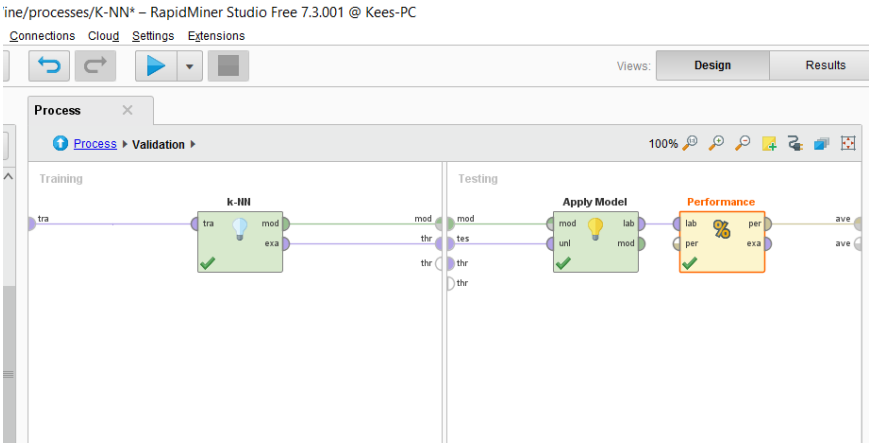




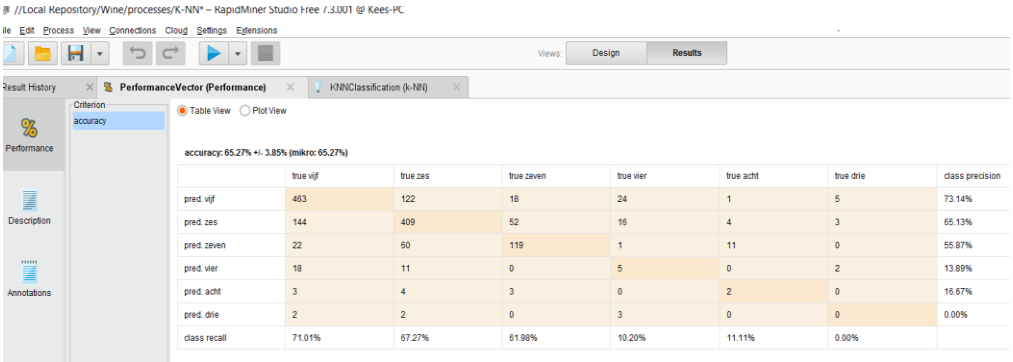
I think that there are relative a few outliers per class. So it may be correct values from special cases. Maybe it's wise to let him.

The two main subclasses of supervised data mining, classification and regression, are distinguished by the type of target. Regression involves a numeric target while classification involves a categorical target. In this case we have a categorical target, so we will use k-NN, naïve Bayes and logistic regression en calculate te accuracy of the classification





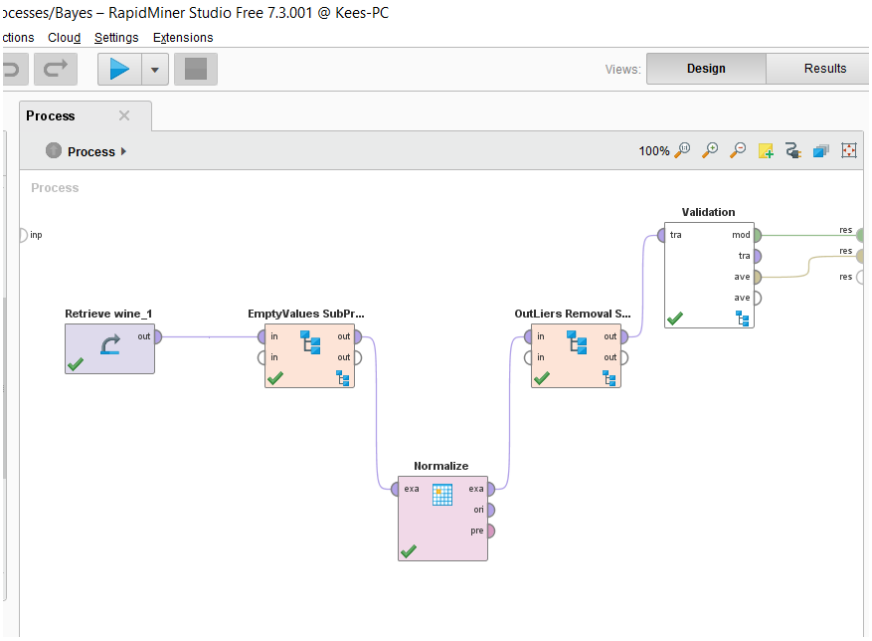
The first part of the validation block is the training. The second part of the validation block is the testing. On the training side a block must be placed that represents a classification algorithm. In this case the k-nearest Neighbour algorithms are selected from the operators
The performance block is added to provide the accuracy of the classification.

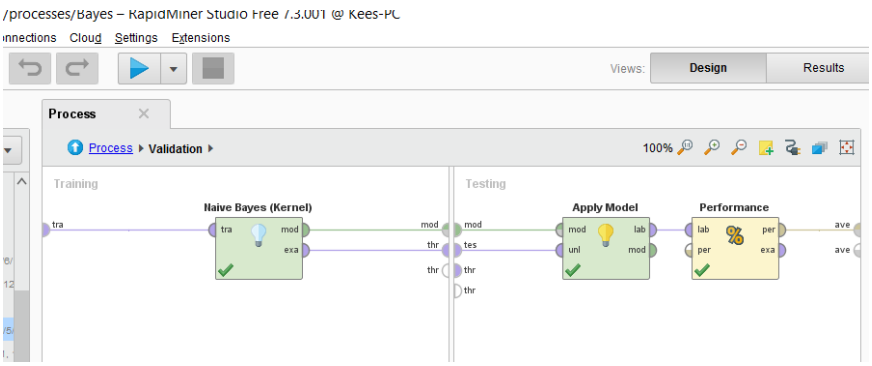


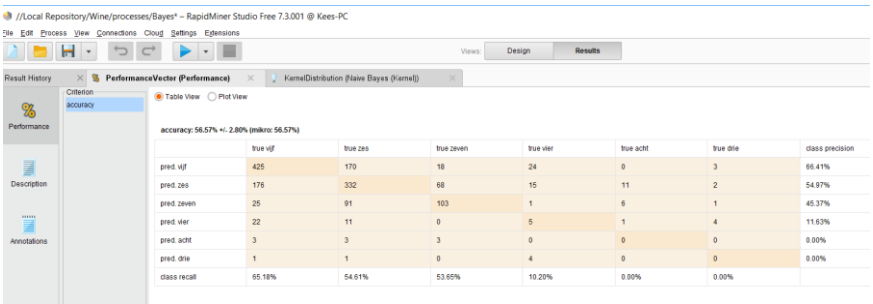
	true vijf	true zes	true zeven	true vier	true acht	true drie	class precision
pred. vijf	463	122	18	24	1	5	73.14%
pred. zes	144	409	52	16	4	3	65.13%
pred. zeven	22	60	119	1	11	0	55.87%
pred. vier	18	11	0	5	0	2	13.89%
pred. acht	3	4	3	0	2	0	16.67%
pred. drie	2	2	0	3	0	0	0.00%
class recall	71.01%	67.27%	61.98%	10.20%	11.11%	0.00%	

This result shows that a k-NN model is quite effective on the Wine dataset. It also shows the result in terms of a confusion matrix. It shows how well a given algorithm performs with respect to a class in the dataset selected. The selected k-NN procedure produce very few misclasifications

Bayes:







With respect to k-NN, naïve Bayes applies a simplistic probabilistic approach to analyse the data. We can compare the confusion matrix from k-NN and naïve Bayes.

4. Question 3. Data analysis Part 1.
   If you have an association rule mining problem:

   a. Explain all the steps needed in the analysis. (300 words max.)

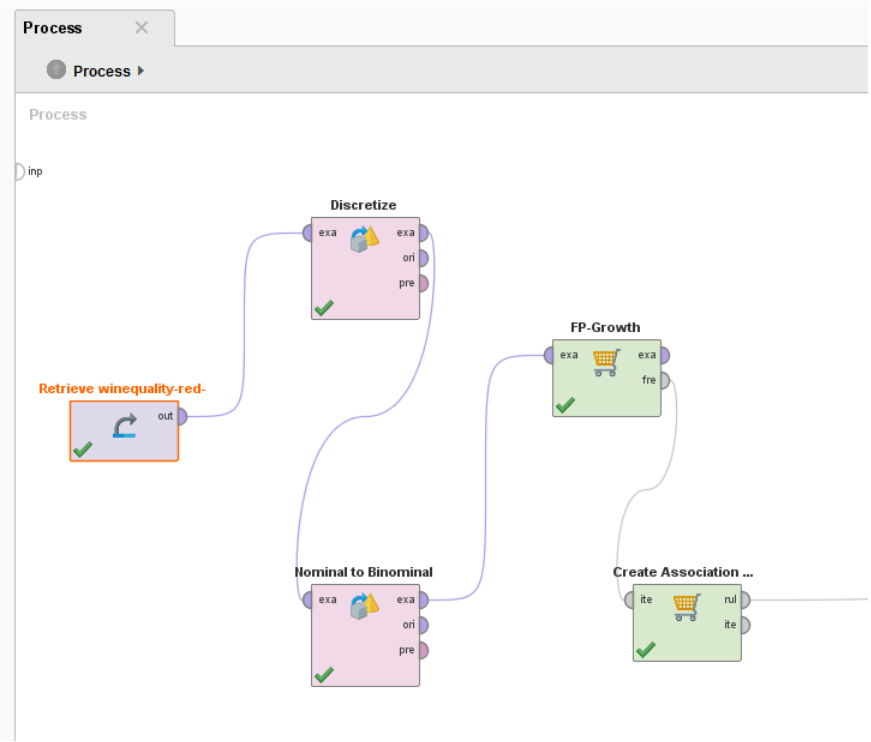
The problem of association rule mining is defined as:
Let I = {i1, i2, …..in}
Let D = {t1, t2, ….tn}

Each transaction in D has a unique transaction ID and contains a subset of the items in I.
A rule is defined as an implication of the form:

X => Y, where X, Y is in the collection I. So X=> ij for ij element I

Every rule is composed by two different sets of items, also known as itemsets, X and Y, where X is called antecedent of left-hand-side (LHS) and Y consequent or right-hand-side (RHS).

Associating rule mining is a suite of techniques for finding associations between items or attributes of a record in a data warehouse. In order to analyse the database using Rapidminer, we use the following workflow.



Association rule mining algorithms need the data to be discretised in order to work, we apply the discretisation to the data in the Wine dataset. The FP-growth algoritm is probably the most efficient itemset discovery algorithm available to date. FP-growth does not work on every kind of data though; it requires the data to be in binomial format. So, in addition to discretising the attributes, the discretised attribute must also be transformed into True/False values using the 'Nomial to Binomial'block of RapidMiner.

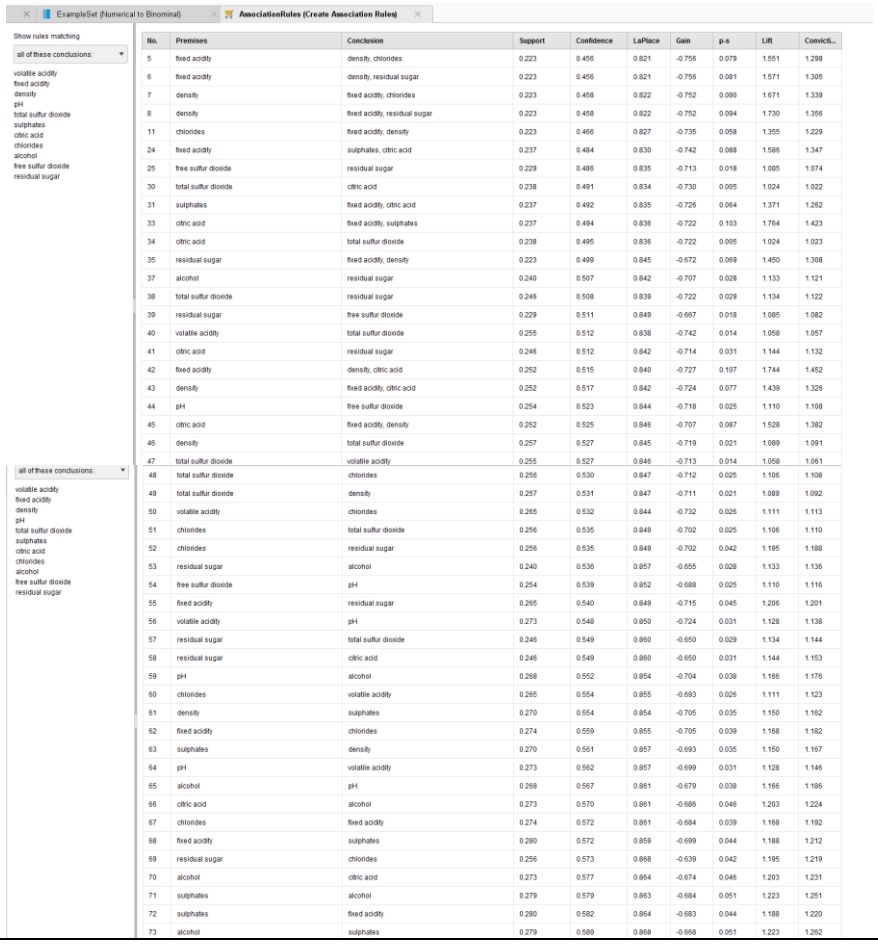
How FP-growt works:
In the first pass, the algorithm counts occurrence of items (attribute-value pairs) in the dataset, and stores them to 'header table'. In the second pass, it builds the FP-tree structure by inserting instances. Items in each instance have to be sorted by descending order of their frequency in the dataset, so that the tree can be processed quickly. Items in each instance that do not meet minimum coverage threshold are discarded.

If many instances share most frequent items, FP-tree provides high compression close to tree root.

Recursive processing of this compressed version of main dataset grows large item sets directly, instead of generating candidate items and testing them against the entire database. Growth starts from the bottom of the header table (having longest branches), by finding all instances matching given condition. New tree is created, with counts projected from the original tree corresponding to the set of instances that are conditional on the attribute, with each node getting sum of its children counts. Recursive growth ends when no individual items conditional on the attribute meet minimum support threshold, and processing continues on the remaining header items of the original FP-tree.

Once the recursive process has completed, all large item sets with minimum coverage have been found, and association rule creation begins

b. Explain the meaning of the found itemsets. (300 words max.)

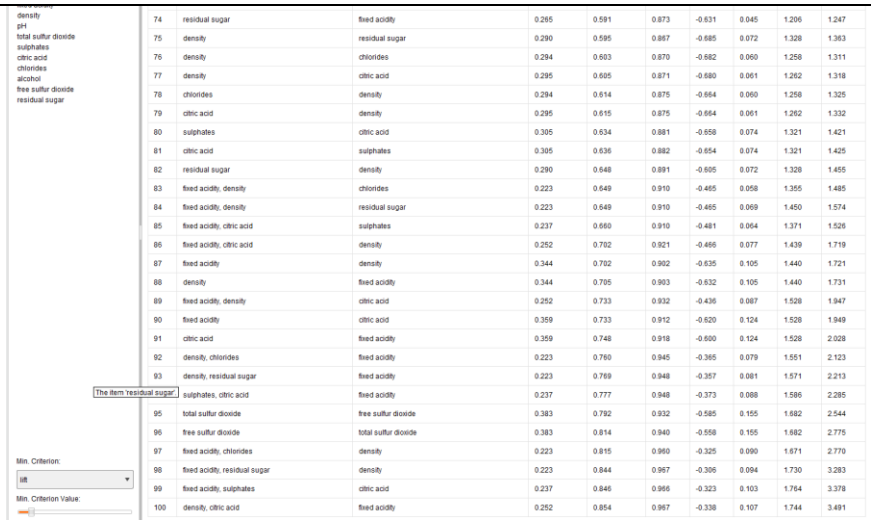

ExampleSet (Numerical to Binominal) AssociationRules (Create Association Rules)

Show rules matching

all of these conclusions:

volatile acidity
fixed acidity
density
pH
total sulfur dioxide
sulphates
citric acid
chlorides
alcohol
free sulfur dioxide
residual sugar

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convicti...
5	fixed acidity	density, chlorides	0.223	0.456	0.821	-0.756	0.079	1.551	1.298
6	fixed acidity	density, residual sugar	0.223	0.456	0.821	-0.756	0.081	1.571	1.305
7	density	fixed acidity, chlorides	0.223	0.458	0.822	-0.752	0.090	1.671	1.339
8	density	fixed acidity, residual sugar	0.223	0.458	0.822	-0.752	0.094	1.730	1.356
11	chlorides	fixed acidity, density	0.223	0.466	0.827	-0.735	0.058	1.355	1.229
24	fixed acidity	sulphates, citric acid	0.237	0.484	0.830	-0.742	0.088	1.586	1.347
25	free sulfur dioxide	residual sugar	0.229	0.486	0.835	-0.713	0.018	1.085	1.074
30	total sulfur dioxide	citric acid	0.238	0.491	0.834	-0.730	0.005	1.024	1.022
31	sulphates	fixed acidity, citric acid	0.237	0.492	0.835	-0.726	0.064	1.371	1.262
33	citric acid	fixed acidity, sulphates	0.237	0.494	0.836	-0.722	0.103	1.764	1.423
34	citric acid	total sulfur dioxide	0.238	0.495	0.836	-0.722	0.005	1.024	1.023
35	residual sugar	fixed acidity, density	0.223	0.499	0.845	-0.672	0.069	1.450	1.308
37	alcohol	residual sugar	0.240	0.507	0.842	-0.707	0.028	1.133	1.121
38	total sulfur dioxide	residual sugar	0.246	0.508	0.839	-0.722	0.029	1.134	1.122
39	residual sugar	free sulfur dioxide	0.229	0.511	0.849	-0.667	0.018	1.085	1.082
40	volatile acidity	total sulfur dioxide	0.255	0.512	0.838	-0.742	0.014	1.058	1.057
41	citric acid	residual sugar	0.246	0.512	0.842	-0.714	0.031	1.144	1.132
42	fixed acidity	density, citric acid	0.252	0.515	0.840	-0.727	0.107	1.744	1.452
43	density	fixed acidity, citric acid	0.252	0.517	0.842	-0.724	0.077	1.439	1.326
44	pH	free sulfur dioxide	0.254	0.523	0.844	-0.718	0.025	1.110	1.108
45	citric acid	fixed acidity, density	0.252	0.525	0.846	-0.707	0.087	1.528	1.382
46	density	total sulfur dioxide	0.257	0.527	0.845	-0.719	0.021	1.089	1.091
47	total sulfur dioxide	volatile acidity	0.255	0.527	0.846	-0.713	0.014	1.058	1.061
48	total sulfur dioxide	chlorides	0.256	0.530	0.847	-0.712	0.025	1.106	1.108
49	total sulfur dioxide	density	0.257	0.531	0.847	-0.711	0.021	1.089	1.092
50	volatile acidity	chlorides	0.265	0.532	0.844	-0.732	0.026	1.111	1.113
51	chlorides	total sulfur dioxide	0.256	0.535	0.849	-0.702	0.025	1.106	1.110
52	chlorides	residual sugar	0.256	0.535	0.849	-0.702	0.042	1.195	1.188
53	residual sugar	alcohol	0.240	0.536	0.857	-0.655	0.028	1.133	1.136
54	free sulfur dioxide	pH	0.254	0.539	0.852	-0.688	0.025	1.110	1.116
55	fixed acidity	residual sugar	0.265	0.540	0.849	-0.715	0.045	1.206	1.201
56	volatile acidity	pH	0.273	0.548	0.850	-0.724	0.031	1.128	1.138
57	residual sugar	total sulfur dioxide	0.246	0.549	0.860	-0.650	0.029	1.134	1.144
58	residual sugar	citric acid	0.246	0.549	0.860	-0.650	0.031	1.144	1.153
59	pH	alcohol	0.268	0.552	0.854	-0.704	0.038	1.166	1.176
60	chlorides	volatile acidity	0.265	0.554	0.855	-0.693	0.026	1.111	1.123
61	density	sulphates	0.270	0.554	0.854	-0.705	0.035	1.150	1.162
62	fixed acidity	chlorides	0.274	0.559	0.855	-0.705	0.039	1.168	1.182
63	sulphates	density	0.270	0.561	0.857	-0.693	0.035	1.150	1.167
64	pH	volatile acidity	0.273	0.562	0.857	-0.699	0.031	1.128	1.146
65	alcohol	pH	0.268	0.567	0.861	-0.679	0.038	1.166	1.186
66	citric acid	alcohol	0.273	0.570	0.861	-0.686	0.046	1.203	1.224
67	chlorides	fixed acidity	0.274	0.572	0.861	-0.684	0.039	1.168	1.192
68	fixed acidity	sulphates	0.280	0.572	0.859	-0.699	0.044	1.188	1.212
69	residual sugar	chlorides	0.256	0.573	0.868	-0.639	0.042	1.195	1.219
70	alcohol	citric acid	0.273	0.577	0.864	-0.674	0.046	1.203	1.231
71	sulphates	alcohol	0.279	0.579	0.863	-0.684	0.051	1.223	1.251
72	sulphates	fixed acidity	0.280	0.582	0.864	-0.683	0.044	1.188	1.220
73	alcohol	sulphates	0.279	0.589	0.868	-0.668	0.051	1.223	1.262



density
pH
total sulfur dioxide
sulphates
citric acid
chlorides
alcohol
free sulfur dioxide
residual sugar

The item 'residual sugar'

Min. Criterion:
lift
Min. Criterion Value:

74	residual sugar	fixed acidity	0.265	0.591	0.873	-0.631	0.045	1.206	1.247
75	density	residual sugar	0.290	0.595	0.867	-0.685	0.072	1.328	1.363
76	density	chlorides	0.294	0.603	0.870	-0.682	0.060	1.258	1.311
77	density	citric acid	0.295	0.605	0.871	-0.680	0.061	1.262	1.318
78	chlorides	density	0.294	0.614	0.875	-0.664	0.060	1.258	1.325
79	citric acid	density	0.295	0.615	0.875	-0.664	0.061	1.262	1.332
80	sulphates	citric acid	0.305	0.634	0.881	-0.658	0.074	1.321	1.421
81	citric acid	sulphates	0.305	0.636	0.882	-0.654	0.074	1.321	1.425
82	residual sugar	density	0.290	0.648	0.891	-0.605	0.072	1.328	1.455
83	fixed acidity, density	chlorides	0.223	0.649	0.910	-0.465	0.058	1.355	1.485
84	fixed acidity, density	residual sugar	0.223	0.649	0.910	-0.465	0.069	1.450	1.574
85	fixed acidity, citric acid	sulphates	0.237	0.660	0.910	-0.481	0.064	1.371	1.526
86	fixed acidity, citric acid	density	0.252	0.702	0.921	-0.466	0.077	1.439	1.719
87	fixed acidity	density	0.344	0.702	0.902	-0.635	0.105	1.440	1.721
88	density	fixed acidity	0.344	0.705	0.903	-0.632	0.105	1.440	1.731
89	fixed acidity, density	citric acid	0.252	0.733	0.932	-0.436	0.087	1.528	1.947
90	fixed acidity	citric acid	0.359	0.733	0.912	-0.620	0.124	1.528	1.949
91	citric acid	fixed acidity	0.359	0.748	0.918	-0.600	0.124	1.528	2.028
92	density, chlorides	fixed acidity	0.223	0.760	0.945	-0.365	0.079	1.551	2.123
93	density, residual sugar	fixed acidity	0.223	0.769	0.948	-0.357	0.081	1.571	2.213
94	sulphates, citric acid	fixed acidity	0.237	0.777	0.948	-0.373	0.088	1.586	2.285
95	total sulfur dioxide	free sulfur dioxide	0.383	0.792	0.932	-0.585	0.155	1.682	2.544
96	free sulfur dioxide	total sulfur dioxide	0.383	0.814	0.940	-0.558	0.155	1.682	2.775
97	fixed acidity, chlorides	density	0.223	0.815	0.960	-0.325	0.090	1.671	2.770
98	fixed acidity, residual sugar	density	0.223	0.844	0.967	-0.306	0.094	1.730	3.283
99	fixed acidity, sulphates	citric acid	0.237	0.846	0.966	-0.323	0.103	1.764	3.378
100	density, citric acid	fixed acidity	0.252	0.854	0.967	-0.338	0.107	1.744	3.491

AssociationRules

```
Association Rules
[volatile acidity] --> [residual sugar] (confidence: 0.443)
[density] --> [free sulfur dioxide] (confidence: 0.451)
[pH] --> [sulphates] (confidence: 0.452)
[sulphates] --> [pH] (confidence: 0.456)
[fixed acidity] --> [density, chlorides] (confidence: 0.456)
[fixed acidity] --> [density, residual sugar] (confidence: 0.456)
[density] --> [fixed acidity, chlorides] (confidence: 0.458)
[density] --> [fixed acidity, residual sugar] (confidence: 0.458)
[fixed acidity] --> [total sulfur dioxide] (confidence: 0.462)
[chlorides] --> [free sulfur dioxide] (confidence: 0.466)
[chlorides] --> [fixed acidity, density] (confidence: 0.466)
[free sulfur dioxide] --> [density] (confidence: 0.467)
[total sulfur dioxide] --> [fixed acidity] (confidence: 0.468)
[total sulfur dioxide] --> [sulphates] (confidence: 0.470)
[fixed acidity] --> [alcohol] (confidence: 0.471)
[sulphates] --> [total sulfur dioxide] (confidence: 0.473)
[volatile acidity] --> [density] (confidence: 0.473)
[free sulfur dioxide] --> [chlorides] (confidence: 0.474)
[volatile acidity] --> [free sulfur dioxide] (confidence: 0.478)
[sulphates] --> [free sulfur dioxide] (confidence: 0.479)
[pH] --> [total sulfur dioxide] (confidence: 0.480)
[total sulfur dioxide] --> [pH] (confidence: 0.482)
[density] --> [volatile acidity] (confidence: 0.483)
[fixed acidity] --> [sulphates, citric acid] (confidence: 0.484)
[free sulfur dioxide] --> [residual sugar] (confidence: 0.486)
[citric acid] --> [chlorides] (confidence: 0.486)
[chlorides] --> [citric acid] (confidence: 0.487)
[alcohol] --> [fixed acidity] (confidence: 0.487)
[free sulfur dioxide] --> [sulphates] (confidence: 0.490)
[total sulfur dioxide] --> [citric acid] (confidence: 0.491)
[sulphates] --> [fixed acidity, citric acid] (confidence: 0.492)
[residual sugar] --> [volatile acidity] (confidence: 0.493)
[citric acid] --> [fixed acidity, sulphates] (confidence: 0.494)
[citric acid] --> [total sulfur dioxide] (confidence: 0.495)
[residual sugar] --> [fixed acidity, density] (confidence: 0.499)
[free sulfur dioxide] --> [volatile acidity] (confidence: 0.506)
[alcohol] --> [residual sugar] (confidence: 0.507)
[total sulfur dioxide] --> [residual sugar] (confidence: 0.508)
[residual sugar] --> [free sulfur dioxide] (confidence: 0.511)
[volatile acidity] --> [total sulfur dioxide] (confidence: 0.512)
[citric acid] --> [residual sugar] (confidence: 0.512)
[fixed acidity] --> [density, citric acid] (confidence: 0.515)
[density] --> [fixed acidity, citric acid] (confidence: 0.517)
[pH] --> [free sulfur dioxide] (confidence: 0.523)
[citric acid] --> [fixed acidity, density] (confidence: 0.525)
[density] --> [total sulfur dioxide] (confidence: 0.527)
[total sulfur dioxide] --> [volatile acidity] (confidence: 0.527)
```

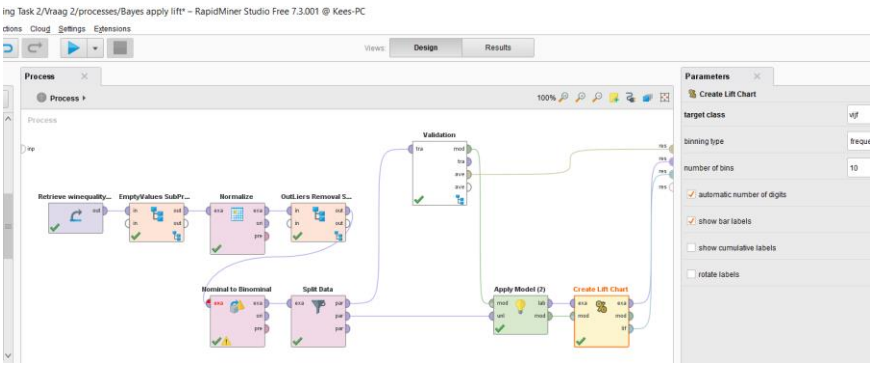
Nb. There are more associationRules.

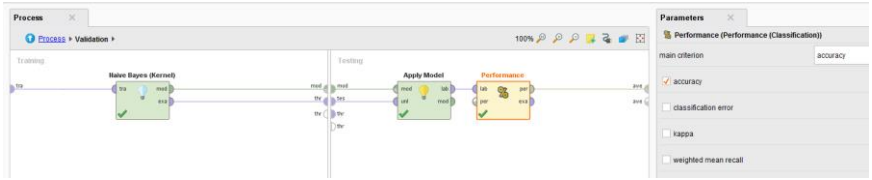
Lets take the following rule.

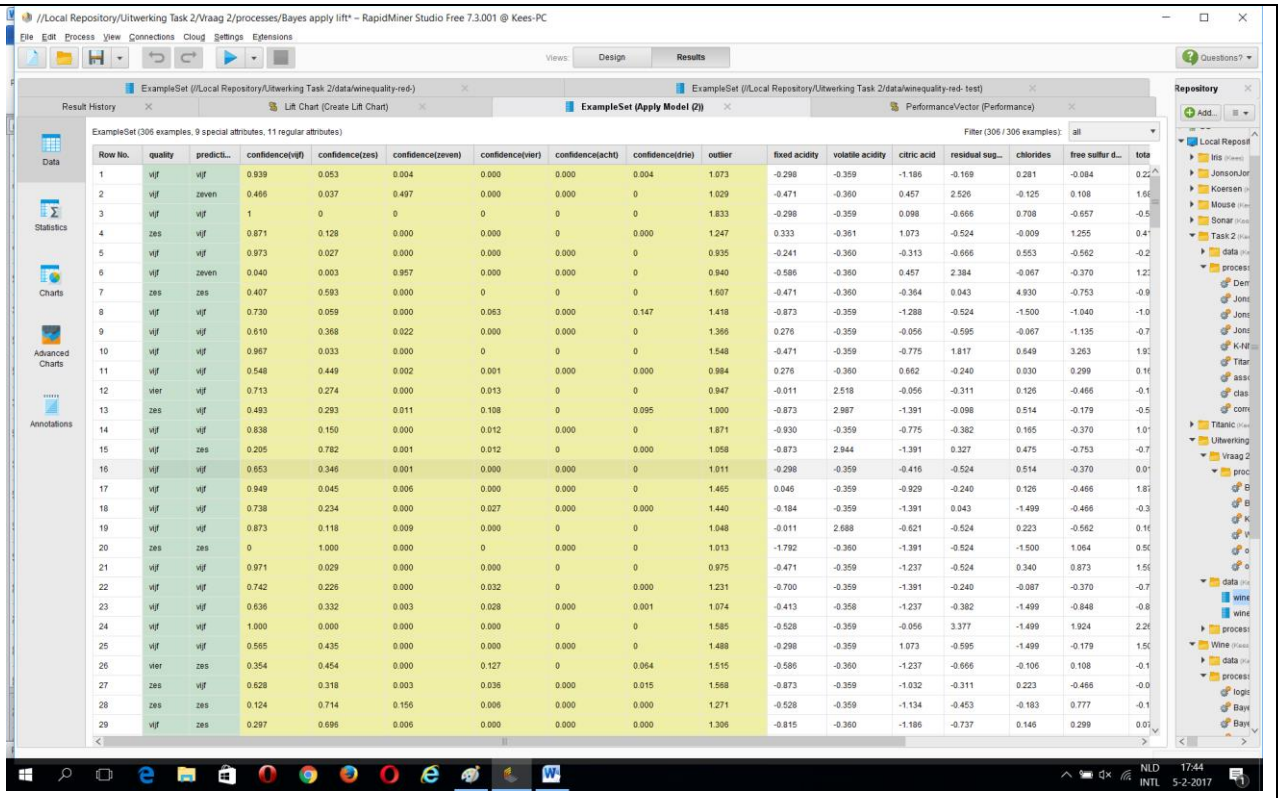
No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convicti...
98	fixed acidity, residual sugar	density	0.223	0.844	0.967	-0.306	0.094	1.730	3.283

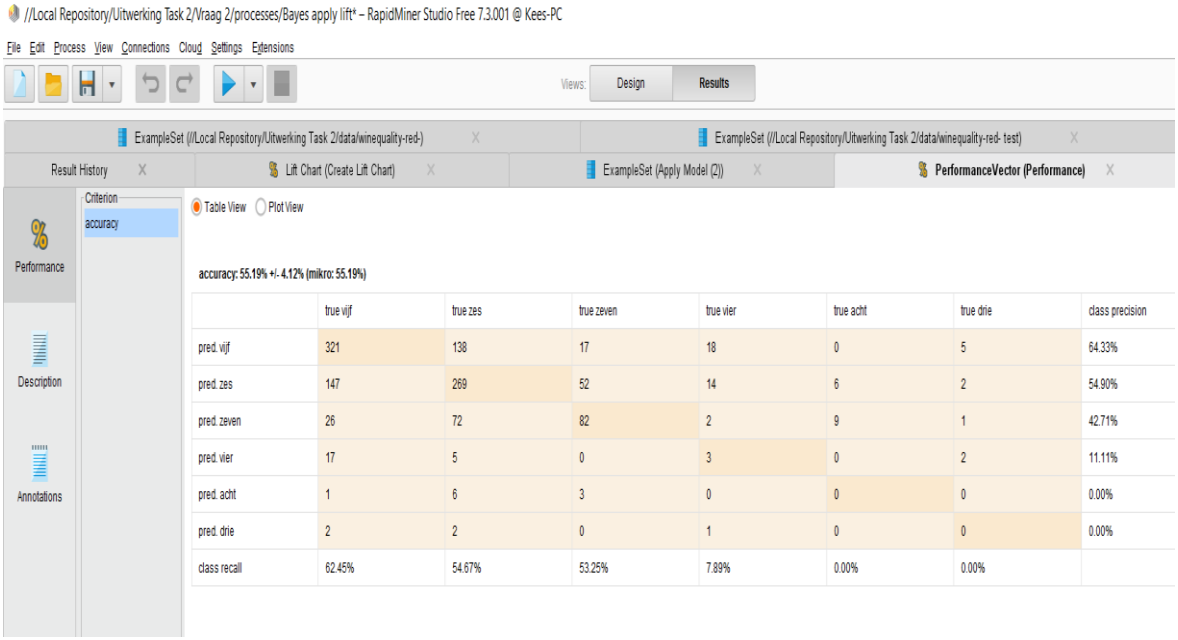
The algoritm discover that wine with fixed acidity and residual sugar, had a support of 22,3 % (0,223), with the support indicating how frequently that particular rule has influence on the density. The confidence is an indicator of how often the rule has been found to be true. The lift of a rule is the ratio of the observed support to that expected if X and Y where independent.

c. Select one of the categorical variables in the dataset and apply a classification algorithm of your choice selected from those studied in the course. Report the confusion matrix obtained and explain the result. Alternatively, take a subset of the continuous attributes in your data set and apply a clustering algorithm, explaining the results (300 words max.). Provide screenshots of the workflows.

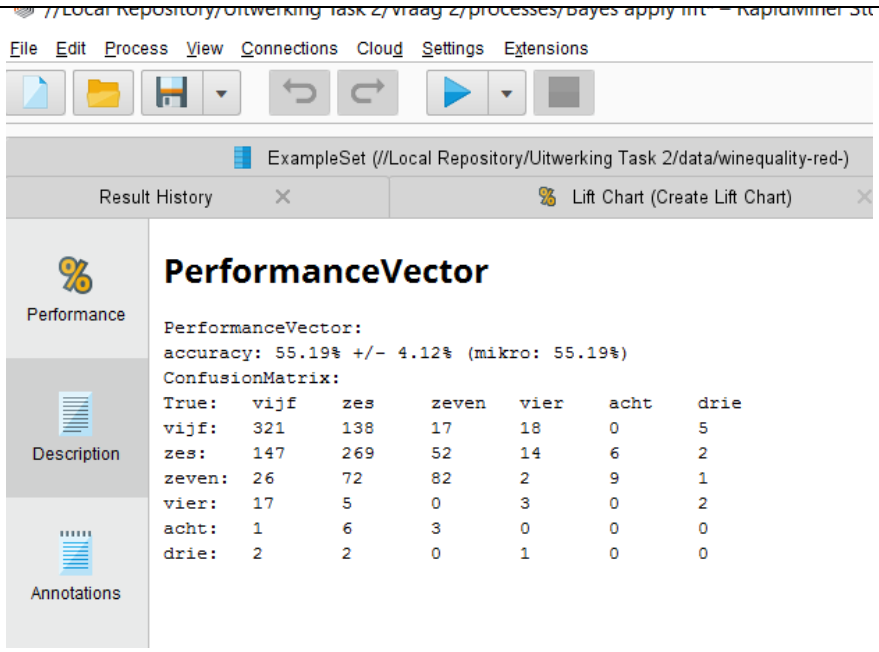


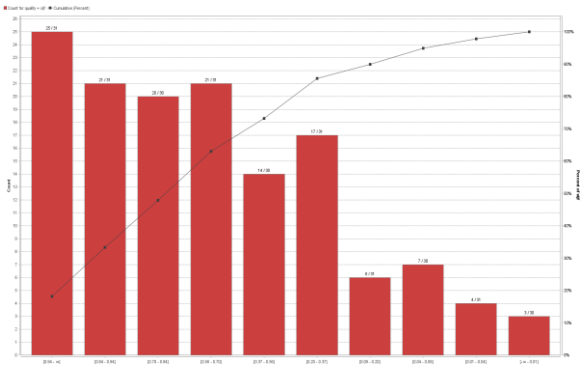






	true vijf	true zes	true zeven	true vier	true acht	true drie	class precision
pred. vijf	321	138	17	18	0	5	64.33%
pred. zes	147	269	52	14	6	2	54.90%
pred. zeven	26	72	82	2	9	1	42.71%
pred. vier	17	5	0	3	0	2	11.11%
pred. acht	1	6	3	0	0	0	0.00%
pred. drie	2	2	0	1	0	0	0.00%
class recall	62.45%	54.67%	53.25%	7.89%	0.00%	0.00%	





A confusion matrix illustrates the accuracy of the solution to a classification problem.

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The table shows the confusion matrix for a class classifier.

The entries in the confusion matrix have the following meaning:

1. a is the number of correct predictions that an instance is negative,

2. b is the number of incorrect predictions that an instance is positive,

3. c is the number of incorrect of predictions that an instance negative, and

4. d is the number of correct predictions that an instances positive [9].

Some standards and terms:

1. True positive (TP): If the outcome from a prediction is p and the actual value is also p,

then it is called a true positive.

2. False positive (FP): However if the actual value is n then it is said to be a false positive

3. Precision and recall: Precision is the fraction of retrieved instances that are relevant, whilerecall is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance. Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. Recall is nothing but the

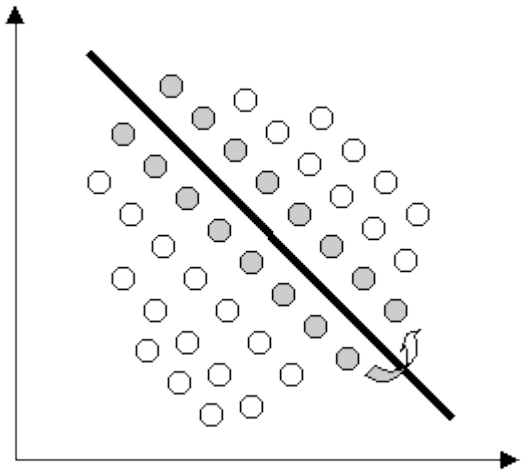
In this case the prediction of vijf and six are good. Zeven matig and the prediction of four, acht and seven are bad.

5. Data analysis Part 2.
If in answering Question 3c. you applied a clustering (classification) algorithm, now apply a classification (clustering) algorithm and explain the results (400 words max.). Provide screenshots of the RapidMiner workflows.

Iám a bit unclear about this question. Clustering and Clasification are two different methods of machine learning and two distinct groups of algoritm. For example a clustering algoritm could use k-means to segment the database, whereas K-nn algorithm can be used for a classification task. We can build a hybrid of clustering and classification but the two are distinct type of analysis.

There I have in used a have used a classification algoritm, I supposed now to use clustering.

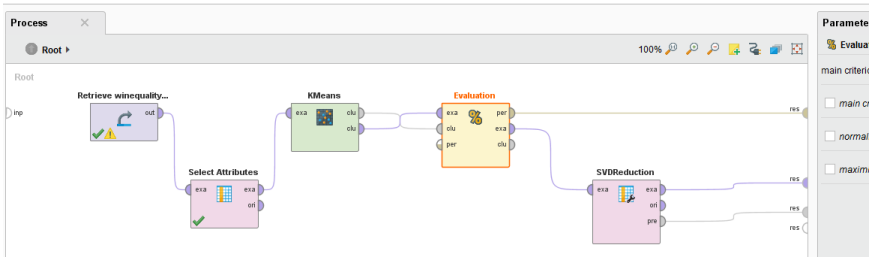
Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.
A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".
A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

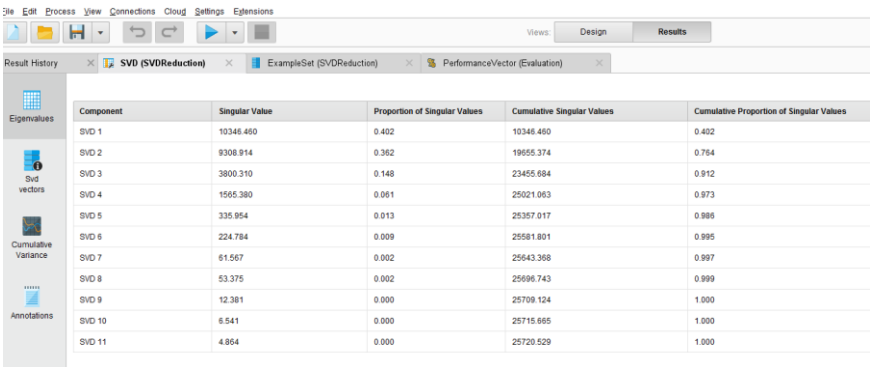


The Goals of Clustering
So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to

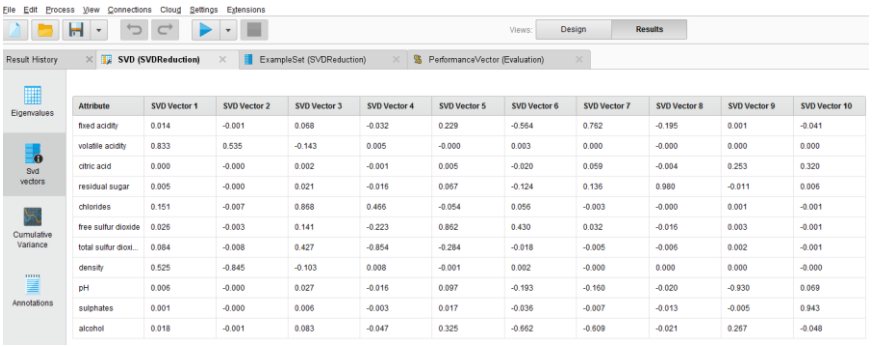
decide what constitutes a good clustering? It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances. However, even in this case the Euclidean distance can sometimes be misleading. Figure shown below illustrates this with an example of the width and height measurements of an object. Despite both measurements being taken in the same physical units, an informed decision has to be made as to the relative scaling

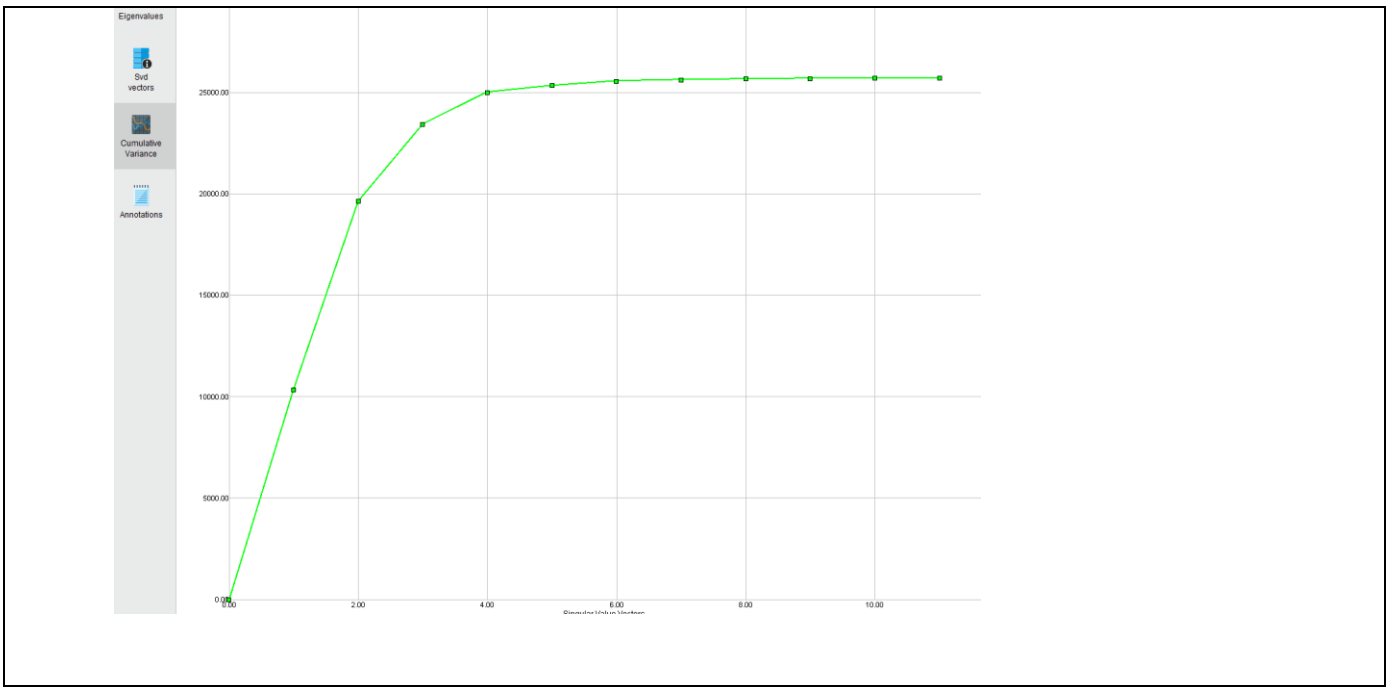


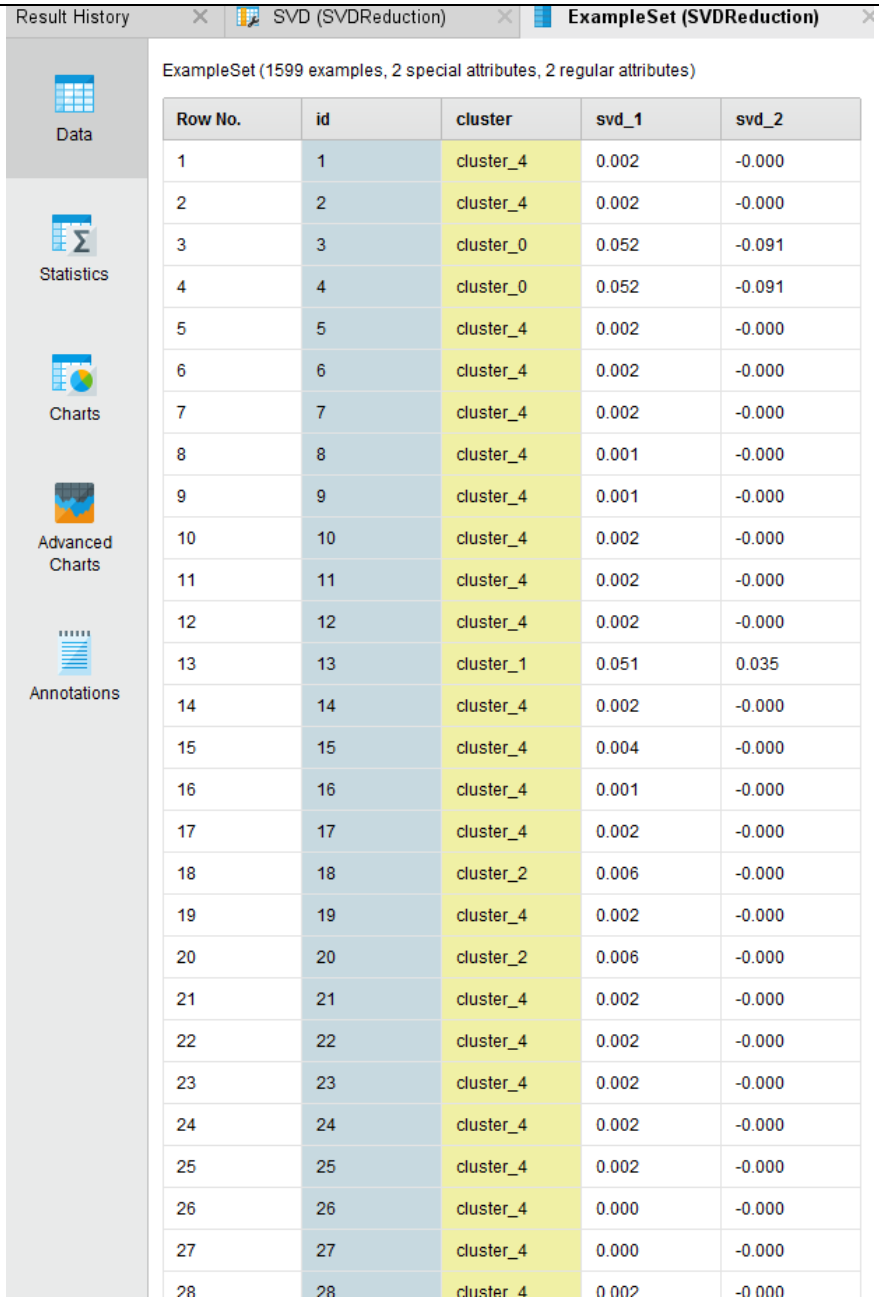


Component	Singular Value	Proportion of Singular Values	Cumulative Singular Values	Cumulative Proportion of Singular Values
SVD 1	10346.460	0.402	10346.460	0.402
SVD 2	9308.914	0.362	19655.374	0.764
SVD 3	3800.310	0.148	23455.684	0.912
SVD 4	1565.380	0.061	25021.063	0.973
SVD 5	335.954	0.013	25357.017	0.986
SVD 6	224.784	0.009	25581.801	0.995
SVD 7	61.567	0.002	25643.368	0.997
SVD 8	53.375	0.002	25696.743	0.999
SVD 9	12.381	0.000	25709.124	1.000
SVD 10	6.541	0.000	25715.665	1.000
SVD 11	4.864	0.000	25720.529	1.000

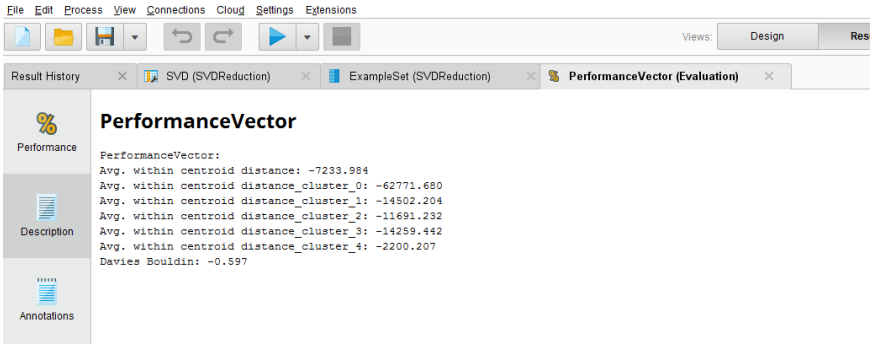


Attribute	SVD Vector 1	SVD Vector 2	SVD Vector 3	SVD Vector 4	SVD Vector 5	SVD Vector 6	SVD Vector 7	SVD Vector 8	SVD Vector 9	SVD Vector 10
fixed acidity	0.014	-0.001	0.068	-0.032	0.229	-0.564	0.762	-0.195	0.001	-0.041
volatile acidity	0.833	0.535	-0.143	0.005	-0.000	0.003	0.000	-0.000	0.000	0.000
citric acid	0.000	-0.000	0.002	-0.001	0.005	-0.020	0.059	-0.004	0.253	0.320
residual sugar	0.005	-0.000	0.021	-0.016	0.067	-0.124	0.136	0.980	-0.011	0.006
chlorides	0.151	-0.007	0.868	0.466	-0.054	0.056	-0.003	-0.000	0.001	-0.001
free sulfur dioxide	0.026	-0.003	0.141	-0.223	0.862	0.430	0.032	-0.016	0.003	-0.001
total sulfur dioxi...	0.084	-0.008	0.427	-0.854	-0.284	-0.018	-0.005	-0.006	0.002	-0.001
density	0.525	-0.845	-0.103	0.008	-0.001	0.002	-0.000	0.000	0.000	-0.000
pH	0.006	-0.000	0.027	-0.016	0.097	-0.193	-0.160	-0.020	-0.930	0.069
sulphates	0.001	-0.000	0.006	-0.003	0.017	-0.036	-0.007	-0.013	-0.005	0.943
alcohol	0.018	-0.001	0.083	-0.047	0.325	-0.662	-0.609	-0.021	0.267	-0.048


Eigenvalues
Svd vectors
Cumulative Variance
Annotations
25000.00
20000.00
15000.00
10000.00
5000.00
0.00
0.00
2.00
4.00
6.00
8.00
10.00



Row No.	id	cluster	svd_1	svd_2
1	1	cluster_4	0.002	-0.000
2	2	cluster_4	0.002	-0.000
3	3	cluster_0	0.052	-0.091
4	4	cluster_0	0.052	-0.091
5	5	cluster_4	0.002	-0.000
6	6	cluster_4	0.002	-0.000
7	7	cluster_4	0.002	-0.000
8	8	cluster_4	0.001	-0.000
9	9	cluster_4	0.001	-0.000
10	10	cluster_4	0.002	-0.000
11	11	cluster_4	0.002	-0.000
12	12	cluster_4	0.002	-0.000
13	13	cluster_1	0.051	0.035
14	14	cluster_4	0.002	-0.000
15	15	cluster_4	0.004	-0.000
16	16	cluster_4	0.001	-0.000
17	17	cluster_4	0.002	-0.000
18	18	cluster_2	0.006	-0.000
19	19	cluster_4	0.002	-0.000
20	20	cluster_2	0.006	-0.000
21	21	cluster_4	0.002	-0.000
22	22	cluster_4	0.002	-0.000
23	23	cluster_4	0.002	-0.000
24	24	cluster_4	0.002	-0.000
25	25	cluster_4	0.002	-0.000
26	26	cluster_4	0.000	-0.000
27	27	cluster_4	0.000	-0.000
28	28	cluster_4	0.002	-0.000



Task 2 1 of 1 Ronald koeman