#YAP470-Project

MBTI-PERSONALITY-PREDICTION

The repository for our Machine Learning Project.

Besides installing the necessary libraries from our requirements.txt, there is not anything else to do to run our code.
-Deployment file can be run with the command: streamlit run App_MBTI_Prediction.py

Important Files:

The repository contains two python executables:

-Yap470_Project_Myers-Briggs.ipynb (Main Project)

-App_MBTI_Prediction.py (Deployment Project)

An "mlruns" folder containing previous mlflow loggings,

Two csv files:

mbti_1.csv (Original Dataset)

pre_runs.csv (Logging table of older mlflow-runs that had to be deleted from the "mlruns" folder due to a memory exception.)

The serialized XGBoost pipeline file. ("XGboost_pipeline.pkl")

-The code consists of roughly 5 parts.

1-)Data Analysis Stage

-A SEED value was implanted to get same values for every user.

-Dataset checked for null values

-Labels were plotted to spot class imbalance

2-)Preprocessing Stage

-First, the noisy text corpus was cleansed of all non-word elements.

-Then, words containing label names were removed to prevent potential leakage.

-Stop-Words(Meaningless words like the, i, a etc.) were removed.

-Remaining words were lemmatized and transferred into a new column.

-As the last step, this new column is given to a TFIDF vectorizer and each word becomes a feature.

3-)Model Training Stage

-This stage consists of 6 model-train functions, one for each model.

-Every function takes all the necessary parameters to be trained and tuned under any desired condition.

-The data is split separately inside the functions.

-SelectKBest algorithm is used for dimension reduction AFTER data-splitting.

-Mlflow logging is called upon all the parameters, results and models.

4-)Hyper-Parameter Tuning Stage

-Again, this stage consists of 6 tuning functions, one for each model.

-These functions just train the models with different parameters and print the results.

5-)Miscellaneous

-There is an attempt on a Grid_Search algorithm after the previous stage but due to some unresolved issues, the algorithm had to be ran with a single core and unsurprisingly it drove the computer into crashing.

-Pipeline for XGBoost is created.

-MLflow ui can be accessed from the last cell.