CS4100 March Madness Prediction Tool: AI-Powered Tournament Outcome Predictor

Team Members and the Division of Labor

- **Kai Webber**: Data collection and preprocessing - responsible for gathering historical NCAA data, cleaning datasets, feature engineering, and ensuring data quality
- **Arjun Avinash**: Model development and architecture - responsible for designing, implementing, and optimizing the machine learning models for prediction
- **Akash Alaparthi**: Model evaluation and visualization - responsible for creating metrics to evaluate model performance, building an interactive visualization system, and researching more effective models

Problem Description

The NCAA March Madness basketball tournament is known for its unpredictability and exciting upsets. Our project aims to create a machine learning system that can predict game outcomes and identify potential upsets better than traditional methods like seeding or basic statistical approaches. Given the multi-stage, single-elimination tournament structure with 68 teams, this creates a complex prediction problem with significant public interest.

Technical Problem Statement

Input:

- **Team Identification & History:** Unique Team IDs, Team Names, and seasons as Division-I teams.
- **Compact Game Results:** Historical outcomes (since 1985 men, 1998 women) for regular season, NCAA tournaments, and secondary tournaments (winners, losers, scores, overtime, locations).
- **Detailed Box Scores:** Game-level team statistics (since 2003 men, 2010 women): field goals, three-pointers, free throws, rebounds, assists, steals, blocks, turnovers, personal fouls.
- **Tournament Seeding & Brackets:** NCAA tournament seeds, regional identifiers, play-in games, and round-by-round tournament structures.
- **Season Dates & Regions:** Season reference dates (DayZero), standardized DayNum for games, and regional tournament naming (Regions W, X, Y, Z).

- **Public Team Rankings:** Massey Ordinal rankings (men's data since 2003), including systems like Pomeroy, Sagarin, RPI, ESPN.
- **Coach Information:** Head coach tenure periods, including mid-season coaching changes.
- **Conference Data:** Historical conference affiliations, descriptions, and conference tournament results.
- **Secondary Tournaments:** Teams, results, and tournament identifiers (NIT, WNIT, CBI, CIT, Vegas 16, The Basketball Classic).
- **Game Geography:** City and state information for all games (since 2010), linking CityIDs to specific games.

Output:

- Game-by-game win probabilities for potential matchups
- Upset likelihood indicators
- Predicted bracket outcomes
- Visualizations showing confidence levels and key factors
- Explanations for predictions (feature importance)

Ideal Project Outcome

We aim to develop a system that can:

1. Predict NCAA tournament outcomes with accuracy superior to baseline methods
2. Identify potential upsets with higher precision than seed-based predictions
3. Provide interpretable results explaining why certain teams are favored
4. Generate complete tournament brackets with confidence metrics
5. Offer a user interface for exploring different scenarios and predictions

Algorithms and Methods

We plan to use and compare multiple approaches:

- Logistic regression as a baseline model
- Random forests
- Neural networks for complex pattern recognition
- Monte Carlo simulations for generating full bracket probabilities
- Possibly look into XGBoost (extreme gradient boosted) regressions

Libraries, Platforms, and Learning Requirements

We will need to learn and utilize:

- **Data Collection**: NCAA APIs, web scraping tools, sports data repositories, and Kaggle NCAA data sets
- **Data Processing**: Pandas and NumPy
- **Modeling**: TensorFlow, scikit-learn, and possibly others as we progress through the project
- **Visualization**: Matplotlib, Seaborn, and potentially more interactive ones as we progress

Dataset

We will use multiple data sources:

- Historical NCAA tournament results (1985-present)
- Regular season game statistics from NCAA and sports databases
- Team and player statistical profiles from reliable sports data repositories
- Kaggle's March Machine Learning Mania datasets
- Additional data like which stadium the game is being played in and travel distance

Halfway Milestone (Due 4/11)

By the halfway point, we aim to:

1. Complete data collection and preprocessing
2. Implement baseline prediction models (logistic regression, simple decision trees)
3. Establish evaluation metrics and testing framework
4. Create an initial model evaluation on historical tournaments

Weekly Plan

Week 1 (3/12-3/18):

- review of existing approaches
- Data collection strategy and initial gathering
- Project setup (GitHub repo, development environment)

Week 2 (3/19-3/25):

- Complete data collection and initial cleaning
- Exploratory data analysis to identify key predictive features

Week 3 (3/26-4/1):

- Implement baseline models
- Create evaluation framework
- Initial testing on historical tournaments

Week 4 (4/2-4/8):

- Implement more advanced models
- Begin comparative analysis between models
- Start implementing the visualization aspects

Week 5 (4/9-4/15):

- Complete milestone deliverables
- Refine models based on initial performance

Week 6 (4/16-4/22):

- Complete model optimization
- Finalize visualization and design ideas
- Run the full model on historical tournaments for validation

Week 7 (4/23-4/24):

- Final polishing and report writing
- Documentation and code cleanup
- Prepare final presentation materials

Rationale for Team Size

Our team of three is justified for this project because:

1. The data collection and preprocessing component requires significant effort due to the diverse sources and the need for feature engineering
2. The modeling component involves implementing and comparing multiple sophisticated approaches
3. The visualization and user interface are essential for making the project accessible and useful
4. The scope includes not just prediction but also interpretation and scenario exploration

All three team members will contribute to implementation, analysis, and documentation, but the clear division of primary responsibilities ensures efficient project management while tackling a problem that would be challenging for a smaller team to complete within the time constraints.

Kaggle Competition and Dataset

https://www.kaggle.com/competitions/march-machine-learning-mania-2025/overview/

Possible Supplementary Dataset

https://evanmiya.com/?player_ratings