ML Project Report:
Personality Cluster Prediction

Nipun Verma (IMT2023591)
Sahil Singh (IMT2023521)
Mohammed Ibrahim (IMT2023112)

GitHub Repository

Background and Problem Statement

Human personality affects decision-making, social behaviour and long-term psychological patterns. Understanding personality clusters is useful in psychology, recommender systems, learning platforms and user modelling. In this project, the objective is to predict an individual's **personality cluster** (Cluster A–E) based on demographic, behavioural, environmental and activity-related information.

The dataset originates from the Kaggle competition *"Multidimensional Personality Cluster Prediction"*. Unlike the previous founder-retention project, here the target is **multiclass** (5 classes) and the features represent aspects of identity, upbringing, activity patterns and guidance-seeking tendencies.

Three model families were evaluated:

1. **Logistic Regression** (multinomial) as an interpretable baseline.
2. **Support Vector Machine (RBF kernel)** as a classical nonlinear classifier.
3. **Neural Network (Tabular MLP)** as a flexible deep learner.

The goal is both predictive performance and an understanding of which factors correlate most strongly with personality types. As in the previous project, the report focuses on careful preprocessing, feature representation and interpretation of model behaviour.

Dataset Description

The training file contains 1,913 rows and 23 columns after cleaning. The target variable is:

$$\texttt{personality\_cluster} \in \{\texttt{Cluster\_A}, \texttt{Cluster\_B}, \texttt{Cluster\_C}, \texttt{Cluster\_D}, \texttt{Cluster\_E}\}.$$

The class distribution is:

$$Cluster\_A : 85$$
$$Cluster\_B : 220$$
$$Cluster\_C : 306$$
$$Cluster\_D : 328$$
$$Cluster\_E : 974$$

This imbalance is substantial: Cluster E constitutes more than half the dataset. Therefore, macro F1 is used throughout.

Features fall into three categories.

Nominal features

- `identity_code`
- `cultural_background`
- `hobby_engagement_level`
- `creative_expression_index`

Ordinal features Values represent ordered ratings or scores:

- `upbringing_influence` (0–4)
- `external_guidance_usage` (0–3)
- `support_environment_score` (0–4)
- `physical_activity_index` (0–4)

Numeric features

- `age_group`
- `focus_intensity`
- `consistency_score`

These are inherently numeric and require scaling but no encoding.

Data quality is generally clean. Missing values exist but are easily imputed.

EDA and Preprocessing

Handling Missing Values

The preprocessing strategy mirrors the founder-retention report methodology:

- **Ordinal + numeric** columns: median imputation.
- **Nominal** columns: mode imputation.

Median is used for numeric-like columns because `focus_intensity` shows right-skewed distributions; mean imputation would shift values in those tails.

A global consistency check confirmed that after imputation, both train and test contained no missing values.

Encoding Strategy

Given the small number of categorical features, one-hot encoding was used only for nominal features. Ordinal features were left as integers to preserve ordering.

This results in a compact representation: the final encoded feature matrix contains only 17 dimensions. This is advantageous for classical models such as SVM and Logistic Regression.

Scaling

All encoded features were standardised:

$$x' = \frac{x - \mu}{\sigma}.$$

Scaling is necessary because numeric features such as `focus_intensity` span large ranges, while one-hot encoded columns are binary. SVMs in particular are sensitive to unscaled inputs.

Sanity Checks

EDA included:

- Target distribution plots.
- Boxplots of ordinal and numeric features.
- Countplots of nominal variables.
- Correlation heatmaps among numeric/ordinal features.

Two dominant observations emerged:

1. `focus_intensity` and `consistency_score` vary significantly and appear correlated with cluster identity.
2. Nominal features such as `identity_code` and `cultural_background` show mild but non-negligible associations with clusters.

Feature Importance and Mutual Information

To quantify how strongly each feature relates to the personality cluster, two analyses were performed:

Cramér's V for Nominal Features

Cramér's V captures association strength between nominal variables and the categorical target. The strongest nominal associations were:

- `hobby_engagement_level`
- `creative_expression_index`
- `identity_code`

Though values were moderate (≈ 0.05–0.08), these features still provide useful separation among clusters.

Pearson Correlation for Numeric/Ordinal Features

The largest correlations were:

$$\texttt{consistency\_score} : 0.726, \qquad \texttt{focus\_intensity} : -0.138, \qquad \texttt{support\_environment\_score} : -0.126.$$

`consistency_score` exhibits a remarkably strong linear correlation with personality cluster (after encoding clusters numerically). Such a high absolute correlation suggests that consistency is one of the central traits distinguishing personality groupings.

Mutual Information

Mutual information ranked features as follows:

$$\texttt{focus\_intensity} > \texttt{consistency\_score} \gg \text{others}.$$

This again confirms that behavioural intensity and internal consistency are the primary drivers.

Modelling Strategy

The dataset was split:

$$\text{Train} : 80\% \text{ (1530 rows)}, \qquad \text{Validation} : 20\% \text{ (383 rows)},$$

with stratification to preserve class proportions.

Three model families were trained and compared, exactly as in the founder-retention project.

Macro F1 was the evaluation metric.

Model 1: Logistic Regression

Motivation

Multinomial Logistic Regression provides a clean baseline: simple, interpretable and reflective of linear separability in the encoded feature space.

Hyperparameters

The model used L2 regularisation, and the best parameters via 3-fold GridSearch were:

$$C = 0.01, \quad \text{solver} = \texttt{saga}.$$

Validation Performance

$$\text{Macro F1} = 0.478.$$

The confusion matrix shows Logistic Regression especially struggles with minority classes (Clusters A, B, C). Cluster E is predicted well due to its dominance.

The linear boundary is insufficient to carve out the nuanced separation required for rare clusters.

Model 2: Support Vector Machine (RBF Kernel)

Motivation

SVMs with RBF kernels often excel on small- to medium-sized tabular datasets. They can capture nonlinear boundaries without needing deep architectures.

Hyperparameter Search

GridSearch over (C, γ) yielded:

$$C = 3, \quad \gamma = 0.01, \quad \text{kernel} = \texttt{rbf}.$$

Validation Performance

$$\text{Macro F1} = 0.6065.$$

This is a noticeable improvement over Logistic Regression. SVM particularly increases recall for mid-frequency classes (Clusters B, C, D), though Cluster A remains difficult due to extreme scarcity.

Model 3: Neural Network (Tabular MLP)

Architecture

A compact MLP was used:

- Input dimension: 17
- Hidden layers: 128, 64
- Activation: ReLU
- Dropout: 0.2
- Optimiser: Adam

This is essentially the same architecture used for founder retention but adjusted for 5-class output.

Training Behaviour

The NN steadily improved over epochs and achieved:

$$\text{Best Validation Macro F1} = \mathbf{0.636}.$$

This is the best performance among all models. The NN learned the strong nonlinear relationship between consistency/focus and cluster identity more effectively than SVM.

Confusion Matrix

The NN predicts Cluster E with high precision and recall, and improves recall for Clusters B–D relative to SVM. Cluster A remains the hardest class due to only 17 validation samples.

Model Comparison

Model	Macro F1	Notes
Logistic Regression	0.478	Linear baseline; weak on rare classes
SVM (RBF)	0.6065	Strong nonlinear model; costly to train
Neural Network	**0.636**	Best performance overall

Table 1: Comparison of model performance on the validation set.

The NN achieves the highest macro F1 despite its simplicity. Its ability to model smooth nonlinearities and interactions gives it an advantage on this task.

Design Choices and Alternatives

Encoding

Ordinal features were intentionally *not* one-hot encoded. Preserving order is crucial for personality-relevant scores (guidance usage, upbringing influence, etc.).

Scaling

StandardScaler was essential for both SVM and NN stability.

Macro F1 vs Overall Accuracy

Accuracy is high (≈ 0.75) for all models, but heavily biased by Cluster E. Macro F1 provides a more truthful assessment of overall performance.

Additional Model: CatBoost Classifier

Although the primary comparison in this report focuses on Logistic Regression, SVM and a Neural Network, an additional experiment was conducted using **CatBoost**, a modern gradient-boosted decision tree algorithm particularly well-suited for tabular datasets. CatBoost handles categorical variables natively through ordered target statistics and generally requires minimal preprocessing.

Training Setup

A 5-fold stratified cross-validation procedure was used. The model was trained with the following key settings:

- Loss function: `MultiClass`
- Iterations: 4000
- Learning rate: default (adaptive)
- Depth: 6

- Early stopping: enabled based on validation score

Across folds, CatBoost consistently achieved perfect training accuracy after a few hundred iterations, indicating strong fitting capacity. Early stopping prevented excessive overfitting by selecting the iteration with the highest validation macro F1 score.

Validation Results

The per-fold best macro F1 scores were:

Fold 1: 0.6004

Fold 2: 0.5826

Fold 3: 0.5928

Fold 4: 0.5977

Fold 5: 0.5669

CatBoost achieves an overall **out-of-fold macro F1 of 0.5886**, which places it between SVM (0.6065) and Logistic Regression (0.478), but below the Neural Network (0.636). Interestingly, CatBoost often reached its best validation performance early (typically between 150–300 iterations), after which overfitting caused gradual degradation in macro F1. This behaviour is consistent with the model's ability to tightly fit minority classes but struggle to generalise across all five personality clusters.

Conclusion

This project demonstrates that personality clusters can be predicted reasonably well from demographic and behavioural features, provided that preprocessing and encoding are handled with care.

Key findings:

- **Consistency score** and **focus intensity** are dominant predictors.
- Logistic Regression is insufficient due to linear separability limitations.
- SVM offers meaningful improvements but at the cost of computation.
- A simple Neural Network achieves the best overall performance with Macro F1 = 0.636.