Model Development Report

Multi-Class Classification of 16 MBTI Personality Types

1. Description

Four machine learning models were trained to classify 16 MBTI personality types using 60 survey questions. The dataset contains ~66,000 samples with personality trait responses as features.

2. Models Trained & Rationale

Model	Rationale
XGBoost	Gradient boosting ensemble; excellent for tabular data with complex patterns
Random Forest	Bagging ensemble; robust to noise, provides feature importance
Logistic Regression	Linear baseline; interpretable coefficients, multinomial softmax
LDA	Dimensionality reduction classifier; finds class-separating linear combinations

3. Hyperparameter Choices

Model	Key Parameters
XGBoost	n_estimators=500, learning_rate=0.1, max_depth=6, subsample=0.8, early_stopping=15
Random Forest	n_estimators=100, max_depth=20, min_samples_split=5, max_features=sqrt
Logistic Regression	multi_class=multinomial, solver=lbfgs, max_iter=1000, C=1.0
LDA	solver=svd, n_components=15 (auto), tol=1e-4

4. Training Iterations

Model	Iterations/Trees	Notes
XGBoost	~100-150 (early stopped)	Stopped early from max 500 based on validation loss
Random Forest	100 trees	Full ensemble trained in parallel
Logistic Regression	~200-400 iterations	Converged before max_iter (1000)
LDA	Single pass	Analytical solution via SVD, no iterations needed

5. Validation Strategy

Stratified Hold-Out Split (70/15/15): Training (70%, ~46,200 samples), Validation (15%, ~9,900 samples), Test (15%, ~9,900 samples). Stratification maintained class proportions. All models used identical splits (random_state=42) for fair comparison. Validation set used for early stopping and model selection.

6. Training Results

Model	Test Accuracy	Macro F1	Train Accuracy	Overfit Gap
XGBoost	98.22%	0.9822	99.95%	1.73%
Random Forest	97.57%	0.9757	99.99%	2.42%
Logistic Regression	91.90%	0.9189	92.04%	0.14%
LDA	90.56%	0.9055	90.72%	0.16%

Best Model: XGBoost achieved highest test accuracy (98.22%) with minimal overfitting.

7. Training Time & Computational Resources

Model	Training Time	Hardware Utilization
XGBoost	~2-3 minutes	Multi-core CPU (n_jobs=-1)
Random Forest	~1-2 minutes	Multi-core CPU (n_jobs=-1)
Logistic Regression	~30-60 seconds	Multi-core CPU (n_jobs=-1)
LDA	~5-10 seconds	Single-core (analytical solution)

Environment: Python 3.x with scikit-learn, XGBoost on standard desktop hardware. Total training time: ~5 minutes for all four models.

8. Key Findings

• **Ensemble methods outperform linear models:** XGBoost and Random Forest achieved >97% accuracy vs ~91% for linear models.
• **Gradient boosting is optimal:** XGBoost's sequential error correction captures subtle personality patterns best.
• **Linear models generalize better:** Smaller train-test gap (0.14-0.16%) but lower overall accuracy.
• **Efficient training:** All models trained in under 5 minutes combined on standard hardware.