**AI ML Internship Log**

# Day 10 -Advanced Preprocessing, EDA for Imbalance & Final Model Evaluation

# Date - 28 June 2025

# Team Role - Member

# Project Title - Personality Prediction from Social Media

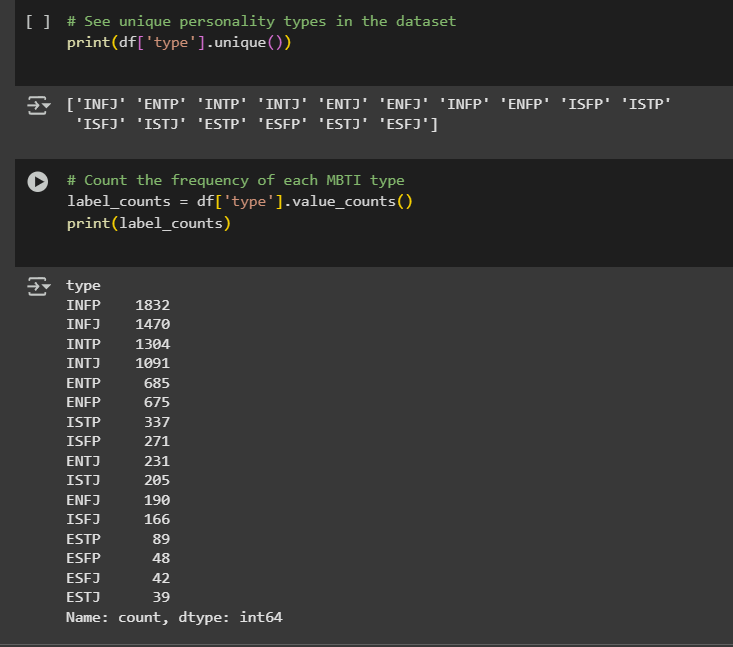
✅ **What I Did Today:**

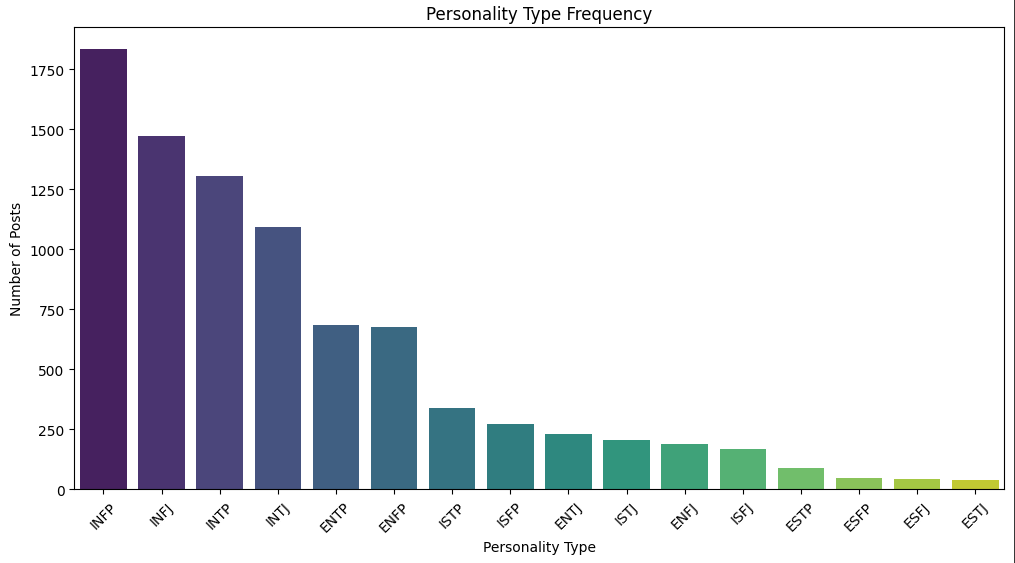
* ✅ Implemented **advanced preprocessing**:
  + Lowercasing
  + Removing URLs
  + Removing punctuation
  + Lemmatization
* ✅ Updated the dataset with cleaned\_posts.
* ✅ Reran all three models — Logistic Regression, Linear SVM, and SVM with RBF Kernel — on the cleaned data.
* ✅ Compared model performance **before and after preprocessing**.
* ✅ Performed **Exploratory Data Analysis (EDA)** on the target labels (type) to detect **class imbalance**.

📊 **EDA Insights on Imbalanced Dataset:**

* Total 16 MBTI personality types.
* **INFP**: 1832 samples (most frequent)
* **ESTJ**: 39 samples (least frequent)
* Huge imbalance — some classes have 40x more data than others.
* This imbalance was **negatively impacting model fairness**, especially macro F1-score.

**Code Screenshots:**

****

****

**📈 Model Performance Comparison:**

| **Model** | **Accuracy** | **Macro F1** | **Weighted F1** |
| --- | --- | --- | --- |
| Logistic Regression | 0.6414 → 0.6438 | 0.43 → 0.44 | 0.62 → 0.63 |
| Linear SVM | 0.6155 → 0.6126 | 0.47 → 0.47 | 0.61 → 0.61 |
| SVM RBF | 0.6403 → 0.6340 | 0.46 → 0.47 | 0.63 → 0.62 |

🔍 **Key Takeaways:**

* Accuracy remained nearly the same.
* Macro F1 improved slightly — a sign that rare classes are better predicted.
* Accuracy drop (~0.006) is an acceptable trade-off for improved fairness.
* Advanced preprocessing helped remove noise, but imbalance remains the bottleneck**.**

**🧠 Self Reflection:**

* I had doubts early in the day, wondering if this was leading anywhere.
* But the results taught me a **valuable lesson** — even small improvements in macro F1 **mean big impact** in real-world problems.
* I now understand:
  + Why **accuracy isn't everything**.
  + The **value of fairness** in imbalanced data.
  + How to **improve a model’s performance** with both preprocessing and strategy.

**📅 Status:** A technically rich day, tough mentally but filled with **real ML insight**. I grew as an ML learner.