Day 8 – 2 July 2024

Training Day 8 Report

Date: 2 July 2024
Topic: Data Cleaning and Manipulation in Pandas

Summary:
On Day 8, we focused on **data cleaning and manipulation**, which is a crucial step before building any Machine Learning model. The trainer explained that real-world datasets often contain **missing values, duplicate entries, incorrect data types, and inconsistent formats**, and it is essential to clean the data for accurate analysis and modeling.

We started by learning to **handle missing values** using `dropna()` to remove rows with missing data and `fillna()` to replace missing values with specific numbers or statistics like mean or median. Next, we learned to **remove duplicate entries** using `drop_duplicates()` to ensure the dataset is clean and unique.

We also explored **renaming columns** with `rename()` for better readability and changing **data types** using `astype()` to ensure the data is suitable for analysis. The trainer demonstrated **applying functions to columns** using `apply()` for transformations such as converting scores to grades or normalizing values.

Finally, we used the `groupby()` function to **aggregate data** by categories, such as calculating average scores for each class or department. We practiced combining all these operations on sample datasets to make them ready for visualization and Machine Learning tasks.

Key Learnings:

- Handled missing values using `dropna()` and `fillna()`.
- Removed duplicate entries for data consistency.
- Renamed columns and changed data types.
- Applied functions on columns for transformations.
- Used `groupby()` to aggregate data effectively.

Conclusion:
Day 8 emphasized the importance of data cleaning and manipulation. I now feel confident in preparing datasets, which is a critical step for accurate Machine Learning modeling