

https://dspace.org/

Record linkage to de-duplicate sex-worker registers at the sex-work clinics in Zimbabwe using supervised learning techniques

Musemburi, Sithembile

2021
https://hdl.handle.net/10539/33606

Downloaded from a DSpace repository. https://dspace.org/

A B S T R A C T

Background

The Centre for Sexual Health, HIV and AIDs Research Zimbabwe implements the national sex work program in Zimbabwe on behalf of the Ministry of Health and Child Care and the National AIDs Council since 2009. The program offers free clinical services to sex workers through the 'Sisters with a voice' clinics. Sex workers are registered into the link log register and demographic information is collected at first visit. No identification is required to confirm the identity of the sex worker during registration. Due to stigma and criminalisation of sex work, sex workers sometimes use pseudo names at registration which they are likely to forget on their next visit. This has led to duplication as clients are assigned a sisters' number in chronological order but there is no unique identifier to uniquely identify sex workers as they are registering.

Aim

This study aimed to apply Naïve Bayes Classifier and Support Vector Machine as supervised machine learning approaches to match and de-duplicate records in the link log and demographic data sets for female sex workers registered at Sisters' clinics in Zimbabwe. This information is key in enumerating the number of unique individual sex workers that were registered at the clinic between 2017 and 2019. This would help the program to come up with improved monitoring, interventions, strengthen key health priorities, and inform policy and practice. Furthermore, the study also aimed to ascertain the accuracy rate of repeat visits by sex workers to the Sisters' clinics and develop an optimum framework using improved supervised machine learning techniques as alternatives to conventional probabilistic record linkage techniques.

Methods

The study applied the Python record linkage toolkit to pre-process, index and link the demographic and link log data sets. The study used 85% of the data for training the algorithms and 15% of the data for testing and validation. Support Vector Machine and Naïve Bayes Classifier algorithms were applied on the linked dataset and results of the matching were compared in terms of scalability, accuracy and F1 score. Performance evaluation and validation was done to measure the Precision, Recall, Accuracy and F1 score of the algorithms.

Results

The study results showed that Support Vector Machine performs better than Naïve Bayes Classifier in record linkage. This study managed to de-duplicate the data and ascertain rate of repeat visits, from the 40 507 sex workers who were registered in the demographic data between 2017 and 2019. Furthermore this study showed an 8% duplicate rate in the records suggesting that 8% of the clients had been to the clinic before. The Support Vector Machine and Naïve Bayes Classifier algorithms were fit on the test data and Support Vector Machine outperformed Naïve Bayes Classifier with a Precision of 95,5%, Recall of 1 and Accuracy of 99,9% and an F1 score of 0.9778 respectively.

Conclusion

The study results showed that Support Vector Machine performs better than Naïve Bayes Classifier in record linkage. The Precision and Accuracy for Support Vector Machine and Naïve Bayes Classifier was above 90%. The Support Vector Machine model could be implemented on the sex work program data routinely to de-duplicate data and get accurate statistics for reporting and contributing to the size estimation of sex workers. The proportion of male and transgender sex workers in the program data can be a proxy to estimating the number of male sex workers and transgender sex workers in Zimbabwe as there is a gap in program data for male sex workers and transgender sex workers.

Although there was a limitation in sample size as records of clients registered before 2017 had missing data in the variables of interest so they could not be included in this analysis, the results show that there is a gap in programming for male sex workers and transgender sex workers. There is need to intensify efforts and capacitate community-based organisations to come up with intervention programmes to increase the uptake of health services by male and transgender sex workers so that the country can achieve its 95-95-95 goal of having 95% people tested for HIV and knowing their status, 95% of the HIV positives on treatment and 95% of those on treatment being virally suppressed.