

Node-C Technologies, Inc

Department of Research and Development

Quarter IV, FY 2019

Project II.a Vol.I Description/Methodology/Results

Principle Investigator(s): Richard Harry, Mohamed Gaber

*K-Means Machine Learning Approach for Multi-Variable Water Treatment Plant Fault Prediction and Process Optimization*

\*Note: The following document contains confidential information that is property of Node-C Technologies and noted affiliates. The information presented in this archive is for research purposes only and should not be taken as a marketable product.

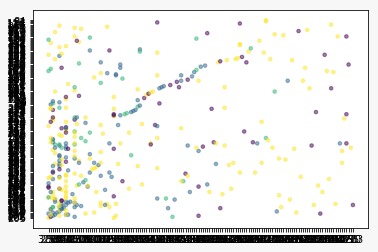
Methodology Questions

1. How many classes are we looking for in order to design the clusters?
2. How does knowing the class of each day help in predicting the faults
3. Do we need to run a model that includes all 38 variables, or are there more important variables than others?
4. How would you classify the four classes (Low, Low-Mid, Mid-High, High)?

* D-12/4/91
* D-16/2/90
* D-15/1/91
* D-19/2/90

Since this data is not labeled (no classification for any point) then we need to implement unsupervised machine learning. For that, we need to decide how many classes we are looking for so that I can design the clusters. So how many classes are you trying to get?

I have already built a clustering model using K-means approach. I used 4 clusters (arbitrarily) and I received a Silhouette Coefficient score of 0.502 which is a good score for the model



\*Important Notes for Paper Construction\*

Background (Methodology)

Silhouette Coefficient for Clustering

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.[1]

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.