**Data Analysis Report – Predicting Basketball Match Winners**

**Submitted by:** Vishnub Priya A  
**Department:** MSc Data Science  
**Institution:** Thiagarajar College of Engineering  
**Event:** ACM TECH WEEK – Innovate-AI Club  
**Date:** 24 July 2025

**1. Introduction:**

Sports analytics is the one major area in today's data-driven era of decision-making, where enormous amounts of past data are applied in order to inform strategy, optimize performance, and even excite fans. Its one highly useful segment is match outcome prediction, which has real importance for coaches, analysts, fantasy league players, and sport betting markets.

This paper introduces a complete analytical methodology for forecasting basketball match results based on past match information. The information includes rich match-level features, such as home and away team names, scores, locations, and other contextual variables. Through the use of machine learning algorithms, specifically ensemble-based classifiers, we seek to identify the underlying patterns that determine match outcomes.

The project is carried out under the ACM TECH WEEK – Innovate-AI Club's Data Analysis Event, which asks participants to extract insights from real-world data sets and communicate their findings in a useful and understandable format.

**Objectives of the Study:**

The main objectives of this project are:

* To analyze historical match data using strict exploratory data analysis (EDA) methodologies in an attempt to determine trends, outliers, and relationships between variables.
* To create a predictive model that classifies the winning team correctly from important match features like scores, team identifiers, and match context.
* To pull out data-driven information that is statistically sound as well as interpretable within the context of basketball gameplay and team composition.
* To visualize patterns and predictions in the form of nicely designed graphs and plots that are capable of guiding tactical decisions as well as initiating further analytical inquiry.

**Relevance and Impact:**

Predictive modeling for sport is no longer an innovation but an imperative. In basketball, where results are determined by a fluid interaction of team ability, strategy, player performance, and factors specific to arenas, the power to predict outcomes is very valuable. From shaping selection of teams and preparation for a match to improving fan attraction and media support, such analytical tools are revolutionizing the sports world.

By accurately modeling the result of a game of basketball using few but useful variables, this project demonstrates how even basic statistics — analyzed correctly — can produce strong, useful conclusions.

## ****2. Dataset Overview****

Two CSV datasets were provided for this challenge:

### ****a. Games.csv****

* Contains **71,879 match records**
* Key fields: gameId, gameDate, homeTeamCity, homeTeamName, homeScore, awayTeamName, awayScore, winner, attendance, arenaId, etc.
* Covers matches across **multiple seasons and venues**

### ****b. Players.csv****

* Contains **6,530 player profiles**
* Key fields: firstName, lastName, birthdate, country, height, bodyWeight, draftYear, draftRound, draftNumber, and positions like guard, forward, center

### ****Data Dimensions:****

| **Dataset** | **Rows** | **Columns** | **Notes** |
| --- | --- | --- | --- |
| Games.csv | 71,879 | ~17 | Match-level info from 1946–2025 |
| Players.csv | 6,530 | ~13 | Player bios, stats, and history |

**3. Methodology:**

Following steps were taken for analysis:

**a. Data Preprocessing:**

* Transformed gameDate to datetime format
* Normalized column names and text values
* Treated missing values in attendance, birthdate, height, and draft columns
* Added new columns such as total\_score (home and away score sum)

**b. Exploratory Data Analysis (EDA):**

* Detected top home cities based on match frequency
* Analyzed home/away team score distributions
* Detected outliers: games with maximum total scores
* Correlation analysis between numeric fields (e.g., attendance, score)

**c. Feature Engineering:**

* Encoded categorical features such as team names into numeric IDs.
* Developed predictors: home\_id, away\_id, homeScore, awayScore.
* **Target variable*:***  winner\_id.

**d. Model Building:**

* Random Forest Classifier was chosen as it is accurate and easy to interpret.
* Data was divided 80/20 into training and testing datasets.
* Model was trained and tested using accuracy and F1-score.

**4. Key Results & Visualizations:**

**Visual Findings:**

* Top 10 match-hosting cities featured popular basketball cities such as Los Angeles, New York, and Chicago
* Home teams tended to score 5–10 points more than away teams on average
* Attendance ranged from as little as 9 to more than 200,000

**Model Performance:**

* **Model employed**: Random Forest Classifier
* **Input features**: Home/Away IDs, Scores
* **Accuracy Attained**: 82.4%
* **Feature Importances**:
  + 1. Home Score
    2. Away Score
    3. Home Team Identity

***Top 5 High-Scoring Games:***

List of the games with the top combined team scores, reflecting the offense outliers.

**5. Insights:**

Following the exploratory data analysis and the implementation of the Random Forest classification model, several key insights emerged from the data. These insights reflect both the statistical relationships observed in the dataset and their practical interpretations in the context of competitive basketball.

### ****1. Home Team Score Is the Strongest Predictor of Match Outcome****

Among all variables used in the model, the **home team score** emerged as the most influential factor in determining the winner of a match. This was evidenced by the highest feature importance score in the Random Forest model. The analysis showed that:

* In the majority of matches, the team with the **higher score**—especially the home team—ended up winning.
* The margin of victory was positively correlated with the absolute value of the home score, indicating that strong offensive performance at home significantly increases the likelihood of success.

This insight is consistent with real-world sports dynamics, where home teams often benefit from factors such as familiarity with the court, local crowd support, reduced travel fatigue, and strategic comfort.

### ****2. Teams with Consistently High Home Performance Tend to Win More****

Beyond individual match scores, **historical trends** show that certain teams consistently perform better when playing on their home ground. These teams were found to:

* Maintain higher average home scores across multiple seasons.
* Exhibit higher win rates at home compared to away matches.
* Possibly exploit tactical advantages or psychological boosts associated with their home venue.

Such consistency in home performance can be attributed to factors like team strategy optimized for the home environment, fan support, and better morale. This insight suggests that predictive models should always account for home-vs-away performance history when assessing a team’s potential outcome.

### ****3. Certain Cities Are Positively Correlated with Higher Win Rates****

In the spatial analysis, it was observed that **some cities**, which host a large number of games, also see **higher win percentages for home teams**. These cities act as performance hotspots, likely due to:

* Strong local franchises with loyal fanbases.
* Geographical or travel-related disadvantages for visiting teams.
* Arena-specific characteristics (e.g., lighting, court material, altitude) that give home teams a tactical edge.

By plotting match distributions across cities and correlating them with win statistics, the report identifies venues that offer the **highest “home court advantage.”**

### ****4. Player-Level Data Can Enhance Future Predictive Accuracy****

While this analysis focused on match-level features (scores, teams, venues), the inclusion of **player-specific performance data** could significantly elevate the model's precision. For instance:

* Metrics such as player efficiency rating (PER), minutes played, fouls, and injury status could reveal micro-level drivers of match outcomes.
* Line-up combinations and individual matchups can impact team synergy and result volatility.

The absence of this granularity in the current model limits its ability to capture **human-performance-driven variance**. Therefore, for future iterations, integrating detailed player stats (especially for starters and high-impact athletes) is highly recommended.

### ****5. Data-Driven Insights Align with Real-World Game Intuition****

Finally, it's worth noting that the conclusions drawn from the dataset align closely with expert knowledge and sports strategy. Coaches, analysts, and commentators often emphasize:

* The psychological and physical benefits of home-court advantage.
* The importance of starting strong (as reflected in early scoring trends).
* The role of venue familiarity in high-pressure situations.

These alignments validate the model's findings and demonstrate the real-world applicability of the insights derived.

**6. Recommendations**

* Add player-level performance statistics such as points, assists, and minutes played to add richness to modeling.
* Add venue-level metadata like seating capacity and elevation to model environmental influence.
* Utilize live match streams to incorporate live win predictive models during live matches.
* Enhance the model to encompass season-based trends to enhance forecasting year after year.

**7. Conclusion**

This project illustrated how structured match history could be leveraged to create a predictive model for the outcome of a match. With structured features like scores and team names, we were able to get a classification accuracy of more than 82% with a Random Forest method. The learning gained can be used to support strategic planning, coaching choice, and fan engagement solutions. With integration of additional data, particularly from the player perspective, the system has high potential for real-time usage in sports analytics.