Project Final Report

Matthew Sadecki

November 30, 2018

Introduction:

To begin analyzing the dataset, I opened the excel spreadsheet and took note of the different attributes. I noticed that there was a total of 7043 records and 21 attributes. I noticed there was an customerID attribute and decided to eliminate it as I did not see how it would help with predicting wether a customer churns. Next I decided to test different hypotheses that came to mind on a very small sample size. An example of a quick prediction I made was after looking at a very small sample of records I found that the payment method attribute could potentially be a good predictor of whether the customer churns or not. While quickly scanning the data set I tried to see if I could find any missing data or data that was not formatted the same way as other data in the same column. After scanning the data quickly I did not find missing data or data that was formatted differently from other data in the same column.

Clustering the data: (See appendix 1 and 2 for cluster outputs and knime nodes setup)

Next, I tested out two clustering algorithms; k-means and hierarchical clustering for the monthly charges and tenure attributes. I wanted to see if there is a relationship between monthly charges and the tenure attribute. I also wanted to see what a k-means cluster would output (less computationally intensive) versus a hierarchical clustering algorithm (more computationally intensive). To cluster the data using k-Means I used the k-Means node and cluster-assigner node.

To cluster the data using hierarchical clustering I used the hierarchical clustering node. I configured the hierarchical clustering node so that it found clusters using euclidean distance and using average linkage. I decided to use Euclidean distance because it takes every variable into account and doesn't remove redundancies. I chose average linkage because this would generate clusters that generalize the dataset the best. I decided to test outputting 2 clusters, then 3 clusters, up to 5 clusters. I then decided to look at the outputs to see which seemed reasonable and could be justified. I thought that two clusters when using both algorithms produced results indicating there may be too few clusters. In the case of k-means I think it would be unreasonable to assume that there is only two different levels of pricing. In the case of hierarchical clustering I think it is somewhat reasonable to assume there maybe a cluster in the top right corner representing the best customers. The customers in the top right stay with the service provider for a long time and choose the most expensive pricing packages. In the case of 3 clusters for k-means it seemed more reasonable and justifiable. There appeared to be a low monthly charge level from about $20-$40, a medium monthly charge level from about $40-$70, and a high monthly charge level from $70+. In the case of 3 clusters for hierarchical clustering there appeared to be a cluster in the top right, bottom right, and top/bottom left. I thought that this was unreasonable because if there is a high-paying/long-tenure customer and low-paying/long-tenure customer, then surely there must be a high-paying/short-tenure customer and low-paying/short-tenure customer. Therefore I think a hierarchical clustering of 4 cluster would be more reasonable than 3 clusters. In the case of 4 clusters using k-means I found this to be reasonable and justifiable because it could be possible that there are 4 different monthly charges levels. In the case of 4 hierarchical clusters I found this to be the most likely scenario. There was a cluster for high-paying/long-

tenure customers and low-paying/long-tenure customers, in addition to high-paying/short-tenure customers and low-paying/short-tenure customers. In both k-means and hierarchical clustering I thought that while five clusters maybe possible, it probably wouldn't help with prediction and may yield lower accuracy. I thought that the output of the hierarchical clustering algorithm would be a better predictor than the k-means algorithm. I thought this because the k-means clustering algorithm clustered customers based on monthly charge ranges while the hierarchical clustering algorithm clustered customers based on wether the customer was good or bad (for example high paying/high tenure vs low paying/low tenure). Therefore I decided to proceed with my analysis based on the results from the hierarchical clustering algorithm.

Decision Tree: (See appendix 3 for knime nodes setup and examples of predictive attribute decision trees. See appendix 4 for examples of eliminated attributes)

Next, I used the decision tree learner node and decision tree predictor node to try and find which attributes are good for predicting wether or not a customer churns. I used the column filter node to run the analysis with one attribute at a time versus the churn attribute. I found that the contract duration attribute was the most predictive in determining wether a customer will churn. Only 11.3% of the customers that had a one year contract churned while only 3.2% of the customers with a two year contract churned. The contract attribute also showed that 43.1% of customers who had a month-to-month contract churned. Therefore the customers who had a month-to-month contract were the most likely to churn (this was not surprising). I continued looking at attributes one at a time versus the churn attribute and found that the internet service, payment method, and whether or not a customer has online security were good predictors of

wether a customer will churn. Comparing attributes one at a time against the churn attribute using a decision tree I was successfully able to eliminate the Gender, PhoneService, MultipleLines, StreamingTV, and StreamingMovies attributes. I was able to eliminate these attributes because they did not help narrow down which types of customers are likely to churn. This was due to the having results where roughly 50% of the customers for a particular attribute churned (for example: gender attribute) or there was no difference when an attribute was used (for example: multiple lines attributes had roughly the same probabilities for customers that churned and customers that didn't churn). Finally I decided to run the decision tree predictor with all the selected/predictive attributes to try and learn more about why customers were leaving. I found that it was difficult to determine an exact reason when looking at the whole tree because sometimes an attribute was useful for certain branches of the tree while in other branches the attribute was not useful (I used the gain ratio option in the decision tree predictor). For example in some branches the senior citizen attribute was useful while in others the dependents attribute was more useful.

To learn more about why customers were leaving I decided to focus my analysis on customers who had month-to-month contracts. I found that on average I was able to predict with 76-77% accuracy why a customer churned when using the contract, payment method, and internet service attributes. I found that customers who had a month-to-month contact, who were using fibre internet, and paid with an electronic cheque were the most likely to churn. I found that it was difficult to further narrow down why customers were leaving due to leafs in the decision tree showing 50-50 or 60-40 percent splits for predicting if a customer churned given an attribute.

Normalization and Random Forest/Support Vector Machine (See appendix 5 for random forest and support vector machine nodes setup. See appendix 5 for final random forest prediction accuracy and confusion matrix):

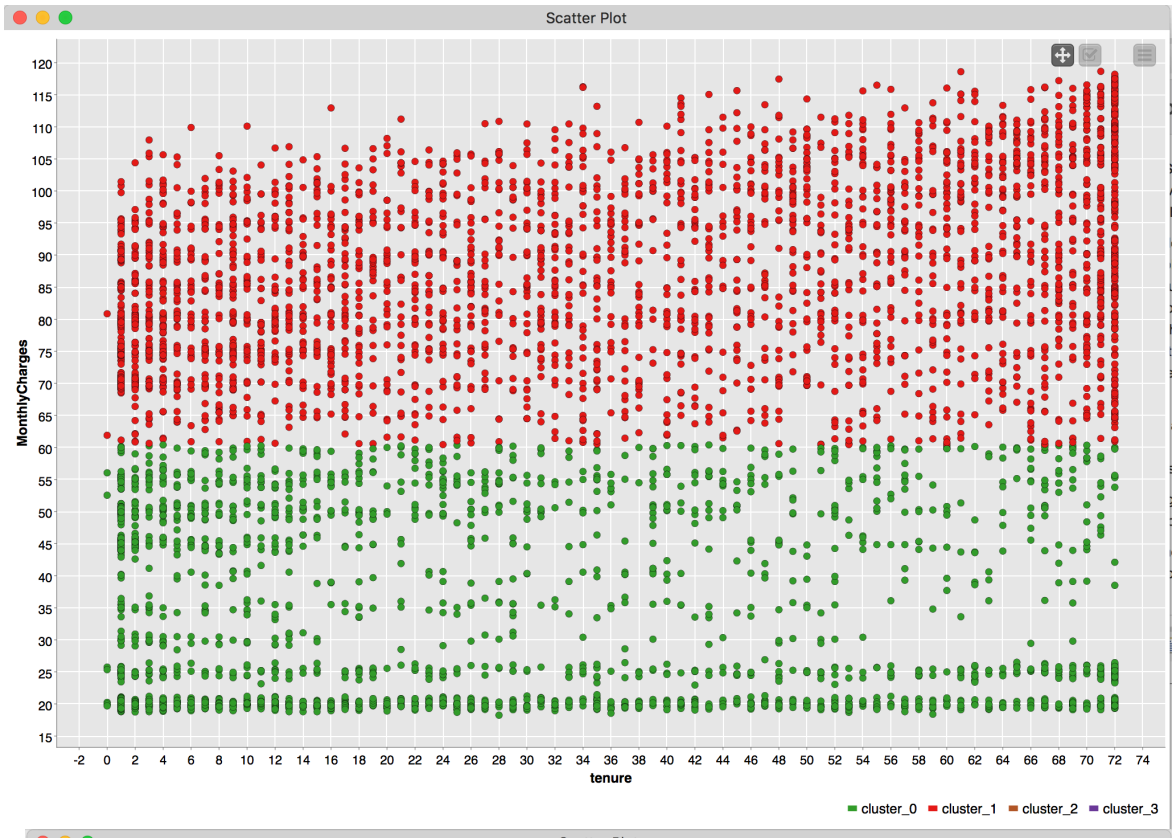
Next, I decided to normalize the month charges and tenure attributes as this will likely yield better results. I decided to normalize the two attributes using the z-score method so that it is easier to detect outliers. I then utilized the x-partitioner and x-aggregator nodes for cross validation. I tried using both the random forest predictor and support vector machine learner nodes to see which method would yield the best results (both methods seemed like valid solutions for the final prediction). I found that the polynomial and rbf options yielded the best results (polynomial and rbf methods had on average equal accuracy) when using the support vector machine prediction method. I found it difficult to explore different options such as using different kernels in the support vector machine as it was computationally intensive. I found that a random forest yielded slightly better accuracy results for predicting wether a customer churns and was significantly less computationally intensive. Therefore I eliminated the support vector machine predictor and used the random forest predictor for fine tuning. I was surprised to find that when using cluster as an attribute did not improve the accuracy of the random forest or support vector machine. In the end I was able to predict with on average 80% accuracy wether or not a customer churned.

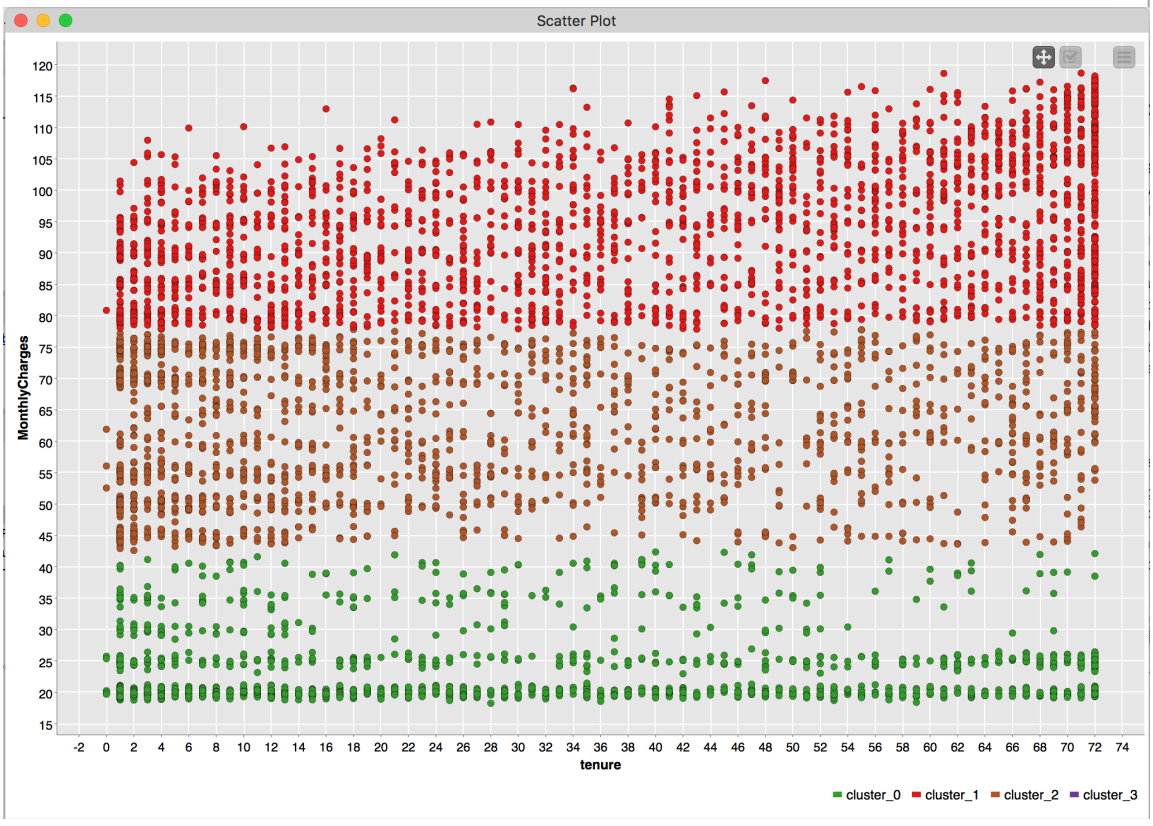
Conclusion:

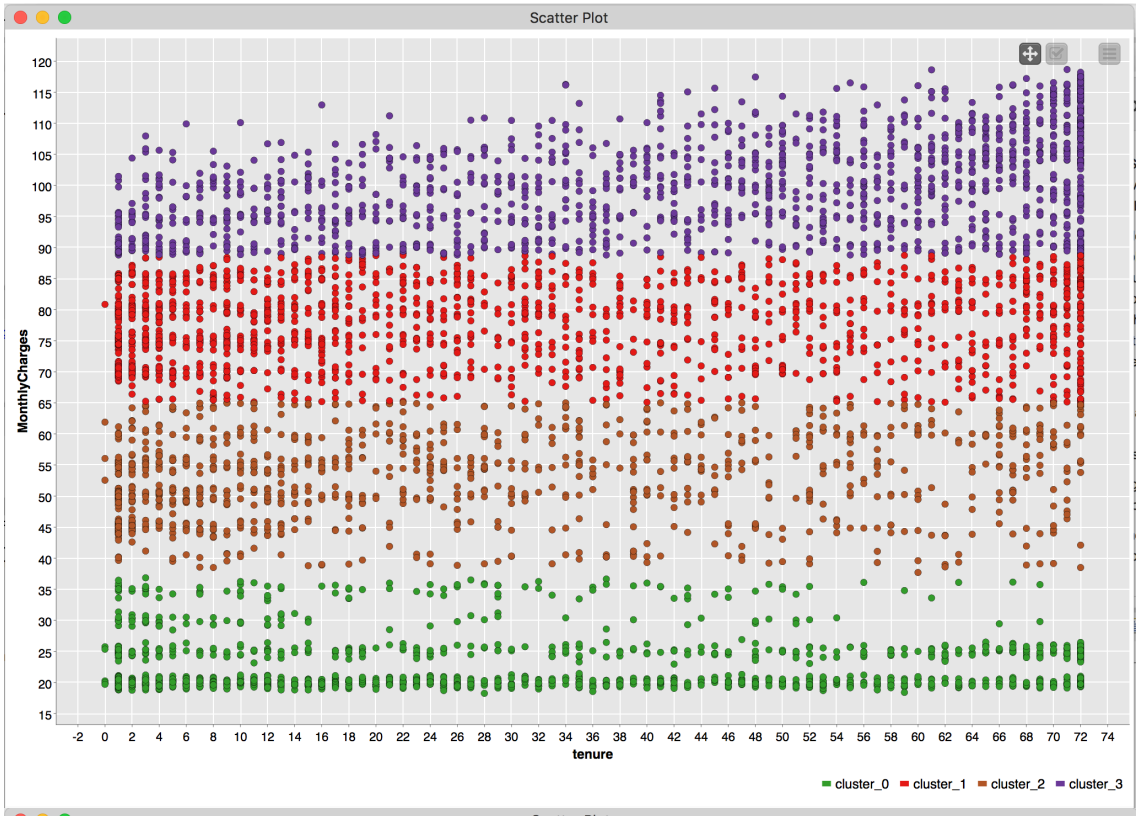
In conclusion, I found that customers were leaving the service provider mainly due to the duration of their contract. Customers who had month-to-month contracts and were paying with

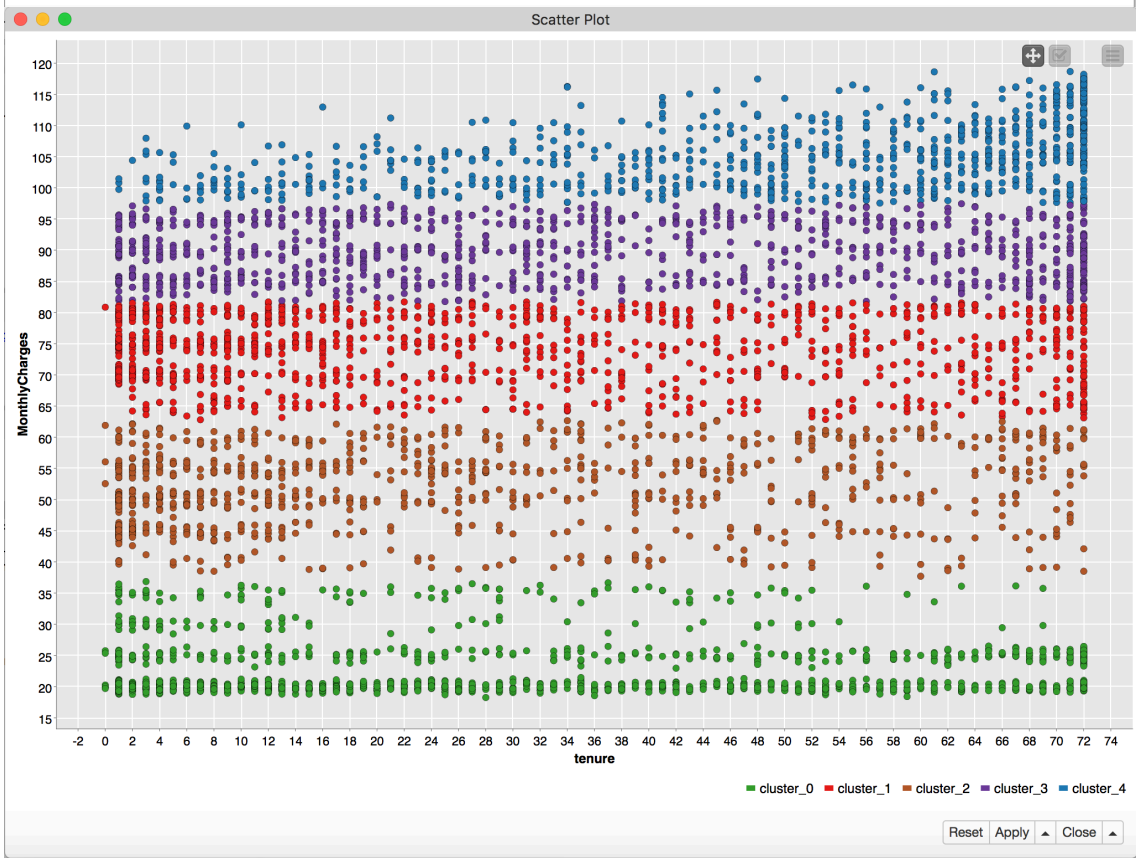
electronic checks or had fibre optic internet were the most likely to churn. Looking back I would have liked to work more with the hierarchical clustering algorithm to see if I could achieve better accuracy results. In addition, I would have liked to work with the target shuffling node to see if I can further eliminate attributes and get a better idea of why customers are leaving. Finally I would have liked to experiment with the overlapping penalty, and parameters for each kernel when using the support vector machine predictor. Overall I enjoyed the analyzing the data, however, I would have liked to achieve closer to 90% prediction accuracy. I think that it is possible to achieve such a high accuracy if more clever data analysis choices are made in determining whether or not a customer churns.

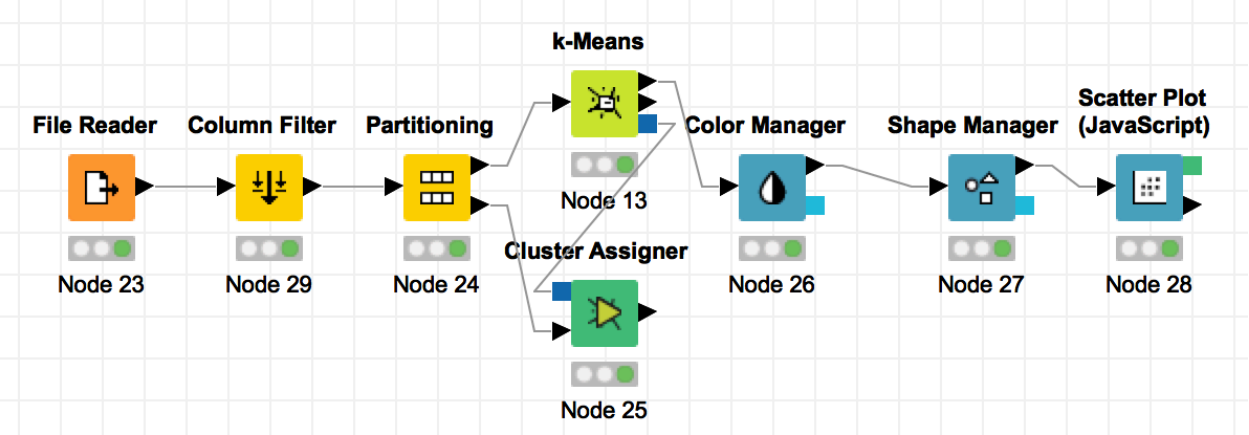
Appendix 1 - K-Means Clustering



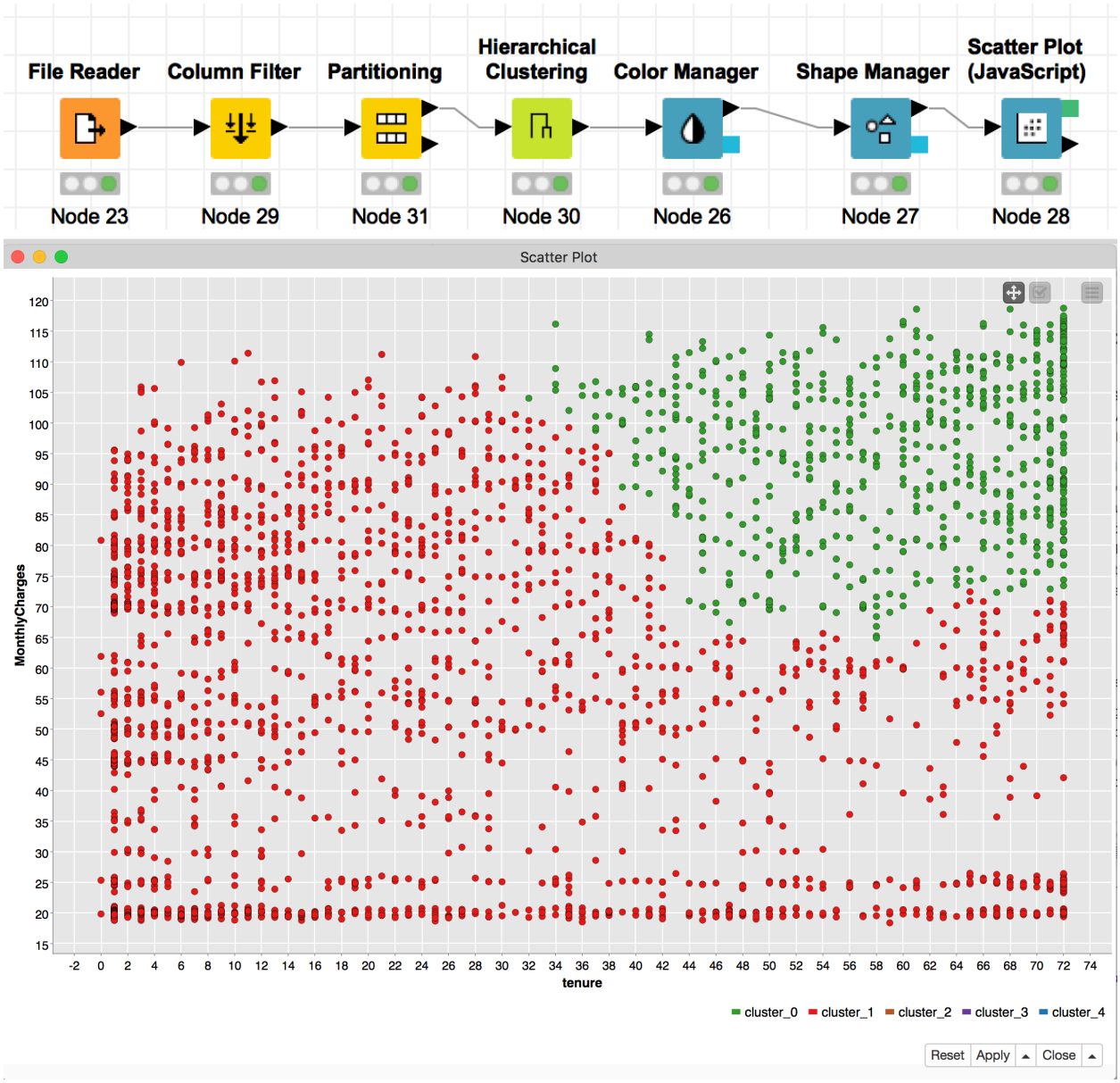


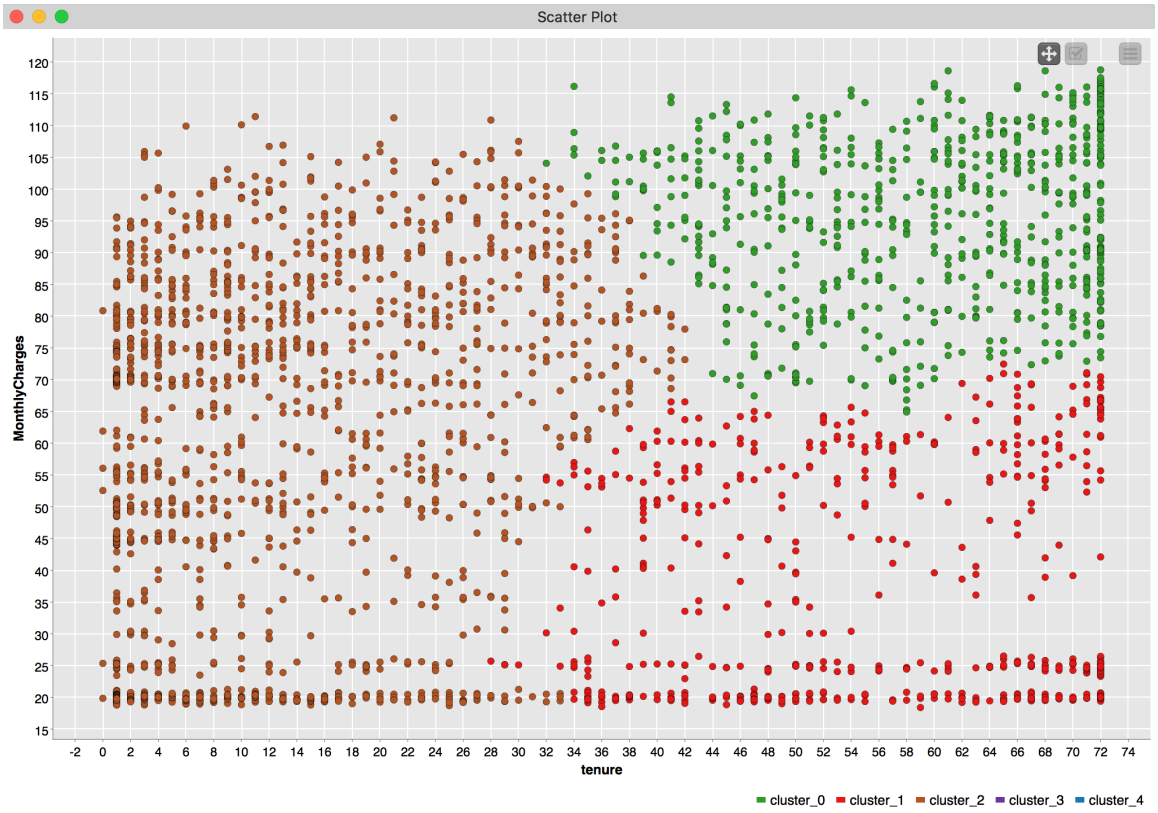

Scatter Plot
MonthlyCharges
tenure
cluster_0
cluster_1
cluster_2
cluster_3

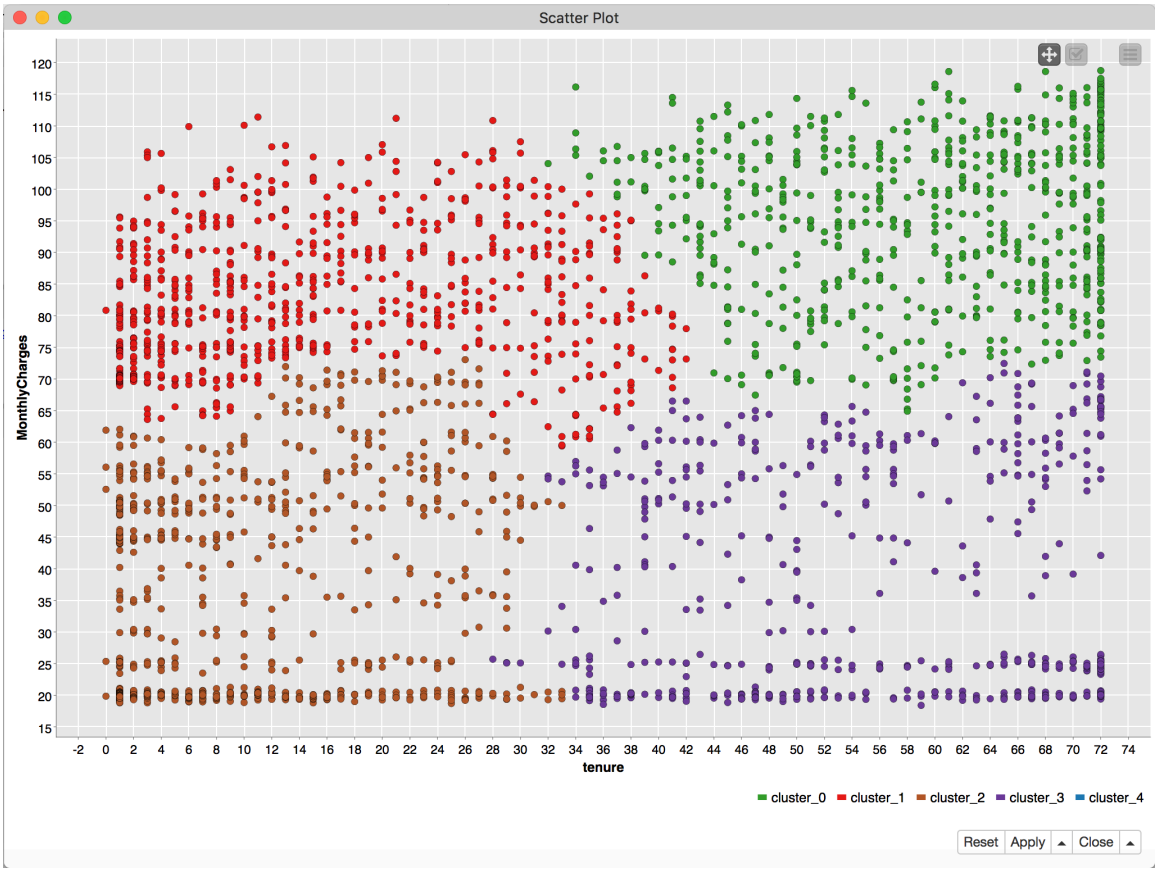

Scatter Plot
MonthlyCharges
tenure
cluster_0
cluster_1
cluster_2
cluster_3
cluster_4
Reset
Apply
Close

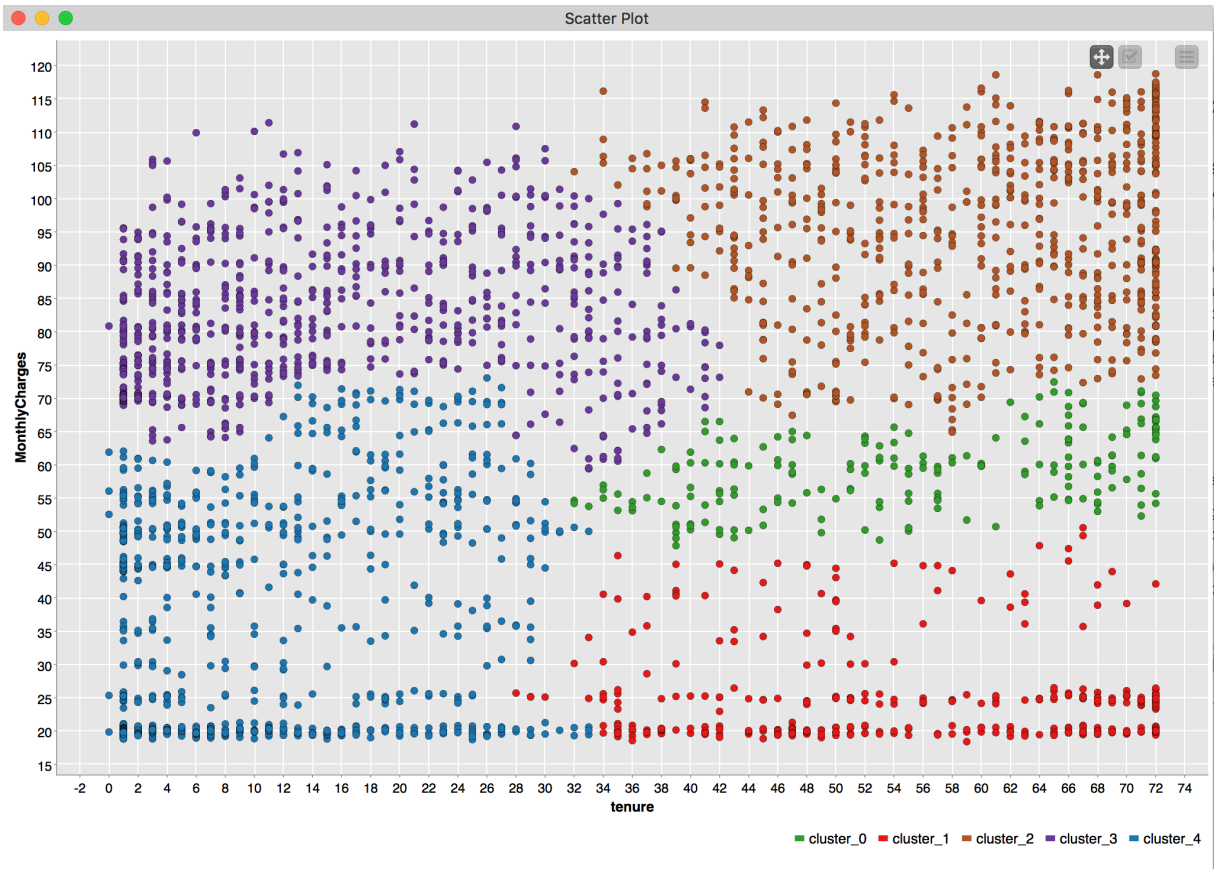

k-Means
File Reader
Column Filter
Partitioning
Color Manager
Shape Manager
Scatter Plot (JavaScript)
Node 13
Cluster Assigner
Node 23
Node 29
Node 24
Node 26
Node 27
Node 28
Node 25

Appendix 2 - Hierarchical Clustering

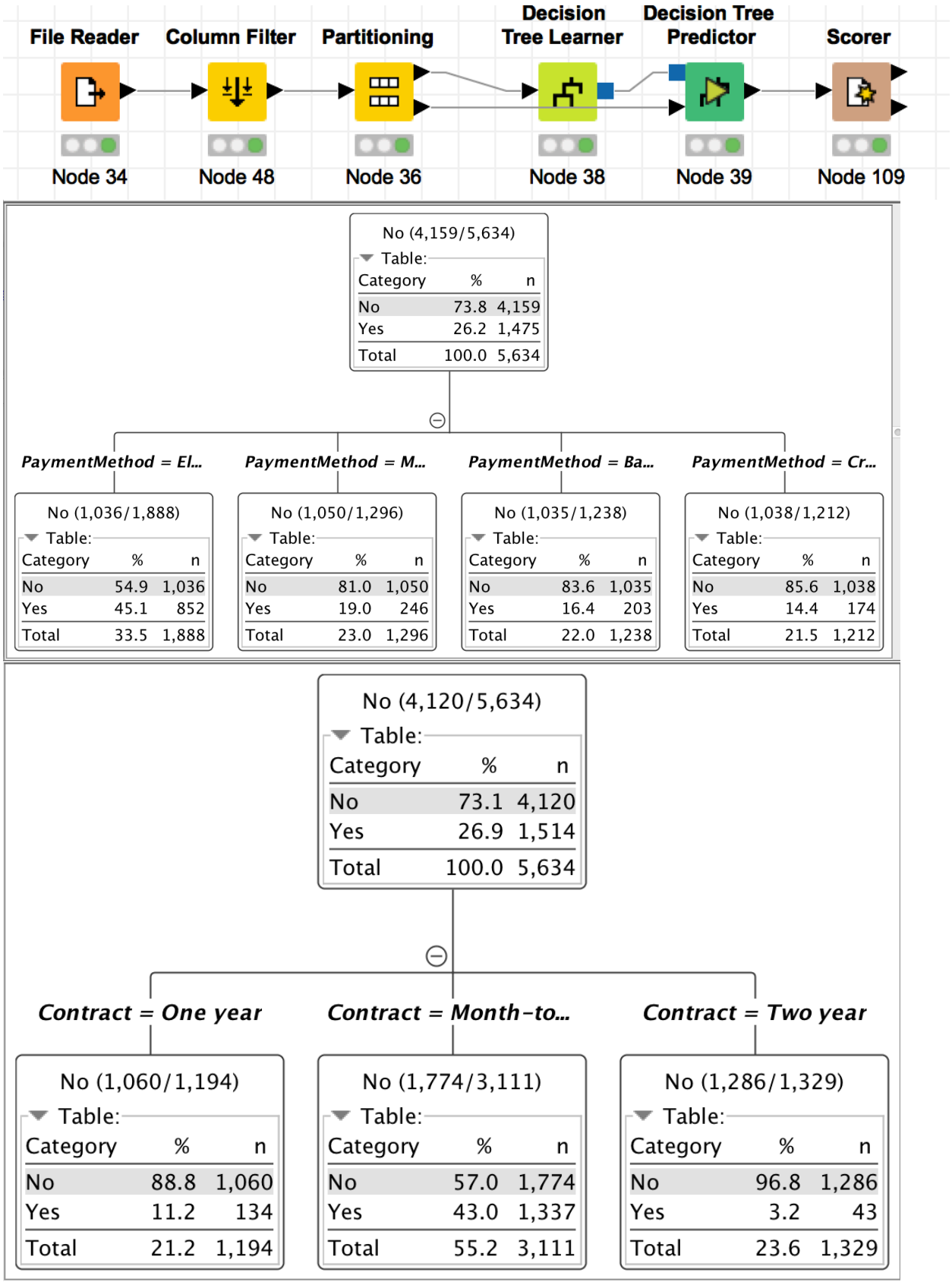



Scatter Plot
MonthlyCharges
tenure
cluster_0 cluster_1 cluster_2 cluster_3 cluster_4

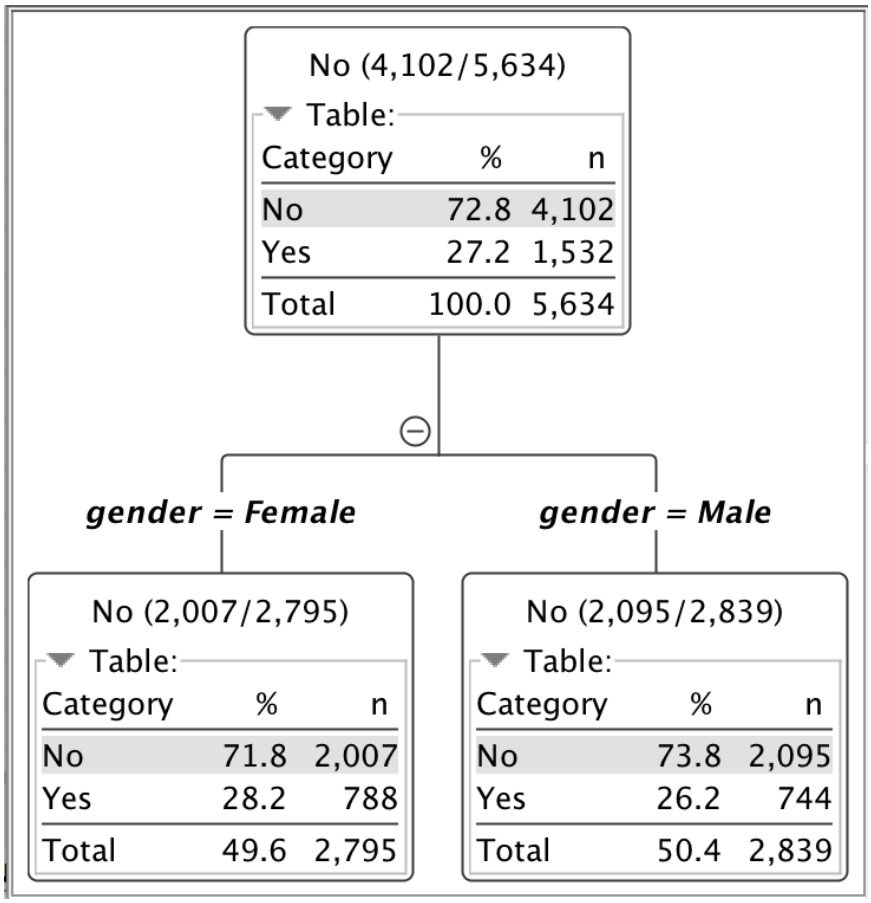

Scatter Plot
MonthlyCharges
tenure
cluster_0 cluster_1 cluster_2 cluster_3 cluster_4
Reset Apply Close

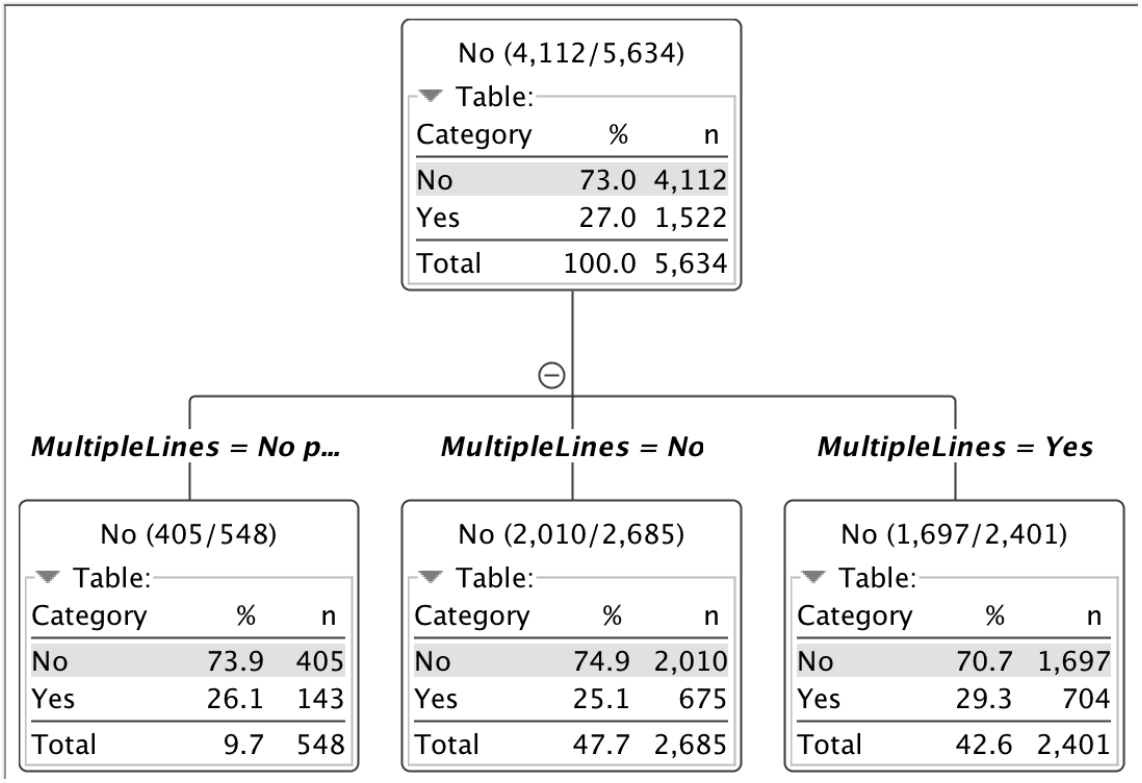

Scatter Plot
MonthlyCharges
tenure
cluster_0
cluster_1
cluster_2
cluster_3
cluster_4

Appendix 3 - Decision Tree

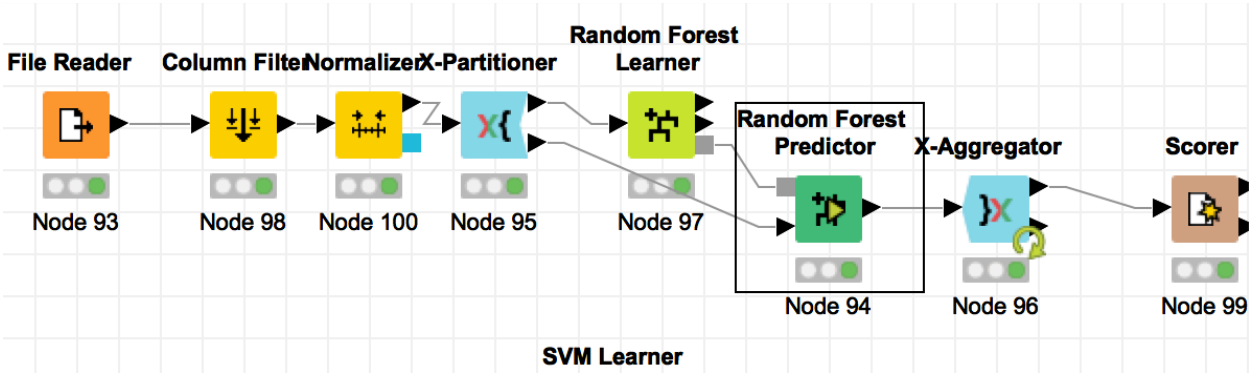


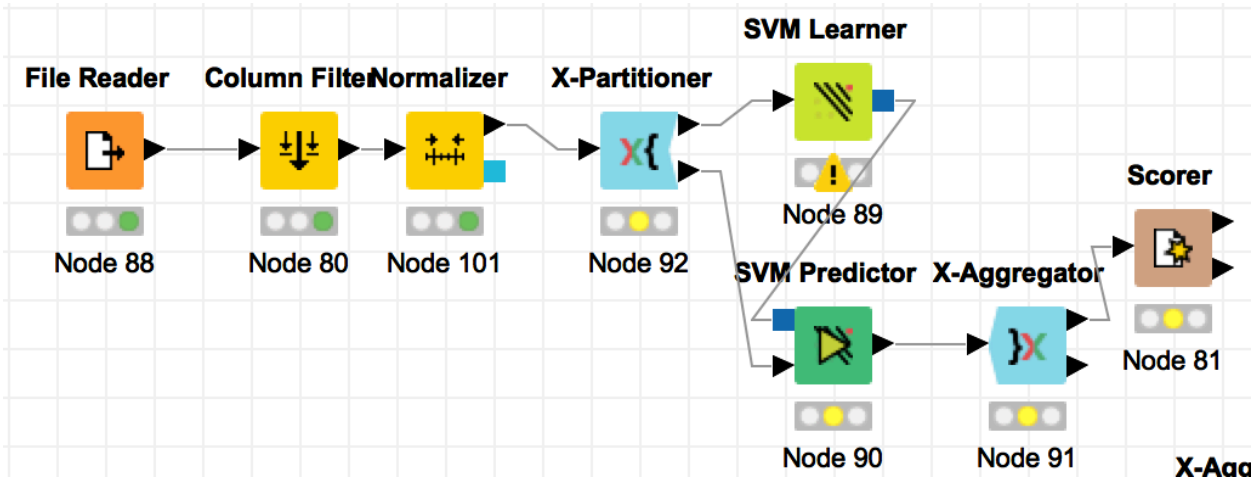
Appendix 4 - Decision Trees





Appendix 5 - Random Forest and Support Vector Machine





Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifity	F-me...	Accur...	Cohen...
Yes	754	724	3747	409	0.648	0.51	0.648	0.838	0.571	?	?
No	3747	409	754	724	0.838	0.902	0.838	0.648	0.869	?	?
Overall	?	?	?	?	?	?	?	?	?	0.799	0.442

Missing Value