**ENGLISH PREMIER LEAGUE**

**MATCH PREDICTION**

**A MINI PROJECT REPORT**

***Submitted by***

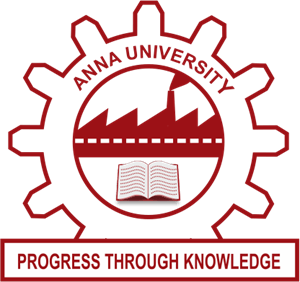
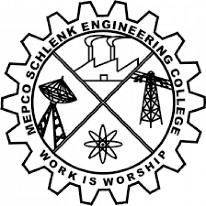
**Jayavelan S (Reg. No. 9517202309041)**

**Nandhagopal V (Reg. No. 9517202309074)**

**Sanjay K N (Reg. No. 9517202309101)**

***in partial fulfillment for the requirement of the course***

**23AD453 – MINI PROJECT II: DATA ANALYTICS**

*** ***

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**MEPCO SCHLENK ENGINEERING COLLEGE (AUTONOMOUS)**

**SIVAKASI**

**APRIL 2025**

**BONAFIDE CERTIFICATE**



This is to certify that it is the bonafide work done by Jayavelan S (9517202309041), Nandhagopal V (9517202309074), Sanjay K N (9517202309101) for the course **23AD453 - Mini Project – II: DATA ANALYTICS** work entitled **“English Premier League Match Prediction”** at Department of Artificial Intelligence & Data Science, Mepco Schlenk Engineering College, Sivakasi during the year 2024 – 2025.

………………………… ……………………………

***Faculty In-Charge Head of the Department***

Submitted for the Practical Examination held at Mepco Schlenk Engineering College, Sivakasi on ……………………….

………………………… ……………………………

***Internal Examiner - I Internal Examiner - II***

**ACKNOWLEDGEMENT**

First and foremost, we express our deepest gratitude to the **LORD ALMIGHTY** for His abundant blessings, which have guided us through our past, present, and future endeavors, making this project a successful reality.

We extend our sincere thanks to our management, esteemed Principal, **Dr. S. Arivazhagan, M.E., Ph.D.,** for providing us with a conducive environment, including advanced systems and well-equipped library facilities, which have been instrumental in the completion of this project.

We are profoundly grateful to our Head of the Artificial Intelligence and Data Science Department, **Dr. J. Angela Jennifa Sujana, M.Tech., Ph.D.,** for granting us this golden opportunity to work on a project of this significance and for her invaluable guidance and encouragement throughout.

Our heartfelt appreciation goes to our project guides, **Dr. R. Nirmalan, M.Tech., Ph.D.,** Assistant Professor, and **Mrs.A. Revatheeswari, M.Tech., Ph.D.,** Assistant Professor, Department of Artificial Intelligence and Data Science for their unwavering support, insightful suggestions, and expert guidance. Their experience and encouragement have been vital in shaping our project and helping us bring forth our best.

We also extend our sincere gratitude to all our faculty members, lab technicians, and our beloved family and friends, whose timely assistance and moral support played a pivotal role in making this mini-project a success.

**ABSTRACT**

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Title of Table** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title of Figure** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**List of Symbols and Abbreviations**

|  |  |
| --- | --- |
| **Symbol / Abbreviation** | **Description** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**SYNOPSIS**

**Module 1: Problem Statement Identification**

This module is dedicated to understanding the problem at hand and structuring data effectively for analysis. It emphasizes data warehouse schemas, which form the backbone of organizing and managing large datasets. By leveraging the Star, Snowflake, and Constellation schemas, students can structure their data efficiently. Additionally, the module covers the analytical operations of Online Analytical Processing (OLAP), which facilitates multidimensional data analysis, aiding in problem scoping and decision-making.

**Module 2: Data Acquisition**

This module focuses on gathering data from various sources and preparing it for analysis. Students will explore different dataset types, including structured (tabular data), semi-structured (JSON, XML), and unstructured (text, images, audio). The module ensures proper categorization of attributes into numerical, categorical, ordinal, and nominal data. Further, students will learn to identify independent (features) and dependent (target) attributes, derive new attributes when necessary, and perform feature selection to retain only the most relevant attributes for analysis. The ability to source and organize high-quality data is fundamental to ensuring meaningful insights in later stages.

**Module 3: Preprocessing of Datasets**

Before applying machine learning algorithms, raw data must undergo preprocessing to enhance its quality and usability. This module teaches students to clean data by handling missing values, detecting and removing outliers, and eliminating duplicate records. Using Weka, students will perform data normalization to scale features within a standard range, standardization to achieve a consistent mean and variance, and transformation techniques such as one-hot encoding and discretization. Additionally, dimensionality reduction techniques, like Principal Component Analysis (PCA), will be explored to optimize model performance while reducing computational complexity.

**Module 4: Model Generation**

Building predictive models is the core objective of this module. It covers different machine learning techniques, including association rule mining, regression, classification, and clustering models. Students will work with algorithms like Apriori and FP Growth for association rule mining, various regression models (linear, multiple linear, polynomial, logistic), and classification models such as Decision Trees (ID3) and Naïve Bayes. For clustering, methods like k-means, Agglomerative Clustering, and DBSCAN will be applied. This module provides a hands-on approach to selecting and implementing suitable models for diverse data-driven problems.

**Module 5: Testing of Model**

Once a model is built, evaluating its performance is crucial to ensure its reliability and accuracy. This module introduces various evaluation metrics tailored for different models. For classification models, key metrics include accuracy, precision, recall, F1-score, and ROC-AUC. Regression models will be evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared. Clustering models will be assessed using the Silhouette Score, Dunn Index, and Davies-Bouldin Index, while association rule mining will involve Support, Confidence, and Lift. Additionally, visualization techniques such as confusion matrices, ROC curves, scatter plots, and heatmaps will help interpret model results effectively.

**Module 6: Deployment**

The final step in the data science workflow is deploying the developed models for practical use. This module focuses on finalizing the trained models, integrating them into a user-friendly interface, and ensuring their seamless deployment. Students will learn how to prepare the final model, develop a functional UI for interaction, and host the model on appropriate platforms. This stage bridges the gap between theoretical learning and real-world implementation, allowing users to interact with the model in a meaningful way.

**Table of Contents**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chapter No.** | **Contents** | | | | **Page No.** |
|  | Abstract | | | |  |
|  | List of Tables | | | |  |
|  | List of Figures | | | |  |
|  | List of Symbols and Abbreviations | | | |  |
|  | Synopsis | | | |  |
| **1** | **INTRODUCTION** | | | |  |
|  | 1.1 | Overview of the Project | | |  |
|  | 1.2 | Objectives | | |  |
|  | 1.3 | Scope of the Project | | |  |
|  | 1.4 | Tools & Programming Languages Used | | |  |
| **2** | **PROBLEM STATEMENT IDENTIFICATION** | | | |  |
|  | 2.1 | Understanding the Problem | | |  |
|  | 2.2 | Data Warehouse Design | | |  |
|  | 2.2.1 | | Star Schema | |  |
|  | 2.2.2 | | Snowflake Schema | |  |
|  | 2.2.3 | | Constellation Schema | |  |
|  | 2.3 | OLAP and Analytical Operations | | |  |
| **3** | **DATA ACQUISITION** | | | |  |
|  | 3.1 | Data Collection Methods | | |  |
|  | 3.2 | Types of Data | | |  |
|  | 3.3 | Attribute Categorization and Classification | | |  |
|  | 3.3.1 | | Independent vs. Dependent Attributes | |  |
|  | 3.3.2 | | Feature Selection and Extraction | |  |
| **4** | **DATA PREPROCESSING** | | | |  |
|  | 4.1 | Data Cleaning | | |  |
|  | 4.1.1 | | Handling Missing Values | |  |
|  | 4.1.2 | | Removing Outliers and Duplicates | |  |
|  | 4.2 | Data Normalization and Standardization | | |  |
|  | 4.3 | Data Transformation | | |  |
|  | 4.3.1 | | Encoding Categorical Data | |  |
|  | 4.3.2 | | Feature Selection Techniques | |  |
| **5** | **MODEL GENERATION** | | | |  |
|  | 5.1 | Association Rule Mining | | |  |
|  | 5.1.1 | | Apriori Algorithm | |  |
|  | 5.1.2 | | FP-Growth Algorithm | |  |
|  | 5.2 | Regression Models | | |  |
|  | 5.2.1 | | Linear Regression | |  |
|  | 5.2.2 | | Multiple Linear Regression | |  |
|  | 5.2.3 | | Logistic Regression | |  |
|  | 5.3 | Classification Models | | |  |
|  | 5.3.1 | | | Decision Tree Classifier |  |
|  | 5.3.2 | | | Naïve Bayes Algorithm |  |
|  | 5.3.3 | | | K-Nearest Neighbor (KNN) Classification |  |
|  | 5.4 | Clustering Models | | |  |
|  | 5.4.1 | | | k-means Algorithm |  |
|  | 5.4.2 | | | Agglomerative Clustering |  |
|  | 5.4.3 | | | DBSCAN Clustering |  |
| **6** | **TESTING OF MODEL** | | | |  |
|  | 6.1 | Evaluation Metrics | | |  |
|  | 6.1.1 | | | Association Rule Mining Evaluation |  |
|  | 6.1.2 | | | Regression Model Evaluation |  |
|  | 6.1.3 | | | Classification Model Evaluation |  |
|  | 6.1.4 | | | Clustering Model Evaluation |  |
|  | 6.2 | Results Visualization | | |  |
|  | 6.2.1 | | | ROC Curve |  |
|  | 6.2.2 | | | Confusion Matrix |  |
|  | 6.2.3 | | | Scatter Plots and Heatmaps |  |
| **7** | **DEPLOYMENT** | | | |  |
|  | 7.1 | Finalizing the Model | | |  |
|  | 7.2 | User Interface (UI) Integration | | |  |
|  | 7.2.1 | | | Tools Used |  |
|  | 7.2.2 | | | Deployment |  |
| **8** | **CONCLUSION** | | | |  |
|  | 8.1 | Summary of Findings | | |  |
|  | 8.2 | Challenges Faced | | |  |
|  | 8.3 | Future Enhancements | | |  |
| **9** | **REFERENCES** | | | |  |
| **10** | **APPENDICES** | | | |  |

**CHAPTER 1**

**INTRODUCTION**

Predicting the winner of a football tournament is a tiresome task that involves analyzing large and complex datasets comprising historical game statistics, team performance metrics, and contextual factors. Our project aims to address this challenge by designing a predictive system capable of forecasting game outcomes and determining the overall tournament winner. To structure and manage the data effectively, the project leverages data warehouse schemas and Online Analytical Processing (OLAP) techniques for multidimensional analysis and decision-making.

**1.1 Overview of the Project**

The project focuses on predicting the winning team of English Premier League by creating a virtual tournament of 8 teams using historical game data and machine learning techniques. By analyzing team performance metrics, game contexts, and other relevant features, the project aims to provide accurate predictions for individual games and simulate the progression of an entire tournament to identify the ultimate winner.

**1.2 Objectives**

1. Game-Level Prediction:

Develop a machine learning model to predict the outcome (win/loss) of individual football games using historical data.

2. Tournament Simulation:

Simulate the progression of a football round-of-8 tournament by iteratively predicting game outcomes and advancing winners through the bracket.

3. Insights and Interpretations:

Uncover the key factors influencing game outcomes and tournament success, providing actionable insights for teams, analysts, and fans.

**1.3 Scope of the Project**

The scope of this project is to develop a predictive system capable of forecasting the winner of a basketball tournament. The system will analyze historical game data, team performance metrics, and contextual factors using machine learning and data analytics techniques. The project aims to address both game-level outcome predictions and tournament-level winner forecasting, providing insights and tools for stakeholders such as sports analysts, teams, and fans.

**1.4 Tools & Programming Languages Used**

This project utilizes a combination of tools and programming languages to ensure efficient data analysis, visualization, model development, and deployment. Below is an overview of the tools and programming languages used:

1. Python

Python serves as the core programming language for this project. Its extensive ecosystem of libraries makes it ideal for tasks such as data preprocessing, feature engineering, model development, and deployment. Key Python libraries used include pandas for data manipulation, numpy for numerical computations, scikit-learn for machine learning, and matplotlib and seaborn for data visualization.

2. Jupyter Notebook

Jupyter Notebook provides an interactive environment for performing data analysis and visualization. It is extensively used in this project for exploratory data analysis, iterative development, and documentation.

3. Orange

Orange is an open-source data visualization and analysis tool that supports visual programming. It is used for creating workflows, visualizing data, and experimenting with machine learning models in a user-friendly, drag-and-drop interface.

4. Weka

Weka, a collection of machine learning algorithms for data mining tasks, is employed in this project for advanced analytics and experimentation with various machine learning models.

5. Flask

Flask is a Python-based framework used to create work flow between different webpages.

6. Html & CSS

Html is used to create a webpage with its skeleton and CSS is used to style the webpage for better UI experience

**CHAPTER 2**

**PROBLEM STATEMENT IDENTIFICATION**

**2.1 Understanding the Problem**

The primary challenge in this project is predicting the winning team of a football tournament. This involves analyzing historical basketball game data and using it to forecast future outcomes, both at the individual game level and the tournament level. To address this problem comprehensively, we must break it into several components:

1. Data Understanding

The project relies on historical football game data, which includes team performance metrics (shots and betting ratings), contextual factors (home/away games), and game outcomes (scores, win/loss). Understanding the structure, quality, and completeness of this data is critical for building an effective prediction model.

1. Defining the Objective

The objective is twofold:

* Predict the outcome of individual football games based on historical data.
* Use these predictions to simulate the progression of a tournament and identify the ultimate winner.

1. Key Challenges

Data Quality and Availability: Ensuring the dataset is clean, complete, and representative of the factors affecting game outcomes.

* Feature Selection: Identifying the most relevant features (e.g., offensive/betting ratings, field foul counts, home-stadium advantage) that influence game results.
* Tournament Dependencies: Accounting for the fact that the outcome of one game influences subsequent matches in the tournament bracket.
* Model Accuracy: Balancing the precision and generalization of the predictive model to ensure robust predictions across different teams and tournaments.

1. Approach to the Problem

Data Structuring: Organize the data using appropriate schemas (e.g., Star Schema, Snowflake Schema) for efficient storage and analysis.

* Data Analysis: Use tools like Jupyter Notebook for exploratory data analysis and visualization to understand trends and patterns.
* Machine Learning: Train a predictive model using tools like Python, Orange, Weka, and scikit-learn to forecast individual game outcomes.
* Tournament Simulation: Develop a framework to simulate tournament brackets using the game-level predictions.
* Deployment and Visualization: Use Flask for deploying the model as an interactive application and Plotly for creating dashboards to display insights.

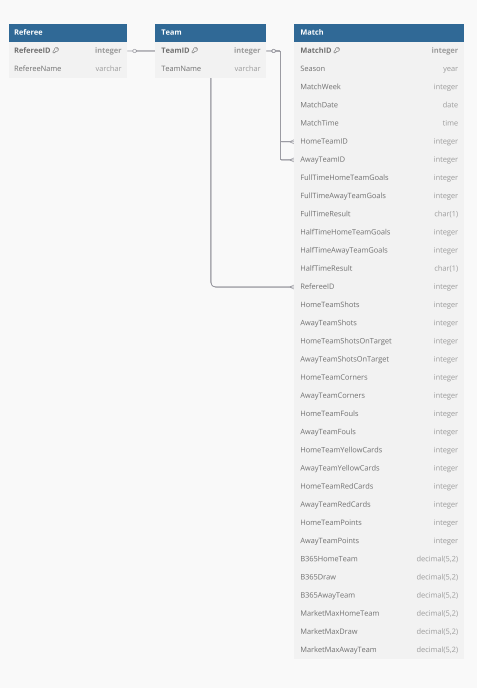
1. Outcome

The project aims to provide a robust system for predicting football tournament winners. It also seeks to deliver valuable insights into the factors driving team success, which can be used by analysts, teams, and fans to enhance their understanding of the game.

**2.1 Data Warehouse Design**

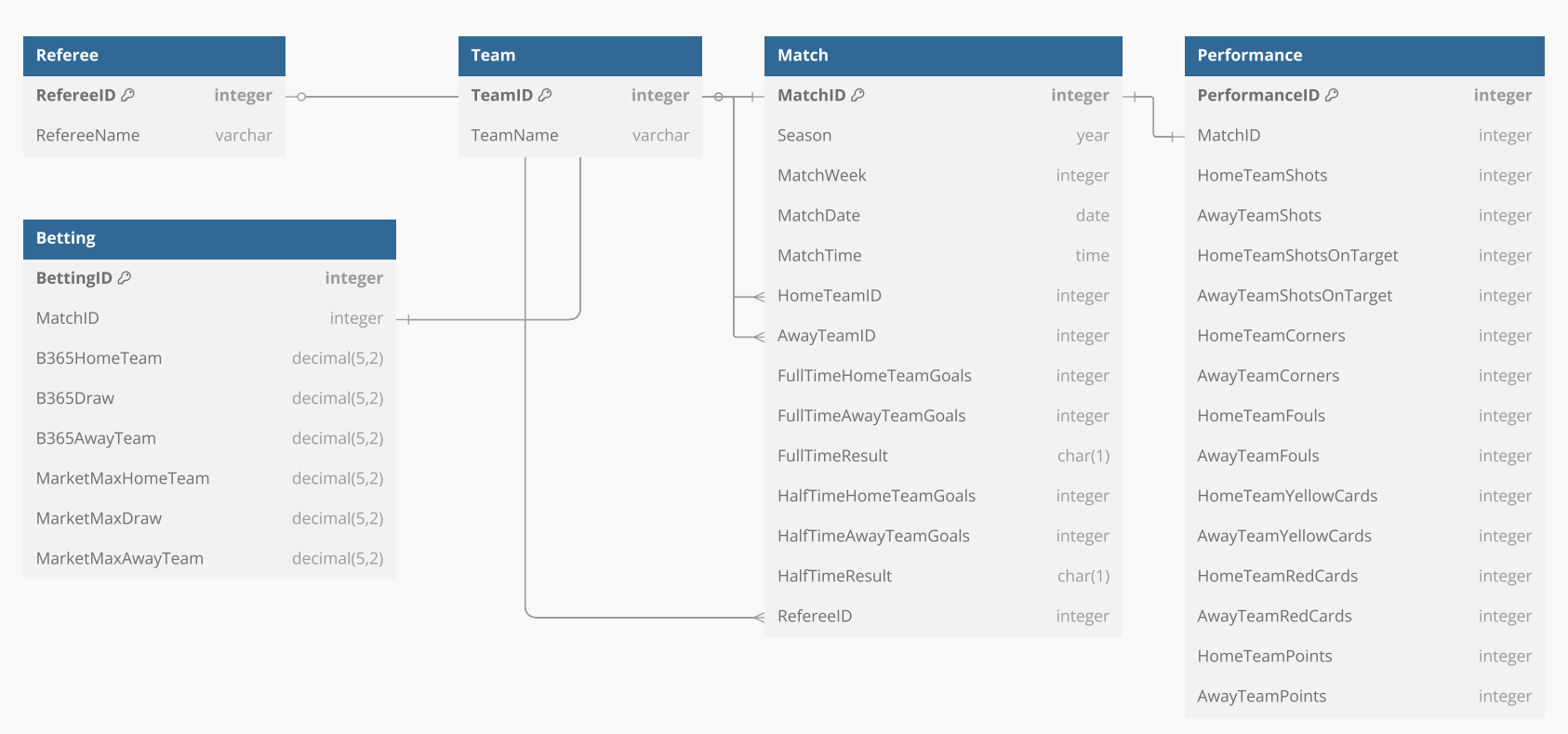
A well-designed data warehouse is essential for storing and managing historical basketball data efficiently. This project incorporates three popular schema designs—Star, Snowflake, and Constellation—to organize the data for analytical purposes. These schemas facilitate Online Analytical Processing (OLAP) operations and make the data accessible for multidimensional analysis, enabling better decision-making and accurate predictions.

**2.2.1 Star Schema**

****

**Fig 2.1 Star schema**

**2.2.2 Snowflake Schema**

****

**Fig 2.2 Snowflake schema**

**2.3 OLAP and Analytical Operations**

Online Analytical Processing (OLAP) is a powerful data analysis technique that enables multidimensional exploration of data and supports decision-making processes. In this project, OLAP operations are used to analyze basketball game data effectively, aiding in identifying patterns, trends, and insights crucial for predicting tournament outcomes. Below is an overview of the OLAP operations and their applications in this project:

1. Slice and Dice

Slice and dice operations are used to focus on specific subsets of data.

Slice: Extracts a single layer of data along a specific dimension. For example, analyzing team performance for a particular season or focusing on home games only.

Dice: Extracts a sub-cube of data by applying multiple filters. For instance, analyzing games where the home team won and attendance was above a certain threshold.

2. Roll-Up and Drill-Down

These operations allow aggregation and detailed exploration of data.

Roll-Up: Aggregates data to a higher level of abstraction. For example, rolling up game-level data to calculate team performance metrics at the season level.

Drill-Down: Allows detailed analysis by breaking data down to a finer level. For instance, drilling down from season-level metrics to individual game statistics.

3. Pivot (Cross-tabulation)

Pivoting reorients the data to view it from different perspectives.

Example: Comparing offensive and defensive ratings across teams over multiple seasons by pivoting the data to focus on these metrics.

**CHAPTER 3**

**DATA ACQUISITION**

**3.1 Data Collection Methods**

The success of this project relies heavily on the quality and accuracy of the data collected. To ensure a robust dataset for predicting basketball tournament outcomes, the following data collection methods have been employed:

1. Historical Game Data

Historical data from past football games is the backbone of this project. This data includes detailed information about team performance, game context, and outcomes. It is collected from publicly available sports databases, APIs, and official league statistics. Sources may include websites like EPL, Kaggle datasets, or sports analytics platforms.

2. Data Updation

As premier league runs all through the year, data needs to be updated constantly. Our dataset is updated after each matchweek by volunteer with real-time data.

3. Public APIs

Several sports-related APIs, such as those provided by ESPN, EpicSports, or other football statistics platforms, are used for real-time and historical data acquisition. These APIs offer structured data, including metrics like team offense/defense ratings, field goal percentages, and game attendance.

4. Manual Data Entry

In cases where automated methods are not feasible, manual data entry is employed. This involves collecting specific datasets, such as past tournament brackets or unique game metrics, by manually entering them into structured formats like spreadsheets or databases.

5. Data Integration

Data from various sources is integrated to create a comprehensive dataset. This involves combining game-level data, team performance metrics, and contextual factors. Data integration tools and techniques, such as ETL (Extract, Transform, Load) processes, are used to ensure consistency and accuracy across sources.

**3.2 Types of Data**

The dataset used in this project includes various types of data that are crucial for building predictive models and conducting effective analysis. Below is a detailed description of the types of data used:

1. Game-Level Data

This type of data captures information about individual basketball games, including their context and outcomes.

Examples:

* Date and time of the game.
* Home team and away team identifiers.
* Referee of the game.
* Foul and Cards.
* Final scores of the home and away teams.
* Win/loss result for each team.
* Betting and other odds

2. Team Performance Data

Team performance data includes metrics that describe how teams perform both offensively and defensively. These metrics are key predictors for game outcomes.

Examples:

* Shots (Shots on target).
* Foul Rating (Red and yellow cards).
* Corners (Close chances).
* Betting odds.

3. Contextual Data

This type of data adds context to the games and tournaments, helping to improve prediction accuracy.

Examples:

* Home-stadium advantage (whether the game is being played at the home team's venue).
* Historical matchups and rivalry data between teams.
* Referee (Different decisions are made).

4. Tournament Data

Data specific to tournaments provides information on the structure and flow of the tournament.

Examples:

* Bracket structure.
* Knockout vs. round-robin format.
* Seeding and ranking of teams.

5. Derived Metrics

Derived or computed metrics are created based on raw data to enhance the analysis.

Examples:

* Comparing each team’s chances created (e.g., comparing a team's offensive rating with their opponent's).
* Win probabilities derived from machine learning models.
* Momentum metrics (e.g., recent winning streaks, rivalry advantages).

**3.3 Attribute Categorization and Classification**

Organizing and categorizing attributes in the dataset is crucial for building a robust predictive model. This section outlines the classification of attributes and their roles in the project.

**3.3.1 Independent vs. Dependent Attributes**

Attributes in the dataset are classified into two main categories based on their role in the predictive model:

1. Independent Attributes (Features):

These attributes serve as inputs to the model and are used to predict outcomes. Examples include:

* Team Performance Metrics: Shots, Shots on target, Corners.
* Game Context: Home/Away Status, Date, Fouls.
* Historical Data: Win/Loss Streak, Head-to-Head Records.

2. Dependent Attributes (Targets):

These attributes represent the outcomes the model aims to predict. Examples include:

* Game Result: Win/Loss for each team.
* Final Scores: Home and Away team scores.
* Tournament Winner: The team that progresses through the bracket to win the tournament.

This distinction helps in defining the problem and structuring the dataset appropriately for machine learning.

**3.3.2 Feature Selection and Extraction**

Feature selection and extraction are critical steps in preparing the dataset for predictive modeling. These processes ensure that only the most relevant and informative attributes are used, improving model performance and reducing complexity.

1. Feature Selection:

This process identifies the most relevant features from the dataset. Techniques used include:

* Correlation Analysis: Identifying features with strong correlations to the target variable, such as Offensive Ratings and Win Outcomes.
* Statistical Tests: Using methods like Gini Index for categorical features or ANOVA for numerical features to determine significance.
* Feature Importance Metrics: Leveraging algorithms like ID3 to rank features based on their importance in prediction.

2. Feature Extraction:

Feature extraction involves creating new features from the existing data to better capture relationships and improve model accuracy. Examples include:

* Derived Metrics: Calculating Offensive-Scoring tendency Differential to quantify team performance.
* Win Probability: Using historical data to estimate the likelihood of winning.
* Momentum Features: Capturing recent form, such as a team’s win streak or average margin of victory in the last five games.