Datasheet for Body Fat Dataset*

Predictive Modeling of Body Fat Percentage Using Body Measurements

Wendy Yuan

November 29, 2024

This datasheet describes the Body Fat Dataset from Brigham Young University. It includes measurements like weight, height, and body circumferences from 250 men. The data helps predict body fat percentage using simple methods instead of expensive equipment. It is publicly available but only represents men and lacks diversity.

Motivation

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.*

   - This dataset was created to analyze body composition and predict body fat percentage using simpler, more accessible measurements. Accurate %body fat measurement can be complex and costly, so this dataset helps in creating prediction models using other physical measurements.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

   - Developed by Brigham Young University's (BYU) Human Performance Research Center under the direction of Mark Ricard.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

   - BYU Human Performance Research Center, details available at: https://www.byu.edu/chhp/intro.html#lrc

4. *Any other comments?*

   - The dataset is relevant for health, fitness, and statistical research.

---

*Code and data are available at: https://github.com/kiwindyy/Body-Fat

Composition

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

   - Each instance represents measurements of a male participant, including body density, body fat percentage, weight, height, and various circumferences (e.g., neck, chest, abdomen).

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

   - 250 male participants.

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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset is a controlled sample of men of various ages. It is not randomized or geographically diverse.

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

   - Raw numerical data, including weight (in pounds), height (in inches), and circumferences (in inches).
   - Body fat percentage, a value derived from body density.

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

   - Percentage Body Fat (%BF), accurately measured using specialized equipment.

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

   - No missing data is evident in the dataset.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - Variables like weight, abdomen circumference, and height have measurable connections with %body fat.

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

   - The dataset does not come with predefined splits, leaving users the flexibility to create their own training, validation, and test sets.

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

   - Minor variations in measurements may exist but are not documented.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, 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 (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? 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 and can be accessed directly via DASL (Ricard n.d.)

11. *Does the dataset contain data that might be considered confidential (for example, 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.*

    - The dataset is anonymized, containing no personally identifiable information.

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.*

    - None.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - TBD

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

    - The dataset represents males, with no explicit data on race, ethnicity, or geographic distribution.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race 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.

- Not possible.

16. *Any other comments?*

- Contains health-related data (e.g., %body fat) but no personal or private data.

Collection process

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

- Measurements were collected through standard anthropometric techniques, supplemented with accurate %body fat calculations based on body density.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

- Likely involved common health assessment tools and equipment.

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

- A non-random, controlled sample of male participants from BYU's study pool.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

- Researchers at BYU Human Performance Research Center.

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

- Unspecified but likely during a single study period.

6. *Were any ethical review processes conducted (for example, 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.*

- Ethical review details are not provided.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - Data was collected directly from study participants.

8. *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.*

   - Likely obtained as part of the study; details not included.

9. *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.*

   - Not explicitly mentioned.

10. *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).*

    - Not explicitly mentioned.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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.*

    - None conducted or documented.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, 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 remaining questions in this section.*

   - No significant preprocessing; raw data is presented for analysis.

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

   - Provided in its entirety.

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

   - DataDesk integration for visualization and analysis.

Uses

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

   - Statistical modeling of body fat percentage.
   - Regression modeling and variable selection analysis.

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.*

   - https://github.com/kiwindyy/Body-Fat

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

   - Machine learning for health and fitness prediction.
   - Biostatistical research and educational tools.

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 dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

   - Dataset is limited to male participants.
   - Results cannot be generalized to the entire population without validation.

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

   - Misuse for discriminatory purposes. Use in contexts that promote harmful body image stereotypes.

Distribution

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

   - Available via DASL (Ricard n.d.) and DataDesk integration.

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

- Downloadable from DASL in .txt format.

3. *When will the dataset be distributed?*

   - It is a public dataset

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 explicit restrictions noted.

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.*

   - None mentioned.

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.*

   - Not applicable.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - Maintained by DASL (Ricard n.d.) and BYU Human Performance Research Center.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   - Director: Mark Ricard
   - Contact: (801) 378-8958

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

   - None documented.

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

   - Dataset is static, with no planned updates.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the 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.*

- Retained indefinitely for public use.

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 dataset consumers.*

   - Only the current version is available.

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 dataset consumers? If so, please provide a description.*

   - Contributions or augmentation not supported.

References

Ricard, Mark. n.d. "Bodyfat." Brigham Young University; BYU Human Performance Research Center. https://dasl.datadescription.com/datafile/bodyfat/?_sf_s=body+fat&_sfm_cases=4+59943.