The Artificial Intelligence for maintenance 4.0: The Unsupervised machine learning applied to Maintenance schedule optimization based on pumps bearing Remaining useful life prediction

Author: Dr. Eduardo Calixto. ecduardocalixto.com

Introduction

The Machine learning is an Artificial Intelligence subject that uses data about individuals' populations, disease, animals and plant samples for research as well as system´s equipment/component and even finance data to perform cluster, classification and prediction based on mathematic models, mostly supported by computer algorithms.
Concerning the maintenance engineering, the machine learning application is applied for the equipment criticality classification, failure regression predictions and the most advanced maintenance strategy, so called Prognostic Health Management, that aims to define equipment Remaining Useful Life (RUL) and State of Health (SoH) based on online monitoring data or non-destructive test by measuring the stressor factors such as vibration, voltage, temperature, humidity and other physical parameter that lead equipment degrade to functional failure.
The concepts behind machine learning is to use the set of knowing the data to cluster, classify or predict future variables response. In order to classify and predict the data response, the machine learning model divides the dataset in the training data (~70% of dataset) and test data (~30% of dataset). The further step is to apply an algorithm to training data for the learning process and then to come out with a model. The model will be applied to the test data and the verification of the result take place. If the result is satisfactory, the new data set can use the same model defined based on the previous dataset to make predictions and if the result is satisfactory the model is validated. The Machine Learning can be performed by different methods that will be explained further. The general steps of the machine learning process are described in figure 1.

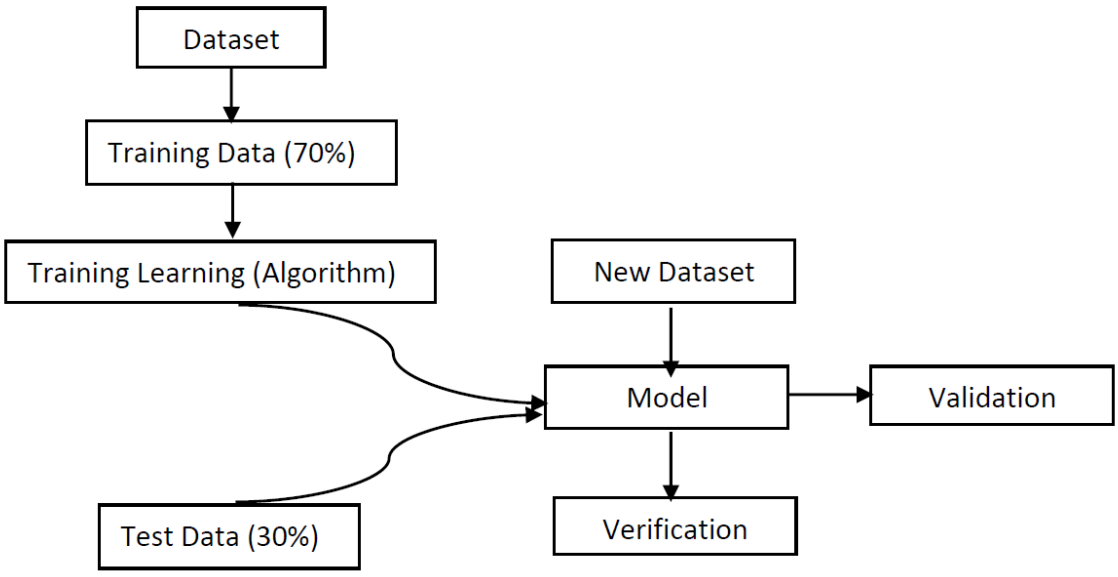

Figure 1: General Machine Learning Steps

The machine learning can basically be divided into unsupervised and supervised models based on different methods application as will be described in the next items.

2 Unsupervised Machine Learning

The Unsupervised Machine Learning aims to define a pattern in the set of data without previous knowledge of data features. Therefore, the first understanding of your data set can start by applying the Unsupervised Machine Learning methods to understand the how your dataset can be organized and if there´s a pattern of such dataset based on their independent variables.
The concepts behind Unsupervised Machine Learning is cluster a set of data without knowing previous classification or any information about the data. In order to cluster the data, the Unsupervised Machine Learning models the dataset and try to organize it in a cluster. The further step verifies the result based on error and finally, If the result is satisfactory, the new dataset can be used the model defined based on the previous dataset and then the model is validated. The general steps of the machine learning process are described in figure 2.

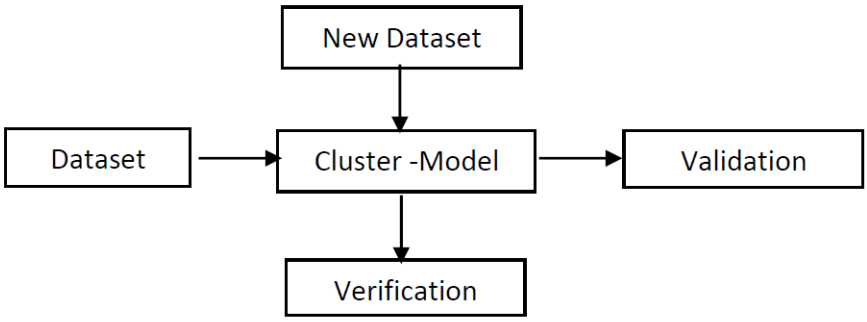

Figure 2: General Unsupervised Machine Learning Steps

Concerning the maintenance engineering, the type of data related to equipment encompasses physical characteristics as well as performance, cost of operations, cost of preventive maintenance, corrective maintenance, spare parts cost. Therefore, by defining some of such variables, it´s possible to group equipment with similar characteristics. In fact, the data organization based on its common features is the concept behind the clustering data that is the result of such unsupervised machine learning methods as shows the figure 2.

The most common used Unsupervised Machine Learning models used for clustering are:

- Principal Component Analysis;
- Multidimensional Scaling;
- K - Means ;
- Gaussian Mixture;
- Hierarchical Clustering;
- Neural Network Self-Organized Map.

The ***Principal Component Analysis (PCA)*** method aims to reduce the number of variables from multidimension space (multiple independent variables) to a one or a two-dimensional space (independent variables) and define the variables that more influence has in data variability.
The ***Multidimensional Scaling*** (MDS) aims to reduce the number of variables from multidimension space (multiple independent variables) to a one or a two-dimensional space (independent variables) and define the variables that more influence has in data variability. However, the MDS enables the reduction of the complexity of visualization of the multidimensional space in one or Bi-Dimensional space concerning the pairwise distance between the points
The ***K-Means*** aims to organize a set of data point in k different clusters considering the k different centroids and group the data closest to each k centroids.
The ***Gaussian Mixture*** aims to organize data based on the different Gaussian distribution of each cluster that each data fit better in. The main assumption of this model is that all K cluster are the K dataset clusters normal distributed.

The ***Hierarchical Clustering*** aims to cluster the dataset by defining the pairwise distance between the points and bason on such distance group the data in different clusters levels.
The ***Neural Network Self-Organized Map (SOM)*** cluster the data based on the weight defined by the Neural Network to each Neuron data considering the closest dataset point to each neuron data. The principles of SOM are, basically, given an observed data, the closest neuron is defined, then the neuron is moved closer to the observed data. The aftermath, the process repeats again by selecting a random data from the dataset.
The figure 3 shows the main concept of organizing a dataset in clusters that can be applied for maintenance planning, optimization considering the result of Remaining Useful Life (RUL) and Degradation Progression Signature (DPS) of a group of equipment.

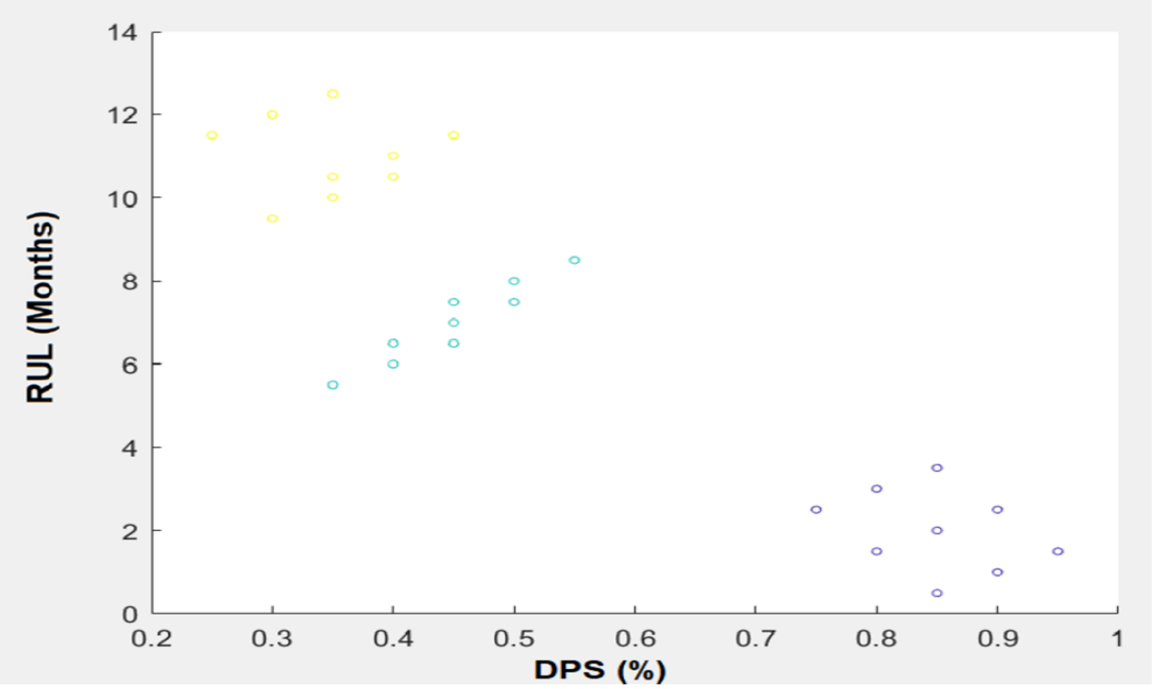

Figure 4 3: Maintenance Planning cluster based on RUL and DPS

3 - Supervised Machine Learning Classification (SMLC)

The Supervised Machine Learning aims to classify of predicting data based on previous knowledge of a set of data features. Therefore, the SML can be divided in two categories such as Classification and Regression.
The Supervised Machine Learning Classification (SMLC) has the main objective to classify data based on the features pre-observed in a dataset. Therefore, it´s applied to classify new equipment criticality, risk, high performance, bad actors, and other classification applied in the maintenance field. The main advantage of such approach is to classify a huge number of data based on the SMLC models enable to save the huge amount of time dedicated to such activities as well as to define alarms alert based on equipment and component criticality classification. The most commons SMLC models are the following:

- K-Nearest Neighbor (KNN);
- Decision Tree Classification;
- Naïve Bayes;
- Linear Discriminant Analysis;
- Support Vector Machine;
- Neural Network Classification;
- Logistic Regression Classification.

The general steps of Supervised Machine Learning Classification are described in figure 4-4.

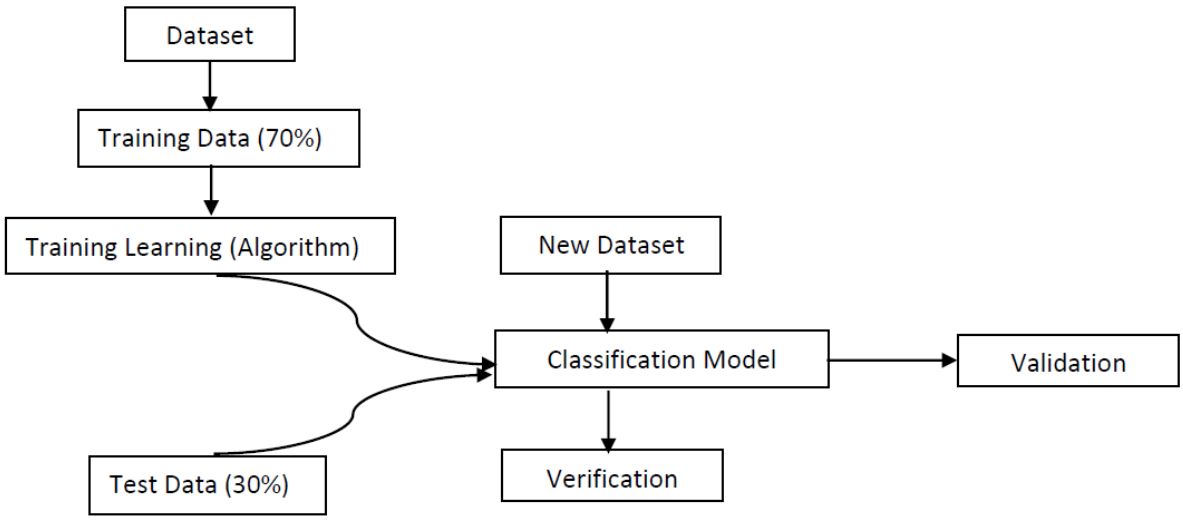

Figure 4: General Supervised Machine Learning Classification Steps

Concerning the maintenance application, the equipment characteristic such as criticality, risk, high performance, bad actors, and other classifications enable the maintenance leaders prioritize the equipment. In addition, the SMLC also enable to define alert limit levels to be part of to the PHM program. The figure 5 shown an example of SMLC applied to equipment criticality classification based on the level of Remaining Useful life (RUL) and vibration (RMS). This example shows the definition of different level of criticality, such as critical (C), Moderate (M) and Low (L).

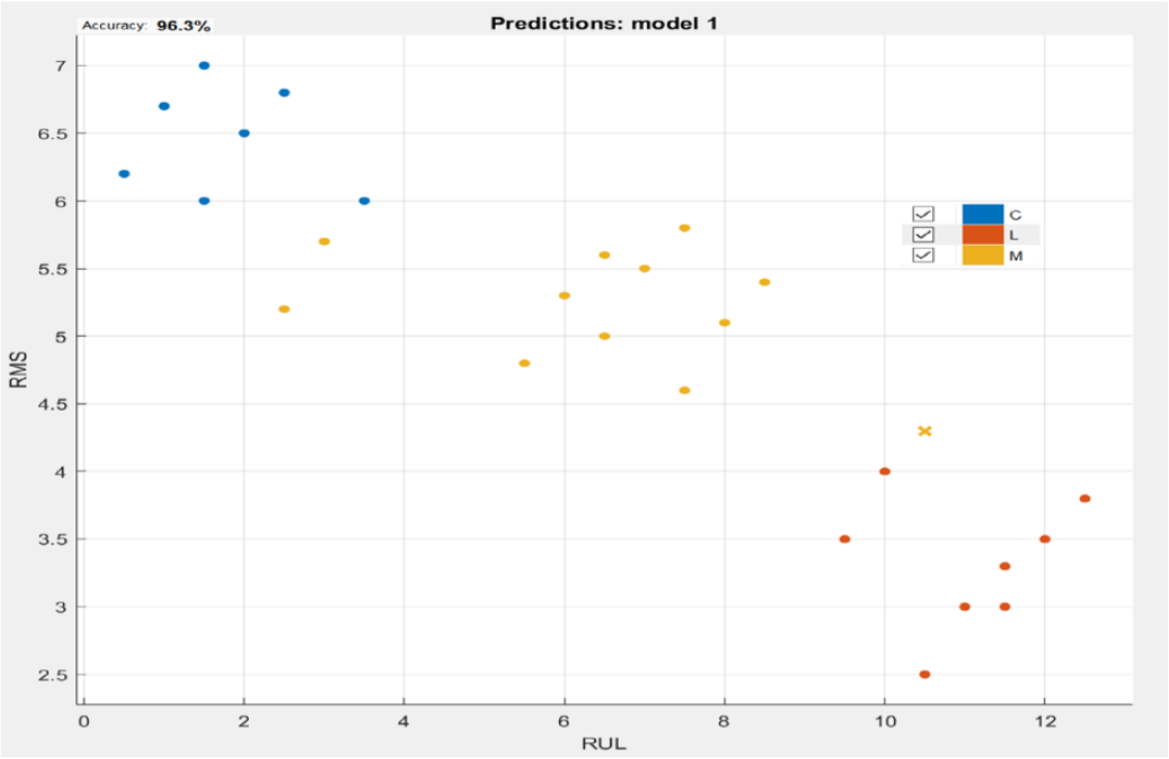

Figure 5: Equipment Classification based on RUL and RMS(mm/s)

4 Supervised Machine Learning Regression (SMLR)

The Supervised Machine Learning Regression (SMLR) has the main objective to predict and forecast future values of dependent variable data based on the independent variables that are equipment/component features of pre-observed dataset. Therefore, it´s applied to predict process variable's value, Remaining Useful Life (RUL), State of Heath (SoH) and other parameters used in the maintenance field. The main advantage of such approach is to predict the parameter values automatically based on current historical data. Therefore, the SMLR models enable to save the huge amount of time dedicated to such activities as well as to link sensor data to PHM models to predict equipment and component RUL and SoH. The most common SMLR models are the following:

- Linear Regression;
- Ridge and Lasso Regression;
- Stepwise Linear Regression;
- Gaussian Regression;
- Decision Tree Regression;
- Support Vector Machine Regression,
- Neural Network Regression (NNR).

The general steps of SMLR process are described in figure 6.

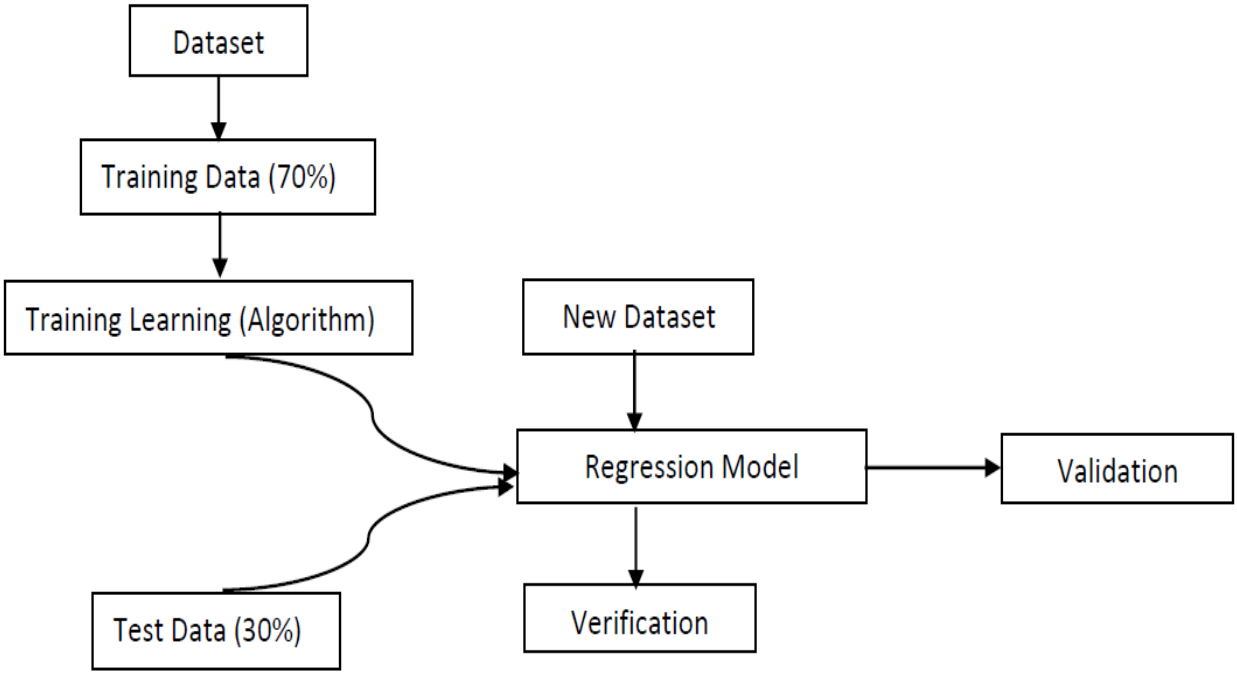

Figure 6: General Supervised Machine Learning Regression Steps

Concerning the maintenance application, the equipment independent variable such as temperature, humidity, rotation, vibration and others variables measured by the sensor as well as degradation measurement such as corrosion, erosion and crack thickness measured by Non-Destructive Test can be used as input for RUL, SoH and other parameter prediction. The figure 7 shows an example of RUL prediction based on temperature degradation (TDPS).

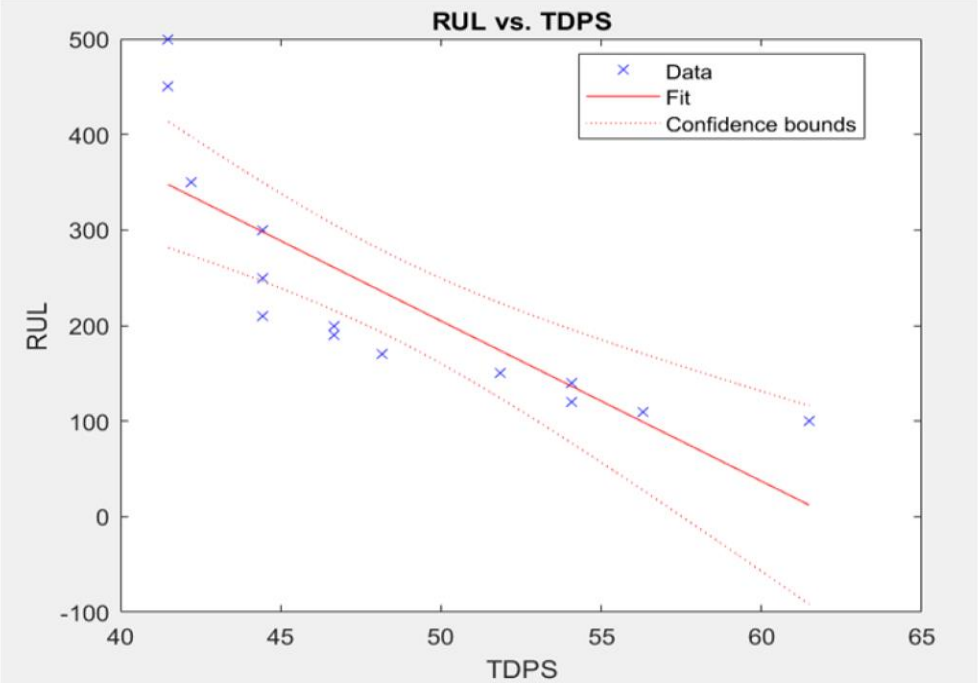

Figure 7: RUL regression

Before finishing the theoretical part of this paper, its important to highlight one important concept applied to Machine learning, that is the overfitting. As we mentioned before, when performing Supervised Machine Learning classification or regression, it's necessary to split the dataset in training (≃70%) and test data (≃30%). Whenever the model fits too much of the data with very high accuracy or very low error, we say the model is overfitted. Since the main objective is to create a general model, that can be applied to new data, the overheating is not good for creating a general model.
The other important concept is the Bias and Variance. The Bias is the inability of a Machine learning model to capture the true relationship between the variables (dependent and independent). The Variance is the difference between training data and test data (or new data) fitting. The overfitting model present low Bias but high variance. The Underfitting is the contrary of overfitting concepts and in this case the model presents high Bias but low variance.
The trade-off between simple and complex classification and regression models is to find the model that performs results with low Bias and low variance.
In fact, as much as possible, we try to find the model with low bias and low variance, but its usual to have some error in the classification and regression prediction.
The next chapters will give a deep explanation about the different Machine Learning methods with examples applied to the maintenance field for the reader clear understanding.

5 - The K-Means Unsupervised Machine Learning methods

The K-Means clustering method is an Unsupervised Machine Learning method for data clustering. The K-Means objective is to organize a set of data point in k different clusters considering the k different centroids and group the data closest to each k centroids. The K means algorithm follows the further steps:

- To define the value of K, that means, the number of clusters that the data set will be organized.
- To define a random centroid point to start the clustering process.
- To calculate the minimum distance between the closest points, close to centroids.
- To organize the dataset point based on the nearest points closest to each centroid.
- Update the centroid point based on average positions of each group of data.
- To define the new centroids and run the process again. Compare the minimum distance between the centroids and the points inside each cluster.

In order to define the distance among the k neighbors point, the distance can be calculated based on Euclidian distance, Manhattan Distance or Minlowski distance as the following equations:

$$Euclidian\ distance = \sqrt{\sum_{i=1}^{K}(X_i - Y_i)^2}$$

or

$$Manhattan\ distance = \sum_{i=1}^{K}|X_i - Y_i|$$

or

$$Minlowski\ distance = \left(\sum_{i=1}^{K}(|X_i - Y_i|)^q\right)^{1/q}$$

To assign n data points $(x_1, x_2, x_3 \ldots . x_n)$ to a j clusters $(c_1, c_2, c_3 \ldots . c_j)$

By defining the center of the cluster μ_j for the specific j^{th}cluster

Where:

$$\mu_j = \frac{1}{|c_j|}\sum_{x_i \varepsilon c_j} x_i$$

Then, minimize the distance between the data point and the cluster center represented by the function J

$$J = \sum_{j=1}^{k}\sum_{x_i \varepsilon c_j}\left\|x_i - \mu_j\right\|^2$$

This process will repeat based on the number of interactions defined or when the center of the cluster does not change significantly anymore.
The best centroid points are the one that the cluster data has the minimum distance to each centroid and chose the centroids with the minimum distance as shows figure 8.

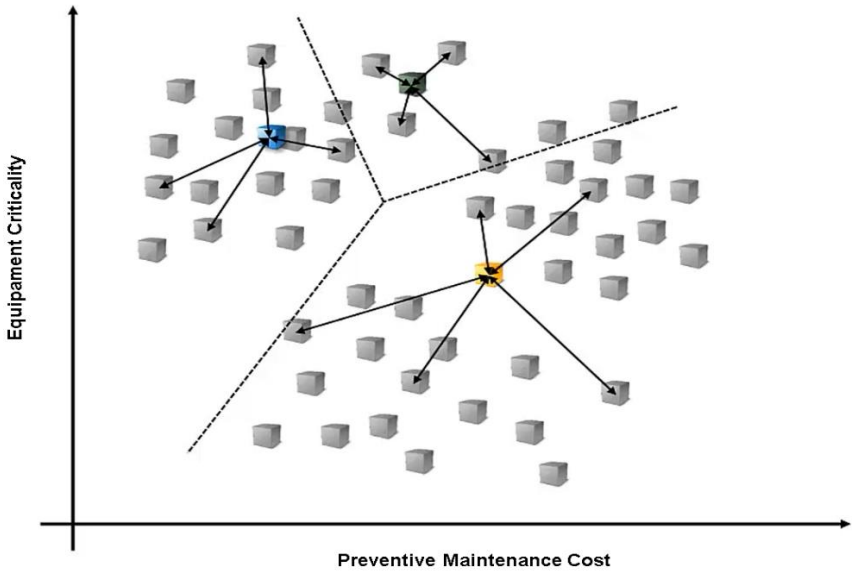

Figure 8: K-Means Centroids (Source: Adapt from MATLAB)

In order to practice the K-Means concepts, we can group the equipment of a different classes in different class based on the criticality or even define the time to perform preventive interventions based on Prognostic Health Management prediction from a group of equipment with different Remaining useful life.

Concerning the PHM, the K means cluster can be a very good solution for this type of problem, where a group of equipment need to be planned for preventive intervention based on the result of the RUL prediction as a result of PHM assessment in several equipment. The table 5-3 shows the summary of the twenty-seven pumps with different Root Means Square (RMS - mm/s – vibration velocity), Degradation Progression Signature (DPS) and Remaining Useful Life (RUL).

Table 1: Pumps Bearing Vibration degradation

	RMS (mm/s)	DPS (%)	RUL (Months)
Pump 1	3.000	0.400	11.000
Pump 2	3.500	0.300	12.000
Pump 3	4.000	0.350	10.000
Pump 4	5.100	0.500	8.000
Pump 5	5.500	0.450	7.000
Pump 6	5.300	0.400	6.000
Pump 7	5.700	0.800	3.000
Pump 8	6.500	0.850	2.000
Pump 9	6.700	0.900	1.000
Pump 10	2.500	0.350	10.500
Pump 11	3.000	0.250	11.500
Pump 12	3.500	0.300	9.500
Pump 13	4.600	0.450	7.500
Pump 14	5.000	0.400	6.500
Pump 15	4.800	0.350	5.500
Pump 16	5.200	0.750	2.500
Pump 17	6.000	0.800	1.500
Pump 18	6.200	0.850	0.500
Pump 19	3.300	0.450	11.500
Pump 20	3.800	0.350	12.500
Pump 21	4.300	0.400	10.500
Pump 22	5.400	0.550	8.500
Pump 23	5.800	0.500	7.500
Pump 24	5.600	0.450	6.500
Pump 25	6.000	0.850	3.500
Pump 26	6.800	0.900	2.500
Pump 27	7.000	0.950	1.500

In order to implement the K-Means method the MATLAB software is applied. The K-Means cluster result is demonstrated in the figure 5-10 where the data is organized is K= 3 clusters.

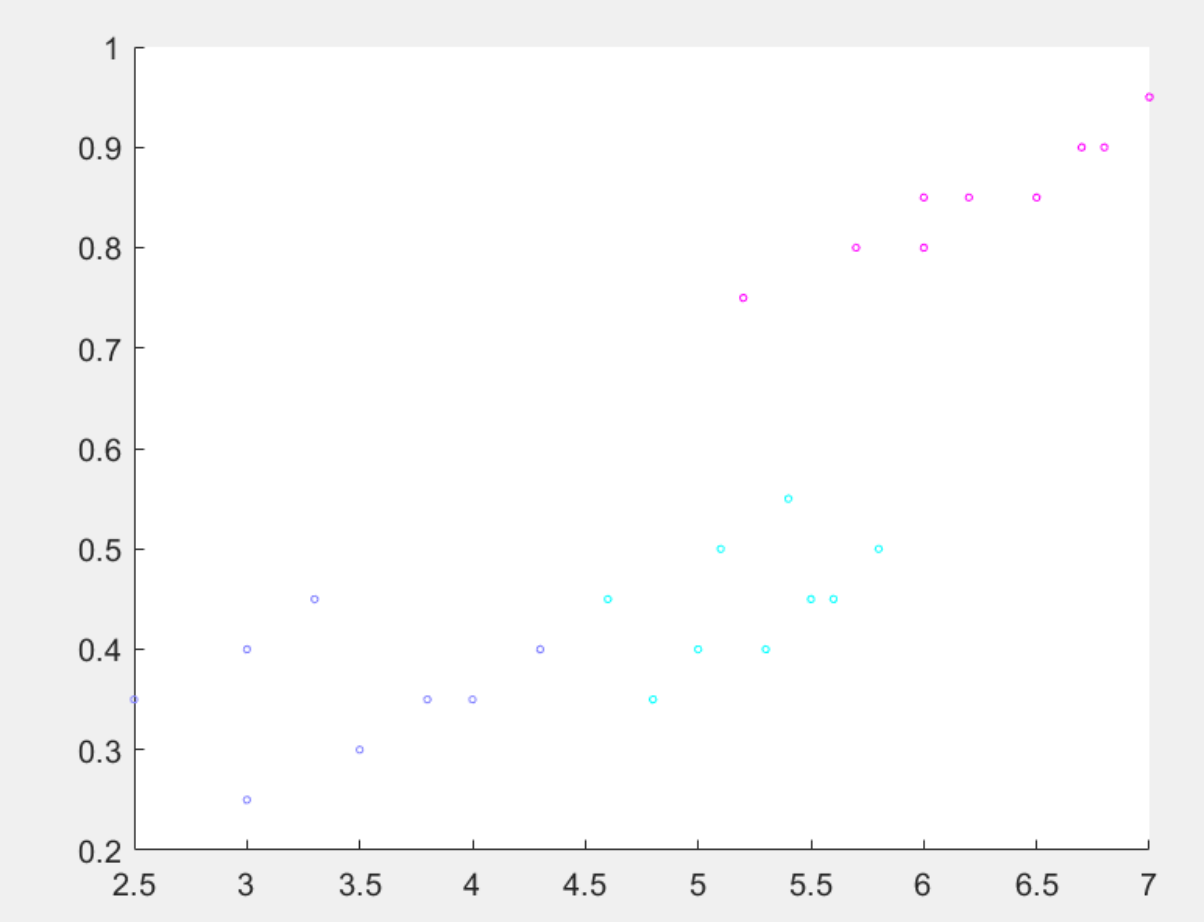

Figure Error! No text of specified style in document.**-1: K-Means Cluster Result**

The K-Means cluster enable a fast and precise cluster of the group of pumps with different values of RMS, DPS and RUL described. The figure 5-10 shows the different clusters by different colours such as pink, dark blue and light blue. The cluster on the top on the right side marked by pink colour represents the pumps with higher RMS (5.5≤RMS≤7) and highest DPS(0.75≤RMS≥0.95). The cluster in the middle marked by light blue colour represent the pumps with medium RMS(4.5≤RMS≥6) and medium DPS(0.35≤RMS≥0.4). The cluster in the left low corner marked by dark blue colour represent the pumps with the lowest RMS (2.5≤RMS≤4.5) and lowest DPS(0.25≤RMS≥0.45). Therefore, the K-means methods, clustering demonstrate that the most critical point is the pumps cluster in the right high corner with the highest RMS and DPS.

Therefore, the maintenance expert can get in the conclusion that the pumps that belongs to these clusters have more critical bearing degradation and will be the first to be prioritized for a preventive intervention. In this case, the K-Means methodology enables to give an automatic response of the group of equipment that needs intervention considering the RMS, DPS and RUL. In this case, each cluster will have the preventive intervention time based on the lower RUL of the pump in the cluster to minimize the risk of late intervention after equipment failure.

Since the PHM results applied to the pumps RUL prediction (or other equipment) is totally dynamic on time, whenever new values of RMS, DPS and RUL comes out, the K-Means will cluster the equipment in different groups and set up a new time of preventive interventions for each group of equipment in each cluster. Now imagine how the challenge is for the maintenance team to plan hundreds of equipment under the PHM program that's will produce different RUL every day. The K-Means cluster solve this issue and make easier the maintenance team maintenance planning in daily basis. Therefore,

the K-means algorithm in the end need to be integrated with the Asset Management Intelligence that will be described in another paper.

6 - Conclusion

After implementation of the Unsupervised Machine Learning method K-Means cluster we get into some conclusion concerning the advantages and drawbacks are the following:

The K-Means advantages are:

- Simple understanding and application.
- K-means works fine when the cluster data is circular.
- Good result conversion.
- Easy to adapt to new examples of the dataset.

The K-Means drawbacks are:

- The initial K values have a high influence on the clustering result.
- K-means does not account for variance.
- K-means does not work fine when the cluster data is not circular.
- K-means tells us what data point belong to which cluster, but won't provide us with the probabilities that a given data point belongs to each of the possible clusters.
- Centroids can be influenced by outliers,
- Outliers might get their own cluster instead of being ignored.
- In higher dimensional space the centroid definition becomes more complex and unclear.

In general, the advantages of Machine Learning application are:

- Classification of data very fast based on the model defined by knowledge of previous dataset classification. Such classification will save the huge effort to classify data based on the risk priority number in daily basis as well as other type or criteria of classification.
- Prediction of Remaining Useful life based on past data to anticipate the future failure when equipment is operating over their operation design limits condition and degrade before the expected time.
- To group equipment in different clusters based on their common features and enable the preventive maintenance plan optimization, overhaul plan optimization and priority defined for group of critical equipment.
- In recognition of image of damage equipment gives an alert for the control of production lines or even for maintenance and inspection teams when cameras are installed and are integrated with deep machine learning solutions.

The disadvantage of Machine Learning lies on it complexity and high demand of different knowledge such as:

- Requires a deep mathematical background to understand the machine learning methods.
- Requires a programming skill to use Machine Learning software such as MATLAB or even a deep programming skill to create algorithms in different languages such as Python, R and others.
- Requires a maintenance knowledge to have a clear understanding about the variable to be measured by sensors, installed the sensor properly, understand the sensor data and
- Requires a clear understanding about maintenance critical parts, failure modes, cause to create a proper classification label.

References

Amnon Shashua. Introduction to Machine Learning. 67577 - Fall, 2008. School of Computer Science and Engineering. The Hebrew University of Jerusalem. Jerusalem, Israel.

Calixto. E. Gas and Oil Reliability Engineering: Modelling and Simulation. Second edition, Elsevier ISBN: 9780123919144 – (Release in 26 May2016). http://store.elsevier.com/Gas-and-Oil-Reliability-Engineering/Eduardo-Calixto/isbn-9780128054277/.

Calixto, Eduardo. Integrated Asset Integrity Management: Risk Management, Human Factor, Reliability and Maintenance integrated methodology applied to subsea case. ESREL 2015.Zürich,Switzerland. 2015 Taylor & Francis Group, London, ISBN 978-1-138-02879-1

Calixto. E. Artificial Intelligence for Maintenance 4.0. First edition, AMAZON ISBN: ISBN: 9798672601021 (Release in 02 August 2020).

Daniel Géron. Machine Learning for Beginners: A Step-by-Step Guide to Learning and Mastering Machine Learning for Absolute Beginners with Real Examples.

EFNMS Asset Management Survey. ESREDA conference. Porto,Portugal, May 2013 http://www.efnms.org/mod/newsarchiv/view/cp-m10/newsmeldung-28

Guidelines for Vibrations in Reciprocating Compressor Systems, European Forum Reciprocating Compressors (EFRC), 2009.

Goodman, Douglas. Prognostics and Health Management: A Practical Approach to Improving System Reliability Using Condition-Based Data (Wiley Series in Quality & Reliability Engineering). 2019.

ISO 55002 standard, 2014. http://www.bsigroup.com/en-GB/search-results/?q=ISO+55002.

Jos Luis Rojo-lvarez; Manel Martnez-Ramn; Jordi Muoz-Mar; GustauCamps-Valls, ”Support Vector Machine and Kernel Classification Al-gorithms,” in Digital Signal Processing with Kernel Methods, IEEE,2018, pp.433-502.[17]

PAS 55 standard, 2008. http://www.bsigroup.com/en-GB/search-results/?q=PAS+55

Paul Wilmott Machine Learning: An Applied Mathematics Introduction (English Edition).Pnada Ohana Publishing, 2019.

Phil Kim, MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence. APRESS. Juni 2017.

Stuart Russell und Peter Norvig. Artificial Intelligence: A Modern Approach, Global Edition Pearson. Mai 2016

https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw

https://www.youtube.com/channel/UCNU_lfiiWBdtULKOw6X0Dig

https://www.youtube.com/channel/UCOgKur3rytBtcKkNrBMiRZQ

https://www.enkelt.co.uk .2017.

https://www.ridgetopgroup.com/products/advanced-diagnostics-and-prognostics/sentinel-motion/railsafe-integrity-analysis-system/

https://www.maximocon.com, 2018.