Match score dataset for team ball sports

T. N. Alleck T. Giovannelli L. N. Vicente R. Mitchell O. Remen

April 8, 2024

Abstract

In this data article, we present a dataset containing match scores from major international competitions for 12 popular team ball sports: basketball, cricket, field hockey, futsal, handball, ice hockey, lacrosse, roller hockey, rugby, soccer, volleyball, and water polo. The dataset was obtained by web scraping data available on Wikipedia pages and includes the following information related to individual matches: names of the two opposing teams, their respective scores, and the name of the winning team. Our match score dataset provides researchers in the field of sports analytics with valuable data that can be used to compute team statistics, develop team ranking and rating systems, infer patterns and trends in a team's performance across the years, build predictive models to forecast the outcome of future matches, and evaluate the performance of machine learning algorithms.

Keywords— Sports analytics, Team ball sports, Match scores

Specifications table

Subject area	Data Science
Specific subject area	Big Data Analytics, Sports Analytics
Data format	Raw, Filtered
Type of data	Table
Data collection	Python web-scraping from 147 Wikipedia pages that include match outcomes from different editions of major international competitions selected for 12 team ball sports
Data source location	Lehigh University
Data accessibility	Repository name: Mendeley Data Data identification number: 10.17632/2pt4vmyf27.1 Direct URL to data: `https://data.mendeley.com/datasets/2pt4vmyf27/1`

Value of the data

- The dataset can be used to compute team statistics and develop team ranking and rating systems to evaluate the performance of teams based on their past match scores.
- Statistical and data-driven methodologies can be used to infer patterns and trends in a team's performance across the years.

Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015-1582, USA
Emails: tna324@lehigh.edu, tog220@lehigh.edu, lnv@lehigh.edu, rgm424@lehigh.edu, orr224@lehigh.edu

- Researchers in the field of sports analytics can use the dataset to build a predictive model to forecast the outcome of future matches and evaluate the performance of different machine learning algorithms.

1 Background

In the last decade, there has been increasing research interest in the analysis of team ball sports [6]. Various avenues have been explored. Some articles use match score data to identify sports with the most random outcomes [1, 4]. Others aim to determine scoring patterns across different sports [5]. Finally, there are articles focused on predicting the outcome of sports matches using machine learning algorithms [2, 3].

The information contained in the match score dataset presented in this article is highly valuable for researchers in the field of sports analytics. By using and exploring the dataset, researchers can compute team statistics (such as the number of matches won and lost by a team) and develop team ranking and rating systems to evaluate the performance of teams based on their past match scores. By analyzing historical match score data, researchers can gain insights into the strengths and weaknesses of each team across different editions of a competition. Statistical and data-driven methodologies offer the opportunity to infer patterns and trends in a team's performance across the years. The dataset can also be used to build predictive models aimed at forecasting the outcome of future matches. By training machine learning algorithms on historical match score data, such models can predict which team is likely to win upcoming matches. Additionally, the dataset enables machine learning practitioners to evaluate the performance of different machine learning algorithms in predicting match outcomes, allowing for the selection of the most effective algorithms for this task.

2 Data description

The dataset we present in this article contains match score data from major international competitions across 12 team ball sports: basketball, cricket, field hockey, futsal, handball, ice hockey*, lacrosse, roller hockey, rugby, soccer, volleyball, and water polo. For each sport, the dataset includes the following information related to individual matches: names of the two opposing teams, their respective scores, and the name of the winning team. Table 1 provides the official names of the competitions selected for each sport, all of which are men's events, and the total number of matches across all the years of the editions of each competition. The complete table with the edition years for each selected competition is provided in Table 3 of Appendix A. Tables 4–7 present the number of matches and teams for each edition. Some popular team ball sports like American football, baseball, and tennis were omitted from our dataset due to either the absence of international competitions or the limited size of their teams compared to the sports included in our paper.

Table 2 introduces some general notation that will allow us to formally describe the match score dataset. Denoting as $\mathcal{S}$ the set of sports, let $\mathcal{E}_s$ be the set of editions for the competition selected for sport $s \in \mathcal{S}$ in Table 1 (one can think of such a set as a set of edition years) and let $\mathcal{P}_e^s = \{p_1, p_2, \ldots, p_{N_e^s}\}$ be the set of all teams playing in edition $e \in \mathcal{E}_s$, where N_e^s is the total number of teams in that edition. We will denote as $\mathcal{M}_e^s \subseteq \mathcal{P}_e^s \times \mathcal{P}_e^s$ the set of matches in edition e, represented as a set of tuples (i, j), where i and j are opposing teams belonging to $\mathcal{P}_e^s$. For any sport $s \in \mathcal{S}$, edition $e \in \mathcal{E}_s$, and match $(i, j) \in \mathcal{M}_e^s$, let $\text{score}_{ij}^{s,e}(i)$ denote the score that team i obtained when playing against team j in edition e (for example, if "5-6" is the outcome of the match (i, j), then $\text{score}_{ij}^{s,e}(i) = 5$ and $\text{score}_{ij}^{s,e}(j) = 6$) and let $\text{winner}_e^s(i, j) \in \{i, j, \text{Draw}\}$ represent the winner of the match, where Draw denotes that the match resulted in a tie. We have

$$\text{winner}_e^s(i,j) = \begin{cases} i, & \text{if } \text{score}_{ij}^{s,e}(i) > \text{score}_{ij}^{s,e}(j), \\ j, & \text{if } \text{score}_{ij}^{s,e}(i) < \text{score}_{ij}^{s,e}(j), \\ \text{Draw}, & \text{otherwise.} \end{cases}$$

*Despite its use of a puck, we classify ice hockey among team ball sports.

Sport	International Competition	Total Number of Matches
Basketball	Summer Olympic Games	572
Cricket	ICC Men's Cricket World Cup	474
Field Hockey	Men's FIH Hockey World Cup	565
Futsal	FIFA Futsal World Cup	412
Handball	Summer Olympic Games	344
Ice Hockey	Winter Olympic Games	544
Lacrosse	World Lacrosse Men's World Cup	306
Roller Hockey	World Skate Roller Hockey World Cup	257
Rugby	Rugby World Cup	325
Soccer	FIFA World Cup	852
Volleyball	FIVB Volleyball Men's World Cup	706
Water Polo	FINA Men's Water Polo World Cup	246
	Total:	5603

Table 1: Major international competitions selected for the team ball sports included in our paper, along with the total number of matches across all the edition years of each competition.

Given a sport $s \in \mathcal{S}$, the match score dataset for each edition $e \in \mathcal{E}_s$ can be represented by the following set

$$\mathcal{D}_e^s = \{(i, j, \text{score}_{ij}^{s,e}(i), \text{score}_{ij}^{s,e}(j), \text{winner}_e^s(i,j)) \mid (i,j) \in \mathcal{M}_e^s\}. \tag{2.1}$$

Note that given a sport s and an edition e, the size of $\mathcal{D}_e^s$ is equal to the size of $\mathcal{M}_e^s$, which can be obtained from the values in the 'Matches' column in Tables 4–7. The size of $\mathcal{P}_e^s$, which is equal to N_e^s, can be obtained from the values in the 'Teams' column in Tables 4–7. Combining the match score datasets in (2.1) across all the edition years of the competition selected for each sport s, one obtains the following dataset

$$\mathcal{D}^s \ = \ \cup_{e \in \mathcal{E}_s} \mathcal{D}_e^s. \tag{2.2}$$

The dataset available in our repository consists of 12 CSV files, each corresponding to a sport $s \in \mathcal{S}$. Such files are named as follows

"match_score_dataset_[SPORT_NAME].csv",

where [SPORT_NAME] denotes the name of a sport (see the 'Sport' column in Table 1). For each sport s, the corresponding CSV file contains the dataset $\mathcal{D}^s$, as defined in (2.2). Each dataset $\mathcal{D}^s$ consists of five features, i.e., 'Team 1', 'Team 2', 'Score 1', 'Score 2', and 'Winner', which are associated with i, j, $\text{score}_{ij}^{s,e}(i)$, $\text{score}_{ij}^{s,e}(j)$, and $\text{winner}_e^s(i,j)$ in (2.1), where $(i,j) \in \mathcal{M}_e^s$.

3 Experimental design, materials, and methods

For each of the 12 team ball sports included in our paper, to populate the datasets (2.1) and (2.2), we collected real data on match scores from available editions of the major international competitions listed in Table 1. In particular, we obtained match score data through Python web-scraping of 147 Wikipedia pages that include match outcomes from the selected editions of the competitions. Table 8 of Appendix B includes links to the web-scraped pages. Most of the major international sports competitions listed in Table 1 include a group stage, where teams are divided into groups and compete against each other within their group to accumulate points and progress in the competition, and a knockout stage (or bracket stage), where teams are eliminated from the competition if they lose a match. The knockout stage typically consists of the following additional phases: rounds of of 16, quarterfinals, semifinals, and

$\mathcal{S}$	Set of sports.
$\mathcal{E}_s$	Set of editions for the competition selected for sport $s \in \mathcal{S}$.
$\mathcal{P}_e^s = \{p_1, p_2, \ldots, p_{N_e^s}\}$	Set of all teams playing in edition $e \in \mathcal{E}_s$.
N_e^s	Cardinality of the set $\mathcal{P}_e^s$.
$\mathcal{M}_e^s \subseteq \mathcal{P}_e^s \times \mathcal{P}_e^s$	Set of matches in edition $e \in \mathcal{E}_s$.
$\text{score}_{ij}^{s,e}(i)$, $\text{score}_{ij}^{s,e}(j)$	Scores obtained by teams i and j when facing each other, where $(i, j) \in \mathcal{M}_e^s$.
$\mathcal{D}_e^s$	Match score dataset for edition $e \in \mathcal{E}_s$, as defined in (2.1).
$\mathcal{D}^s$	Dataset resulting from the union of the datasets $\mathcal{D}_e^s$ for all $e \in \mathcal{E}_s$, as defined in (2.2).

Table 2: Notation.

finals. The FIVB Men's World Cup for volleyball is an exception to this format. In such a competition, teams are divided into two groups. In the first phase, each team plays one match against all other teams in its group. In the second phase, each team plays one match against all the teams in the other group, and the competition champion is determined based on the total number of points and other criteria.

Limitations

We note that due to the complexity of the cricket scoring system, scores for cricket matches are not available in the dataset. However, the dataset does include the name of the winning team for each match, as this information is obtainable from Wikipedia pages related to the ICC Men's Cricket World Cup. In 7 matches, the winning team is not available and is listed as "No result", while in 5 matches, it is labeled "Match abandoned".

Ethics statement

The authors declare that they have read and followed the ethical requirements for publication in Data in Brief and confirm that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Acknowledgments

This work is partially supported by the U.S. Air Force Office of Scientific Research (AFOSR) award FA9550-23-1-0217.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. Ben-Naim, F. Vazquez, and S. Redner. Parity and predictability of competitions. *Journal of Quantitative Analysis in Sports*, 2:1–1, 01 2006.

[2] R. P. Bunker and T. Susnjak. The application of machine learning techniques for predicting match results in team sport: A review. *J. Artif. Int. Res.*, 73, May 2022.

[3] R. P. Bunker and F. Thabtah. A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15:27–33, 2019.

[4] M. J. Lopez, G. J. Matthews, and B. S. Baumer. How often does the best team win? A unified approach to understanding randomness in north american sport. *The Annals of Applied Statistics*, 2017.

[5] S. A. Merritt and A. Clauset. Scoring dynamics across professional team sports: Tempo, balance and predictability. *EPJ Data Science*, 3:1–21, 2013.

[6] H. Sarmento, F.M. Clemente, Afonso J., Araújo D., Fachada M., Nobre P., and Davids K. Match analysis in team ball sports: An umbrella review of systematic reviews and meta-analyses. *Sports Med Open*, 13, May 2022.

A International competitions for the considered team ball sports

Table 3 contains the major international competitions and the corresponding edition years selected for the team ball sports included in our paper. Tables 4–7 present the number of matches and teams per edition for each international competition selected for the team ball sports included in our paper.

B Source of web-scraped data

Table 8 includes links to the Wikipedia pages that were web-scraped to populate our match score dataset.

Sport	Competition	Edition Years
Basketball	Summer Olympic Games	1964, 1968, 1972, 1976, 1980, 1984, 2000, 2004, 2008, 2012, 2016, 2020
Cricket	ICC Men's Cricket World Cup	1975, 1979, 1983, 1987, 1992, 1996, 1999, 2003, 2007, 2011, 2015, 2019
Field Hockey	Men's FIH Hockey World Cup	1971, 1973, 1975, 1978, 1982, 1986, 1994, 1998, 2002, 2006, 2010, 2018, 2023
Futsal	FIFA Futsal World Cup	1989, 1992, 1996, 2000, 2004, 2008, 2012, 2016, 2020
Handball	Summer Olympic Games	1976, 1988, 1992, 1996, 2000, 2008, 2012, 2016, 2020
Ice Hockey	Winter Olympic Games	1924, 1928, 1932, 1948, 1952, 1956, 1960, 1964, 1972, 1976, 1980, 1984, 1988, 2002, 2006, 2010, 2014, 2018, 2022
Lacrosse	World Lacrosse Men's World Cup	1974, 1978, 1982, 1986, 1990, 1994, 1998, 2002, 2006, 2010, 2014
Roller Hockey	World Skate Roller Hockey World Cup	1999, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015
Rugby	Rugby World Cup	1987, 1991, 1995, 1999, 2003, 2007, 2011, 2015, 2019
Soccer	FIFA World Cup	1930, 1934, 1950, 1954, 1958, 1962, 1966, 1970, 1974, 1978, 1982, 1986, 1990, 1994, 1998, 2002, 2006, 2010, 2014
Volleyball	FIVB Volleyball Men's World Cup	1965, 1969, 1977, 1981, 1985, 1989, 1991, 1995, 1999, 2003, 2007, 2011, 2015, 2019
Water Polo	FINA Men's Water Polo World Cup	1979, 1981, 1983, 1985, 1993, 1995, 1999, 2002, 2010, 2014, 2018

Table 3: Major international competitions and corresponding edition years selected for the team ball sports included in our paper.

Basketball		
Edition	Matches	Teams
1964	56	16
1968	56	16
1972	57	17
1976	45	13
1980	61	14
1984	47	13
2000	42	12
2004	68	18
2008	38	12
2012	38	12
2016	38	12
2020	26	12

Cricket		
Edition	Matches	Teams
1975	15	8
1979	15	8
1983	27	8
1987	27	8
1992	39	9
1996	37	12
1999	42	12
2003	54	14
2007	48	16
2011	63	14
2015	49	14
2019	58	10

Field Hockey		
Edition	Matches	Teams
1971	30	10
1973	42	12
1975	42	12
1978	51	14
1982	42	12
1986	42	12
1994	42	12
1998	42	12
2002	72	16
2006	42	12
2010	38	12
2018	36	16
2023	44	16

Table 4: Tables presenting the number of matches and teams per edition for each competition selected for basketball, cricket, and field hockey.

Futsal		
Edition	Matches	Teams
1989	40	16
1992	40	16
1996	40	16
2000	40	16
2004	40	16
2008	56	20
2012	52	24
2016	52	24
2020	52	24

Handball		
Edition	Matches	Teams
1976	32	12
1988	36	12
1992	38	12
1996	38	12
2000	44	12
2008	42	12
2012	38	12
2016	38	12
2020	38	12

Ice Hockey		
Edition	Matches	Teams
1924	16	8
1928	16	11
1932	12	4
1948	36	9
1952	36	9
1956	36	13
1960	30	9
1964	28	8
1972	25	11
1976	27	12
1980	34	12
1984	34	12
1988	34	13
2002	27	14
2006	37	12
2010	26	12
2014	30	12
2018	30	12
2022	30	12

Table 5: Tables presenting the number of matches and teams per edition for each competition selected for futsal, handball, and ice hockey.

Lacrosse		
Edition	Matches	Teams
1974	6	4
1978	6	4
1982	6	4
1986	6	4
1990	10	5
1994	15	6
1998	28	11
2002	38	18
2006	72	22
2010	56	34
2014	63	40

Roller Hockey		
Edition	Matches	Teams
1999	36	12
2001	21	15
2003	24	16
2005	24	16
2007	24	16
2009	24	16
2011	24	16
2013	32	16
2015	48	16

Rugby		
Edition	Matches	Teams
1987	31	16
1991	32	16
1995	31	16
1999	41	20
2003	48	20
2007	8	8
2011	48	20
2015	48	20
2019	38	17

Table 6: Tables presenting the number of matches and teams per edition for each competition selected for lacrosse, roller hockey, and rugby.

Soccer		
Edition	Matches	Teams
1930	18	13
1934	17	16
1938	18	15
1950	22	13
1954	26	16
1958	35	16
1962	32	16
1966	32	16
1970	32	16
1974	38	16
1978	38	17
1982	52	24
1986	52	24
1990	52	24
1994	52	24
1998	64	32
2002	64	32
2006	64	33
2010	64	32
2014	80	32

Volleyball		
Edition	Matches	Teams
1965	35	11
1969	35	11
1977	42	12
1981	28	8
1985	28	8
1989	28	8
1991	48	12
1995	66	12
1999	66	12
2003	66	12
2007	66	12
2011	66	12
2015	66	12
2019	66	12

Water Polo		
Edition	Matches	Teams
1979	28	8
1981	28	8
1983	28	8
1985	28	8
1993	12	8
1995	16	8
1999	20	8
2002	20	8
2010	22	8
2014	22	8
2018	22	8

Table 7: Tables presenting the number of matches and teams per edition for each competition selected for soccer, volleyball, and water polo.

Sport	Link	[YEAR] Range
Basketball	https://en.wikipedia.org/wiki/Basketball_at_the_[YEAR]_Summer_Olympics	1964–1976
	https://en.wikipedia.org/wiki/Basketball_at_the_[YEAR]_Summer_Olympics_%E2%80%93_Men%27s_tournament	1980–2020
Cricket	https://en.wikipedia.org/wiki/[YEAR]_Cricket_World_Cup	1975–2019
Field hockey	https://en.wikipedia.org/wiki/[YEAR]_Men's_Hockey_World_Cup	1971–2023
Futsal	https://en.wikipedia.org/wiki/[YEAR]_FIFA_Futsal_World_Cup	1989–2020
Handball	https://en.wikipedia.org/wiki/Handball_at_the_[YEAR]_Summer_Olympics_%E2%80%93_Men%27s_tournament	1976–2020
Ice Hockey	https://en.wikipedia.org/wiki/Ice_hockey_at_the_[YEAR]_Winter_Olympics	1924–2022
Lacrosse	https://en.wikipedia.org/wiki/[YEAR]_World_Lacrosse_Championship	1974–2014
Roller Hockey	https://en.wikipedia.org/wiki/[YEAR]_Rink_Hockey_World_Championship	1999–2009
	https://en.wikipedia.org/wiki/[YEAR]_FIRS_Men%27s_Roller_Hockey_World_Cup	2011–2015
Rugby	https://en.wikipedia.org/wiki/[YEAR]_Rugby_World_Cup	1987–2019
Soccer	https://en.wikipedia.org/wiki/[YEAR]_FIFA_World_Cup	1930–2014
Volleyball	https://en.wikipedia.org/wiki/[YEAR]_FIVB_Volleyball_Men%27s_World_Cup	1965–2019
Water Polo	https://en.wikipedia.org/wiki/[YEAR]_FINA_Men%27s_Water_Polo_World_Cup	1979–2018

Table 8: Links to the web-scraped Wikipedia pages, with [YEAR] indicating the year of the edition of a competition (see the 'Editions' column in Table 3).