Datasheet: Panda babyhjk

Original authors/creators: Jeff Sackamnn
Organization: Heavy Topspin, The tennis abstract blog
Source: Github (https://github.com/JeffSackmann/tennis$_{a}tp)$

Kristian van Kuijk i6214757
Pavan Aakash Rangaraj i6214168
Boris Borisov i6211839

I. Motivation For Datasheet Creation

A. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

The datasheet was created for non-commercial purpose (publicly open dataset) and has become a go-to source for anyone interested in tennis statistics.

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

We do not have information about previous use of the dataset for research.

C. What (other) tasks could the dataset be used for?

The dataset could also be used to analyse the evolution of tennis from 1968-2022. As it contains information about the players from different eras, score sheets, game times, game statistics and several other factors during a game we could extract various patterns and trends in the game of tennis.

D. Who funded the creation dataset?

The tennis abstract blog community funded the dataset creation.

E. Any other comment?

Details about the dataset's author: https://www.tennis.com/news/articles/30-love-a-little-website-for-big-data

II. Datasheet Composition

A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

In our dataset, we have different csv files containing each year of tournament in tennis. There is also a text file with description of each feature.

B. How many instances are there in total (of each type, if appropriate)?

The features are mainly described in text format. Small subset of the features are described in integers. [1], [2]

C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

There are around 50 features in each csv file.

D. Is there a label or target associated with each instance? If so, please provide a description.

Each instance represents a professional tennis match.

E. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Information are missing for the detailed statistics (aces, first serve point won etc.) for instances that refer a tennis game fro m more than a decade ago. Hence, we will filter the dataset for our analysis.

F. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No.

G. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset references all professional tennis games.

H. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No.

I. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

No.

K. Any other comments?

No.

III. Collection Process

A. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

We do not entirely know the procedures involved in data collection, but we imaging there was a lot of manual work done to retrieve match info from the late 70's.

B. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

All the data was directly observable.

C. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

For our dataset we only look at the matches from 2020-2022.

D. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

No information.

E. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataframe expands over the timeframe of 1968-2022.

IV. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No data preprocessing has been done.

B. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

No.

C. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No.

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

Cannot be answered since there is no processing procedure.

E. Any other comments

None.

V. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The dataset is available in a Github Repo linked in the introduction.

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

Tennis databases, files, and algorithms by Jeff Sackmann / Tennis Abstract is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Based on a work at https://github.com/JeffSackmann.

In other words: Attribution is required. Non-commercial use only.

C. Are there any copyrights on the data?

No copyrights.

D. Are there any fees or access/export restrictions?

No

E. Any other comments?

None.

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

Jeff Sackmann is supporting the dataset.

B. Will the dataset be updated? If so, how often and by whom?

The dataset is updated weekly with the latest tennis matches.

C. How will updates be communicated? (e.g., mailing list, GitHub)

The updates are communicated via GitHub.

D. If the dataset becomes obsolete how will this be communicated?

This would then be communicated by Jeff Sackmann website.

E. Is there a repository to link to any/all papers/systems that use this dataset?

No.

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

It is not possible to extend yourself the dataset.

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

There is no information on previously conducted ethical review processes.

B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No, the dataset is about publicly telecasted games thus there is no confidentiality infringement

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

No, the dataset is predominantly tennis statistics.

D. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset is about games between two players where all their tennis stats including age, gender and nationality is provided.

E. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The dataset identify sub-populations using age and gender.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Yes, we can identify the tennis stars based on their name, age and gender.

G. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No, the dataset does not reveal anything that can be used for discrimination.

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data collection method is unknown.

I. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Unknown.

J. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Unknown.

K. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Unknown.

L. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown.

M. Any other comments?

None.

References

[1] Frank Mittelbach, Michel Gossens, Johannes Braams, David Carlisle, and Chris Rowley. *The LaTeX Companion*. Addison-Wesley Professional, 2 edition, 2004.

[2] Leslie Lamport. *LaTeX: a Document Preparation System*. Addison Wesley, Massachusetts, 2 edition, 1994.