Datasheet for 'Toronto Outbreak Trends'*

Sindhu Priya Mallavarapu

April 18, 2024

This paper analyzes the disease outbreak patterns across the city of Toronto in the years between 2020 - 2024 based on location and models future outbreak trends. It is found that the highest number of disease outbreaks within these five years happened Long-Term Care Facilities. The results of this paper can help government officials see where more healthcare funding and stricter safety practices are needed.

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

Motivation

1. *For what purpose was the dataset created?*
    - The dataset was created to analyse disease outbreak trends in Toronto given the type of healthcare facility and year.
2. *What does this study want to help with/achieve?*
    - The study aimed to understand where medical funding and additional healthcare workers are needed.

Composition

1. *What datasets are included?*
    - The raw dataset includes data from Open Data Toronto (Gelfand 2022) package in R (R Core Team 2022).

Data Collection

1. *Where is the data collected from?*

*Code and data are available at: https://github.com/MSindhuPriya/toronto_outbreaks

- The text of "Pride and Prejudice" was retrieved using the `opendatatoronto` package (Gelfand 2022) in R (R Core Team 2022), which provides access to Open Data Toronto datasets.

2. *How much of the data is used?*

   - Out of the given data just outbreak setting and the year of the start of the outbreak was used.

Data Cleaning and Analysis

1. *How was the raw data altered and changed?*

   - The raw data variable names were cleaned. Additional variables that counted the number of outbreaks to happen each year at each setting were also created.

2. *How was the cleaned data analysed?*

   - The frequency the outbreaks based on year and setting were counted. Then we analyzed trends and created a model to predict the frequency of outbreaks.

Conclusion

1. *What did the study conclude?*

   - The study concluded that the highest number of outbreaks took place in Long-Term Care Facilities regardless of year. Additionally the highest number of outbreaks happened in the year 2022 regardless of setting. It is interesting to not that Transitional Care facilities had the least amount of outbreaks regardless of year.

Applications

1. *How can I use this dataset?*

   - The dataset can be used to analyse the outbreak trends in Toronto with respect to type of healthcare facility and year.

2. *How is this useful in terms of application?*

   - This information can be used to understand where funding needs to be given and where we needed scricter health and safety policies. This can help us aid our medical system so that it can run more smoothly.

Distribution (Reproducibility)

1. *Can I access the dataset used in this study and/or use it to reproduce?*

   - Yes, the dataset is available for download at https://github.com/MSindhuPriya/toronto_outbreaks. Researchers interested in using the dataset for their own analyses can access it too.

Maintenance

Note - Any updates or corrections to the dataset will be documented and made available to users if necessary. The creators of this paper can be contacted for further information at sindhupriya.mallavarapu@mail.utoronto.ca.

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.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.