Explanation of the Deduplication Process

First, I converted the original .parquet file into a .csv file to make it easier to work with and to be able to open it in Excel, since I am more familiar with that format.

After manually analyzing the features, I concluded that the most important ones are “unspsc”, “product_name”, “intended_industries”, “applicability”, and “brand” because many products showed similarities when searching through the dataset.

Then, I searched online for a solution to group products based on those features and found the concept of a "signature", which is used to group rows that appear to represent the same product. So, I created a new feature called “signature_components” in the form of a dictionary with the selected features.

Initially, I created a script to deduplicate the dataset, but the result was not satisfactory (only about 2,000 rows were removed). I deduced that this was due to many NULL values, so grouping didn’t happen even when the other features matched. I researched possible solutions to this issue and found two:

1) Fuzzy Matching: where I could manually introduce a threshold to determine similarity. However, this didn’t help with cases where one feature was NULL and the other wasn’t.
2) Soft Matching: which does not require all features to be identical, allowing for a margin of error. Here, I decided that at least 4 out of the 5 manually selected features must match. I chose this method despite the increased risk of false positives.

Then I iterated through the entire dataset, row by row, and grouped rows where Soft Matching detected at least 4 out of 5 matching features.

Next, for each group, I concatenated the remaining features as follows:

- If the column was a string or list, I concatenated the values.
- For numeric columns, I kept the first valid value.

Finally, I created a new deduplicated .csv file based on the concatenated groups. This final version contains approximately 6,000 fewer rows than the original.