Study of clustering methods of Linac outages: K-Means vs G.M.M.

Aleksandar Dyulgerov

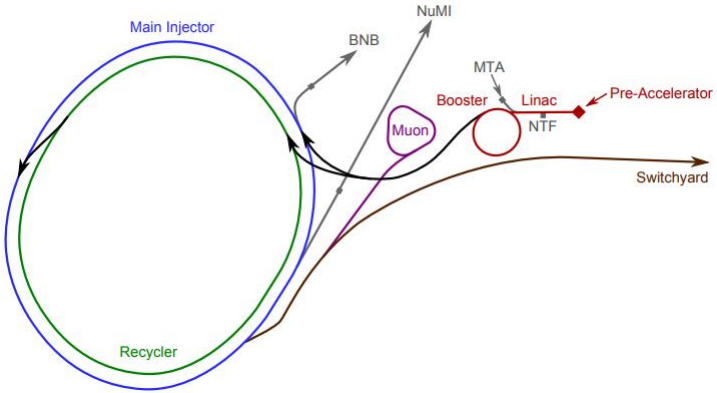
CCI Project Presentation

10 August 2022

Supervisors: Jason St. John, Brian Schupbach

Linear Accelerator

- The Linac (Linear accelerator) accelerates H- ions to more than 70% the speed of light, at 400 MeV kinetic energy.
- This is achieved with a carefully tuned series of RF (Radio Frequency) accelerating cavities.
- For the RF stations and support systems, the Linac presents about 2800 control system devices, monitored and operated by the Main Control Room.



Linac within Fermilab's accelerator complex
https://operations.fnal.gov/rookie_books/concepts.pdf

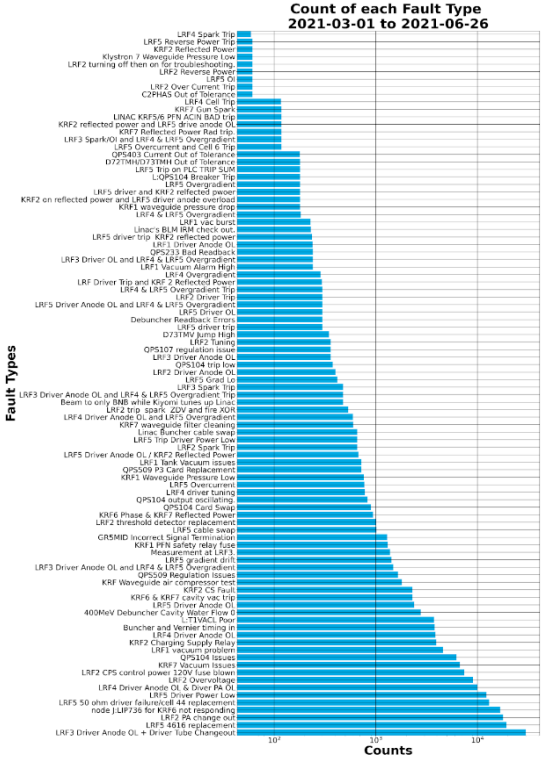
Linear Accelerator, Photo Credit: Reidar Hahn

Fermilab

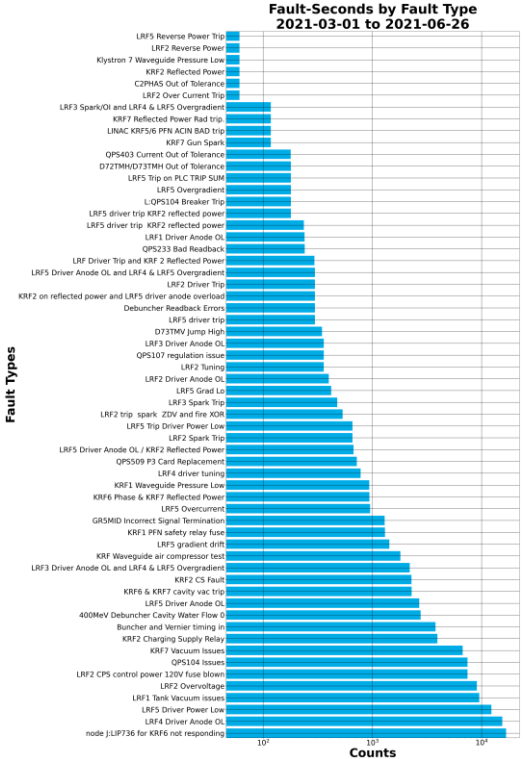
L-CAPE Project

- The L-CAPE project is a collaboration between PNNL (Pacific Northwest National Laboratory) and Fermilab.
- L-CAPE has been utilizing training of Long-Short-Term-Memory AutoEncoders to identify precursor to outages.
- The goal is to predict upcoming outages, how long an outage will be, and to automatically identify the type outage.

Data Cleaning and Interpretation



Count of each Fault Type: Before Data Cleaning



Count of each Fault Type: After Data Cleaning

Fermilab

Methods: UMAP

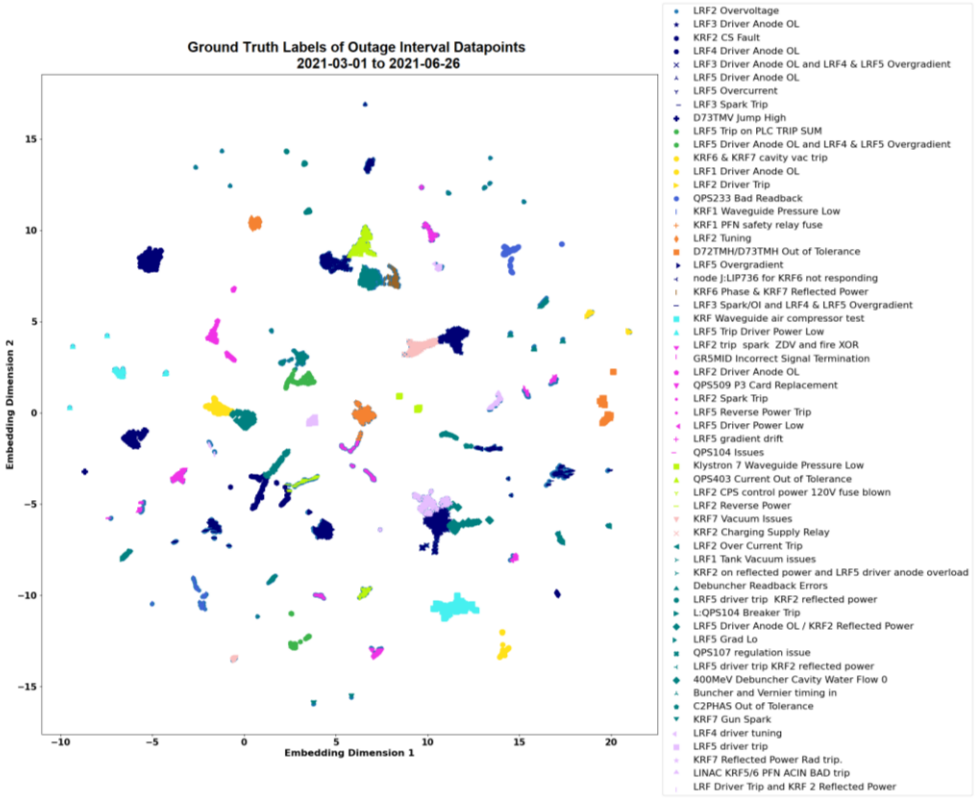
- UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) is a manifold learning and dimension reduction algorithm.
- Dimensional Reduction through UMAP allows to project all of the outage points into a two-dimensional space keeping their nearby associations.
  → Our machine state is 240 dimensions (24 devices * 10 one-second samples)

Why UMAP?

→ The output of UMAP enables studying and defining the boundaries of clusters to predict future outages.

→ Dimensional reduction is needed to avoid the Curse of Dimensionality (Clustering is ineffective in high dimensional spaces because everything is nearby in some direction)

Ground Truth Labels



Ground Truth Labels shown using color and marker

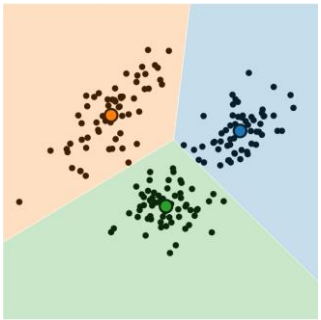
Fermilab

Methods: K-Means

- K-Means: partitions data into non-overlapping K set number of clusters.

  Steps Involved:
  Initialize K number of centroids randomly and partition data.
  Numerous reiterations occur until cluster assigning no longer changes:
  1. Assigning data points to the closest centroids using Euclidean distance.

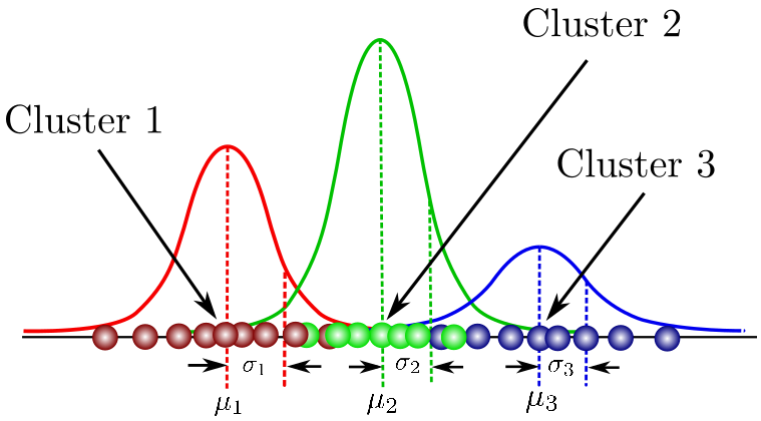
  2. Recomputing the centroid of each cluster by calculating the mean of the datapoints.



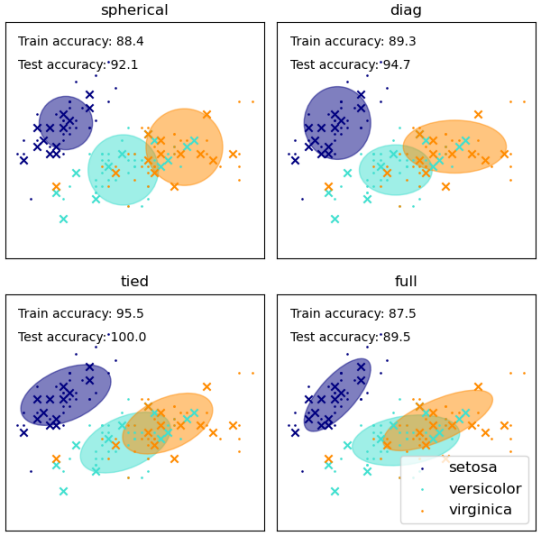
K-Means Voronoi Plot
https://antoinebrl.github.io/blog/kmeans/

Methods: Gaussian Mixture Model (G.M.M.)

- G.M.M. is a soft clustering algorithm
- Contains a finite K number of gaussians where each data point is assigned a probability of the gaussian distribution it falls under.
- The G.M.M. algorithm relies on two parameters: the mean and covariance.



Example 1: 1D G.M.M. clustering
https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95



Example 2: 2D G.M.M. clustering
https://scikit-learn.org/stable/modules/mixture.html

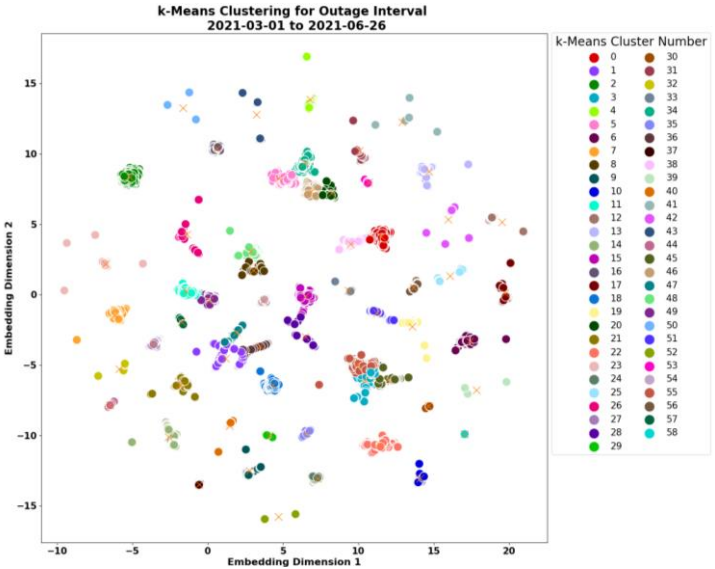
Fermilab

Methods: Homogeneity and Completeness Metrics

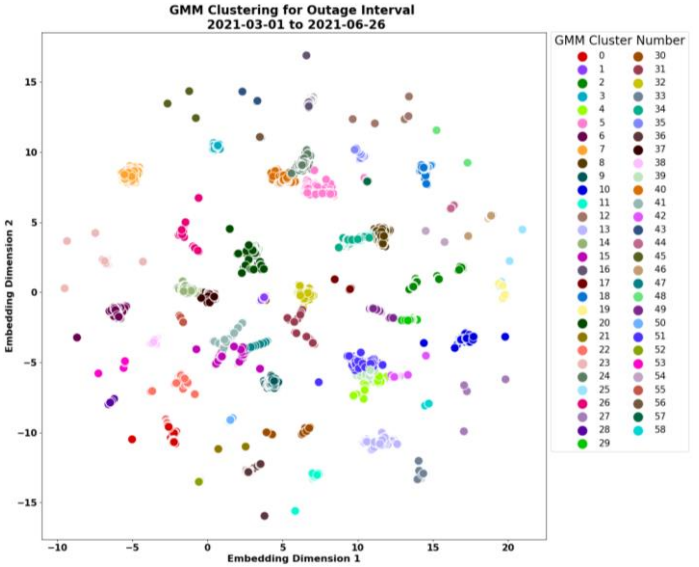
- To evaluate cluster performance, homogeneity and completeness scores were deployed.
- Homogeneity shows what amount of predicted clusters contain only members of a single class.
- Completeness measures if all members of a class are assigned to the same cluster.
- Both metrics use a scale of 0 to 1 where a higher score is better.

Results: Clustering

- In the dimensionally reduced UMAP preclustering plane for Outage Interval, G.M.M and K-Means both separately generated and assigned 59 unique cluster numbers to each individual outage interval data point.



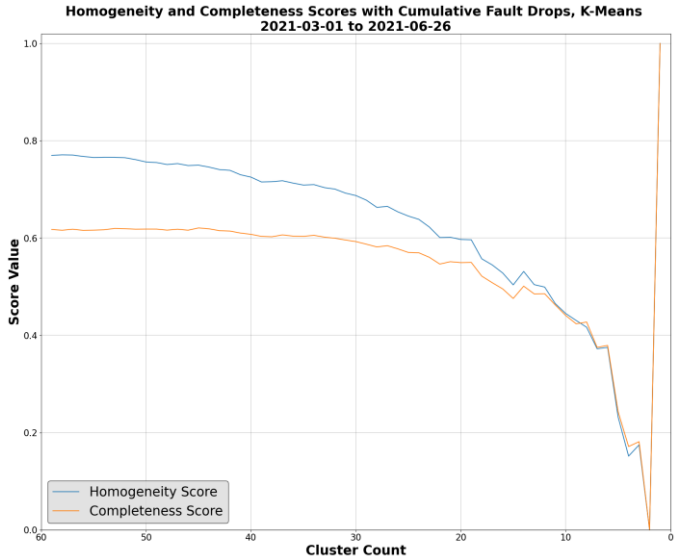
K-Means Clustering of Outage Interval



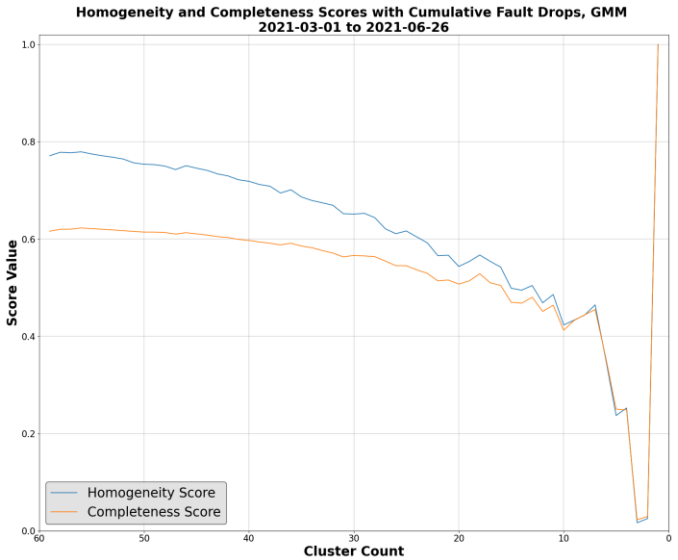
G.M.M. Clustering of Outage Interval

Results: Metrics

- Investigating the possibility that dropping the rarest fault types and re-clustering could improve performance.
- Left and right Figs. were produced by iteratively dropping the rarest fault types and re-clustering.



K-Means Performance



G.M.M. Performance

Future Work

- Truth labels need to be matched to their predicted labels.
- Future outage faults could then be given a predicted label and reassessed.
- UMAP: Sensitivity to hyperparameters to explore: random seed, n_neighbors, min_dist, n_components, and metrics.

Acknowledgements

Thank you to my supervisors for their help in my research this summer, and L-CAPE for letting me be part of their team and contributing to machine learning at Fermilab.

And thank you to the CCI committee for their role in making this opportunity possible.

Thank you