Datasheet for Police Race and Identity Based Data - Arrests and Strip Searches for the city of Toronto*

Akshat Aneja

December 3, 2024

Extract of the questions from Gebru et al. (2021).

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.*
    - The dataset was created to investigate arrest patterns in Toronto, focusing on racial representation and potential disparities.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was likely compiled by researchers or academic institutions studying criminal justice or racial equity, potentially in collaboration with public entities.
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.*
    - The funding source is not explicitly stated.
4. *Any other comments?*
    - None

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.*

*Code and data are available at: https://github.com/Akshat211202/bias-in-arrests-toronto

- Each instance represents an individual arrest event, with associated demographic, geographic, and legal attributes.

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

   - The dataset contains thousands of arrest records (specific count to be derived from the full document).

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).*

   - It is unclear if the dataset includes all arrests or represents a sample.

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.*

   - Each instance includes details such as race, gender, arrest location, type of offense, and socioeconomic indicators of the area.

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

   - Yes, labels may include arrest outcomes, offense categories, or demographic classifications.

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.*

   - Some records may have missing or incomplete demographic or arrest details.

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.*

   - Relationships such as correlations between demographic variables and arrest outcomes may be implied but not explicitly documented.

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.*

   - No specific splits are provided, but the dataset could be partitioned by time, geography, or demographics.

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

    - The dataset may contain biases or inconsistencies due to data collection methods.

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 but may reference external demographic or geographic datasets for enrichment.

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.*

    - No personally identifiable information is included; all data appears to be anonymized.

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, though the results could highlight sensitive systemic issues.

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.*

    - Yes, it identifies sub-populations by race, gender, and location.

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.*

    - No, all data appears to be anonymized.

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.*

    - Yes, data includes variables like race, gender, and arrest location.

16. *Any other comments?*

    - None

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.*

    - Data was likely acquired from public law enforcement records and demographic databases.

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?*

    - Data was collected through administrative and observational means, potentially including manual and automated processes.

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

    - Unclear; it is not specified if this is a sample or comprehensive dataset.

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)?*

    - Likely government or research personnel; details on compensation are unavailable.

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.*

    - Data covers multiple years, though the specific timeframe is not stated.

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.*

    - Not explicitly mentioned.

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 obtained indirectly from public records and administrative datasets.

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.*

    - No evidence suggests individuals were notified.

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.*

    - No evidence of explicit consent.

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).*

    - No evidence suggests a consent mechanism was in place.

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.*

    - Not mentioned in the document.

12. *Any other comments?*

    - None

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.*

    - Details about preprocessing or cleaning are not explicitly provided.

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.*

    - Not specified.

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

- Not mentioned.

4. *Any other comments?*
    - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Yes, for analyzing arrest patterns and racial disparities.
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.*
    - Not specified.
3. *What (other) tasks could the dataset be used for?*
    - Policy evaluation, geographic analyses, and socioeconomic impact studies.
4. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - Should not be used for purposes that perpetuate biases or stigmatize communities.
5. *Any other comments?*
    - None.

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.*
    - Likely available through research publications or institutional platforms.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - Via academic platforms or government repositories.
3. *When will the dataset be distributed?*
    - Distribution likely coincides with the research publication timeline.
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.*

- Not explicitly stated, but typically subject to institutional guidelines.

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.*

   - Not 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 mentioned.

7. *Any other comments?*

   - None.

Maintenance

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

   - Likely the authors or affiliated academic institutions.

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

   - Contact information likely provided in the corresponding research paper.

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

   - Not mentioned.

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)?*

   - Updates may occur, but no schedule is specified.

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.*

   - Not mentioned.

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.*

- Not mentioned.

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.*

    - Not mentioned.

8. *Any other comments?*

    - None.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.