CISC251 Data Analysis Final Report

Alex Panfilov

Table of Contents

Intro & Purpose

Source: CISC251 project.html

Drug producers and wholesalers add extre chemicals to their product to increase their volume and so make larger profits. (Retail drug dealers do this too, but that's not our concern here.) As drugs pass through the pipeline from drug producing countries (South America, Central Asia) to Western countries, adulterants can be added at several stages. The pattern of these adulterants behaves like a kind of fingerprint, capturing the provenance of each shipment.

This can be used by law enforcement to understand the way in which drugs are handled. If multiple shipments all have the same pattern of adulterants, then they were presumably sourced from the same place. If all the patterns are different, then the pipelines are complex and overlapping.

This dataset describes the chemical composition of the adulterants added to drug shipments that were interdicted by Customs. Each row of the dataset results from the chemical composition of a single shipment. Each column corresponds to one adulterant whose presence is checked for by Customs.

The primary goal of the analytics is:

- To cluster the samples according to their adulterant pattern, and see what this reveals about the possible pathways by which the drugs arrived at this country's border.
- To reinforce any conclusions drawn from these clusterings by trying to predict shipment label from adulterant profile.

Properties of the Dataset

Initial Analysis of Chemical Presence

From looking at the upper and lower bounds of each chemical, the following was observed:

The following 7 chemicals were not present in any of the shipments:

- Dextromorphan
- Theophylline
- Creatinine
- Ketamine
- Quinine
- Strychnine
- Thebaine

At first, I considered disincluding these chemicals from further analysis by filtering out those columns in KNIME. However, I decided not to do so, because I wanted to build predictors which would be able to analyze origins of other shipments too. Removing these chemicals would result in clustering algorithms and predictors potentially making wrong predictions when ignoring these values if they are non-zero for some other shipments.

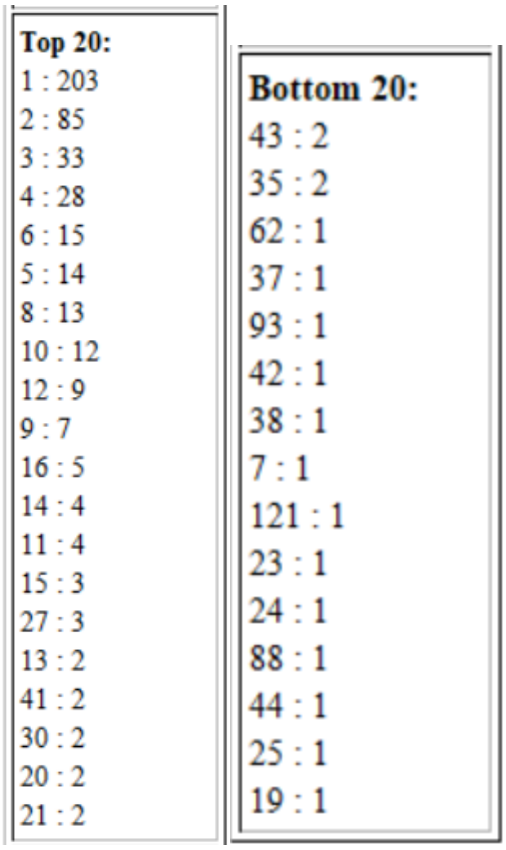
Hence out of 57 chemicals, 7 of them are not present in any of the shipments.

One chemical was present in all shipments in some amount: Truxilline.

Seizure Names

Upon trying to set the seizure names as the row ids in KNIME, I discovered that some of the seizure names were repeated. I used the Value Counter to count the number of times each of the Seizure Names were repeated. For a total of 2474 shipping records, there were 465 unique Seizure Names. The amount of records with identical Seizure Names ranged from 1 to 121.

From the Statistics View, I obtained the mean, which was 5.3204, and the standard deviation, which was 10.9053. The large difference between the maximum value, 121, and the mean, as well as the large standard deviation implied that most of the seizure names actually only repeated once or twice. Upon examining the distributions of values, this was confirmed: For 203 of the 465 unique Seizure Names (43.66%), there was only 1 shipment with that Seizure Name. For 85 of the Seizure Names (18.28%), there were 2 shipments with that Seizure Name. Below is a diagram detailing the distribution of the amount of shipments per Seizure Name.



Note that the "Bottom 20" of the Statistics View only contains 15 items. This is because there were only 35 different amounts of shipments with the same seizure name. Furthermore, 78% of the seizure names had 5 or less shipment records in the dataset.

However, it will be interesting to see if the remaining 22% of Seizure Names that have multiple shipments with that name will result in clustering. A strong correlation between the Seizure Names and the adulterant pattern would imply that the drugs seized in the same investigation have come from the same place, whereas a weak correlation or no correlation would imply that the drugs seized in the same instance were from different places.

Clustering

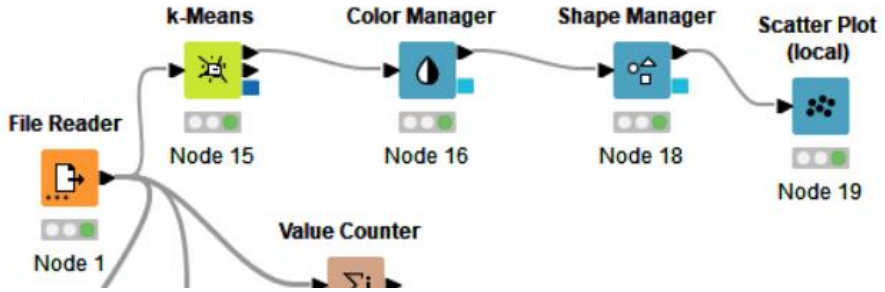
The purpose of this step is to find similarity between the records. First, this requires the decision of what qualifies similarity.

K-means Clustering

I started with k-means clustering because it would consider all attributes when evaluating similarity. I wanted to find what kind of attributes would have a big impact on which cluster a datapoint would end up in, and which had less impact. I would then be able to compare these findings to other clusterings to note similarities and draw overall conclusions.

I used 5 clusters because it resulted in roughly the same number of datapoints per cluster, which was interesting. Visualizing the clusters onto a colour-coded, shape-coded scatter plot and going through the

different options for the x and y axis resulted in some interesting findings. Here is the KNIME workflow I used:



For the scatterplot analysis, I will use the following legend:

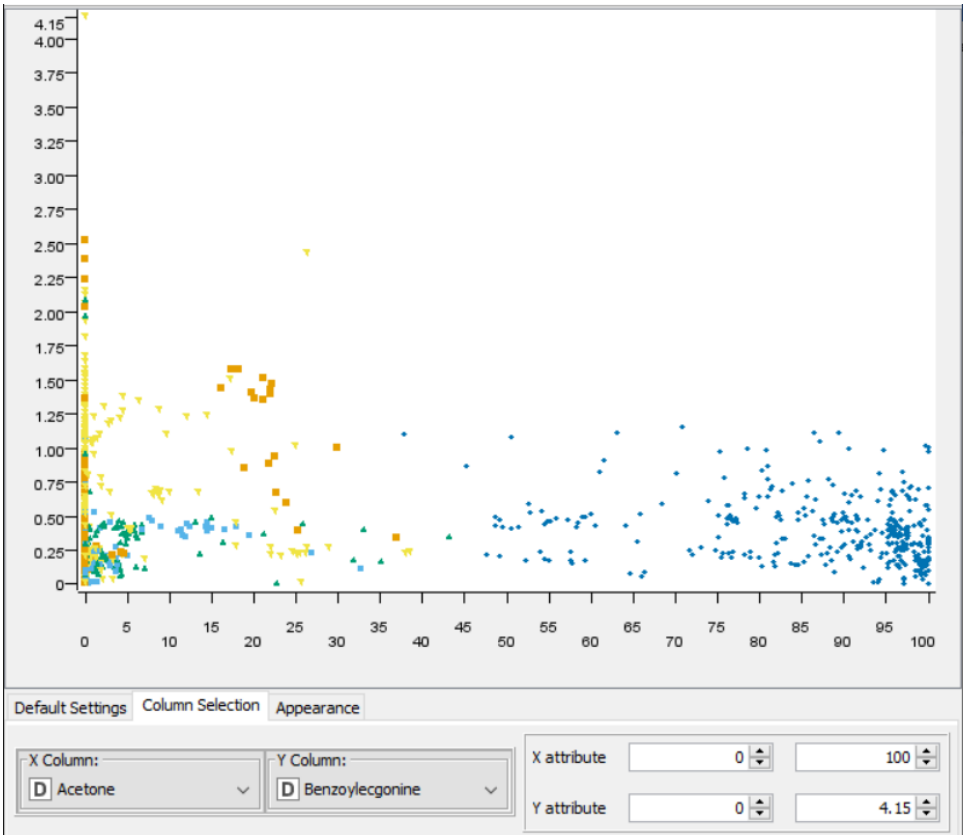
Cluster 1 – Orange Rectangle (490 items total)

Cluster 2 – Blue Circle (587 items total)

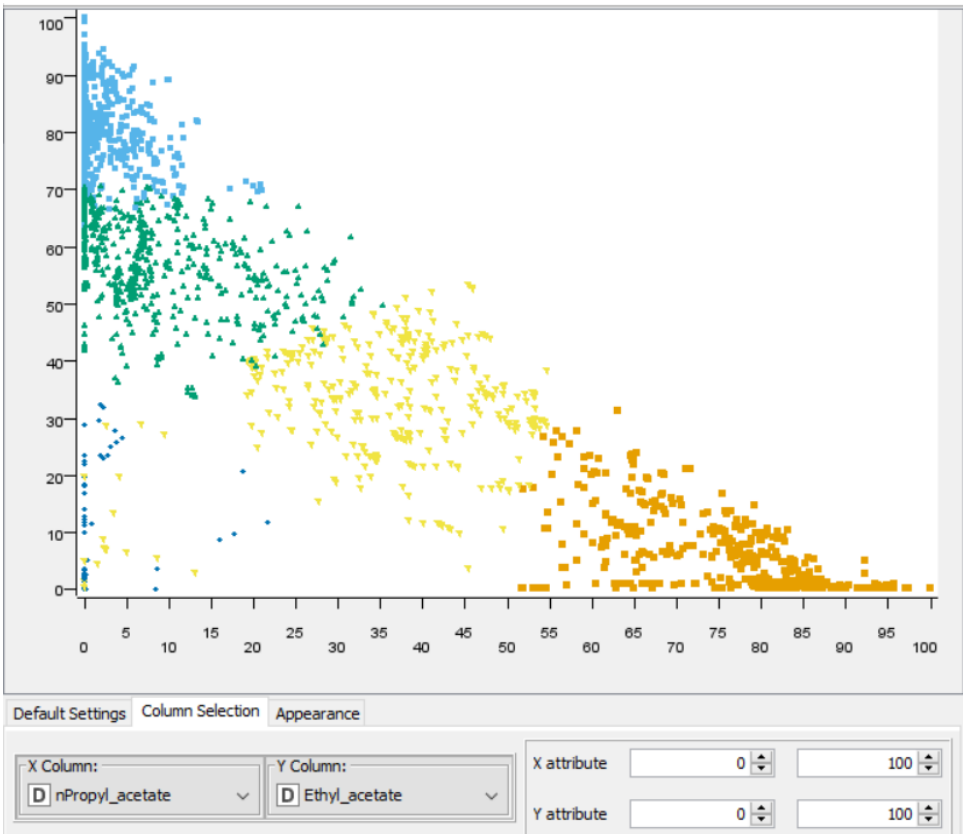
Cluster 3 – Green Triangle (539 items total)

Cluster 4 – Yellow Reverse Triangle (505 items total)

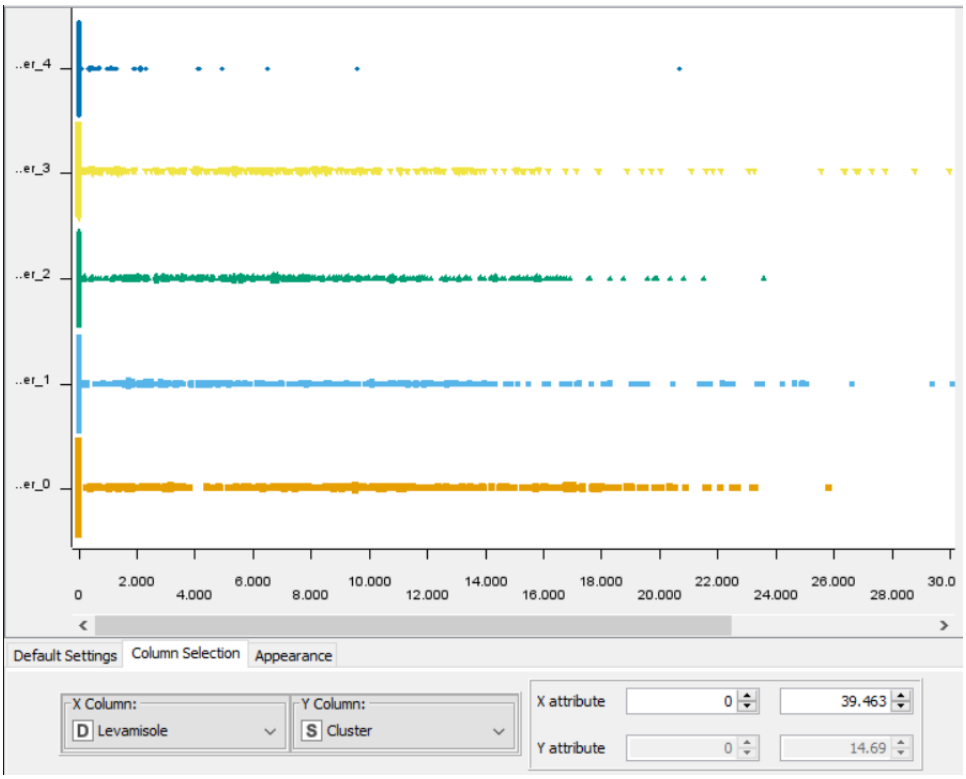
Cluster 5 – Navy Diamond (353 items total)



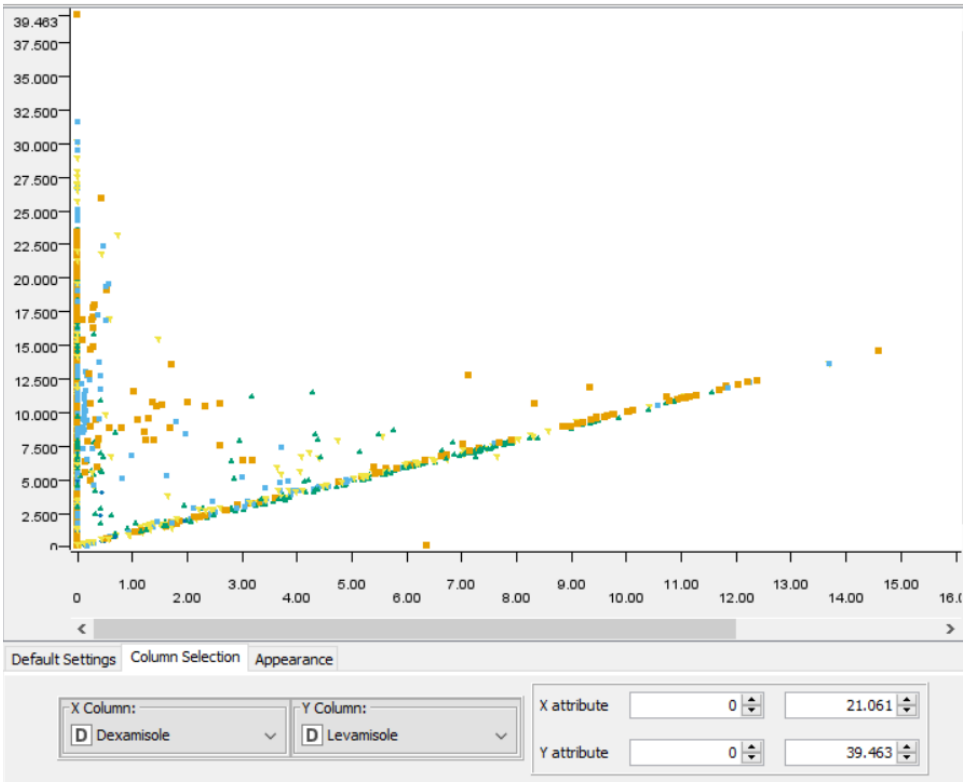
In the graph above, Acetone is the X axis, and almost the entirety of Cluster 5 has values above 45, whereas all the other clusters are almost entirely below 45. The graph shows that cluster 5 is differentiated by a high acetone amount in the shipment.



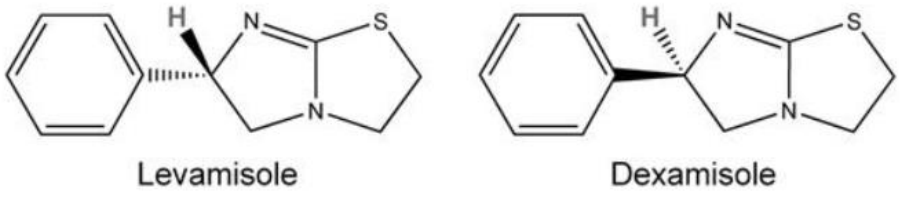
This has n-propyl acetate as the X axis and ethyl acetate as the Y axis. There is a negative linear correlation in the data points, where high n-propyl acetate correlates to low ethyl acetate, and low n-propyl acetate correlates to high ethyl acetate. Furthermore, the clusters are clearly distinguishable in their characteristics: Cluster 2 has low ethyl acetate and high n-propyl acetate, cluster 3 also has low ethyl acetate but the n-propyl acetate is lower than in cluster 2. Cluster 4 has mid to low ranges of both n-propyl acetate and ethyl acetate. Cluster 1 has high ethyl acetate and low n-propyl acetate. Cluster 5 has both low ethyl acetate and low n-propyl acetate.



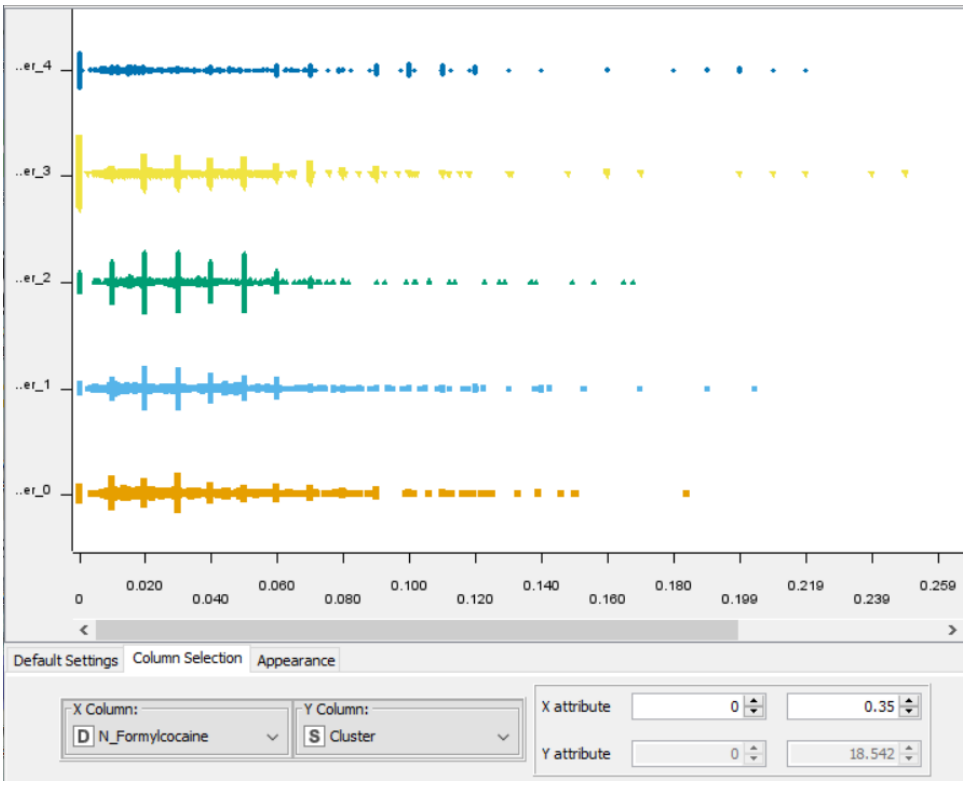
There are also chemicals that do not have much of an impact on this clustering. An example is Levamisole, as the graph above shows that most of the datapoints in each cluster have close to zero levamisole. The datapoints which are non-zero are mostly evenly distributed in clusters 1-4, while cluster 5 has less datapoints that have nonzero levamisole.



This graph has Dexamisole in the X axis and Levamisole in the Y axis. Although there are no clear-cut patterns related to the 5 clusters that k-means algorithm produced, there is a very strong linear pattern visible. Most of the datapoints fall into one of 2 categories: One, they have close to zero dexamisole and some amount of levamisole, which can range from close to zero to about 40% (mostly the datatpoints have under 25% levamisole). Two, they have a very strong linear correlation between dexamisole and levamisole that looks to be 1:1. Looking into this, I found that levamisole and dexamisole are often both used as an adulterant for cocaine (1, 2). Furthermore, sometimes tests fail in differentiating between the two (1). Below are the structure diagrams for both chemicals. They are very structurally similar.



Source (1)



This graph has n-formyl cocaine in the X axis and clusters in the Y axis. All of the clusters have spikes at certain quantities, which is very interesting. I was not able to find any other chemical that this correlated to.

Based on the scatterplots, these are the chemicals that made a big difference in clusterings:

- Acetone
- Methylisobutenylketone (lots of non-zero values only in cluster 5)
- Toulene (some different distributions)
- MEK (very little non-zero MEK in cluster 5)
- Mesitylene (lots of non-zero only in cluster 5)
- Istobutyl acetate (very little non-zero in cluster 5)
- Benzene (very little non-zero in cluster 5)
- Tropa cocaine (very little non-zero in cluster 5)
- Trimethoxycocaine (very little non-zero in cluster 5)
- Truxiline (very little non-zero in cluster 5)
- Trans-cinnamoyl cocaine
- Cis- cinnamoyl cocaine
- Norcocaine
- Ecgonine methyl ester (a significant amount of non-zeroes only in cluster 1)
- n-propyl acetate and ethyl acetate
- n-formyl cocaine

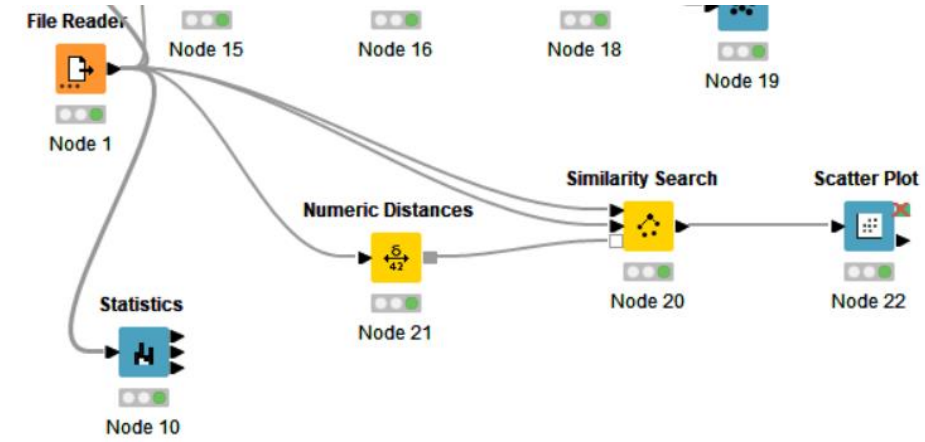
Based on the scatterplots, these are the chemicals that did not make a big difference in clusterings:

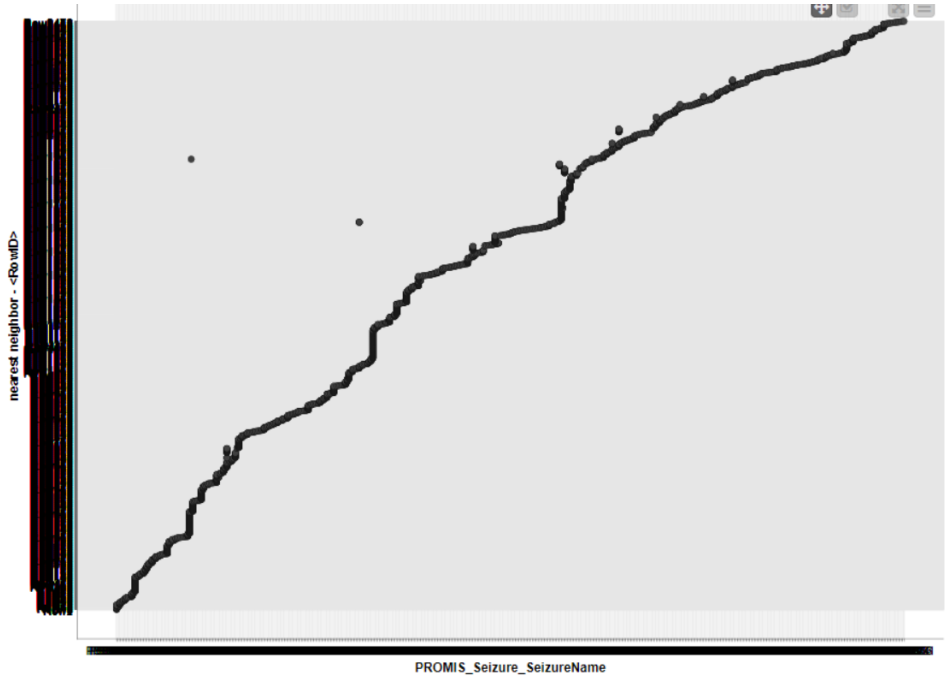
- Levamisole, Dexamisole
- The 8 sugars
- The 7 chemicals which are zero for all shipments
- Hydroxyzine
- Diltiazem
- Creatine
- Paracetamol
- Phenacetin
- Phenobarbitol
- Caffeine
- Benzocaine
- Aspirin
- Methylbenzoate
- O Xylene
- MP Xylene
- Methylene Chloride
- MIBK
- Nicotinamide
- Procaine
- Lignocaine
- Dexamisole (barely any in cluster 5 though)
- Methyl acetate
- Hexane
- Chloroform
- Acetonitrile
- Benzoylecgoine
- Ecgonine

Euclidian Distance Similarity Search

Next I decided to use Euclidian Distance similarity search to see if I would be able to find anything in regards to the seizure names similarity. Euclidian distance seemed appropriate because it assumes all similarities in different attributes are equally important, and there is not enough evidence about if different attributes have more or less impact on similarity. In addition, Euclidian distance places a lot of emphasis on attributes which are more numerically different. This seems like a logical thing to do for this dataset, since more difference in attributes likely would imply different origins; big differences in the adulterant composition likely means the chemicals had different pathways. However, this poses the danger that it will imply big difference in pathways if different values, which may not be true.

Here is the KNIME workspace I used, and the resulting scatterplot I got:





This graph has Seizure Names in the X axis and the nearest neighbour row ID in the Y axis. This graph is not perfectly linear, which is expected since some of the x values would have multiple y values (since some seizure names had multiple datapoints). However, there was nothing else that I could extract from this data so I stopped this path of analysis.

Expectation Maximization Clustering

I decided to try clustering algorithm because there were likely to be distributions for the amounts of adulterant in shipments that the data followed. I thought that this technique would be able to reveal more patterns about how different chemicals impacted shipment. The clustering groups were very different from the k-means clustering. I decided to use 5 clusters as well to compare to the k-means clustering, which resulted in a distribution of:

Cluster 1 – 270 items (red)

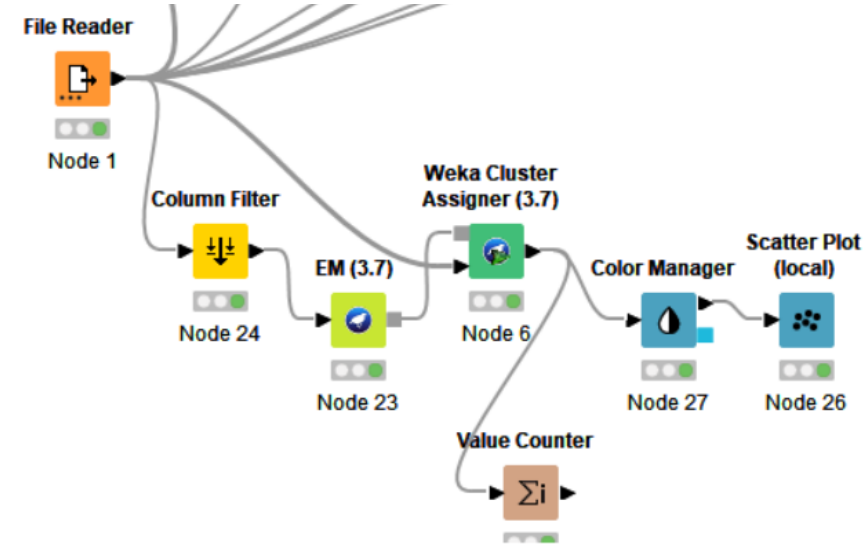
Cluster 2 – 364 items (maroon)

Cluster 3 – 1274 items (purple)

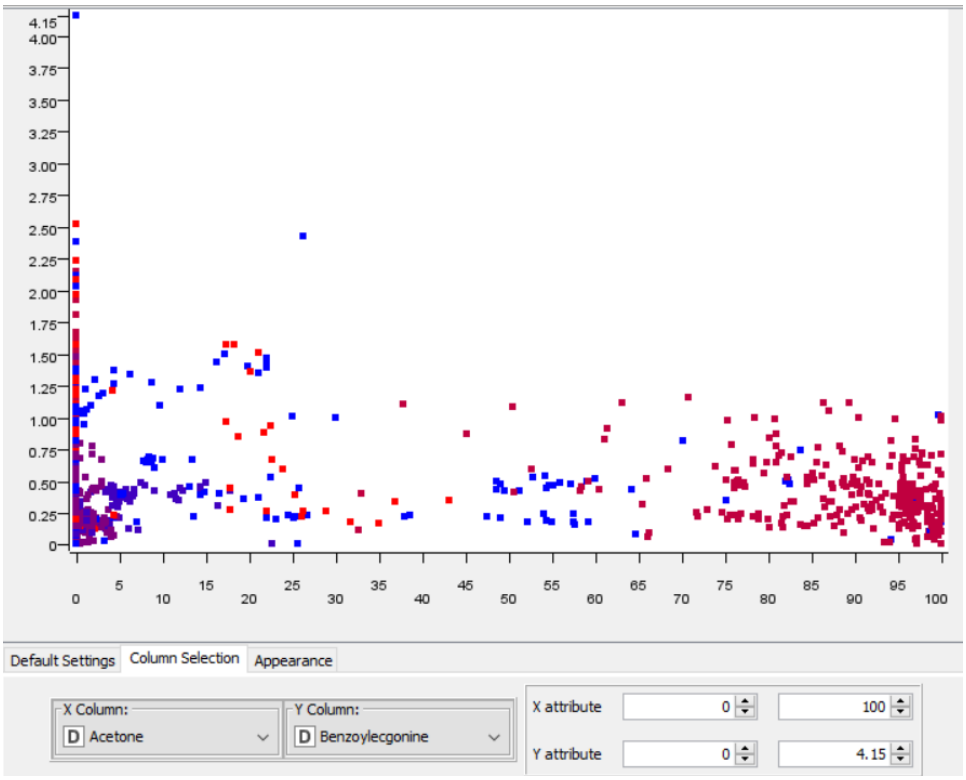
Cluster 4 – 362 (navy)

Cluster 5 – 204 (blue)

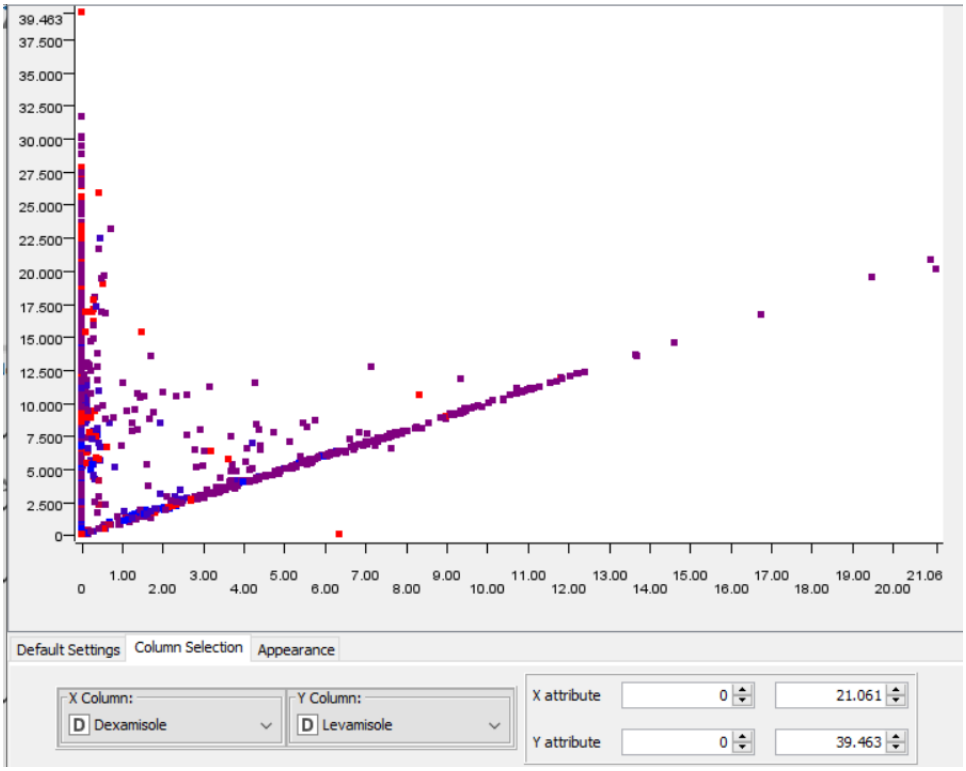
Here is my KNIME workflow:



Here are some interesting findings:



In the graph above, Acetone is the X axis and Benzolygonine is in the Y axis. This is the same graph as in the k-means clustering section, but using EM clustering. It is visible that one cluster has most of its datapoints above 60, whereas the other clusters are almost entirely below 60. In the k-means clustering, the split was closer to 45, whereas in EM clustering it is closer to 60. However, there once again is a cluster that is differentiated by a high acetone amount in the shipment, hence I would deduce that Acetone is a big telling point about the adulterant composition and process.



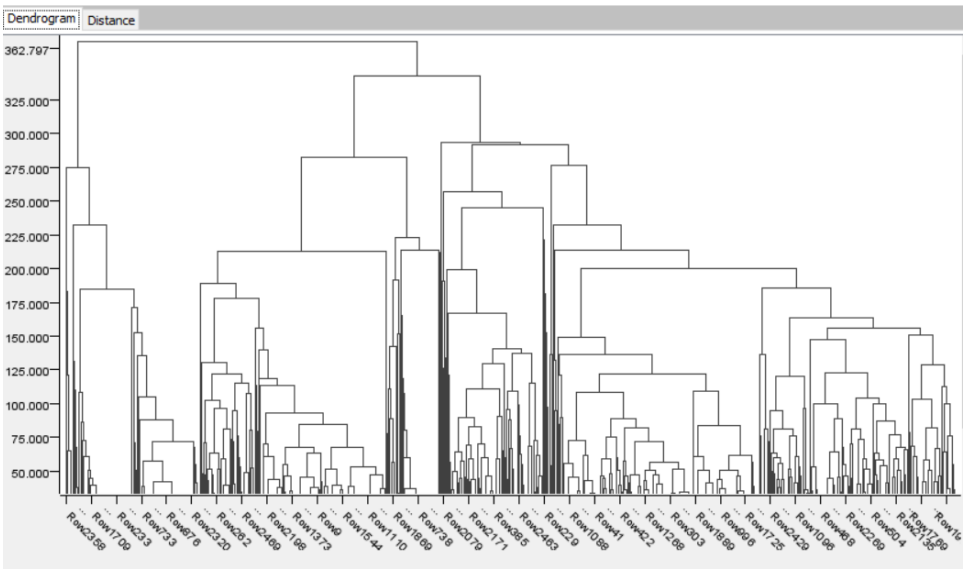
This graph has Dexamisole in the X axis and Levamisole in the Y axis, as there was in the k-means clustering. This time, most of the data points in the linear part of the graph are from the same cluster 3. However, cluster 3 contains 51.5% of all the data points, hence it is not very telling.

Hierarchical Clustering

I decided to use hierarchical clustering, particularly the bottom up approach because of the earlier observations about seizure names. I wanted to explore whether the datapoints with the same seizure names would end up in the same clusters.

My prediction was correct and the data points with the same seizure names really did end up in the same clusters. I used 20 clusters, which was a lot less than there were unique seizure names. However, the items with the same seizure names were in the same clusters. Furthermore, they would generally be grouped together within the clusters.

Here is the dendrogram produced in the prediction:



Predictors

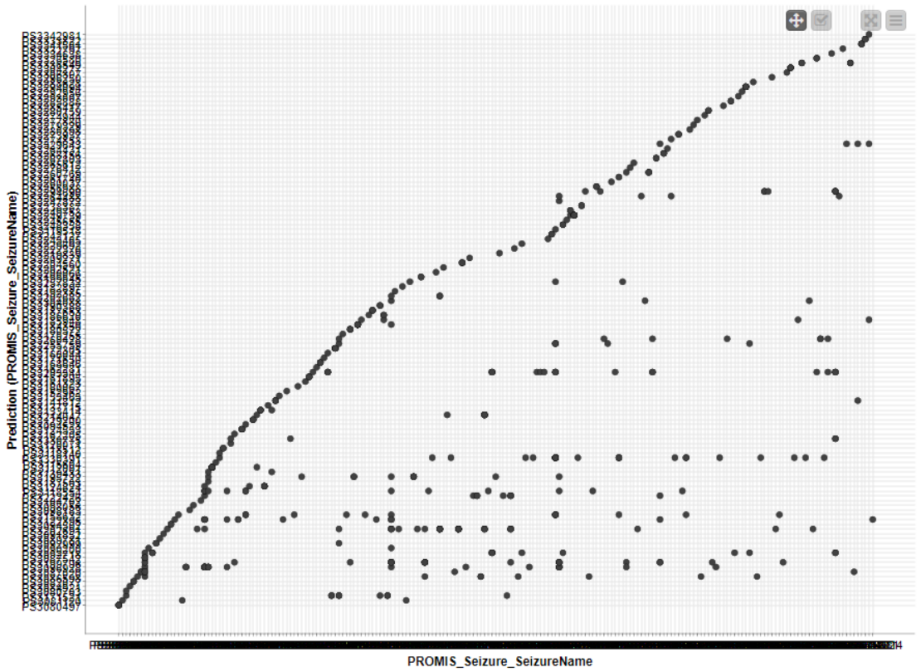
Bayesian Predictor

The Bayesian Predictor uses all the attributes available to make the prediction, hence I decided to use it to make a model that would predict the seizure name of a shipment. This realistically was a very difficult task since a lot of the seizure names only had 1 or 2 records. The results were not very accurate, as expected. The predictor had an accuracy of 48.081%. However, this is better accuracy than simply guessing at random for each of the 465 unique seizure names, the accuracy of which would have been 1/465.

Here is the confusion matrix:

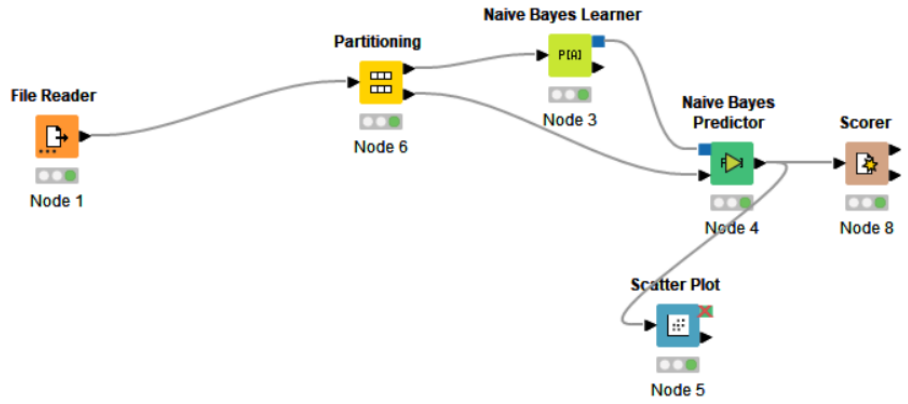
Correct classified: 238	Wrong classified: 257
Accuracy: 48.081%	Error: 51.919%
Cohen's kappa (κ): 0.476%	

Here is the scatterplot of prediction versus the real Seizure Name:



The linear pattern supports the clustering findings from above that showed that the shipments had very similar chemical compositions.

Here is my KNIME workflow:



Using the Bayesian predictor on the k-means clustering also was not very effective. Here is the confusion matrix:

Cluster \P...	cluster_0	cluster_1	cluster_2	cluster_4	cluster_3
cluster_0	75	14	0	0	0
cluster_1	0	123	0	3	0
cluster_2	7	90	3	2	0
cluster_4	0	0	0	73	0
cluster_3	19	48	0	38	0

Correct classified: 274 Wrong classified: 221

Accuracy: 55.354% Error: 44.646%

Cohen's kappa (κ): 0.432%

Random Forest

I used the random forest predictor on the k-means clustering because the model would show how any of the data points can be assigned to a cluster. This would allow for further analysis of what chemicals make a bigger difference.

The model was very accurate. Here is the confusion matrix:

Cluster \P...	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
cluster_0	111	0	0	2	0
cluster_1	0	107	0	0	0
cluster_2	0	3	90	2	1
cluster_3	0	0	1	104	0
cluster_4	0	0	0	0	74

Correct classified: 486 Wrong classified: 9

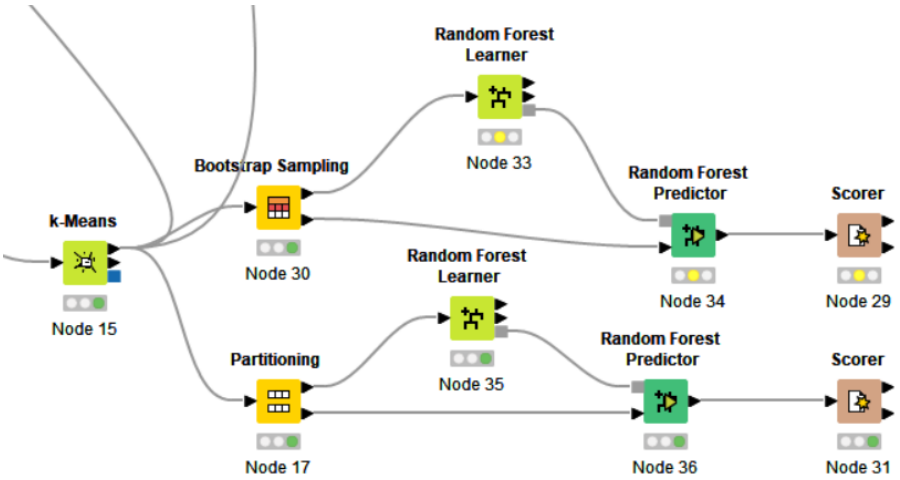
Accuracy: 98.182% Error: 1.818%

Cohen's kappa (κ): 0.977%

The decision tree learner statistics showed that the following chemicals made a big impact on the clustering divisions:

- Acetone
- n-propyl acetate
- Ethyl acetate
- Trixilline
- Methylisobutenylketone

Here is my KNIME workflow:



Conclusions

The analysis of the dataset showed that the following chemical amounts likely distinguish between different shipment paths:

- Acetone
- n-propyl acetate
- Ethyl acetate
- Trixilline
- Methylisobutenylketone

The shipments which were seized in the same instance had very similar adulterants, meaning that they travelled through the same pipeline of shipments.

The analysis performed on the clustering of Dexamisole and Levamisole showed that they were oftentimes both be added to a chemical. Certain sources (1) also suggest that it is possible that the chemical analysis for them was not accurate, and they could be confused for the other.

There were also chemicals that revealed very minimal information about the chemicals.

Citations

1. John F. Casale, Valerie L. Colley, Donald F. LeGatt, Determination of Phenyltetrahydroimidazothiazole Enantiomers (Levamisole/Dexamisole) in Illicit Cocaine Seizures and in the Urine of Cocaine Abusers via Chiral Capillary Gas Chromatography–Flame-Ionization Detection: Clinical and Forensic Perspectives, *Journal of Analytical Toxicology*, Volume 36, Issue 2, March 2012, Pages 130–135, https://doi.org/10.1093/jat/bkr025
2. Madry MM, Kraemer T, Baumgartner MR. Cocaine adulteration with the anthelminthic tetramisole (levamisole/dexamisole): Long-term monitoring of its intake by chiral LC-MS/MS

analysis of cocaine-positive hair samples. Drug Test Anal. 2019 Mar;11(3):472-478. doi: 10.1002/dta.2505. Epub 2018 Oct 17. PMID: 30239147.