Datasheets for datasets – Project 2 deliverable

Srivasa Pranathy Mutcherla - A20555875

Kiran Velamati - A20525555

- **Why you selected the dataset –** We wanted to choose a dataset that when analyzed would also have societal benefits. Chicago data portal proved to be the best place to do so, and hence we selected a dataset that covered the educational performance of 500+ public schools in Chicago for the year 2011-2012
- **How did you acquire the dataset –** We utilized the Chicago data portal to determine the dataset, and it's inbuilt offering to export the dataset to CSV

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**.
It is primarily designed to provide detailed insights into the academic performance and school environment across different schools within the CPS system. The dataset includes comprehensive details like school demographics, student academic achievement, school culture and climate, graduation rates, faculty information, and parental involvement, which all contribute to the overarching goal of enhancing educational quality and outcomes in Chicago Public Schools.

2. **Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?**
The dataset on Chicago Public Schools (CPS) Progress Report Cards for 2011-2012 is created and maintained by the Chicago Public Schools system itself. CPS is a large public school district that oversees all public schools in Chicago, Illinois, USA. This dataset is typically compiled by the Office of Accountability within the Chicago Public Schools system, or a similar department responsible for monitoring and evaluating school performance metrics. The creation of such datasets is part of the CPS's ongoing efforts to evaluate educational progress and ensure transparency in reporting school achievements and challenges to the public, parents, students, and other stakeholders. The dataset is made available to the public through the City of Chicago's data portal, which is an initiative by the city to promote open data and transparency for various city departments and services, including education.

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 dataset on the Chicago Public Schools (CPS) Progress Report Cards is funded by the operational budget of the Chicago Public Schools system, which includes state, federal, and local funding sources. There isn't a specific grant associated with the creation of this dataset; it is part of the routine administrative activities of CPS to ensure transparency and accountability in education quality and performance.

4. **Any other comments? Composition. Dataset creators should read through these questions prior to any data collection and then provide answers once data collection is complete. Most of the questions here are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions. Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.**
This dataset was not written by people, it is more of a statistical representation of the performance of public schools in Chicago.

5. **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.**
The instances in the Chicago Public Schools (CPS) Progress Report Cards dataset for 2011-2012 represent individual schools. Each instance is a detailed report card providing data on various aspects of school performance and characteristics, including academic performance metrics, student demographics, school culture and climate, staff information, and parental involvement. There are no multiple types of instances; each instance focuses solely on one school.

6. **How many instances are there in total (of each type, if appropriate)?**
566 schools are covered.

7. **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 Chicago Public Schools (CPS) Progress Report Cards dataset for 2011-2012 typically includes data on all public schools within the CPS system, not just a sample. This means each public school that was operational during the 2011-2012 academic year should have a corresponding entry in the dataset. It encompasses various types of schools, including elementary, middle, and high schools across Chicago.

The representativeness of the dataset in terms of providing a complete picture of the school system is generally validated by the CPS itself, which is responsible for maintaining accurate and up-to-date records of all its schools.

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

The data set consists of structured data with 79 columns. Each cell contains either text, number or location datatypes. A few fields are missing in some of the columns i.e.; blank spaces.

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

In the context of the Chicago Public Schools (CPS) Progress Report Cards dataset for 2011-2012, there isn't typically a "label" or "target" in the traditional sense used in predictive modeling or supervised machine learning. Instead, each instance (school) in the dataset includes various metrics and indicators that can be analyzed or assessed on their own merits. These metrics could be discretized to help with a target for each instance.

10. **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.**

Few metrics are not assessed because of limitations at that time considering data collection issues or it may not be deemed necessary to collect from that school etc. This is addressed by placing NDA to such fields. And the other columns which derived from these NDA filled cells are left blank.

11. **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.**

In the Chicago Public Schools (CPS) Progress Report Cards dataset for 2011-2012, the relationships between individual instances (schools) are generally not made explicit. Each instance in the dataset typically stands alone, focusing on providing a comprehensive report on the performance and characteristics of a specific school.

12. 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.

In the Chicago Public Schools (CPS) Progress Report Cards dataset for 2011-2012, there typically aren't predefined recommended data splits like training, validation, and testing sets, which are common in machine learning projects. This is primarily because the dataset is not inherently designed for predictive modeling tasks but rather for assessment, reporting, and analysis of school performance.

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

There are some columns that predominantly contain NDA values (No data available) that prove to make analysis more difficult. But considering the breadth of the data that is available, we can still utilize more information-rich columns to generate network models and insights.

14. 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 primarily self-contained. There is one column that consists of links to the respective schools, and the guarantee that they remain working is based on the team's rate of updating the dataset. The Chicago data portal allows access to the latest version of the dataset to ensure that utmost workability is achieved. There are no restrictions available on the links that have been provided, as they are meant for open, public access.

15. 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.

This dataset does not contain any PPIDs (Publisher Provided Identifiers) or Personal Identifiers (PIDs) as it covers school level performance metrics that cannot be tied to any particular individual (student, parent or faculty).

16. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. If the dataset does not relate to people, you may skip the remaining questions in this section.

As this dataset **does not** tie back any individual, it would not be offensive, insulting, threatening or anxiety-inducing.

17. Does the dataset identify any subpopulations (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.

While performance metrics are provided by age, they **do not** lead to identification for any subpopulations.

18. 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.

It is not possible to identify individuals directly.

19. 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.

The dataset does not contain data that might be considered sensitive in any way.

20. Any other comments? Collection process. As with the questions in 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 outlined earlier, the following questions are designed to elicit information that may help researchers and practitioners to create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

There are no personal identifier concerns.

21. 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.

The data is directly observable – in the sense that these metrics were collected by each of the schools, and this was cumulatively presented. This dataset is typically compiled by the Office of Accountability within the Chicago Public Schools system or a similar department responsible for monitoring and evaluating school performance metrics and is presumably verified by them.

22. 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?

While there could be software programs that could help with collecting on a school level, this cumulative dataset could have either been compiled similarly using a software tool or manually curated by humans.

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

Since this dataset is not a sample, the sampling strategy does not come into the picture.

24. 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)?

Since it is a governmental effort, the team working on this could have been compensated for their efforts via state or federal funds.

25. 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.

The dataset focuses on the performance of Chicago public schools from 2011-2012 and was initially published then. It was last updated in 2018.

26. 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. If the dataset does not relate to people, you may skip the remaining questions in this section.

This does not relate to people, but nevertheless, this dataset is owned by Chicago Public Schools who verify the dataset before publication.

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

The data was exported from the Chicago Data Portal, which hosts the original dataset posted by the individuals in question.

28. 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.

As this was collected by the schools about them, there is no data collection problem.

29. 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.

As a team effort, it was a consensual decision.

30. 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).

N/A

31. 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.

Currently, there are no licensing values associated with the dataset, and as an open-source data portal, this is meant to be publicly available to everyone.

32. Any other comments? Preprocessing/cleaning/labeling. Dataset creators should read through these questions prior to any preprocessing, 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-ofwords" is not suitable for tasks involving word order.
The data set given in the portal is straightforward, they defined all the datatypes accurately. No other comments.

33. 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.
Since there were a lot of columns, we had to evaluate the parameters that would help in our data model and discretize those parameters to have a clean dataset. We also took care of missing (NDA – No Data Available) values in this dataset.

34. 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.
Raw data can be always retrieved from the Chicago Data portal. Here is the link additionally press here

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

Microsoft Excel.

36. Any other comments? Uses. The following 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.
No other comments for this.

37. Has the dataset been used for any tasks already? If so, please provide a description.
We have not used it in any other tasks apart from this academic project.

38. 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.
Everything is on this link here.

39. What (other) tasks could the dataset be used for?
This dataset shows all school level performance data used to create CPS School Report Cards for the 2011-2012 school year. It could also be used to analyze the potential reasons for the performance levels of schools and find areas of improvement the school might have.

40. 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?
As there are no risks associated with identifying individuals or communities from the dataset, there is no immediate harm associated with the same.

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

There is nothing described as such by the curator. However, unethical practices using this data are discouraged.
42. Any other comments? 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.
NO other comments.

43. 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.
It is a publicly available dataset, so any individual or an organization can use it.

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

The dataset doesn’t have a DOI, and it is available on the Chicago Data Portal website.

45. When will the dataset be distributed?

It was first published in November 2011 and was updated until September 2018.

46. 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.

The license for this dataset is unspecified.

47. 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.

There are no restrictions imposed on the dataset.

48. 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.

As this is a publicly available dataset, there are no such regulatory restrictions applicable.

49. Any other comments? Maintenance. As with the previous questions, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

No other comments.

50. Who will be supporting/hosting/maintaining the dataset?

The Chicago Public Schools remain owners of the dataset, whereas the Chicago Data Portal will be hosting the dataset.

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

http://www.cps.edu/

Can be contacted through here https://data.cityofchicago.org/nominate

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

Not applicable for this dataset.

53. 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)?

The enlisted plan says that the update frequency will be yearly. It will be managed and maintained by the above-listed team.

54. 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.

It relates indirectly to people, but it doesn't require gathering data about individuals.

55. 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.

The latest version is only accessible through the portal. There is no older version available to the public.

56. 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.

No. This dataset is maintained by the Chicago data portal, they are the only ones to have access to modify or build on this dataset.

57. Any other comments?

No other comments.