Project Overview

This project implements a comprehensive machine learning system to predict Premier League match outcomes using statistical data from the 2022/23 season. The system combines data preprocessing, model training, and a user-friendly GUI application to provide real-time match predictions.

System Architecture

Data Pipeline Workflow

1. Data Preprocessing (preprocess.py)
Clean and transform raw Premier League CSV data

2. Model Training (train_model.py)
Train Gradient Boosting Classifier with hyperparameter optimization

3. GUI Application (gui_app.py)
Interactive prediction interface for end users

Data Preprocessing Module

File: preprocess.py

Handles data cleaning and validation for the Premier League dataset:

- **Data Cleaning:** Removes missing values and ensures numeric data types
- **Attendance Formatting:** Removes commas from attendance figures and converts to integers
- **Data Validation:** Ensures all core statistical columns have valid numeric values
- **Data Export:** Saves cleaned dataset maintaining original CSV structure

Machine Learning Model

File: train_model.py

Model Specifications

- **Algorithm:** Gradient Boosting Classifier
- **Preprocessing:** Standard Scaler normalization
- **Optimization:** Grid Search with 5-fold cross-validation
- **Target:** Binary classification (Home Win vs. Not Home Win)

GUI Application

File: gui_app.py

Interactive desktop application built with Tkinter providing:

Key Features

- **Smart Input Validation:** Auto-calculates away possession to ensure total = 100%
- **Real-time Predictions:** Instant match outcome predictions with confidence scores
- **User-friendly Interface:** Clean, intuitive design with clear result display
- **Error Handling:** Comprehensive input validation and error messaging

Input Fields

- Home team possession percentage
- Away team possession percentage (auto calculated)
- Home team shots on target
- Away team shots on target
- Expected attendance

Model Performance & Evaluation

The system employs robust evaluation methods:

- **Train/Test Split:** 80/20 split with stratification
- **Cross-validation:** 5-fold CV during hyperparameter tuning
- **Metrics:** Accuracy, confusion matrix, classification report
- **Random State:** Fixed seed (42) for reproducible results

Usage Instructions

Setup Process

1. Ensure Premier_League.csv dataset is in the project directory
2. Run preprocess.py to clean and prepare the data
3. Execute train_model.py to train and save the model
4. Launch gui_app.py to start the prediction interface

Making Predictions

1. Enter home team possession percentage
2. Input shots on target for both teams
3. Specify expected attendance
4. Click "Predict" to get match outcome probability

Project Benefits

- **Data-Driven Insights:** Leverages comprehensive 2022/23 season statistics
- **Automated Pipeline:** End-to-end machine learning workflow
- **User Accessibility:** Non-technical users can make predictions via GUI
- **Scalability:** Framework can be extended to other leagues/seasons
- **Performance Optimization:** Grid search ensures optimal model parameters

Future Enhancements

- Integration of additional features (weather, referee, form)
- Multi-class prediction (Win/Draw/Loss probabilities)
- Historical match analysis and trends
- Real-time data integration via APIs
- Web-based interface deployment

Conclusion

This Premier League prediction system demonstrates a complete machine learning workflow from data preprocessing through to user-facing application deployment. The combination of robust data handling, advanced modelling techniques, and intuitive user interface creates a valuable tool for football analysis and prediction.

The project showcases best practices in machine learning development including proper data validation, hyperparameter optimization, and user experience design, making it suitable for both educational purposes and practical application.