Datasheet: *FIDE Chess Player Ratings (2010-2024)*

Original authors/creators: **FIDE (International Chess Federation)**

Organization: *FIDE (International Chess Federation)*

Source:

The dataset can be found directly on the FIDE website under their ratings section. Specifically, for the January 2024 data, it is available for download via this URL: https://ratings.fide.com/download.phtml?period=2024-01-01.

This datasheet was edited by: *Pierre Paul Charbonnier, Simon Garland and Filip Straka*

Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

1. **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

   The FIDE Chess Player Ratings dataset records the skills and performance levels of chess players worldwide, facilitating their ranking and matchmaking in official tournaments. It tracks players' progress and fills the gap of a standardized, globally recognized rating system, promoting a fair and competitive chess environment.

2. **Who created this dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)**?

   This dataset is created and maintained by FIDE, the international organization that connects national chess federations worldwide and governs international chess competition. FIDE's Ratings and Statistics department manages the dataset, compiling and updating player ratings based on their performance in sanctioned events.

3. **What support was needed to make this dataset?** (e.g. who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

   FIDE internally supports the dataset creation and maintenance with funding from membership fees, chess tournaments revenues, and sponsorships. It is an independent organization and does not rely on external grants. FIDE may collaborate with partners and institutions for additional resources.

4. **Any other comments?**

   Your Answer Here

Composition

Dataset creators should read through the questions in this section prior to any data collection and then provide answers once collection is complete. Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

1. **What do the instances that comprise the dataset represent (e.g. documents, photos, people, countries)?** Are there multiple types of instances (e.g. movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

   Rated chess games played and rated chess players

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

   There is 1390912 instances in the dataset

3. **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 contains all possible instances

4. **What data does each instance consist of?** "Raw" data (e.g. unprocessed text or images) or features? In either case, please provide a description.

   Various features such as ID number, if it's a player, their name, sex, country, birthday, any professional titles they hold and their rating. For games, it contains how long the game lasted, how many mistakes were made, the participants of the game, the winner of the game and other such details.

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

   The instances are labelled with their International Chess Federation (FIDE) ID numbers.

6. **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.

   Some ratings and games are missing and all subsequent values associated with them are missing as a result due to them being unavailable to the FIDE for an unspecified reason.

7. **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.

8. **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.

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

   No.

10. **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.

    The dataset is self-contained

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

    No.

12. **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.

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

    Yes.

14. **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.

    No.

15. **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.

16. **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.

17. **Any other comments?**

    Your Answer Here

Collection

As with the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals of the prior section, the answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

1. **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.

   The data was directly observed from online chess games. Each game instance includes visible moves, player titles, ratings, and other relevant metadata directly observable from the game interface. This data reflects real-time interactions and decisions made by the players during each game.

2. **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?

   A software program leveraging an internal API was used to collect the data. This program automatically extracted game data, including moves, player ratings, and titles, directly from the online platform where these games were played. The software's accuracy and reliability were validated through a series of tests comparing collected data samples with known game outcomes and player details.

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

   A mix - we want to represent each skill category but within that category - we take a random sample for 4 different age categories (this study will be limited to younger people since older people rarely play online chess but within these categories we assume there will still be notable difference)

4. **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)?**

   Nobody - it was mapped by an internal API

5. **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. Finally, list when the dataset was first published.

   Over the past years and months - the data was collected on a monthly basis.

6. **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.

   No. however - every player (of which the data has been collected) by registering on the site that the data was collected from has consented to the data collection.

7. **Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.

   Yes, the dataset relates to individuals who play chess online. It includes data that could potentially identify players based on their game activity, ratings, and titles.

8. **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 was collected directly from the games played on the online chess platform, using an internal API designed to map and extract game data. So the data is technically obtained via a third-party.

9. **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.

   No - however - the data is open source and there was no obligation to cite these sources, or to otherwise notify the subjects

10. **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.

    No. no consent was needed since this is open source data without the necessity of a citation.

11. **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).

    No

12. **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.

    No. this analysis doesn't directly affect the subjects since the subjects were not aware of this study at the time of providing the data. Moreover - this analysis is not to be made public. Since the test subjects agreed to have this data open-source we assume that the impact analysis was done by the team that published the data and made it open-source in the first place.

13. **Any other comments?**

    NO.

Preprocessing / Cleaning / Labeling

Dataset creators should read through these questions prior to any pre-processing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

1. **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.

   The FIDE Chess Player Ratings dataset undergoes various preprocessing and cleaning steps before it is published. These steps include the removal of inactive players who have not participated in FIDE-rated events for a certain period, updating player ratings based on the latest tournament results, processing missing values in player information such as country or club affiliation, consistency checks to ensure correct alignment of player names, IDs, and ratings, and labeling new instances with unique identifiers and initial ratings based on their performance. However - there are still inaccuracies and missing values present in the datasets.

2. **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.

   While we can't know what data FIDE retains as confidential to both ensure transparency and privacy at the same time, the data made available publicly seems to not have been edited - or very lightly so.

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

   The specific software tools used for preprocessing, cleaning, and updating the FIDE Chess Player Ratings dataset are proprietary to FIDE and its data management team. While the tools themselves are not publicly available, FIDE provides comprehensive documentation on the methodologies and algorithms used for rating calculations and data updates. This documentation can be found on the FIDE website.

4. **Any other comments?**

   No.

Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

1. **Has the dataset been used for any tasks already?** If so, please provide a description.

   The dataset has not been used for any tasks yet to the best of our knowledge. However - since it is open-source data, we can't assume it has not been used.

2. **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

No.

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

   This can serve various purposes. It can be used to analyze the correlation between age and chess performance, study the progression of younger players' skills, etc. The latter (obtained from Lichess API) dataset might be used for various other tasks – machine learning, discovering correlations between different attributes, strategy development, etc.

4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g. stereotyping, quality of service issues) or other undesirable harms (e.g. financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

   Future users should be aware that the dataset exclusively focuses on younger players and may not represent the full spectrum of chess play across all age groups. This limitation could impact studies aiming to generalize findings across the entire population of chess players. To mitigate potential biases, users should consider supplementing this dataset with additional data covering other age groups or explicitly noting the dataset's scope and limitations in their analysis.

5. **Are there tasks for which the dataset should not be used?** If so, please provide a description.

   When using the dataset, avoid harmful purposes like creating stereotypes or invasive analysis. It involves minors, so respect their privacy and avoid revealing personal information.

6. **Any other comments?**

   No.

Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

1. **Will the dataset be distributed to third parties outside of the entity (e.g. company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

   The dataset is made available for public use and viewing.

2. **How will the dataset will be distributed (e.g. tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

   The dataset has been uploaded on the FIDE's official website

3. **When will the dataset be distributed?**

   The dataset has been publicly available since 2001

4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

   No.

5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

   No.

6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

   No.

7. **Any other comments?**

   No.

Maintenance

As with the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

1. **Who is supporting/hosting/maintaining the dataset?**

   The dataset is supported, hosted, and maintained by FIDE, the international governing body for the game of chess. Maintenance tasks include regular updates, quality checks, and responding to queries or concerns related to the dataset.

2. **How can the owner/curator/manager of the dataset be contacted (e.g. email address)?**

   The dataset manager can be contacted through the official FIDE email for ratings inquiries: ratings@fide.com. This contact is available for questions, suggestions, or reporting errors in the dataset.

3. **Is there an erratum?** If so, please provide a link or other access point.

   Yes, any corrections or updates to the dataset are documented in an erratum section on the FIDE ratings website. Users can access this information at https://ratings.fide.com/errata.phtml for the most up-to-date corrections and clarifications.

4. **Will the dataset be updated (e.g. to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g. mailing list, GitHub)?

   The dataset is updated monthly by FIDE to reflect new game results, player entries, and rating adjustments. Any significant changes, corrections, or updates are communicated through the FIDE website and, for registered users, via email *notifications. Users can subscribe to updates through the FIDE newsletter or by following FIDE on social media platforms.*

5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g. were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

   As the dataset consists of publicly available player ratings and is a record of players' performance in chess competitions, there are no fixed limits on the retention of the data. Players are informed upon registration with FIDE that their game results and ratings will be publicly recorded and maintained as part of their competitive history.

6. **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

   Older versions of the dataset are archived and accessible for historical research and analysis. FIDE maintains a comprehensive archive of past ratings, available on the FIDE website. Users are informed of any major changes to data accessibility or format through the FIDE website and direct communications to registered users.

7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

   FIDE supports scholarly research using its datasets, but doesn't have a formal process for external contributions. Researchers can contact FIDE to publish analyses based on the dataset, which will be verified by FIDE's data team for accuracy and consistency with existing data standards.

8. **Any other comments?**

   No.