Analysis of Seized Drug Samples

CISC-235 Final Project

Manas Srinivasaiah 20107447

Introduction:

Illicit drug shipments that are seized at the border are often sent for further analysis. This analysis reveals unique chemical compositions for each sample (chemical fingerprints). This gives clues to where a substance may originate from or how illicit substances are changing chemically overtime. The information can then be used to inform public policy, aid harm reduction efforts, and keep law enforcement informed.

The focus of this report is to look specifically at the chemical fingerprints of seized shipments to understand if anything can be learned as to their origin. Shipments that all share the same pattern of adulterants can then be assumed to come from the same place or passed through a shared point. By analyzing a breadth of shipments we can get information as to the different pipelines and routes the drugs are entering the country from.

The first part of the report will focus on using different clustering techniques. These techniques will be focused on the adulterant chemical profile of each sample from each shipment.

To understand to how effective the clustering process was, the second part of the report will use these clustering classifications in prediction models. For each substance, the amount of each chemical compound present as well as its cluster identification will be used as attribute fields in trying to predict the drug sample's associated shipment.

Overview and Preliminary Analysis:

Data Description:
A list of drug samples. Each sample is associated with a an overall shipment seizure. Each sample has a list of chemicals that were present in the sample and at what amounts.

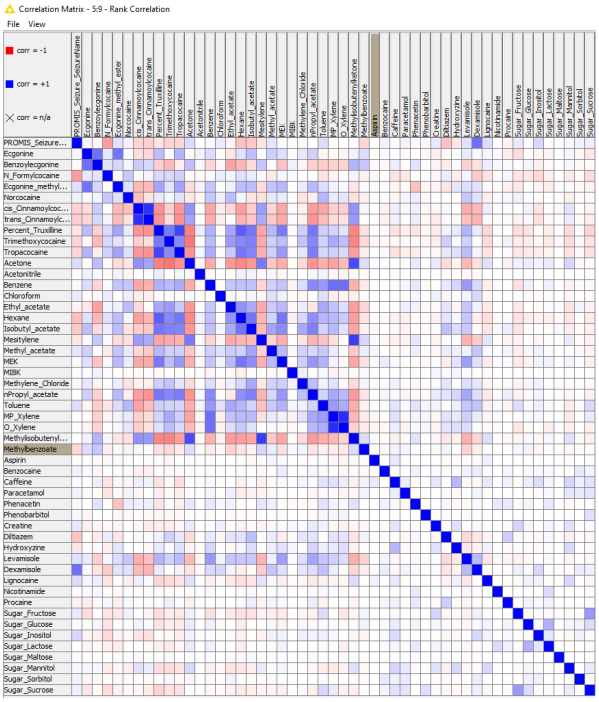
Data Preprocessing/Overview:

The data provided is a sample from 465 unique seizures/shipments. For 195 seizures, only one sample was recovered. There are no missing values, unless those singleton shipment seizures had additional samples that were not provided. I did realize that 7 chemical compounds were not present in any of the collected samples. I have decided to omit them from clustering and the predictive models as it is not a discerning quality for clustering or predicting. The samples were analyzed for 57 chemical compounds, 7 compounds returned 0 values. In total 2474 samples were collected. Every sample had the chemical Truxilline.

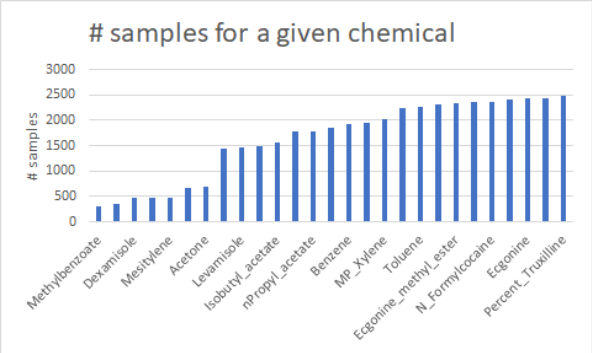
These chemicals had no values:
Quinine
Dextromorphan
Theophylline
Creatinine
Ketamine
Strychnine
Thebaine

	BH	BI	BJ	BK	BL
1	PROMIS_Seizure_Seizure	Count of samples	seizure number	seizures with 1 sample	
2	PS3187838	121	465	195	
3	PS3098390	93	464		
4	PS3239679	88	463		
5	PS3087542	62	462		
6	PS3247872	44	461		
7	PS3199796	43	460		
8	PS3202682	43	459		
9	PS3112001	42	458		
10	PS3111318	41	457		
11	PS3127494	41	456		
12	PS3124824	38	455		

I also conducted a correlation matrix on the data's attributes to discern what attributes may be the most helpful in prediction models and for clustering. Using rank correlation I got the following table:



name of chemical	samples that contained the chemical
Phenobarbitol	1
Nicotinamide	1
Sugar_Maltose	2
Aspirin	3
Acetonitrile	6
Sugar_Sorbitol	6
MIBK	13
Benzocaine	19
Paracetamol	20
Procaine	23
Sugar_Fructose	23
Sugar_Inositol	39
Creatine	50
Sugar_Lactose	50
Chloroform	54
Sugar_Glucose	54
Sugar_Sucrose	59
Lignocaine	88
Sugar_Mannitol	101
Hydroxyzine	117
Caffeine	143
Phenacetin	144
Diltiazem	185
Methylbenzoate	307
Methylisobutenylketone	345
Dexamisole	472
Methylene_Chloride	479
Mesitylene	482
Methyl_acetate	661



This shows chemical compounds at the cutoff that the substance was present in more than 300 samples.

Analysis: It seems like the most useful data is where a lot of samples were collected for a given chemical. At the same time, if it is present in every sample (like Truxilline) it is not very helpful. I have a hunch that the chemicals with less than 300 samples may not matter all that much in clustering. They may hold some importance in the predictive modelling however (especially with the chemicals associated with one sample).

From the correlation matrix I believe that these chemical compounds to be strongly correlated:

1. MP_Xylene and O_Xylene
2. Ethyl_acetate and Hexane and Isobutyl_acetate
3. Tropacocaine and Trimethoxycocaine and Percent_Truxilline
4. cis_Cinnamoylcocaine and trans_Cinnamoylcocaine
5. The group of (Tropacocaine, Trimethoxycocaine, Percent_Truxilline) and (Ethyl_acetate, Hexane, Isobutyl_acetate)
6. Benzoylecgonine and Ecgonine

Negatively Correlated:

1. acetone and the group (Percent_Truxilline, Trimethoxycocaine, Tropacocaine)
2. Methylisobutenylketone and the group of (Percent_Truxilline, Trimethoxycocaine, Tropacocaine)
3. hexane and acetone
4. (cis_Cinnamoylcocaine, trans_Cinnamoylcocaine) and the group of (Percent_Truxilline, Trimethoxycocaine, Tropacocaine)

This information is helpful as a clustering model that puts the strongly correlated groups together is a good sign. If they group together negatively correlated substances, the cluster output may not be reliable. In decision trees or predictive rules based modelling (or even in neural networks), one of the first evaluations could be centered around the negatively correlated attributes.

Clustering:

I attempted the following clustering methods: K-means, Expectation Maximization, Density Based Clustering, and Hierarchical Clustering. For all clustering methods, I normalized the chemicals values as to be more easily compared to one another. I used z-score normalization as the Gaussian would match the Gaussian assumption in EM.

K-Means:

K-means clustering uses the full dataset to cluster the different samples. It creates clusters based on the distance between each record. Some drawbacks to this approach are that it doesn't take into account that some points may be closes to the center of another cluster than their true cluster (clusters of different sizes or non-spherical clusters)

I used this as a preliminary approach to understanding the dataset I was working with. I tried different values of K, but the inputs of 3 clusters, where 27 chemicals were omitted, produced the best results.

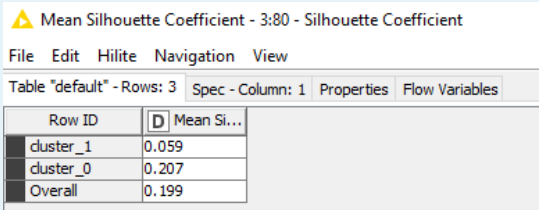
$$\text{Silhouette Score} = \frac{b - a}{max(a, b)}$$

To analyze the iterations of each cluster output, I used the Silhouette Score. Where a is equal to the average intra-cluster distance (between a point in a cluster), and b is equal to the average inter-cluster distance (between clusters). The score ranges from -1 to 1 and a score of 0 indicates indifference, 1 indicates distinguished clusters, while -1 is clusters are nonsense.

Analysis: It seems that having two clusters gives us a better score, however with three clusters, the elements are spread out more evenly. This may indicate that simply basing the clusters off distance may not be the best approach and the actual clusters may be oddly shaped. There seem to be two distinct cluster areas. Within one of these "cluster areas" there are two clusters that may be close together. The hunch is that there are three origin points for the drugs. Something to keep in mind for the predictive models.

50 chemicals considered (ones with values present)

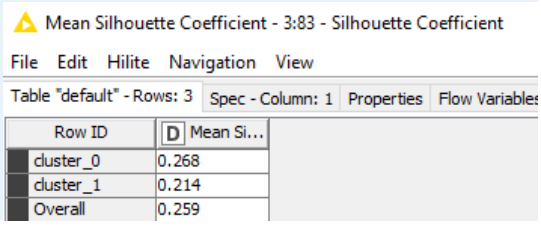
1. k=2



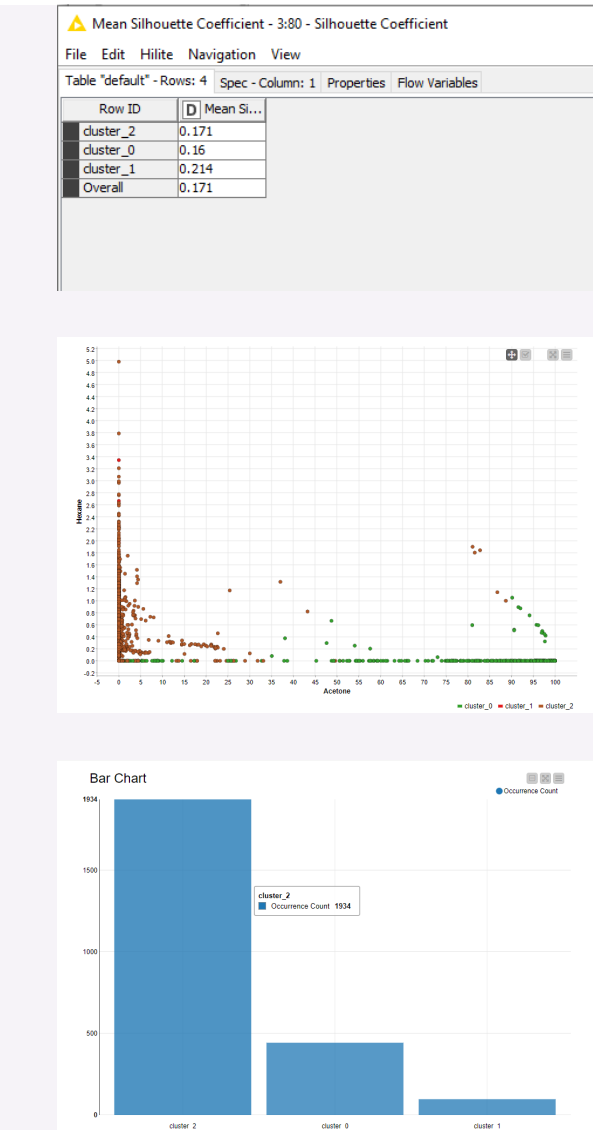
2. k=3

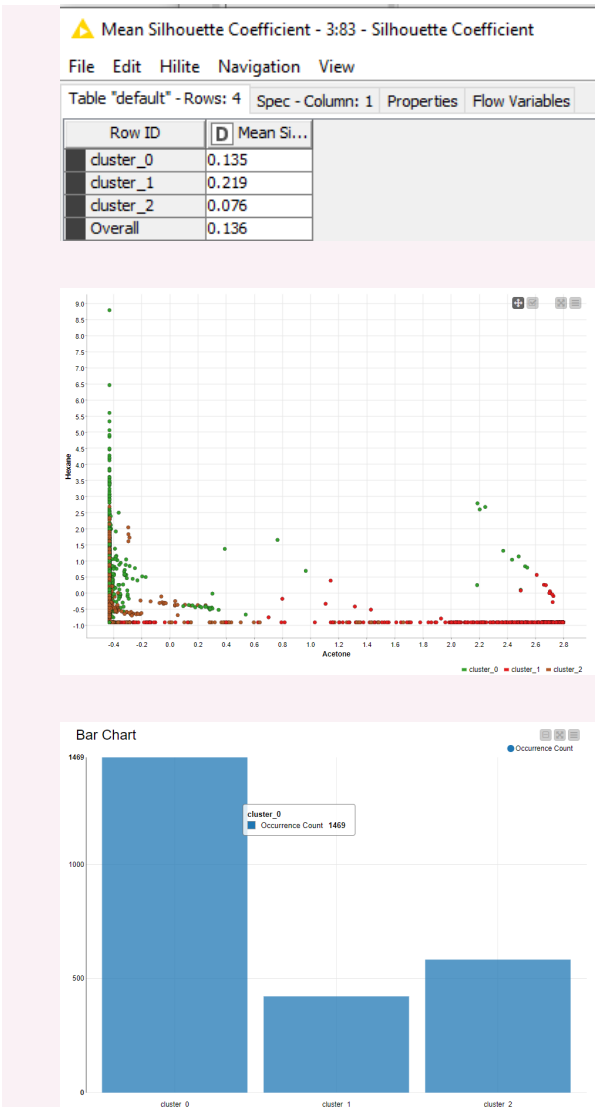
Only Chemicals with samples > 300

1. K=2



2. k=3





3. k=4/k=5 made the coefficient worse.

EM

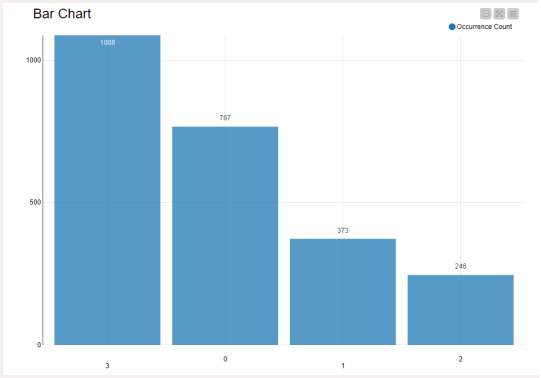
The expectation maximization find the likelihood of a record being in a certain distribution. It is an iterative process by which the mean and variance is revaluated and used to shape the distribution. This process is repeated until you can maximize the likelihood. I used the same silhouette score to measure the effectiveness of this technique.

I used EM because a drug sample could have passed through the same points on their way to the final destination. For example, if one substance started in location X and

another Y, they may have different profiles. But if they both passed through an intermediary point of W and Z onto their final destination, a part of their adulterant profile might be the same. Thus soft clustering might be helpful for accounting for the complex pathways the drugs might have gone through and the possible missing samples from the shipments that were seized.

All chemicals used

1. -1 for all clusters, 300 iterations



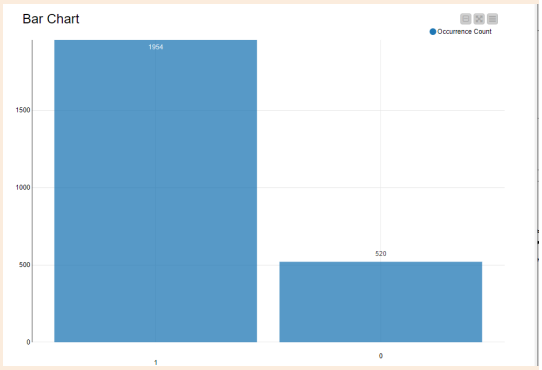
Mean Silhouette Coefficient - 6:82 - Silhouette Coefficient

File Edit Hilite Navigation View

Table "default" - Rows: 5 | Spec - Column: 1 | Properties | Flow Variables

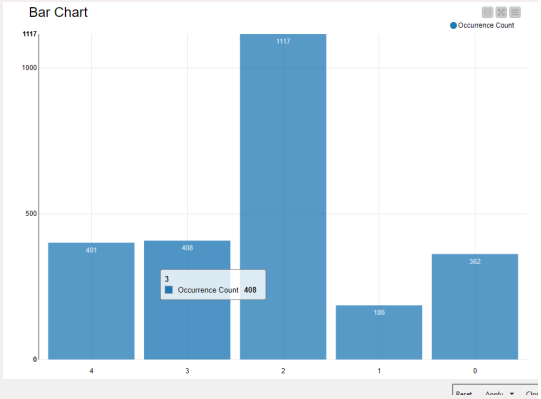
Row ID	Mean Si...
3	0.056
0	0.047
1	0.281
2	-0.218
Overall	0.06

2. 2 clusters



Only Chemicals with samples > 300 (27 chemicals)

1. -1 for all clusters, 300 iterations. When left to find the optimal number of clusters:



Mean Silhouette Coefficient - 6:89 - Silhouette Coefficient

File Edit Hilite Navigation View

Table "default" - Rows: 6 | Spec - Column: 1 | Properties | Flow Variables

Row ID	Mean Si...
4	0.1
3	0.028
2	0.109
1	-0.258
0	0.301
Overall	0.095

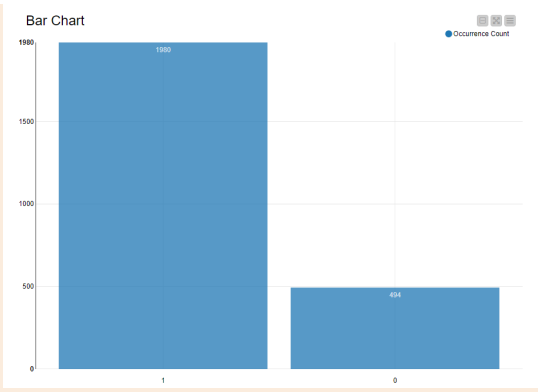
2. 2 clusters

Mean Silhouette Coefficient - 6:82 - Silhouette Coefficient

File Edit Hilite Navigation View

Table "default" - Rows: 3 | Spec - Column: 1 | Properties | Flow Variables

Row ID	Mean Si...
1	0.223
0	0.176
Overall	0.213



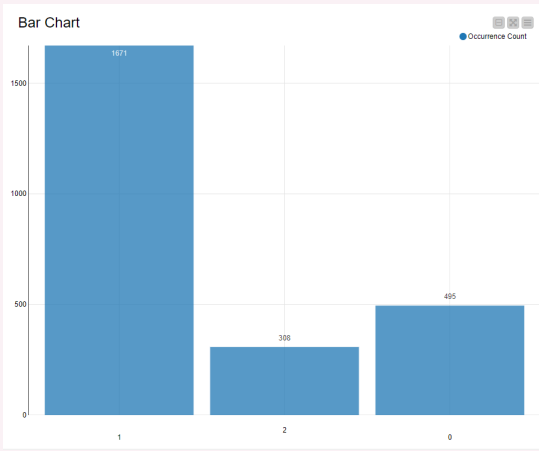
Mean Silhouette Coefficient - 6:89 - Silhouette Coefficient

File Edit Hilite Navigation View

Table "default" - Rows: 3 | Spec - Column: 1 | Properties | Flow Variables

Row ID	Mean Si...
1	0.235
0	0.199
Overall	0.228

3. 3 clusters



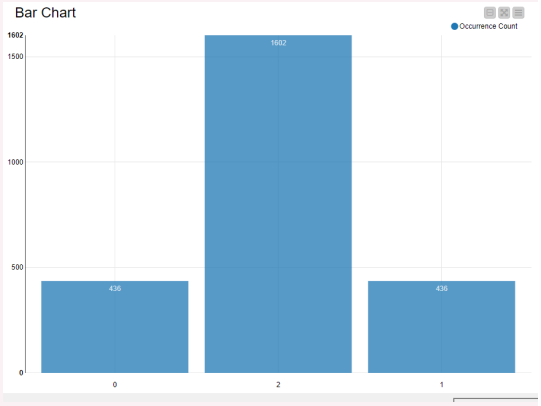
Mean Silhouette Coefficient - 6:82 - Silhouette Coefficient

File Edit Hilite Navigation View

Table "default" - Rows: 4 | Spec - Column: 1 | Properties | Flow Variables

Row ID	Mean Si...
1	0.214
2	-0.148
0	0.152
Overall	0.157

3. 3 clusters



Mean Silhouette Coefficient - 6:89 - Silhouette Coefficient

File Edit Hilite Navigation View

Table "default" - Rows: 4 | Spec - Column: 1 | Properties | Flow Variables

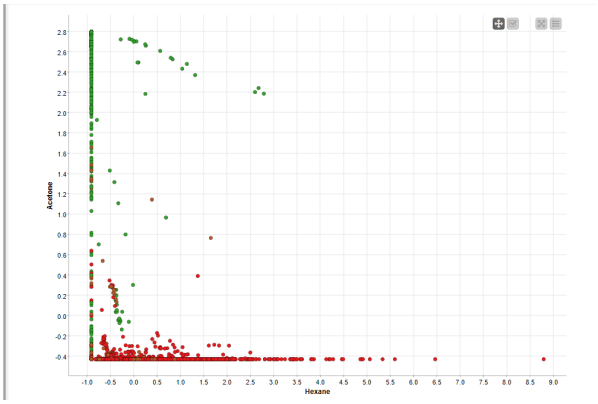
Row ID	Mean Si...
0	-0.14
2	0.208
1	0.171
Overall	0.14

Analysis: I believe that this may be a better model for clustering than k-means, as soft clustering is more appropriate for a context where drug pathways may intersect and chemical profiles may overlap.

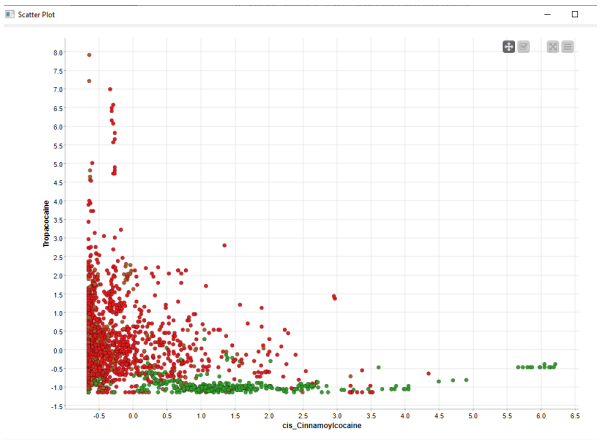
Looking at the silhouette scores, it seems the best way to cluster the samples is according to two major pathways (2 clusters). Additionally, clustering the samples based on the chemicals that were present in more than 300 samples, seems to produce the best clusters (using 27 chemicals). When I used the (-1) for the cluster parameter → letting the program find the number of clusters to maximize the likelihood, it resulted in the lowest score and produced 4 and 5 clusters. When I used 3 as the cluster parameter, I achieved a lower score and a cluster group that shared close to no correlation (a negative score).

Between the 2 and 3 clusters, the results match up well. A cluster of 400 samples seems to come from one area, and the other 2000 samples seem to come from another pathway. When I used 3 clusters, 300-ish of the 2000 samples seem to point to another cluster. The negative score indicates that this sub-cluster is filled with samples that started in the same area as the 2000 samples, but the 300-ish went through a more complicated route/people/organizations. As such those 300 don't share the same chemical profile but rather share the fact that they are all very similar in their difference with the other clusters.

In the scatterplots below, we can see that the third cluster comes from within a bigger cluster. The axis is chemicals that were earlier identified to be negatively correlated with each other. (cluster colors: brown, red, green) The browns also seem to be grouped in one area. I think this further gives credence to their being three clusters.

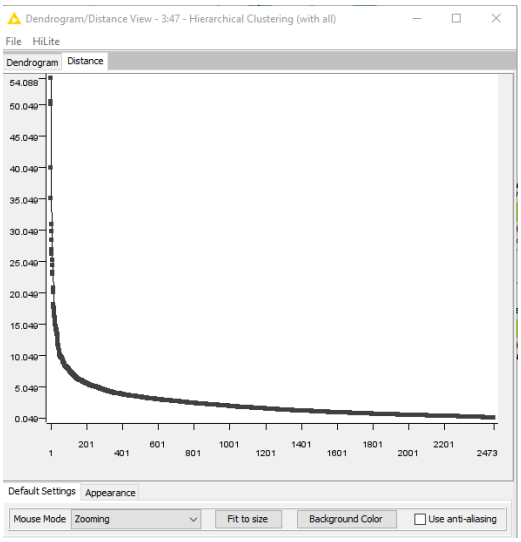


The 3rd cluster of brown seem to be grouped together



Hierarchical Clustering:

This was also a dead end. No matter how I cut the dendrogram, the dataset was so wide that the clusters were always skewed where 1 cluster would have all the elements and the rest would have 3-10 elements. The silhouette scores were impressive (close to 1.0) but the actually clusters themselves were meaningless. This is because of how wide the data set is and how much the profiles overlap.



Shows how wide the data is

Mean Silhouette Coefficient - 3:76 - Silhouette Coefficient

File Edit Hilite Navigation View

Table "default" - Rows: 4 | Spec - Column: 1 | Properties | Flow Variables

Row ID	D Mean Si...
cluster_0	0.991
cluster_1	0.787
cluster_2	0.735
Overall	0.736

this is from the hierarchical clustering, strong scores but meant little when looked at the clusters themselves

Density Based Clustering:

This became a dead end. No matter how much I changed epsilon and minimum points, it was resulted in one giant cluster or a lot of noise.

Overall Takeaways:

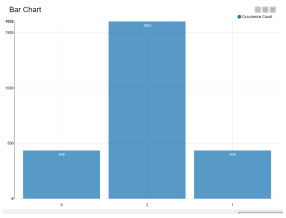
Seems like there are two big pathways for drugs. There also seems to be a major sub-pathway for a portion of the drugs but this is uncertain. It could be either the samples were in more complex pathways or their adulterant profile might be linked to a specific organization/complex pathways.

Going forward with my predictive models, I am choosing to use the cluster labels from the EM model that used only the chemicals present in more than 300 samples. I am using the 3 cluster model, as it resulted in the least negative number for the 3rd cluster $(-0.14 > -0.148)$. The assumption I'm making here is that there are two major

origination points for a majority of our samples. From one of these origination points, a portion of the samples split off in a more complex pathway that is hard to trace. It is almost impossible to know where each drug comes from with certainty but using this three-cluster model we can at least account for a portion of samples that seem to branch off into a complex route.

Predictive Models:

In my predictive models, I used the cluster labels that came from the EM method. The target/predication focus for each model was the shipment from which a sample came from. If my clustering assumptions and the clustering output is good, I would expect the prediction scores to be better than a coin flip (50%), and approaching to at least $\frac{1602+436}{2473} = 80\%$. The 80% is being able to predict the two major groups if we say that the third group is composed of complex/undiscernible pathway drugs. The models I will use:



cluster groupings from the EM (assumption use the top 27 chemicals)

- Random Forest as I believe some chemical compounds have more predictive power than the others, and such smaller more frequent trees would help.
- Support Vector Machine as if we say that we cant to ignore the 3rd subgroup, this model would work well in discerning between the two major clusters.
- Neural Network might work well because we need to account for intrinsic chemical finger prints. I would assume that each layer of the network could hone in one the chemical compound that are related to each other in a cluster/related shipment.

SVM

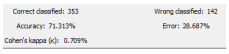
Using support vector machine as a predictor, we are relying on the fact that there are 2 major groups of samples. As a result using SVM in determining a boundary between the

two groups might be a good idea. In this instance we are more interested in the the support vector samples → the drug samples that help define the characteristics of each group rather than every record (that could be involved in a more circuitous route.

Analysis: For this predictor, I used EM clustering with 2 clusters. Letting $C = \text{overlapping penalty}$. It seems that changing the overlapping penalty did not do much after 3.0, and the radial basis function did better than polynomial. In addition, using only 27 chemicals as attributes increased the accuracy of the predictor across the board. There was also a marginal improvement when I used EM clustering with 3 clusters. I think this is further evidence that there appears to be three clusters but the third cluster is not as distinct. The RBF did better because the boundaries are more circular than they are based on a line. This makes sense as many of the chemicals profiles overlap with some chemicals. The predictor overall did very well and hit the suspected accuracy rate of 80%. The samples that were wrongly classified are the same samples that the EM had a hard time clustering, this points to these samples being from a complex pathway.

50 Chemicals considered in Clustering

1. RBF=3.0, C=3.0



2. RBF=4.0, C=4.0

3. RBF =2.0, C=3.0

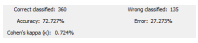
4. RBF =2.0, C=4.0 (75.5%)

50 Chemicals considered in Clustering

1. Polynomial = 2.0, C=3.0 (73.3%)

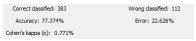
2. Polynomial = 2.0, C=4.0

3. Polynomial = 3.0, C=3.0

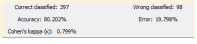


27 Chemicals considered in Clustering

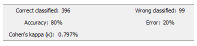
1. RBF=3.0, C=3.0



2. RBF =2.0, C=3.0, (80.2%)



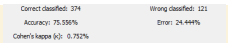
3. RBF =2.0, C=4.0



27 Chemicals considered in Clustering

1. Polynomial = 2.0, C=3.0 (78.38%)

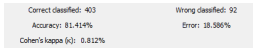
2. Polynomial = 3.0, C=3.0



5. RBF=2.0, C= 2.0

4. Polynomial = 3.0, C=4.0

I then tried using EM 3 clusters with RBF=2.0 and C=4.0 (81.1%), 27 chemicals



Random Forest

In a random forest predictor, we grow k different decision trees with small random subsets of the training records or examples (with a subset of the attributes). It is an iterative process by which we then use all the decision trees, and take either a rank voting or average to get our final prediction. In this way, we have smaller trees that are more distinct, which is better at distinguishing outcomes, and we can get information about which attributes are the best at predicting. This is because the test at the root is most important → if it is repeated in most trees it might be the most discriminatory.

This might be a good predictor for this data as many chemicals could be correlated vs non-correlated. As such, these relationships could be an important test for the tree. Having many subtrees will also help in testing, and placing more importance, on the chemicals that distinguish each cluster group. From my SVM analysis and my prior assumptions, I will use EM clustering based on 3 clusters as an additional attribute for the samples.

Random Forest: *Number of models built is 1000*

Ensemble Predictor: *Number of models built is 1000, gain ratio as splitting criteria, rows with replacement, used same set of attributes*

50 Chemicals considered in Clustering

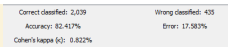
1. Splitting Criteria: gain ratio (82.41%)

27 Chemicals considered in Clustering

1. Splitting Criteria: gain

50 Chemicals considered in Clustering Ensemble Predictor

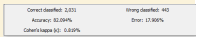
27 Chemicals considered in Clustering Ensemble Predictor



2. Splitting Criteria: information gain

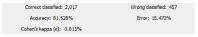
3. Splitting Criteria: gini index

ratio (82.09%)

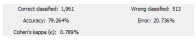


2. Splitting Criteria: information gain

3. Splitting Criteria: gini index

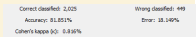


1. fraction data sent (rows) = 0.5, square root attribute sample

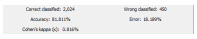


2. fraction data sent (rows) = 0.9, square root attribute sample

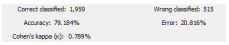
3. fraction data sent (rows) = 0.9, linear fraction attributes = 0.3 (81.85%)



4. fraction data sent (rows) = 0.9, linear fraction attributes =0.6



1. fraction data sent (rows) = 0.5, square root attribute sample



2. fraction data sent (rows) = 0.9, square root attribute sample

3. fraction data sent (rows) = 0.9, linear fraction attributes = 0.3 (81.97%)

4. fraction data sent (rows) = 0.9, linear fraction attributes =0.6

Attribute Statistics:

50 chemicals + Splitting Criteria: gain ratio (random forest)

Sorted Table - 5:62 - Sorter

File Edit Hilite Navigation View

Table "default" - Rows: 28 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	#splits (level 0)	#splits (level 1)	#splits (level 2)	#candidates (level 0)	#candidates (level 1)	#candidates (level 2)	success
Percent_Truxi...	324	256	282	360	745	1387	0.9
nPropyl_acetate	306	441	644	351	765	1428	0.872
Dexamisole	277	352	613	358	738	1479	0.774
Isobutyl_acet...	215	346	548	361	680	1426	0.596
Hexane	158	229	324	348	673	1434	0.454
MEK	169	321	577	375	687	1459	0.451
Methylisobute...	120	251	439	369	713	1440	0.325
Tropacocaine	102	163	224	330	730	1410	0.309
Acetone	61	236	390	341	732	1360	0.179
Trimethoxyco...	52	43	76	358	689	1445	0.145
Ethyl_acetate	53	209	407	370	716	1453	0.143
Levamisole	41	200	445	368	751	1433	0.111
Benzene	38	140	279	367	694	1447	0.104
Benzoylecgon...	27	62	168	346	740	1351	0.078
Winner Cluster	19	36	91	358	700	1436	0.053
Methylbenzoate	17	140	333	353	714	1476	0.048
O_Xylene	7	49	178	374	746	1429	0.019
Toluene	3	45	186	339	707	1401	0.009
Mesitylene	3	4	75	357	746	1390	0.008
Norcocaine	2	49	236	356	687	1436	0.006
MP_Xylene	2	52	155	364	700	1417	0.005
cis_Cinnamoyl...	2	41	183	373	726	1408	0.005
trans_Cinnam...	1	46	188	330	733	1379	0.003
N_Formylcoca...	1	58	214	376	677	1465	0.003
Ecgonine	0	14	112	335	711	1371	0
Ecgonine_met...	0	3	69	377	694	1489	0
Methyl_acetate	0	97	264	352	696	1483	0
Methylene_C...	0	117	293	354	710	1468	0

27 chemicals + Splitting Criteria: information gain ratio (random forest)

Sorted Table - 5:171 - Sorter

File Edit Hilite Navigation View

Table "default" - Rows: 28 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	#splits (level 0)	#splits (level 1)	#splits (level 2)	#candidates (level 0)	#candidates (level 1)	#candidates (level 2)	success
nPropyl_acetate	167	197	283	181	389	725	0.923
Percent_Truxi...	141	112	166	170	373	743	0.829
Dexamisole	147	163	283	193	350	729	0.762
Isobutyl_acet...	100	166	259	173	374	675	0.578
Winner Cluster	107	115	195	196	357	716	0.546
MEK	64	163	264	191	369	668	0.335
Hexane	60	106	169	182	351	719	0.33
Tropacocaine	48	81	114	163	366	691	0.294
Acetone	33	107	199	166	348	786	0.199
Methylisobute...	30	138	233	168	368	741	0.179
Ethyl_acetate	24	72	209	176	376	725	0.136
Trimethoxyco...	24	12	38	191	357	711	0.126
Benzoylecgon...	15	28	84	165	365	737	0.091
Levamisole	13	78	159	186	354	682	0.07
Benzene	10	72	138	195	354	694	0.051
Methylbenzoate	9	71	133	180	373	706	0.05
O_Xylene	4	21	106	171	342	740	0.023
MP_Xylene	2	28	69	188	329	690	0.011
cis_Cinnamoyl...	1	25	114	165	368	702	0.006
N_Formylcoca...	1	42	112	183	358	726	0.005
Ecgonine	0	5	56	162	327	713	0
Ecgonine_met...	0	4	32	181	357	712	0
Norcocaine	0	16	82	173	359	727	0
trans_Cinnam...	0	26	89	168	332	723	0
Mesitylene	0	4	26	195	355	717	0
Methyl_acetate	0	76	156	157	384	720	0
Methylene_C...	0	52	141	203	336	700	0
Toluene	0	20	88	178	329	682	0

Analysis: The gain ratio worked as the best splitting attribute. This is because of how spread out the data was, it was important to understand the intrinsic improvement that comes with each tree. Narrowing the chemical attributes (50→27) considered only marginally decreased accuracy. The random forest node provided better results than the ensemble node. Looking at the two best versions of the predictor, when using all 50 chemicals as attributes only 24 attributes were chosen as the root in 1000 models. When using 27 chemicals as attributes considered, 20 were chosen as the root in 1000 models, and the cluster that they were assigned to became more important. It seems that these top 20 or so chemical compounds are the most discriminatory in discerning a drug's origin. This also aligns with the correlation matrix I did at the start. The tree model also fared better than the SVM as it could place more importance on these top chemicals. This also shows that the clustering by the EM was pretty good in predicting the shipments from which a sample comes from.

Neural Network

Next I wanted to try using a neural network. To do so, I had to normalize first. I had to turn the shipment label into a number between 0 and 465, which I then had to normalize. I believe that this model would work well to pick up the idiosyncrasies of the chemical profiles. This would aid in getting a prediction score of greater than 80%.

Analysis: There seem to be no improvement between the different types of clustering. Additionally, an accuracy of around 80% seems to be the limit

50 Chemicals considered in Clustering

1. 200 iterations, 5 hidden layers, 50 neurons in each layer → 0.788 R^2
2. 200 iterations, 7 hidden layers, 50 neurons in each layer → 0.765 R^2
3. 200 iterations, 5 hidden layers, 100 neurons in each layer → 0.802 R^2

27 chemicals considered in clustering

1. 200 iterations, 5 hidden layers, 100 neurons in each layer → 0.812 R^2
2. 200 iterations, 7 hidden layers, 100 neurons in each layer →0.804%

Without used cluster labels

1. 200 iterations, 5 hidden layers, 100 neurons in each layer → 0.786 R^2

without additional data to address the 3rd cluster group. I suspect that with additional data we can achieve a better accuracy score

Conclusion:

Clustering: From our initial correlation matrix, the assertion that some chemicals are more important in deciding the shipment and clustering was mostly true. In most cases, it provided a marginal improvement. In K-means clustering with k=2, -1 for EM, and a cluster of 2 for EM, we achieved a higher overall silhouette score when considering 27 chemicals. This score always went down, relative to considering all chemicals to cluster, when the cluster was 3.

When looking at the scatterplots, and the counts of the records in the cluster groups, it seems that there are two major clusters of drugs. In other words, there are two locations/pathways that drugs are originating from based on the adulterant profile. There is also a sub-cluster (of around 400) that is either a part of the bigger cluster or is noise (aka drugs from complex pathways).

Even though the 2-cluster model using 27 chemicals produced the best scoring cluster, I chose to use the 3-cluster model (27 chemicals) because I had a hunch that there might be a sub-cluster that we should take into account. This sub-cluster is not a true “cluster” as it had a negative score of -0.14. But this is close to 0 (aka being indifferent). Again pointing to drugs from complex pathways. In total, I believe: 400-500 samples from one origin, 1600-1800 samples from another origin, and 400-500 samples that cannot be identified (unclear if complex pathway or a potential origin point).

Predictors: Most of my predictive models achieved the ability to get 80% accuracy rate. The 80% is significant as it is the total number of samples I was confident I had clustered properly. The 20% error rate corresponds to the inability to categorize the 3 group of samples that may be noise.

The random forest splitting criteria further provided evidence that the clustering was a success. The cluster label for the samples was among the top criteria used in the trees. All the models that used the cluster groupings derived from just 27 chemicals proved to have a better accuracy. This could be because we eliminate the supposed importance that the other chemicals have in clustering output.

The two best predictors I built was a SVM model and a random forest model (using 3 clusters).

1. EM of 2 clusters, SVM, 27 chemicals, RBF =2.0, C=3.0 $\rightarrow$ 80.2% accuracy
2. EM of 3 clusters, SVM, 27 chemicals, RBF =2.0, C=3.0 $\rightarrow$ 81.1% accuracy
3. EM of 3 clusters, Random Forest, 27 chemicals, splitting criteria gain ratio $\rightarrow$ 80.2% accuracy
4. EM of 3 clusters, 27 chemicals, MLP, 200 iterations, 5 hidden layers, 100 neurons per layer $\rightarrow$ 81.2% accuracy

All the predictors capped around 80% accuracy. This is because of the uncertain nature of around 400 elements. The predictors have upheld the output of the clustering I did before as we managed to classify elements that were properly clustered.

Recommendations:

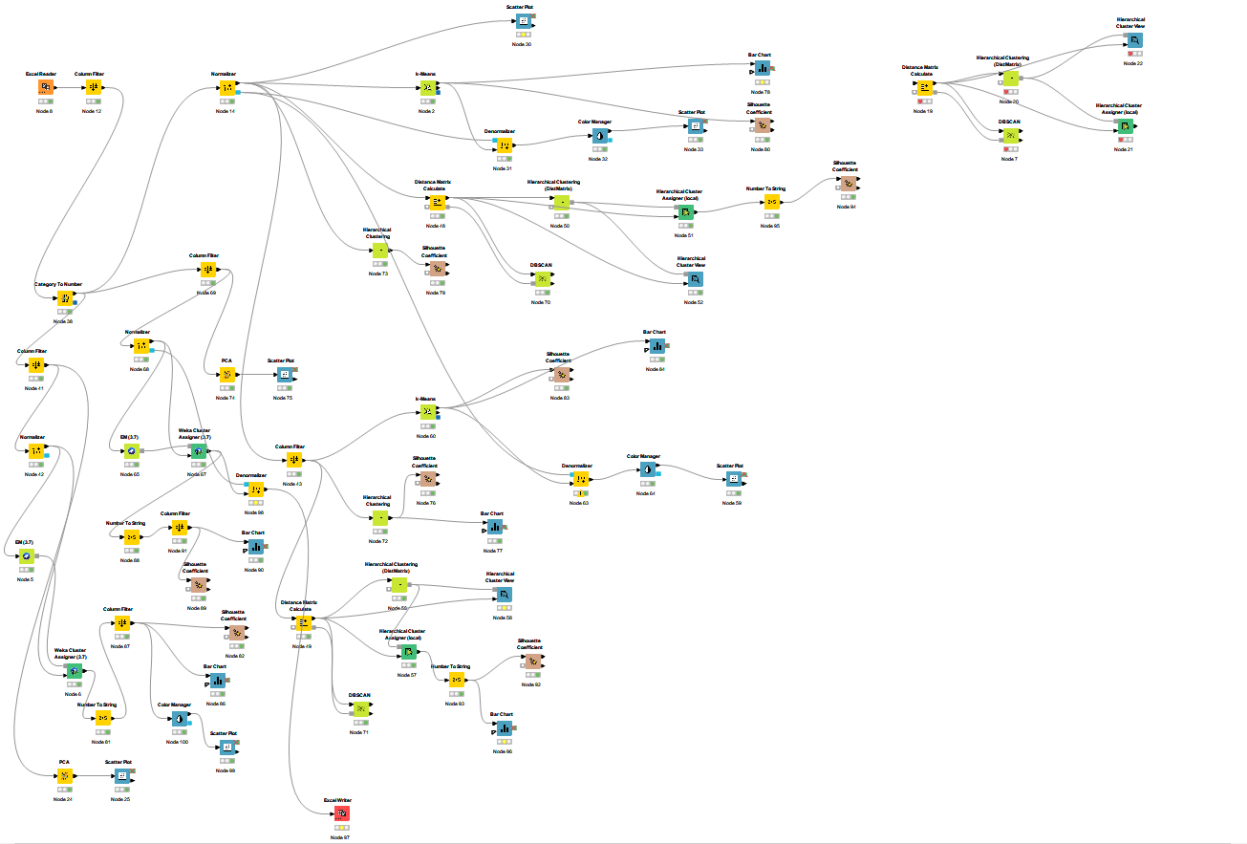
The 2474 samples seem to be lacking in information regarding shipments, and when seized, only one sample was recovered. It would be interesting to see more data regarding additional samples from those singleton seizures (if available). This could help understand the third group I was unable to reliably cluster.

From our random forest models, the top 20-25 chemicals are the most discriminatory in identifying a drug's origin/pathway. Percent_Truxilline, Hexane, nPropyl_acetate, Dexamisole are these chemicals in question to name a few. I would recommend studying these top chemicals to understand where they might come from, how they interact with the human body, and how they interact with the illicit substance. This can inform public policy, harm reduction efforts, and for law enforcement to pinpoint the geographical location of the cluster.

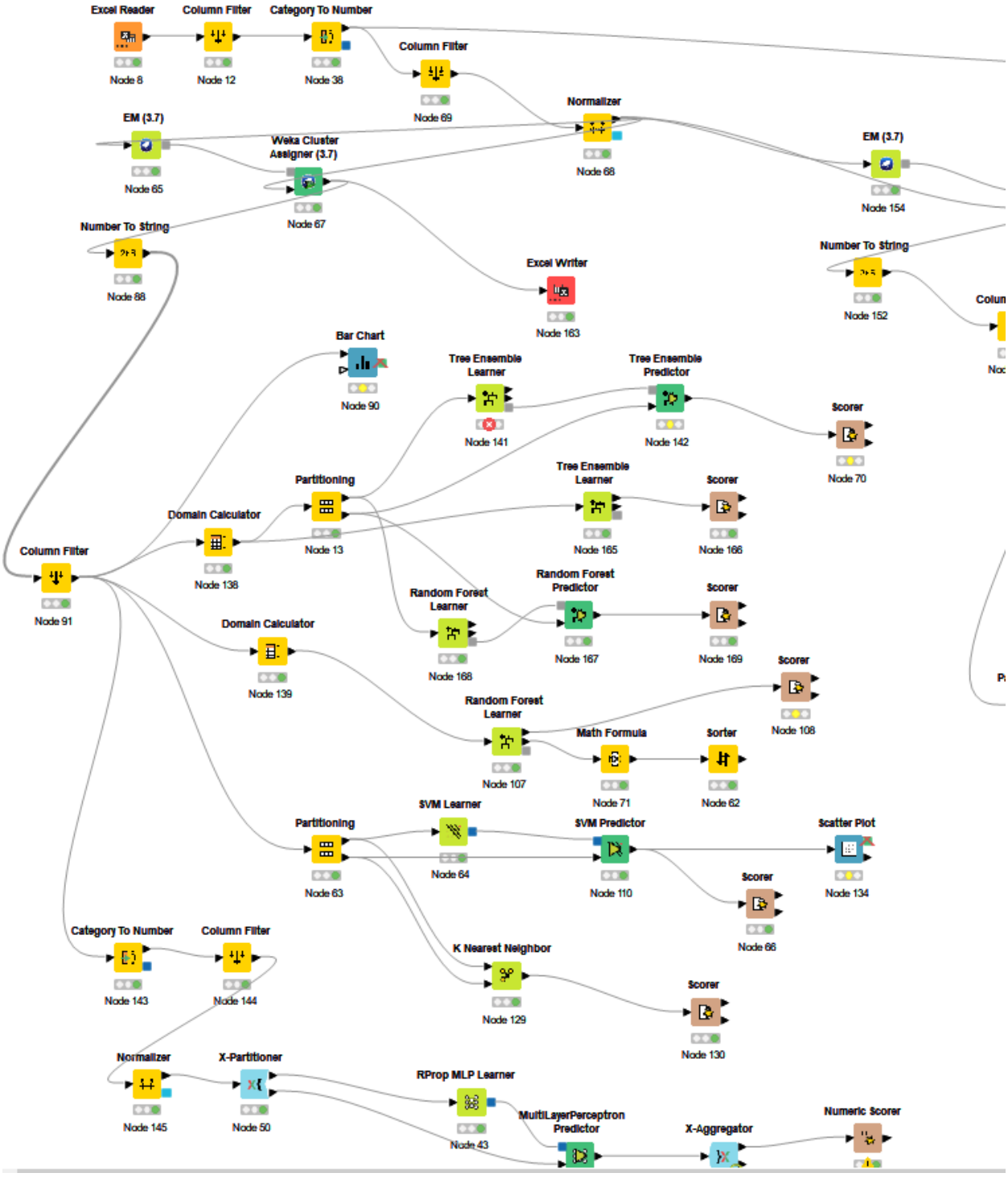
For further modeling refinement, I would look to explore the different dials we can turn in the random forest model and maybe look to adapting SVM for a three-cluster approach. Regarding the neural network, it may be beneficial to train separate models for each of the clusters or use a bigger dataset.

Exhibits:

1. Clustering Workflow



2. Predictors workflow


Excel Reader
Node 8
Column Filter
Node 12
Category To Number
Node 38
Column Filter
Node 69
Normalizer
Node 68
EM (3.7)
Node 65
Weka Cluster Assigner (3.7)
Node 67
EM (3.7)
Node 154
Number To String
Node 88
Excel Writer
Node 163
Number To String
Node 152
Bar Chart
Node 90
Tree Ensemble Learner
Node 141
Tree Ensemble Predictor
Node 142
Scorer
Node 70
Partitioning
Node 13
Tree Ensemble Learner
Node 165
Scorer
Node 166
Domain Calculator
Node 138
Column Filter
Node 91
Random Forest Predictor
Node 167
Random Forest Learner
Node 168
Scorer
Node 169
Domain Calculator
Node 139
Scorer
Node 108
Random Forest Learner
Node 107
Math Formula
Node 71
Sorter
Node 62
SVM Learner
Node 64
Partitioning
Node 63
SVM Predictor
Node 110
Scatter Plot
Node 134
Scorer
Node 66
Category To Number
Node 143
Column Filter
Node 144
K Nearest Neighbor
Node 129
Scorer
Node 130
Normalizer
Node 145
X-Partitioner
Node 50
RProp MLP Learner
Node 43
MultiLayerPerceptron Predictor
X-Aggregator
Numeric Scorer