```
rajatdogra@Rajats-Mac-mini ML_Classification % python3
football_classification_complete.py
============================================================
FOOTBALL MATCH OUTCOME CLASSIFICATION PROJECT
============================================================

1. DATA LOADING AND EXPLORATION
----------------------------------------
✓ Loading local dataset...
✓ Dataset loaded: 67353 matches, 17 columns

Dataset Overview:
- Total matches: 67,353
- Date range: 2024-01-01 05:00:00 to 2026-10-06 18:00:00
- Unique teams: 3747
- Unique leagues: 224
- Unique venues: 3279

Class Distribution:
- Home Win: 26,541 (39.4%)
- Draw: 22,980 (34.1%)
- Away Win: 17,832 (26.5%)

✓ No missing values found

2. FEATURE ENGINEERING
----------------------------------------
✓ Creating basic encodings...
✓ Creating temporal features...
✓ Creating team performance features...
✓ Creating league-specific features...
✓ Creating venue features...
✓ Creating head-to-head features...
✓ Created 26 features total
\nFinal dataset: 67,353 matches with 26 features

4. MODEL TRAINING AND TUNING
----------------------------------------

3. DATA PREPROCESSING
----------------------------------------
✓ Handling class imbalance with SMOTE...
Original distribution:
  - Home Win: 26,541
  - Draw: 22,980
  - Away Win: 17,832
Balanced distribution:
  - Home Win: 26,541
  - Away Win: 26,541
  - Draw: 26,541
✓ Data split: 63,698 training, 15,925 testing samples
```

✓ Selecting top features...
✓ Selected 25 most important features

✓ Training Algorithm 1: Random Forest
  - Tuning hyperparameters...
  - Training completed

✓ Training Algorithm 2: Gradient Boosting
  - Tuning hyperparameters...
  - Training completed

✓ Training Algorithm 3: Neural Network
  - Tuning hyperparameters...
  - Training completed

✓ Training Bonus Algorithm: XGBoost
  - Tuning hyperparameters...
  - Training completed

✓ Training Ensemble: Stacking Classifier
  - Training completed

5. MODEL EVALUATION

✓ Evaluating Random Forest...
  - Accuracy: 0.6287 (62.87%)
  - F1-Score: 0.6267
  - AUC Score: 0.8176

✓ Evaluating Gradient Boosting...
  - Accuracy: 0.6416 (64.16%)
  - F1-Score: 0.6412
  - AUC Score: 0.8265

✓ Evaluating Neural Network...
  - Accuracy: 0.5945 (59.45%)
  - F1-Score: 0.5941
  - AUC Score: 0.7758

✓ Evaluating XGBoost...
  - Accuracy: 0.6484 (64.84%)
  - F1-Score: 0.6476
  - AUC Score: 0.8326

✓ Evaluating Stacking Ensemble...
  - Accuracy: 0.6536 (65.36%)
  - F1-Score: 0.6533
  - AUC Score: 0.8347

6. CREATING VISUALIZATIONS

```
----------------------------------------
✓ Creating model performance comparison...
✓ Creating per-class performance analysis...
✓ Creating feature importance analysis...
✓ All visualizations saved to plots/ directory

7. ERROR ANALYSIS
----------------------------------------
✓ Analyzing errors for best model: Stacking Ensemble
✓ Total misclassified samples: 5516 out of 15925 (34.6%)

Error breakdown by true class:
- Home Win: 1777/5309 errors (33.5% error rate)
  → Most often misclassified as: Away Win
- Draw: 2138/5308 errors (40.3% error rate)
  → Most often misclassified as: Home Win
- Away Win: 1601/5308 errors (30.2% error rate)
  → Most often misclassified as: Home Win

Detailed confusion matrix analysis:
- Home Win → Draw: 713 cases
- Home Win → Away Win: 1064 cases
- Draw → Home Win: 1107 cases
- Draw → Away Win: 1031 cases
- Away Win → Home Win: 970 cases
- Away Win → Draw: 631 cases

8. ETHICAL CONSIDERATIONS
----------------------------------------

ETHICAL CONSIDERATIONS AND BIAS ANALYSIS
========================================

1. DATA BIAS CONSIDERATIONS:
   - Historical bias: The model learns from past match data, which may
reflect
     historical advantages or disadvantages of certain teams or leagues
   - Temporal bias: Older data may not reflect current team performance
   - Geographic bias: Some leagues or regions may be over/under-
represented

2. FAIRNESS CONCERNS:
   - Team fairness: The model should not systematically favor certain
teams
   - League fairness: Performance should be consistent across different
leagues
   - Venue fairness: Home advantage should be modeled appropriately, not
amplified

3. POTENTIAL MISUSE:
   - Gambling: This model should not be used for betting or gambling
purposes
   - Match fixing: Predictions should not influence actual match outcomes
```

```
   - Financial decisions: Should not be used for significant financial
investments

4. TRANSPARENCY:
   - Feature importance is provided to understand model decisions
   - Model limitations are clearly documented
   - Uncertainty in predictions is acknowledged

✓ Ethical considerations documented

9. SAVING RESULTS
----------------------------------------
✓ Comprehensive results saved
\n============================================================
PROJECT COMPLETION SUMMARY
============================================================
✅ Best Model: Stacking Ensemble
✅ Best Accuracy: 0.6536 (65.36%)
✅ F1-Score: 0.6533
✅ AUC Score: 0.8347
✅ Results saved to: results/
✅ Plots saved to: plots/
✅ All project requirements completed successfully!
rajatdogra@Rajats-Mac-mini ML_Classification %
```