Vishnu Kumar D S
Data Analyst

Task 1 Summary Report:

In this Task 1, I was assigned the responsibility of tagging a manufacturing dataset with structured categorical fields including *Root Cause*, *Symptom Condition*, *Symptom Component*, *Fix Condition*, and *Fix Component*, based on free-text entries and a provided taxonomy reference. Although I come from a data science background and not a mechanical domain, I approached the task with curiosity, critical thinking, and a structured analytical mindset.

I began by importing and exploring the dataset using the **pandas** library. Initial data exploration involved understanding column relationships, identifying null values, and cleaning irrelevant or redundant fields. This stage was crucial in forming a conceptual understanding of how the textual fields such as Complaint, Cause, and Correction aligned with the categorical taxonomy. I carefully documented these observations to guide the tagging logic in subsequent steps.

Recognizing that a large portion of the dataset included free-text entries with inconsistent formats, I transformed missing values into clean, preprocessed lists. I used this list to systematically compare entries with the taxonomy using the **fuzzywuzzy** library. This library was chosen for its simplicity and effectiveness in fuzzy string matching—essential when working with categorical but noisy data. I implemented a threshold-based function (≥ 50% similarity) to ensure only relevant matches were considered, as aligned with the task's minimum accuracy requirement.

This function was applied across all key fields, and I cross-validated the output by comparing automated tags with human-readable context. Throughout the process, I maintained a strong focus on logic, pattern recognition, and iterative testing. When ambiguity arose, I made careful notes and flagged observations for potential manual review.

Despite the challenges of domain-specific terminology, I embraced the learning opportunity and continuously refined my methods based on feedback from the data. The result was a clean, tagged dataset with significantly improved structure and coverage. The experience reinforced my ability to adapt, reason through unstructured problems, and bridge technical tools with practical outcomes.