Tennis Match Predictor

Abstract:

This project aims to predict the outcome of professional tennis matches using machine learning techniques and historical match data. By analysing a wide range of statistics—such as player rankings, surface preferences, and in-match performance indicators like aces, break points, and win/loss records—the model learns patterns that influence match results.

The system processes and transforms raw ATP data into a format that's ready for predictive modelling. Multiple machine learning algorithms, including logistic regression, decision trees, and ensemble methods, are used and compared to find the most accurate predictor. The model is trained and tested using matches from past seasons, and its performance is evaluated based on prediction accuracy.

This project is designed for anyone interested in sports analytics, especially tennis fans and data science learners. It not only helps understand which factors most influence match outcomes but also provides a foundation for building smarter, more adaptable prediction systems in the future.

Problem Statement:

Predicting the outcome of professional tennis matches is a complex task due to the large number of variables that influence player performance, such as current form, playing surface, head-to-head history, and match conditions. Traditional ranking systems or simple heuristics often fail to capture the nuanced factors that impact match results.

This project addresses the challenge of building a data-driven system that can accurately predict the winner of a tennis match using historical ATP match data and player statistics. The goal is to analyse and engineer meaningful features from raw match data, select appropriate machine learning models, and evaluate their performance in predicting match outcomes. By doing so, we aim to create a reliable and interpretable tool for tennis analysts, fans, and researchers interested in sports forecasting.

Objectives:

1. Collect and preprocess historical ATP tennis match data.
2. Engineer relevant features that influence match outcomes.
3. Train multiple machine learning models for prediction.
4. Evaluate and compare model performance using accuracy metrics.
5. Visualize insights and model predictions for interpretability.
6. Export the final trained model for deployment or integration.
7. Provide a flexible framework for future improvements and experimentation.

Data Requirements:

1. Historical ATP match data – including match results, player names, dates, and tournament details.
2. Player statistics – such as rankings, height, handedness, and country.
3. Match-level metrics – like number of aces, double faults, break points won/saved, and serve statistics.
4. Surface type – to account for performance differences on hard, clay, and grass courts.
5. Head-to-head records – previous encounters between players, if available.
6. Player form/history – recent match outcomes to capture momentum or slumps.
7. Year-wise data – to enable training and testing across different seasons.