Forecasting Tennis Match Outcomes:

A Machine Learning Comparison with Betting Market Predictions

1. Group Members and Roles:
    a. Luke Feng - Project Manager, Director of Computation
    b. Sam English - Task Manager, Director of Research
    c. Blake Bothmer - Facilitator, Reporter
2. Purpose:
    a. This project explores whether machine learning models can outperform human-based predictions (such as expert analysts or betting markets) in forecasting professional tennis match outcomes.
    b. We aim to analyze historical ATP/WTA match data, focusing on player statistics (serve %, unforced errors, ELO ratings, head-to-head performance) and betting odds to determine:
        i. Can machine learning models trained on player statistics predict match outcomes more accurately than betting odds?
        ii. Which model types (e.g., logistic regression, random forest, XGBoost) perform best?
        iii. How do factors like surface type, recent performance, or fatigue influence accuracy?
    c. We believe the project will be because it connects sports analytics with real-world applications in data-driven forecasting and decision-making
3. Data:
    a. ATP Match Data, all in CSV format

i. https://github.com/JeffSackmann/tennis_atp

1. This contains my master ATP player file, historical rankings, results, and match stats.
2. The player file columns are player_id, first_name, last_name, hand, birth_date, country_code, height (cm).
3. The columns for the ranking files are ranking_date, ranking, player_id, ranking_points (where available). ATP rankings are mostly complete from 1985 to the present.
4. 1982 is missing, and rankings from 1973-1984 are only intermittent.
5. Results and stats: There are up to three files per season: One for tour-level main draw matches (e.g. 'atp_matches_2014.csv'), one for tour-level qualifying and challenger main-draw matches, and one for futures matches.

b. Sports Betting Datasets

i. https://the-odds-api.com/

ii. https://www.kaggle.com/datasets/hakeem/atp-and-wta-tennis-data

c. Current Player Statistics Data source

i. https://www.atptour.com/en/stats/stats-home

1. Will likely have to scrape this data

d. All data will be merged into a unified table of matches with predictors (player and match-level stats) and outcomes (win/loss).

4. Variables: As much as possible, list, and briefly describe, each variable that you plan to incorporate. If you can, be specific about units, scale, etc.
   a. Dependent variable:
      i. Match outcome — binary (1 = win, 0 = loss)
   b. Independent variables:
      i. Serve percentage (% first serves made)
      ii. Aces, double faults
      iii. Unforced errors
      iv. Break points saved/won
      v. Rank difference (numeric)
      vi. ELO rating difference
      vii. Surface type (categorical: clay, grass, hard)
      viii. Recent performance metrics (e.g., wins in last 5 matches)
      ix. Betting odds (implied probability)
5. End Product: Describe what you hope to deliver as a final product. Will it be a Shiny application that will be posted on the Internet? Will it be a GoogleMaps mash-up? Will it be a package that provides an API to a live data source? Will it be a method that draws some statistical conclusions? Will it be a predictive model that forecasts future values of some quantity?
   a. We plan to deliver a predictive model and visualization dashboard comparing model-based predictions to betting market odds.
      i. Specifically, our final product will include:

1. Several predictive models, likely a boosting or bagging algorithm (XGBoost or random forests) but we will also look into using other algorithms such as Logistic Regression
2. Performance comparison visualizations (e.g., ROC curves, calibration plots, SHAP plots).
3. A dashboard hopefully visualizing the two things listed above