**Anamoly Detection**

**Pw Skills**

Q1. \*\*Anomaly Detection\*\*: Anomaly detection, also known as outlier detection, is the process of identifying patterns in data that do not conform to expected behavior. Its purpose is to detect unusual or abnormal instances in datasets, which may indicate potential problems, errors, or interesting observations.

Q2. \*\*Key Challenges\*\*: Key challenges in anomaly detection include defining what constitutes normal behavior, dealing with imbalanced datasets where anomalies are rare, selecting appropriate features and algorithms, and interpreting detected anomalies accurately.

Q3. \*\*Unsupervised vs. Supervised Anomaly Detection\*\*:

- Unsupervised anomaly detection detects anomalies without prior knowledge of normal data, relying solely on the characteristics of the dataset.

- Supervised anomaly detection requires labeled data, with anomalies already identified, to train a model to distinguish between normal and abnormal instances.

Q4. \*\*Main Categories of Algorithms\*\*: Main categories of anomaly detection algorithms include statistical methods, machine learning-based approaches, and hybrid techniques. Statistical methods include z-score, Grubbs' test, and Dixon's Q-test. Machine learning-based approaches include isolation forest, local outlier factor (LOF), and k-nearest neighbors (KNN).

Q5. \*\*Assumptions of Distance-based Methods\*\*: Distance-based anomaly detection methods assume that normal instances are clustered together and anomalies are far from the normal cluster. These methods calculate the distance of each instance from its neighbors or centroids and identify instances with unusually large distances as anomalies.

Q6. \*\*LOF Algorithm\*\*: The Local Outlier Factor (LOF) algorithm computes anomaly scores based on the local density deviation of a data point compared to its neighbors. It measures how much more or less dense a point is compared to its neighbors, with points significantly less dense considered potential anomalies.

Q7. \*\*Parameters of Isolation Forest\*\*: The key parameters of the Isolation Forest algorithm include the number of trees in the forest (n\_estimators) and the maximum depth of each tree (max\_depth). Additionally, other parameters like the subsampling size and random seed may also be relevant.

Q8. \*\*Anomaly Score using KNN\*\*: With K=10, the anomaly score of a data point with only 2 neighbors of the same class within a radius of 0.5 would likely be high, indicating it is an outlier because it has very few similar neighbors within its vicinity.

Q9. \*\*Isolation Forest Anomaly Score\*\*: With 100 trees and an average path length of 5.0 for a data point compared to the average path length of the trees, the anomaly score would depend on how much shorter or longer the path length is compared to the average. Anomalies typically have shorter average path lengths in an isolation forest.