Allocate technician to a task base on:

* **Individual**:

1. **Competency** – did the technician have the skill set/knowledge for the task? Did the technician go through the training?
2. **Experience** – did the technician perform the task before? how many times? How long is the service?
3. **Availability** – is the technician free to perform the task on the specific date?
4. **Efficiency** -
5. **Workload** – is the technician have concurrent task? Is it overload for the technician?
6. **Reliability** – did the technician have any incident/unsafe act? How good is the technician?(from peer to peer)
7. **Condition** – is the technician fit to perform the task? Any past injuries that potentially increase the risk when performing the task?
8. **Problem solving ability** – is the technician able to solve the problem when a challenge arise during the task?
9. **Communication skill** – is the technician able to communicate with the team when the task required more than one person?

* **Workshop**:

1. **Safety** – did the technician familiar and adhere the safety regulation?(might be capture in techRAC)
2. **Tools and equipment** – is the technician familiar with the tools and equipment required for the task?

\_\_\_\_\_ = within our control (we have the data?)

Which depot have problem

Create a risk assessment scoring system:

Likelihood features: (things organisation can’t control)

1. Sleep – did the technician have sufficient rest?
2. Experiences – how many times the technician had perform the same task?
3. Temperature – weather
4. Humidity – depends on the weather?
5. Year of technician – how long the technician in this field
6. Age – how old is the technician?
7. Reliability – peer evaluation result

Severity features:

1. Training – did the technician went through training?
2. Safety brief – is the technician aware of the safety protocol?
3. PPE – did the technician wear standard PPE?
4. Equipment condition – is the equipment quality good to go?
5. Task difficulty level – how difficult is the task?

Scoring formula:

Risk level = likelihood \* severity

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sleep(<7) | Reliability (>=2) | Year of Tech | Experience | Temperature | Age | Difficulty Level |
| 1 | 100 | 100 |  |  |  |  | 100 |
| 2 |  | 75 |  |  |  |  | 75 |
| 3 |  | 50 |  |  |  |  | 50 |
| 4 |  | 25 |  |  |  |  | 25 |

If (sleep >=7 || experience != 0 || temperature !> 31 )

If (difficulty level == 1)

If (

Else

=> 100%

**Unsupervised Machine Learning vs Supervised Machine Learning:**

1. Supervised machine learning: (prediction)
   1. If we have the risk assessment (determine risk score)
      1. Demographics features: age, experience, certification level
      2. Operations feature: hours worked, type of task performed
      3. Performance feature: number of incidents, peer’s score
      4. Environment feature: work conditions, tools and resources
2. Unsupervised machine learning: (find the hidden pattern in the data)
   1. Do not know the risk outcome
   2. Explore the data and identify group might share common risk characteristic
   3. Focus on anomaly detection
      1. To identify outlier technician in their own group. Based on the features below:

* Demographics features: age, experience, certification level
* Operations feature: hours worked, type of task performed
* Performance feature: number of incidents, peer’s score
* Environment feature: work conditions, tools, and resources

**Unsupervised Machine Learning:**

Types:

1. Clustering
   1. Grouping data points into clusters such that points in the same cluster are more similar to each other than to those in other clusters
   2. Types:
      1. K-means: Partitions data into K clusters, where each data point belongs to the cluster with the nearest mean.
      2. Hierarchical: Builds a hierarchy of clusters by either merging smaller clusters into larger ones (agglomerative) or splitting larger clusters into smaller ones (divisive).
      3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Groups data points based on density, identifying clusters of varying shapes and sizes and marking outliers.
2. Dimensionality Reduction
   1. Reduce the number of features while retaining the essential information in the data.
   2. Types:
      1. Principal Component Analysis (PCA): Transforms data into a set of linearly uncorrelated components, reducing dimensionality while preserving as much variance as possible.
      2. t-Distributed Stochastic Neighbor Embedding (t-SNE): Reduces dimensions for data visualization, maintaining the local structure of the data.
      3. Linear Discriminant Analysis (LDA): Projects data onto a lower-dimensional space, maximizing the separability of classes (although LDA is typically used in a supervised context, it can be adapted for unsupervised use).
3. Anomaly Detection
   1. Identifies data points that deviate significantly from the norm, which could indicate outliers
   2. Types:
      1. Isolation Forest: using random splits to isolate points quickly.
      2. Local Outlier Factor: measuring the density deviation of a data point compared to its neighbours.
      3. Autoencoders: identify anomalies based on reconstruction errors.
4. Association Rule Learning
   1. Discovers interesting relationships between variables in large datasets.
   2. Types:
      1. Apriori: Identifies and generates rules based on support and confidence metrics.
      2. Eclat: Uses a depth-first search strategy to find frequent item sets, suitable for datasets with many distinct items.
5. Self-organising Map
   1. Use a grid of neurons to represent the data, with similar data points mapping to nearby neurons. Useful for visualization and clustering.
6. Hidden Markov Model
   1. Assume that the system being modeled is a Markov process with unobserved (hidden) states. Used in time series analysis, speech recognition, and bioinformatics.
7. Gaussian Mixture Model
   1. Fit multiple Gaussian distributions to the data, which can model data points belonging to different subpopulations within an overall population.
8. Non-negative Matrix Factorisation
   1. Useful for uncovering latent features in the data, commonly used in text mining and recommendation systems.

**Why unsupervised machine learning?**

Nature of unsupervised learning allow us to explore without prior knowledge of risk categories. This is very useful to identifying the hidden patterns and grouping within the data that may not be immediately obvious.

Next is unsupervised learning able to adapt to the new data and changing pattern, which is very applicable in our dynamic workshop environment where risks and technician performance may vary over time.

**What did I do?**

Grouping: clustering help in identifying group of technicians with similar risk profiles. These groups can be tailor specific safety measures and task allocations, ensuring that technicians are assigned tasks that match their risk levels. (Idea)

Anomaly detection: by reverse engineering, identifying the outliers, who is ready to perform the tacks. While clustering can highlight to superior that which technician required addition monitoring due to their risk profiles.

**Benefits of using clustering model:**

1. Easy implementation – quick deployment and iterative improvement
2. Interpretation – straightforward to interpret. Current model show the superior which technician is ready for the task.
3. Handling large dataset – handling the dataset efficiently which is crucial when it come to individual technician’s performance, types of tasks and environment factors can be extensive.
4. Flexibility – can be scale to accommodation additional data points or features as more information are available. Clustering ensure the model remain relevant and accurate.

**Comparing to other machine learning:**

1. Supervised Learning Models:
   1. Required labelled training data which may not be readily available in the workshop context.
   2. Overfitting due to Insufficient data
   3. Bias prediction because of the skewed data, which make it hard to interpret the result.
2. Reinforcement Learning:
   1. Involves more complex algorithms and longer training times, making it less practical for quick deployment.
   2. More suited for environments where agents learn through continuous interaction with the environment, which may not align with the project’s goals.
3. Dimensionality Reduction Techniques:
   1. Techniques like PCA (Principal Component Analysis) are more focused on reducing the dimensionality of the data rather than grouping similar data points based on risk.

**Conclusion**

Clustering was chosen for its ability to effectively group technicians based on their risk profiles without the need for labelled data. It is robust against the issues cause by other machine learning. In addition, its simplicity, interpretability, scalability, and adaptability make it an ideal choice for assessing individual risk in a workshop setting. By utilizing clustering, we can ensure that supervisors have actionable insights to allocate tasks efficiently and maintain a safe working environment.

What is unsupervised machine learning algorithm?

* This algorithms discover hidden patterns or data groupings without the needs for human intervention.

What are the types of unsupervised machine learning algorithm are there?

1. Clustering
   1. Data mining technique which groups unlabelled data based on their similarities or differences
   2. Types of clustering:
      1. K-Means
      2. Hierarchical Clustering
      3. Probabilistic
2. Association
3. Dimensionality Reduction