Personality Cluster Prediction

Team Name: Unsupervised Learners Team Members:

▪ R.Sreenivasa Raju (IMT2023122)

▪ U.Trivedh Venkata Sai (IMT2023002)

GITHUB Link: https://github.com/maxinh00000/Multi-Class-Personality

1. Introduction

This project focuses on predicting an individual's personality cluster using various behavioral, lifestyle, and activity-based features. Personality clusters represent groups of individuals exhibiting similar patterns in daily habits, social engagement, and emotional expression.

The task is framed as a multiclass classification problem and evaluated using Macro F1 Score to ensure balanced performance across all classes, especially minority clusters.

The dataset consists of: train.csv, test.csv, submission files

2. Dataset Overview

The dataset contains multiple columns describing:

- Lifestyle attributes, Emotional expression, Emotional expression, Social engagement level, Activity patterns, Stability and environmental support

The initial steps included:

- Reading dataset, checking null values and identifying rows with null values

This helped identify:

- No major missing values
- All features are numerical
- Balanced but slightly skewed class distribution

3.Exploratory Data Analysis (EDA)

EDA involved:

- Understanding feature correlations
- Checking variance
- Identifying dominant behavior patterns

Insights:

- Lifestyle and emotional expression features correlate moderately
- Some features have wide variance, indicating diverse behavior patterns
- The target class has mild imbalance → supports use of Macro F1

4.Feature Engineering

Several new engineered features were created to capture deeper behavioral interactions.

4.1 Lifestyle Balance

df["lifestyle_balance"] = (df["activity_level"] +df["rest_quality"] +df["expression_index"]) / 3

This represents an individual's balance between active lifestyle, rest quality, and expressive behavior.

4.2 Stability Score

df["stability_score"] = (df["support_environment_score"] +df["consistency_score"]) / 2

Measures how stable and supported the environment around the person is.

4.3 Engagement Composite

df["engagement_composite"] = (df["social_activity_level"] +df["communication_score"] +df["hobby_engagement_level"]) / 3

Captures total social & hobby engagement.

4.4 Emotional Strength Score

df["emotional_strength"] = (df["expression_index"] +df["emotional_stability_score"]) / 2

These features improved separability across clusters.

5. Train-Test Split and Preprocessing

Data was split using stratified splitting:

Scaling was applied where needed using StandardScaler or MinMaxScaler (depending on the model).

6. Model Selection and Training

Multiple models were trained:

6.1 Logistic Regression

- Provided baseline performance
- Used after scaling
- Linear boundaries → limited performance

6.2 Random Forest

- Handles non-linear patterns well
- Good first strong model

6.3 XGBoost

- One of the best individual models
- Handles class imbalance with built-in loss functions
- High accuracy and F1

6.4 Support Vector Machine (SVM)

Two variants were tested:

SVM (Linear Kernel)

- Performs well on linearly separable data
- Validation score was similar to Logistic Regression
- Failed to model complex personality boundaries

SVM (RBF Kernel)

- Much better because RBF captures non-linear relationships
- However:
  - Very computationally expensive
  - Slow training due to many samples and features

6.5 Multi-Layer Perceptron (MLP Classifier)

A simple neural network (fully-connected feedforward network) was trained using scikit-learn's MLPClassifier.

- Required feature scaling
- Hidden layer structure like (64, 32) worked fairly well
- Captured some non-linear patterns
- Training took longer
- Tuning learning rate & hidden layers was crucial

7. Final Model — Ensemble of SVM + MLP

Instead of using boosting models, the final model is an ensemble combining SVM (RBF Kernel) and MLP, which blend classification probabilities:

Soft Voting Ensemble

Final prediction:

final_pred = argmax(0.5 * SVM_prob + 0.5 * MLP_prob)

(or weighted based on performance)

Why this works:

- SVM provides excellent margin-based separation
- MLP handles complex non-linear patterns
- Ensemble reduces overfitting
- Ensemble stabilizes predictions across all personality classes
- Improves Macro F1 by leveraging model diversity

8. Conclusion

Through extensive experimentation, the combination of:

- Support Vector Machine (RBF Kernel)
- Multi-Layer Perceptron Neural Network

yielded the best balance of accuracy, generalization, and class-wise fairness (Macro F1 Score).

The SVM + MLP ensemble proved to be a strong model for personality cluster prediction because it merges two very different learning philosophies:

- Max-margin classification (SVM)
- Deep non-linear representation learning (MLP)

This hybrid approach enabled robust predictions across all personality groups.