

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/

RECORD LINKAGE TO DE-DUPLICATE SEX-WORKER REGISTERS AT THE SEX-WORK CLINICS IN ZIMBABWE USING SUPERVISED LEARNING TECHNIQUES



SCHOOL OF PUBLIC HEALTH

FACULTY OF HEALTH SCIENCES

Sithembile Musemburi

Student number: 1179538

A Research Report submitted to the Faculty of Health Sciences in partial fulfilment of the requirements for the degree of Master of Science (MSc) Epidemiology – Public Health Informatics

August 2021

Sithembile Musemburi, August 2021

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

Copyright © University of the Witwatersrand, Johannesburg, South Africa.

All rights reserved. No part of this research report may be stored in a retrieval system, transmitted, or reproduced, in any form or by any means, including but not limited to photocopy, photograph, magnetic or other record, without prior agreement and written permission of the copyright holder.

Supervisors:

Mr Michael Mapundu, Lecturer, University of Witwatersrand
Dr Fortunate Machingura, CeSHHAR Zimbabwe

D E C L A R AT I O N

I hereby declare that this research report is project work carried out by me under the guidance of Mr Michael Mapundu and Dr Fortunate Machingura. I have taken care in all respects to honour the intellectual property rights of others and have acknowledged the contribution of others. I further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university



Parktown, Johannesburg, August 2021 _____

DEDICATION

I dedicate this to my late mother Patricia Zinyemba who gave me a chance to life against all odds, I know she would have been proud of the woman I have become. To my grandmother Ketty Zinyemba, thank you for supporting your teenage daughter and taking over the role of a mother since the day I was born, I will forever be indebted to you.

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.

A C K N OW L E D G M E N T S

I would like to express my gratitude to Mr Michael Mapundu and Dr. Fortunate Machingura for their guidance, supervision and support in working on the project. To Professor Frances Cowan, thank you for creating an environment that promotes learning and for the support you gave me during my studies. I would also like to express my gratitude to CeSHHAR - for providing financial support that enabled me to undergo this master's programme at University of the Witwatersrand, Johannesburg, South Africa and for allowing me to use the sex work program data for my research. My sincere gratitude goes to my "international friend" and sister Edwina Mabuela for opening your house to a stranger and making me feel at home during the school blocks. Last but not least, I would like to thank my husband Wellington Musemburi and my 3 daughters for their support and patience throughout the study period without whom this masters would not have been possible… it was not easy but we made it.

To God be the glory

TABLE OF CONTENTS

D E C L A R AT I O N ..... **iii**

A B S T R A C T ..... **v**

A C K N OW L E D G M E N T S ..... **vii**

LIST OF FIGURES ..... **x**

LIST OF TABLES ..... **xi**

ACRONYMS ..... **xii**

1. INTRODUCTION **1**

1.1 Background to the study 1
1.2 Statement of the Problem 3
1.3 Justification 3
1.4 Motivations 4
1.5 Research Aim 4
1.6 Research Objectives 4
1.7 Outline of the Research Report 5

2. LITERATURE REVIEW **6**

2.1 Introduction 6
2.2 Sex work in Zimbabwe 6
2.3 Record linkage 6
2.4 Gaps in Literature 9

3. METHODOLOGY **10**

3.1 Introduction 10
3.2 Record Linkage Process 10
3.2.1 Data Pre-processing 11
3.2.2 Data cleaning 12
3.2.3 Indexing 13
3.2.4 Classification 14
3.3 Study population and data sources 15
3.4 Sampling 16
3.5 Overview of Record linkage 16
3.6 Support Vector Machines 18
3.7 Naive Bayes Classifier 19
3.8 Evaluation and validation of the classification 20
3.9 Ethical considerations 21

4. RESULTS **22**

4.1 Introduction 22
4.2 Descriptive statistics 22
4.3 Results of the record linkage process 27
4.4 Evaluation of classification results 30

5. DISCUSSION **34**

5.1 Introduction 34
5.2 Descriptive statistics of the data sets 34
5.3 Record linkage 37
5.4 Limitations of the study 38

6. Conclusion and Recommendations **39**

7. REFERENCES **41**

Appendices **44**

Appendix A Plagiarism declaration 44
Appendix B MRCZ Ethics Clearance 45
Appendix C List of variables 46
Appendix D Ethical clearance certificate -Wits 47
Appendix E Approval letter to access CeSHHAR data 48
Appendix F Originality report 49

LIST OF FIGURES

Figure 1.1 Map showing the coverage of the Sisters clinics in Zimbabwe : CeSHHAR (2017) 1
Figure 3.1 Overall Architecture of the Record Linkage Model. Source: (Feng, 2016).....10
Figure 3.2 Work flow diagram..... 110
Figure 3.4 Classification of data by support vector machine (SVM). Source: (García-Gonzalo et.al , 2016) 19
Figure 3.5 Classification of data Naïve Bayes Classifier (NBC): Source (Yang, 2019) 20
Figure 3.3 Confusion matrix 21
Figure 4.1 Distribution of the gender of clients registers at the Sister's clinics 22
Figure 4.2 The distribution of age at first visit for clients registered at the Sisters clinics.....23
Figure 4.3 A comparison of age distribution at first visit for client registered between 2009-2015 and 2016-2019 23
Figure 4.4 Marital status of sex workers at registration.....24
Figure 4.5 Number of clients registered at the Sisters clinics by site. 24
Figure 4.6 Level of education at registration for sex workers registered at the sisters' clinics 27
Figure 4.7 Duplicate records by site 28
Figure 4.8 Duplicate records by age at first visit 29
Figure 4.9 Level of education at registration for duplicate records 30
Figure 4.10 SVM Confusion matrix 31
Figure 4.11 NBC Confusion matrix.....31

LIST OF TABLES

Table 4.1 Table showing number of clients registered by site 26
Table 4.2 Table showing sample matched pairs during record linkage.....28
Table 4.3 Table showing distribution of duplicates by site 29
Table 4.4 Table showing results of the ML algorithms.....31

ACRONYMS

CeSHHAR	Centre for Sexual Health HIV and AIDs Research
CSV	Comma Separated Value
FSW	Female Sex workers
HIV	Human Immuno Virus
MOHCC	Ministry of Health and Child Care
NAC	National AIDS Council
NBC	Naïve Bayes Classifier
STI	Sexually Transmitted Infection
SQL	Structured Query Language
SVM	Support Vector Machine
UIC	Unique Identifier Code
UNAIDS	United Nations Programme on HIV and AIDS
PrEP	Pre-Exposure Prophylaxis

1. INTRODUCTION

This chapter introduces the background of the study, statement of the problem, justification, and motivation for the study. The chapter moves on to consider the contributions of the study to record linkage and de-duplication and gives an outline of the research report. The chapter begins by presenting the background of CeSHHAR, the sex work program and how clients are registered into the program.

1.1 Background to the study

The Centre for Sexual Health, HIV and AIDs Research Zimbabwe (CeSHHAR) implements the national sex work program in Zimbabwe on behalf of the Ministry of Health and Child Care (MoHCC) and the National AIDs Council (NAC) since 2009. The program offers free clinical services to sex workers through the 'Sisters with a voice' clinics. The services offered include, syndromic sexual transmitted infection (STI) treatment, family planning provision, Human Immunodeficiency Virus (HIV) testing, linkage to anti retro therapy (ART) for HIV positive clients and provision of pre-exposure prophylaxis (PrEP) for HIV negative clients. The organisation houses several HIV prevention and research projects. CeSHHAR's work focuses on 5 key functional areas: key populations, sexual reproductive health, masculinities research, health economics studies and children and adolescence. The national sex work program is under the key population's functional area.

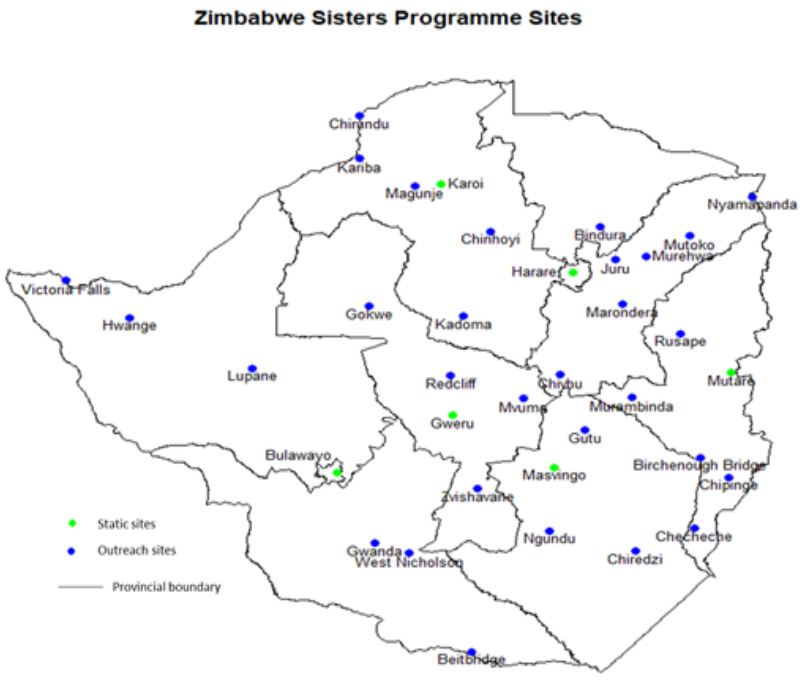


Figure 1.1 Map showing the coverage of the Sisters clinics in Zimbabwe : CeSHHAR (2017)

Figure 1.1 is a map showing the cities and towns where the sisters with a voice clinics are. The sex work program expanded from 5 clinics in 2009 to fifty-seven clinics by 2020 dotted around the country as represented on Figure 1.1. The clinics consist of 4 border sites in Chirundu, Forbes, Beitbridge and Victoria Falls and 6 static sites in the major cities of Harare, Mutare, Gweru, Masvingo, Bulawayo and Karoi. There are twenty-four local outreach sites within the static sites where a team goes into an identified health facility closer to where the clients are and offer services on an agreed day of the week. The 10 static sites (inclusive of the 4 border sites) also offer clinical service once a week to twenty-three highway mobile clinics along the highways of Zimbabwe, where truck drivers pass through to various destinations. The sexual and reproductive health services offered at the clinics are meant to reduce new STIs and HIV infections among sex workers (Ceshhar, 2019).

Recent data shows that 54% of female sex workers in Zimbabwe are living with HIV and the incidence rate is estimated at 10% per year (Ceshhar, 2020). Criminalisation of sex work, stigma and discrimination faced by the sex workers increases vulnerability to HIV as it creates barriers to access health services for sex workers (Ndori, 2018). The sex work program came from the need to have sex work friendly clinics offering Sexual and Reproductive Health Services (SRH) to sex workers to encourage uptake of SRH services among the sex work key population group.

The sex work friendly 'Sisters with a voice' clinics use a link log to register new clients at first visit and assigns them a Sisters' number. Proof of identity is not required during registration. An algorithm for a unique identifier was adopted in 2017 to try and uniquely identify clients accessing services at all the Sisters' clinics and protect client confidentiality by reducing the likelihood of unauthorised personnel from identifying the clients. The unique identifier is dependent on the data supplied by the sex worker as such, it is not fool proof. Previous experience has shown that sex workers change their identity at registration regularly resulting in different Unique Identifier Codes (UICs) being allocated to the same individual.

Female sex workers are sceptical of disclosing their identity as doing so exposes them to threats and risks of violence from law enforcement agents and members of the community given the criminalisation of sex work (Overs, 2002). This results in sex workers using a pseudonym which is a "fictitious name used to conceal someone's real name" (Merriam-Webster, 2019) at registration. The non-requirement of proof of identity at registration and during generation of the Unique Identifier Code (UIC) increases the likelihood of double counting and duplication

at the clinics. Since the names used at registration are sometimes not real, the probability of the sex workers changing the name on next visit is very high leading to the possibility of duplicating clients across the clinics.

To de-duplicate the data, there is need to utilise machine learning matching algorithms to link individuals based on other identifying variables that are collected on both the link log register and the demographic form. The study used Support Vector Machine (SVM) and Naïve Bayes Classifier (NBC) which are supervised machine learning techniques to link and match records between the link log and the demographic data sets.

1.2 Statement of the Problem

Since 2009 when the program started in the 5 clinics, sex workers were registered into the paper-based register called the "link log" using their name, surname, phone number and address before being allocated a Sisters' number, also called the Idlabel. The same sisters' number was to be at all the 'Sisters with a voice' clinics around the country to avoid double-counting. The paper-based system of allocating numbers is prone to duplication as proof of identity is not required for a sex worker to register at the clinic. As a result, sex workers sometimes register using pseudonyms as they are not comfortable with self-identifying due to various factors including stigma and criminalisation. This leads to double-counting sex workers and duplication as a change in the details meant they were reassigned a new number and unique identifier code (UIC) and got counted as a new individual which distorts reported figures.

1.3 Justification

Data that has duplicates can lead to data redundancy, where the data for the same individual would be stored in separate records as different people, resulting in inaccuracies in the analytical processes (Feng, 2016). Duplicates can lead to underestimation or overestimation in reporting which might affect resource allocation to intervention programs targeting key populations. This research aimed to match and de-duplicate records in the link log and demographic data sets for female sex workers and further ascertain the rate of repeat visits. The study used supervised machine learning techniques and a labelled data set with known duplicates for ML training. Supervised machine learning techniques improve record linkage and perform better than unsupervised machine learning (Christen, 2007). Record linkage assists in eliminating redundancy and correctly identifying individuals from the database; thus, it is a crucial step for data integration and data cleaning (Feng, 2016).

The record linkage from this study will assist in identifying duplicate individuals in the database and improve the quality and accuracy of the data. Since the data generated by the program informs policy there is need to use record linkage algorithms to determine whether a particular set of records can be appropriately linked to a specific patient (UNAIDS, 2009). This will enable proper planning and resource allocation in intervention programs targeting sex workers.

1.4 Motivations

This research was motivated by the duplication of records on the link log of sex workers at the Sister's clinic in Zimbabwe. When the link log became electronic in 2017, a dynamic programming approach utilising the Levenshtein algorithm on first name, last name and date of birth was used to check for duplicates within the link log. This algorithm measures the difference between two string sequences by counting the minimum number of single-character edits required to change one word into the other (Setiadi, 2013). The results from the Levenshtein algorithm showed that there were duplicate records in the link log database with Gwanda having the highest proportion of duplicate records at 8.49% followed by Chirundu at 8.43% and other sites having lower rates. The limitation of the Levenshtein algorithm is that it only compared records within one data set (linklog) and no record linkage was done with the demographic data to match and link records. This study used SVM and NBC models in record linkage and data de-duplication and the performance and linkage error rates of these models was evaluated in matching with the known true matches from the test data, these known matches are also known as a gold standard (Harron et al., 2017). Matched and identified duplicate records from the Levenshtein algorithm were used as the gold standard for this study.

1.5 Research Aim

This study aimed to apply SVM and NBC as supervised machine learning techniques to match and de-duplicate records in the link log and demographic data sets for female sex workers and further ascertain the rate of repeat visits by sex workers registered at Sisters' clinic in Zimbabwe. The study also aimed to evaluate the performance of SVM and NBC in record linkages.

1.6 Research Objectives

This study sought to achieve its aim by addressing the following objectives: -

1. Implement a record linkage framework to de-duplicate registers of the female sex workers registered at the Sisters with a voice clinic in Zimbabwe between 2017 and 2019 using supervised machine learning techniques.
2. Enumerate the number of individual sex workers and ascertain the rate of repeat visits by sex workers registered at Sisters with a voice clinics in Zimbabwe.
3. Validate the performance of SVM and NBC against the gold standard thereby identifying the optimum ML algorithm in matching.

1.7 Outline of the Research Report

The study is organised into five chapters with the first one giving the background of the study as above, the remainder of this research project is as follows;

Chapter 2 discusses the relevant literature review of different published articles on record linkage and de-duplication, in-line with the research objectives and questions.

Chapter 3 describes the models and techniques used in this study.

Chapter 4 presents the descriptive analysis of the data and provides a summary of the research findings.

Chapter 5 provides a discussion of the results

Chapter 6 give the conclusions and recommendations based on the analysed data

This chapter discussed the background of the study, research overall aim and related objectives. The next chapter will review related work that has been done in the field of record linkage and de-duplication using machine learning techniques.

2. LITERATURE REVIEW

2.1 Introduction

Previous work has been done on record linkages and de-duplication using both the conventional (probabilistic and deterministic) methods and machine learning techniques. This chapter will review literature on record linkage and de-duplication whilst also looking at literature on sex work in Zimbabwe and the HIV incidence among sex workers. Gaps in literature will be identified and the contribution that this research will make to the field of record linkage and de-duplication will be highlighted with focus on sex workers.

2.2 Sex work in Zimbabwe

Female Sex Workers (FSW) are among key populations who are most at risk of Human Immunodeficiency Virus (HIV) acquisition with 9% of the people living with HIV in the world being sex workers (Avert, 2018). According to (Ceshhar, 2019), engaging sex workers in prevention, treatment and care services, could possibly reduce new STI and HIV infections. Unsafe sex work practices are responsible for 40% of new infections within the population(Overs, 2002). The implementation of intervention programs targeting key populations such as the Sisters with a voice program, were introduced to reduce new infections as sex workers contribute to HIV transmission among the general population.

Analysis of the sex work program data from 2009 to 2017 suggests that sex workers are not coming for regular repeat clinical visits with more than 50% of the sex workers registered at the clinics having visited the clinic only once (Ndori-Mharadze, 2018). Some sex workers might have presented as new clients on repeat visits and may have been assigned duplicate sisters' numbers as a result. The machine learning record linkage approach was used to try and identify repeat visits in the program data and link records in the link log register and the demographic data thereby de-duplicating the data.

2.3 Record linkage

Record linkage is the process of merging information from two or more data sources using common identifiers between the data sources to find possible matches. This is also known as matching and has benefited many research fields by improving data quality, completeness and also assisting with de-duplicating (Zhu, 2015). Different machine learning techniques can be used in record linkage and data de-duplication.

Machine learning is the application and science of algorithms that evolved from artificial intelligence to analyse and make sense of data (Raschka, 2015). It utilises self-learning algorithms to build models from analysing large amounts of data to make predictions (Raschka, 2015). The machine learning approaches that are used in record linkage models are classified into two types which are supervised and unsupervised learning algorithms (Zimmerman and Kang, ND).

Supervised learning uses a labelled data set to train the model and use it for classification of matches and non-matches whereas unsupervised learning does not require training data but looks for similarities in the data to determine if they form a group (Zimmerman and Kang, ND). Support Vector Machine (SVM) and Naïve Bayes Classifier (NBC) are examples of supervised machine learning techniques that have been employed successfully for record pair classification (Christen, 2008) and are believed "to usually achieve better linkage quality compared to unsupervised methods" (Christen, 2007). Record linkages can also be done using probabilistic and deterministic algorithms.

Probabilistic algorithm developed by Fellegi and Sunter (Fellegi and Sunter, 1969), compares vectors for match, non-match or possible match to determine whether a pair of records refer to the same individual (Bruin, 2015). Matching variables are weighted to show how much the variables contribute to the probability of identifying a pair of records as a match or non-match (Kabudula et al., 2014). The probabilistic algorithm is used where there are no unique identifiers or statistical linkage keys and is more accurate than deterministic where there are errors in the variables being used in the record linkage (Europa, N.D).

Previous research studies have used deterministic and probabilistic algorithms which are the two common record linkage algorithms (Dusetzina et al., 2014a). The deterministic algorithm uses the similarity of basic identifiers to link records whereas probabilistic uses probability to determine if the records in the dataset refer to the same individual (Machado, 2004). Deterministic linkage is mostly used where the data set has unique identifiers of personal identifying information on both datasets (Bruin, 2015) or it can use a combination of person identifiers like name, surname as statistical linkage keys (Europa, N.D).

Several demonstration projects of record linkages of health data have been carried out around the world. An analysis of two linkage projects The Netherlands Twin Register (NTR) linked

with the Achmea Health Database (AHD) and the KOALA cohort with the database of the Stichting Farmaceutische Kengetallen (SFK) showed that the quality of information obtained through record linkage complemented existing data obtained from the cohorts databases with a few discrepancies found (Grootheest et al., 2015).

A simulation study was done in Netherlands to compare the performance of probabilistic record linkage against deterministic record linkage. In this study, both methods were used to link data from registries and the results showed that the probabilistic method performed better than the deterministic. Probabilistic method produced the lowest number of false positive links compared to the deterministic strategy across all scenarios (Miranda Tromp, 2011).

In Agincourt, South Africa another study was conducted to assess the feasibility of record linkages with conventional person identifiers. In this study 3 data sets from Agincourt and a health facility in South Africa were linked using a gold standard of 623 records that had been matched using fingerprints. Using the Fellegi-Sunter probabilistic record linkage model, the study evaluated 20 record linkage scenarios after which Sensitivity and Specificity were measured. The results from this study showed that Sensitivity was at 83,6% and Specificity was at 95,1%, a mix of deterministic and probabilistic linkages and use of extended identifiers increased sensitivity to 84.3% and Specificity to 96,9% respectively thus producing better results (Chodziwadziwa W Kabudula, 2014).

Mohammad et al, ND did a study to link prediction using supervised learning. Data about different research publications in the field of biology and computer science was used from Elsevier BIOBASE (http://www.elsevier.com) and Digital Bibliography & Library Project (DBLP) (http://dblp.uni-trier.de/xml/). A comparison of different classes of supervised learning algorithms was made on Accuracy, Precision-Recall, F score, and the mean squared error. The results showed that most classification algorithms can predict links with comparable performances however, SVM outperforms all other algorithms in all performance measures (Mohammad et al, ND).

This study will fill the gap in records linkage and de-duplication of sex workers as there is no data on record linkage and de-duplication of sex work registers. The de-duplicated dataset can be used to generate statistics that might contribute to reliable estimates of sex-worker population in Zimbabwe.

2.4 Gaps in Literature

Data from key populations programming contributes to the population size estimates to inform policies and resource allocation as they are a hard-to-reach population. Although previous work has been done in record linkage and de-duplication using both the conventional statistical methods and machine learning, this study is the first in the domain of sex work and record linkage using machine learning. The aim of this study is to de-duplicate the sex work registers to get the accurate number of sex workers registered at the clinic between 2017 and 2019. The data can also be used to determine the accuracy of repeat visits by sex workers to the Sisters' clinics. If the proportion of duplicate records is small, that could imply that sex workers are not returning for repeat visits. There might be need to explore and find out why sex workers are not returning for services at the clinic and address the issues.

Evidence from previous work has shown that supervised machine learning algorithms are more accurate and reliable in record linkage and de-duplication, as such this study will use SVM and NBC to link and de-duplicate the link log and demographic data for the Sisters with a voice clinics. Results of this study may be used to estimate the population size of sex workers in their diversity. The next chapter will detail the methodology that was used in this study.

3. METHODOLOGY

3.1 Introduction

The chapter details the record linkage process including data cleaning and linking the link log and demographic data set from CeSHHAR Zimbabwe to identify duplicates. The chapter further discusses in detail the methods and tools that were used, the study population, all the steps followed in the linkage process, data sources used and any ethical considerations for this study.

3.2 Record Linkage Process

The Python record linkage toolkit was used for cleaning and standardising the data sets to increase the accuracy of the record linkage (Bruin, 2019b). The record linkage algorithm followed the steps as illustrated in Figure 3.1 below after pre-processing and cleaning the data from both the demographic and the link log data sets.

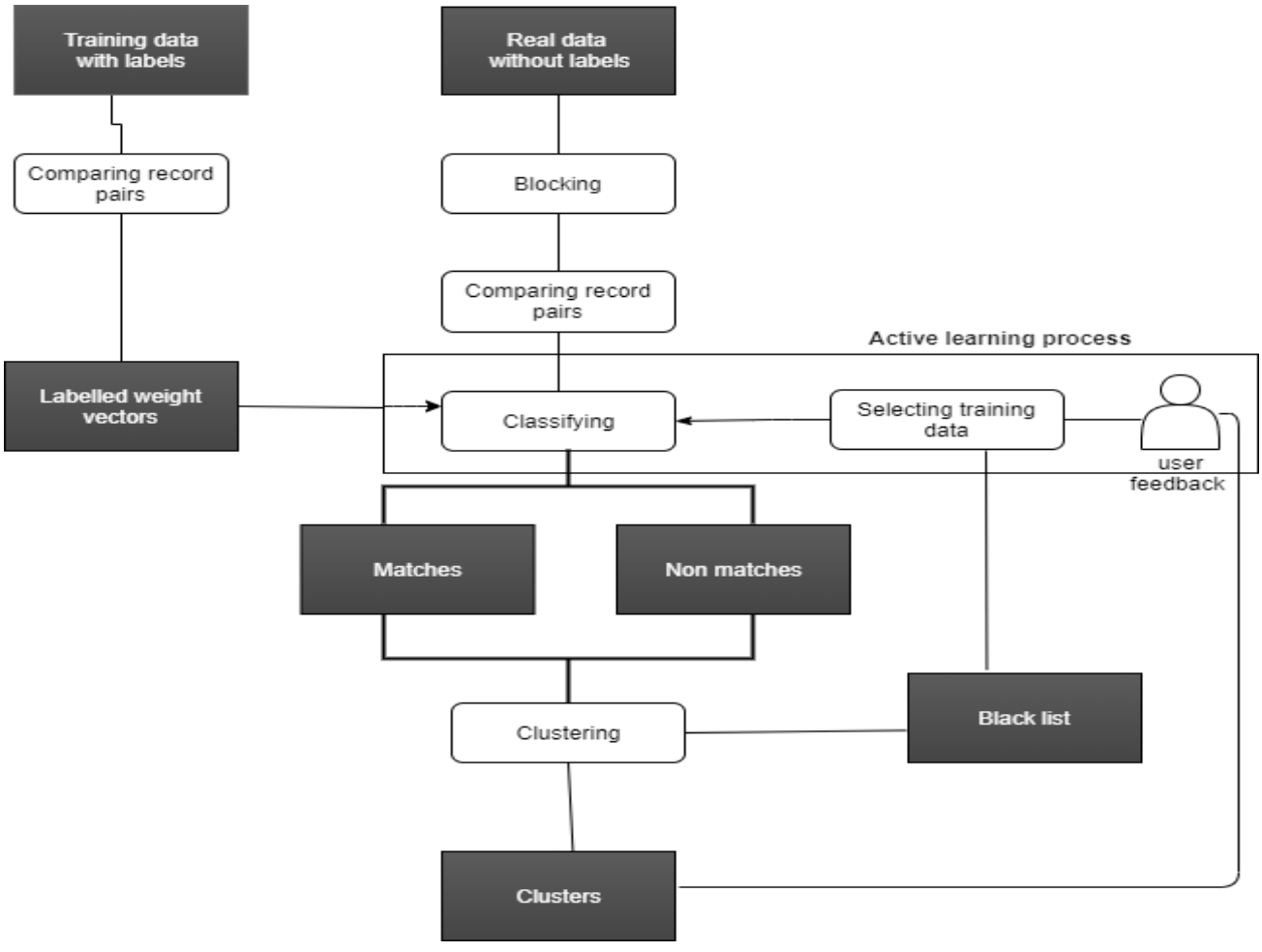


Figure 3.1 Overall Architecture of the Record Linkage Model. Source: (Feng, 2016)

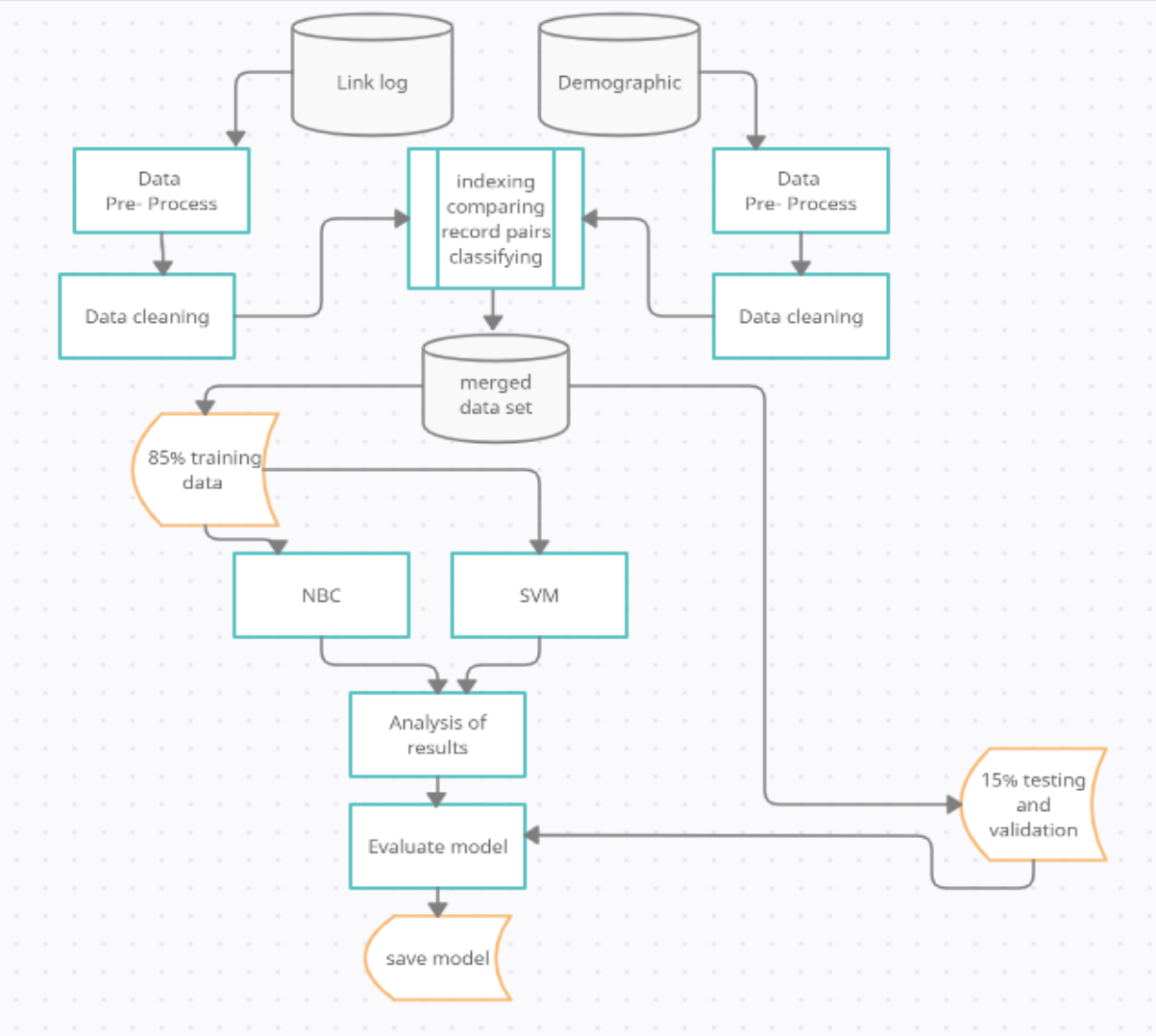


Figure 3.2 Work flow diagram

3.2.1 Data Pre-processing

This is the first step that was done to clean and standardise the data. It involved examining the data quality on completeness, consistency, and accuracy. It involved making sure that the dates are in the same format across the two data sets and same data types. This step was meant to increase the accuracy of the record linkage by minimising false matches as a result of typographical and data entry errors (Dusetzina et al., 2014b). For example, there were cases where age was missing but date of birth was available, in such cases age was calculated from the visit date and date of birth.

The next step was to check for missing values in the data sets. Some columns had missing values ranging between 46% - 86% which were mainly due to changes in variable names between 2009 and 2017 when the program changed data collection systems. A method of imputing was devised to reduce the number of missing values with the aim to keep as much

information as possible to successfully link the two data sets. Due to the change is data collection systems, there were different columns which contained the same information for example `everbeentoclinic` and `haveyoueverbeentoclinc`. Such columns were combined to produce one single column to reduce the proportion of data with missing values.

The following columns were combined:

i. `Haveyoueverbeentoaswclinictoacce and everbeentoclinic`
ii. `Dob and whatisyourdayofbirth.`
iii. `Dob and whatisyourmonthofbirth.`
iv. `Dob and whatisyouryearofbirth.`

3.2.2 Data cleaning

This is a process to correct errors and standardise variables into a consistent form, for example, date of birth format should be consistent across all records in the data sets (Oliveira et al., 2016). The process is also referred to as standardisation where all quantities are changed into common units (for example weights to SI units) (Churches et al., 2002) . Data cleaning also makes sure that categorical variables are represented in the same way in the data sets (Churches et al., 2002), for example, for the variable sex, Female and 2 mapped into F and males and 1 mapped into M (Churches et al., 2002). Data cleaning can be disintegrated into 3 main steps namely; segmentation of the data into specific atomic data elements, transforming atomic elements into their canonical forms and imputation of missing data (Churches et al., 2002)
Some validations were done on the demographic data to make sure the dates are correct and valid. Validations included the following:

i. Checking if days fall in the range 0 and 31 inclusive.
ii. Checking if February dates do not exceed 29.
iii. Checking if month names are valid.
iv. Checking if a year has 4 figures.

Any date that did not meet the four checks was dropped as an invalid date. There are variables that are used to generate the `UIC` that was introduced in 2017. This meant that all data collected before 2017 had missing values for the UIC. The UIC is an identifier which uses an algorithm of joining letters from the mother's name, client's surname, district of birth, date of birth and gender.

All missing values in the gender column for data prior to 2017 were assumed to be females as the program was only serving female sex workers since its inception in 2009. The programme

introduced the gender variable in October 2017 when the scope of the programme increased to cover sex workers in their diversity regardless of their sexual orientation. Visit dates in both data sets were converted to datetime object in the same way the birthdate column was cleaned. Regular expressions were used to clean the visit date in the link log data frame. The column was recorded with the exact time of the visitation. Only the day, month and year was needed, and the regular expressions helped to extract these.

Samples with missing date of births were dropped. The dates were validated in the same way the demographic data dates were validated. The four-way check was used and samples with missing UICs or incomplete data were also dropped. After cleaning and pre-processing, the following columns were deemed to be relevant for our study experiments, `Idlabel, UIC, visitdate, date_of_birth`, and `gender`. The `Idlabel` (a unique chronological number with the clinic suffix allocated to each client) and `UIC` were the unique identifiers used to link the datasets.

3.2.3 Indexing

The indexing module was used to make pairs of records called candidate links or candidate matches and can be used for both linking and duplicate detection (Christen and Goiser, N.D). There are several indexing algorithms available such as blocking and sorted neighbourhood indexing. Blocking is whereby records that are a potential match are grouped to form a block which is then examined to identify the matches (Oliveira et al., 2016). Blocking reduces the amount of pairs that are checked for matches as this methods will block any pairs that are considered as not similar and only measures similarity on pairs considered to be similar (Dannelov, 2018) .Sorted neighbourhood is whereby elements that are similar to each other are sorted closer together and then a key is generated and searches all elements that are close to the key.

The study used the neighbourhood index method to group records that were potential matches in the link log and demographics dataset. The `birthdate` and `visitdate` variables were compared using the deterministic method while the `Idlabel` and UIC were compared using the Jaro Winkler which is a string comparison method. The Jaro Winkler method is ideal for situations where there is a large data set with a lot of spelling mistakes (Bruin, 2019b).

Jaro-Winkler similarity sim_{jw} is obtained by using the Jaro similarity sim_j *as presented on the equation below:-*

$$simj = \begin{cases} 0 & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right) \end{cases} \quad (1)$$

Where m is the number of matches and t is half the number of transpositions of letters to obtain the other strings character ordering.

The Jaro-Winkler similarity corresponds to the following equation:

$$sim_w = sim_j + lp(1 - sim_j) \quad (2)$$

Where l is the length of common prefix starting in the beginning of the string and covers up to 4 characters and p is a scaling factor depending on how much the common prefix should affect the similarity score. The final result is obtained by taking $1-sim_{jw}$.

The study used 85% of the candidate links dataset for training the algorithms and 15% for validation. 85% was for training using the gold standard dataset and 15% was the holdout validation dataset and testing on unseen data. After validation we fitted the models for classification as matches or non matches. The gold standard data of duplicates identified through the Levenshtein algorithm was used as training data and had 3274 records that were identified as duplicate records. This dataset had the following 4 variables, `UIC`, `Idlabel`, `visitdate`, `date_of_birth`, with the variable `Idlabel` as the unique identifiers.

3.2.4 Classification

After the record pair comparison step, all the generated weight vectors from the test dataset were passed to the classifier, whereby all the generated weight vectors from data are classified as either matches or non-matches (Bruin, 2016). There are two categories of text classification approaches, knowledge engineering and machine learning (Ramesh and Sathiaseelan, 2015).

The machine learning approach is used in several domains and is most popular in text mining. The procedure of text classification is to build a classifier based on labelled data and to organise unlabelled data into the pre-specified categories. Different techniques can be used for text classification such as Decision tree, NBC and SVM amongst many, and evidence has shown that SVM performs better than other techniques (Ramesh and Sathiaseelan, 2015).

To address the problem of duplication of CeSHHAR data, record linkage and machine learning algorithms were used to come up with a model to de-duplicate this data. This study used SVM and NBC classifiers to link and de-duplicate the data. The study aimed to evaluate the

performance of these two ML algorithms in de-duplication, thereby comparing and validating their performance against the gold standard so as to select the optimum algorithm for the machine learning model.

The study also utilized the Python record linkage toolkit which is a library to link records and has most of the tools required to link and de-duplicate data (Bruin, 2019b). The study used the following Python libraries;

- Pandas for data analysis
- Numpy for mathematical and logical manipulation on arrays
- Recordlinkage library to link the data sets
- Matplotlib for visualisation of the data through graphs and plots
- Seaborn a library based on Matplotlib was also used for data visualisation
- Scipy for solving mathematical and science problems

SVM and NBC algorithms were implemented separately on the tool kit and the results compared in terms of scalability and Accuracy to come up with an optimum model for de-duplicating the data.

3.3 Study population and data sources

In this study, the population were all sex workers registered at the Sister's with a voice clinic in Zimbabwe between 2009 and 2019. The researcher used secondary data routinely collected by CeSHHAR at their 'Sisters with a voice' clinics. The first data set was extracted from the link log register (where sex workers are registered at first visit) in a CSV format. The second data set was demographic data where basic demographic information like marital status, level of education, religion, number of biological children among other demographic are collected by the clinicians into a separate database at the first visit. The demographic data collected also includes variables that form the UIC as collected in the link log register (it is these common variable between the link log and demographic form that were used to match the records) and these are:-

- `mother's name`
- `client's surname`
- `district of birth`
- `date of birth`
- `gender`

The link log database sits on a Structured Query Language (SQL) 2014 database and was downloaded in Comma Separated Values (CSV) format while the demographic data was downloaded from the coconut cloud server as CSV. The study used data from the 31 sex work clinics (as shown on Figure 1.1) to link and match records using the common variables and where a duplicate exists, the first record was maintained. Coconut is a patient management data collection system that CeSHHAR adopted in 2013 when the program migrated from paper-based data collection tools. The application is supported for use on android based mobile devices where data is stored on the device when working offline for syncing to a cloud server when there is internet connectivity (Kelley and McKay, 2011). The coconut system was adopted as a measure to minimise double-counting sex workers who would have registered at a different clinic by enabling data sharing of patient records across the clinics and to provide continuity of case with some patient history. Data cleaning for duplicates and missing data was done followed by record linkage steps.

3.4 Sampling

This study used a type of purposive sampling method called total population sampling. This type of sampling involves using the whole population that have common traits or exposure to events (Lund Research, 2012). In this study data for all sex workers registers at the sister's clinics between 2009 and 2019 was used. For data collected before 2017, the Idlabel was used to link the records.

3.5 Overview of Record linkage

Record linkages provide a framework for computer-oriented solutions that helps in matching records from two datasets (Fellegi and Sunter, 1969). Record linkage was used to determine if compared data is a match, non-match, or possible match. The two data sets were classified as A and B for matching. Pairs were classified in a product space A × B from the two files into M representing the set of true matches, and U representing the set of true non-matches.

$$R = \frac{P(\gamma|(a,b) \in \Gamma|M)}{P(\gamma|(a,b) \in \Gamma|U)} = \frac{m(\gamma)}{u(\gamma)} \qquad (3)$$

"where γ is an arbitrary agreement pattern in a comparison space Γ" (Winkler, 2014). Γ might be made up of patterns for matches or non-matches and $\gamma \in \Gamma$ might be the frequency with which specific values occur (Winkler, 2014).

Tμ and Tλ are cut off thresholds that are determined by error bounds on matches and non-matches (Winkler, 2014). If $R > T\mu$ the pair is classified as a match, if $T\lambda \leq R \leq T\mu$ the

pair is classified as a possible match for further review and if $R < T\lambda$ the pair is classified as a non-match (Winkler, 2014).

Record linkage algorithms also have a likelihood of false positives where matches are found on data that belongs to different individuals or false negatives where the algorithms fail to match the data that corresponds to the same individual in the database maintaining duplicity (Ansolabehere, 2017).

The model by (Fellegi and Sunter, 1969) accounts for error rates associated with record linkages. The first type of error is when an unmatched pair is classified as a link denoted by A_1 and its probability being:

$$\mu = P(A_1|U) = \sum_{\gamma \in \Gamma} u(\gamma) \cdot P(A_1|\gamma) \quad (4)$$

The second type of error occurs when a matched pair is classified as a non-link denoted by A_3 with the probability being:

$$\lambda = P(A_3|M) = \sum_{\gamma \in \Gamma} m(\gamma) \cdot P(A_3|\gamma) \quad (5)$$

To mitigate against the likelihood of the type 1 and type 2 errors, this study had a robust evaluation and performance testing plan to come up with an optimum algorithm. A comparison of Sensitivity and Specificity scores was done for both algorithms. Sensitivity also called True Positive Rate (TPR) is the proportion of matches that are correctly classified as matched and Specificity also called True Negative Rate (TNR) is the proportion on non-matches that are correctly specified as non-matches (Europa, N.D).

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (6)$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad (7)$$

To validate the record linkage framework, Recall, Precision, and F-score measures were used. Precision is defined as "the proportion of compared record pairs classified as matches that are true matches" (Hand and Christen, 2018), while Recall is the proportion of true matching record pairs that are classified as matches" (Hand and Christen, 2018). The formulas for these validation tests are: -

$$Precision = \frac{True\ positive}{(True\ positive + False\ positive)} \quad (8)$$

$$Recall = \frac{True\ positive}{(True\ positive + False\ negative)} \quad (9)$$

$$Accuracy = \frac{True\ positive + True\ negative}{(True\ positive + False\ positive + False\ negative + True\ negative)} \quad (10)$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad (11)$$

F-score which is the mean of Precision and Recall was used to validate the tests. The value for F-score was high if both Precision and Recall values are high. The F-score "can be seen as a way to find the best compromise between Precision and Recall" (Kabudula et al., 2014).

$$F\ score = \frac{2\ (Precision * Recall)}{(Precision + Recall)} \quad (12)$$

3.6 Support Vector Machines

SVM is a powerful supervised machine learning method usually used to address a binary pattern classification problem." SVM enables the prediction of labels and
aims to create a decision boundary between two classes that enables the prediction of labels from one or more feature vectors" (Huang et al., 2018). The hyperplane is the decision boundary and support vectors are the closest points to the hyperplane (Huang et al., 2018) as illustrated in Figure 3.3 below.

For example, given a labelled training dataset:

$b(x_1, y_1), \ldots, (x_n, y_n), x_i \in R^d$ and $y_i \in (-1, +1)$ (11)

where x_i is a feature vector representation and y_i the class label (negative or positive) of a training compound i.

The optimal hyperplane can then be defined as:

$wx^T + b=0$

where w is the weight vector, x is the input feature vector, and b is the bias.

The w and b would satisfy the following inequalities for all elements of the training set:

$wx_i^T + b \geq +1$ if $y_i=1$

$wx_i^T + b \leq -1$ if $y_i=-1$

The objective of training an SVM model is to find the *w* and *b* so that the hyperplane separates the data and maximizes the margin $1 / \| w \|^2$.

Vectors x_i for which $|y_i| (wx_i^T + b) = 1$ was termed support vector

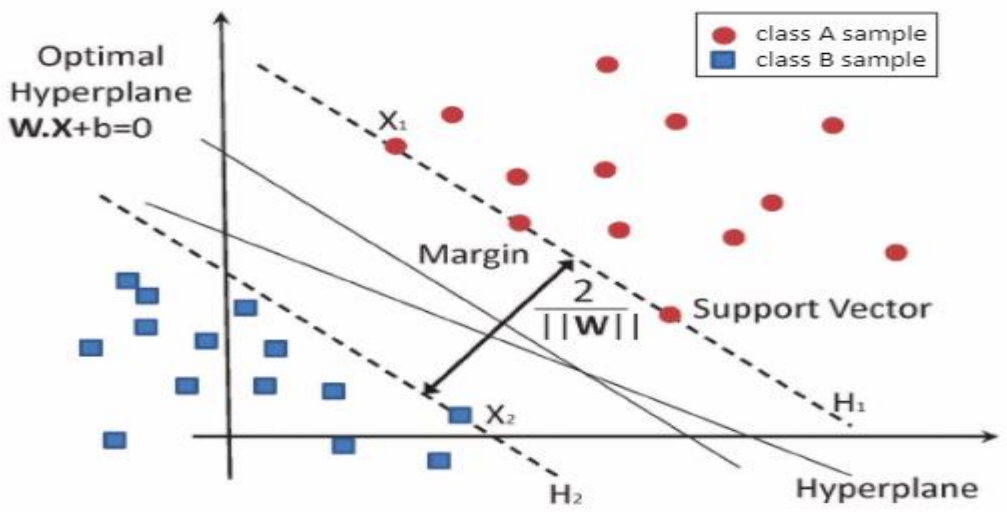


Figure 3.2 Classification of data by support vector machine (SVM). Source: (García-Gonzalo et.al , 2016)

3.7 Naive Bayes Classifier

NBC divides candidate record pairs into matches and non-matches based on probabilistic principles. The NBC method has a close mathematical connection with the Fellegi and Sunter model. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem of conditional independence between every pair given the value of the class variable (Mitchell, 2017). Bayes' theorem states the following relationship as illustrated on Figure 3.4, given class variable y and dependent feature vector x_1 through x_n:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots x_n \mid y)}{P(x_1, \dots, x_n)} \quad (13)$$

Where ***p (x1, x2, …, xn | yi)*** is the probability of a specific combination of features given a class label.

Using the naive conditional independence assumption that

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y), \quad (14)$$

for all i, this relationship is simplified to.

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)} \quad (15)$$

Since $P(x_1, \dots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg \max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y), \tag{16}$$

and we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$ the former is then the relative frequency of class y in the training set.

The different NBCs differ mainly by the assumptions they make regarding the distribution of $P(x_i \mid y)$ (Pedregosa et al., 2011).

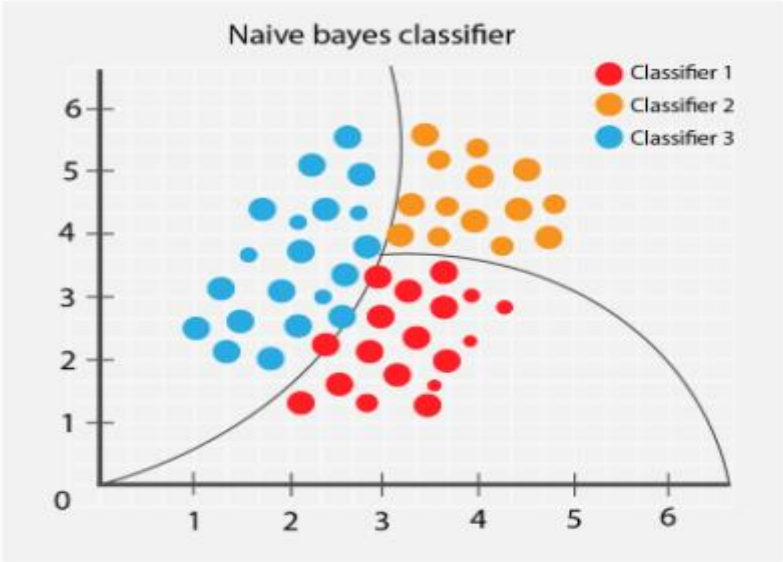


Figure 3.3 Classification of data Naïve Bayes Classifier (NBC): Source (Yang, 2019)

3.8 Evaluation and validation of the classification

Performance validation was done to check for errors on the algorithms. There are two types of errors that can be produced by record linkage, type 1 error where a non-match is classified as a match (False Positive) and a type 2 error where a match is classified as a non-match (False-Negative) (Dusetzina et al., 2014a). The False-Positive and False-Negative rates contribute to the Sensitivity and Precision or Positive Predictive Value (Zhu, 2015).

Confusion matrix is a measure of performance for machine learning algorithms that has 4 different combinations of predicted and actual values as shown in Figure 3.2. Precision, Recall, Specificity and Accuracy can be measured using the matrix. A confusion matrix was run for each of the models to measure and evaluate algorithmic performance.

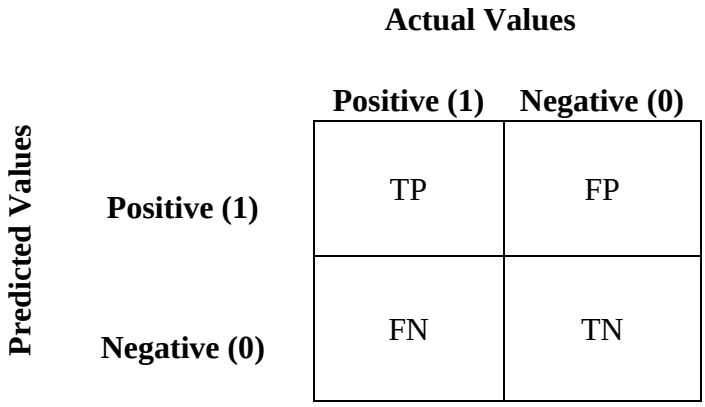


Figure 3.4 Confusion matrix

The program was run several times and measures for classification and clustering were collected at each iteration (Feng, 2016) until weight vectors were correctly classified. The SVM and NBC algorithms was evaluated in terms of computational run time and the Accuracy of the results based of the labelled data was used to select the algorithm that performed better.

3.9 Ethical considerations

The study used approval from host organisation by institutional review boards at the Medical Research Council of Zimbabwe number MRCZ/E/266 which permits secondary analysis of the sex work program data. The protocol for this study was also approved by the University of Witwatersrand Ethics Committee. The researcher had authorization to use the data from the gatekeepers at CeSHHAR Zimbabwe. The data sets had access controls and security measures including the use of passwords and encryption to ensure that only the researcher and authorised persons had access to the data. Both data sets were de identified by removing identifying information of the sex workers to maintain confidentiality.

This chapter looked at the methodology used in the study for record linkage and de-duplication, and how the algorithms were validated to measure their performance. The chapter also detailed the sampling method, the data sources and study population of this study. It concluded by touching on the ethical considerations that were made and measures that were taken to secure the data and protect confidentiality of the participants. The next chapter details the results of the study.

4. RESULTS

4.1 Introduction

This chapter will discuss the results from the record linkage algorithm on the link log and demographic data for CeSHHAR Zimbabwe. The results section will give a more detailed description of the record linkage process results along with tables and graphs to illustrate the findings and present descriptive statistics from exploring the 2 data sets.

The link log register (where sex workers are registered at first visit) is a CSV format and the second data set is the demographic information collected by the clinicians into a separate database at the first visit. Common UICs between the link log and demographic data sets (same information collected on both forms) were used to match the records. The data was collected between 2009 and 2019. The link log database sits on an SQL 2014 database and was downloaded in CSV format while the demographic data was downloaded from a cloud server which houses the sex work program data as described in section 3.3 (Study population and data sources).

4.2 Descriptive statistics

This section gives descriptive statistics of demographic data for the sex work program focusing on age, gender, marital status and level of education of the sex workers registered at the clinics. Descriptive statistics for the link log data focused on numbers registered at the different sites within the sisters program. The statistics mainly looked at the distribution, frequencies and proportions / percentages for these variables to get a better understanding of characteristics of the sex workers registered at the clinics.

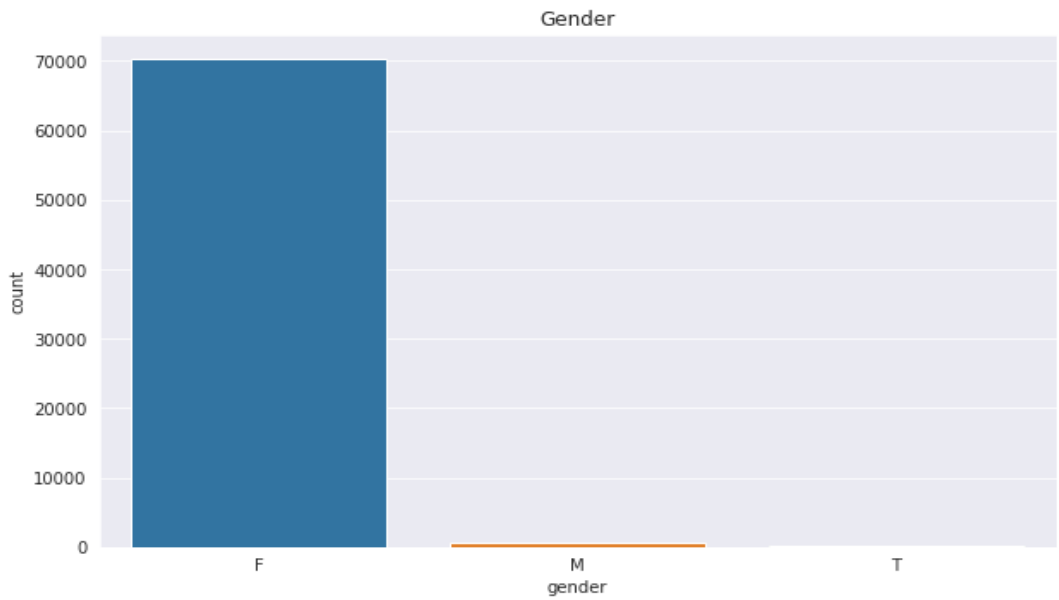


Figure 4.1 Distribution of the gender of clients registers at the Sister's clinics

Figure 4.1 represents the distribution of gender in the demographic data set showing that 98.8% (70275) of the total number of clients registers were females, 0.8% (572) were males while 0.4% (277) were transgender.

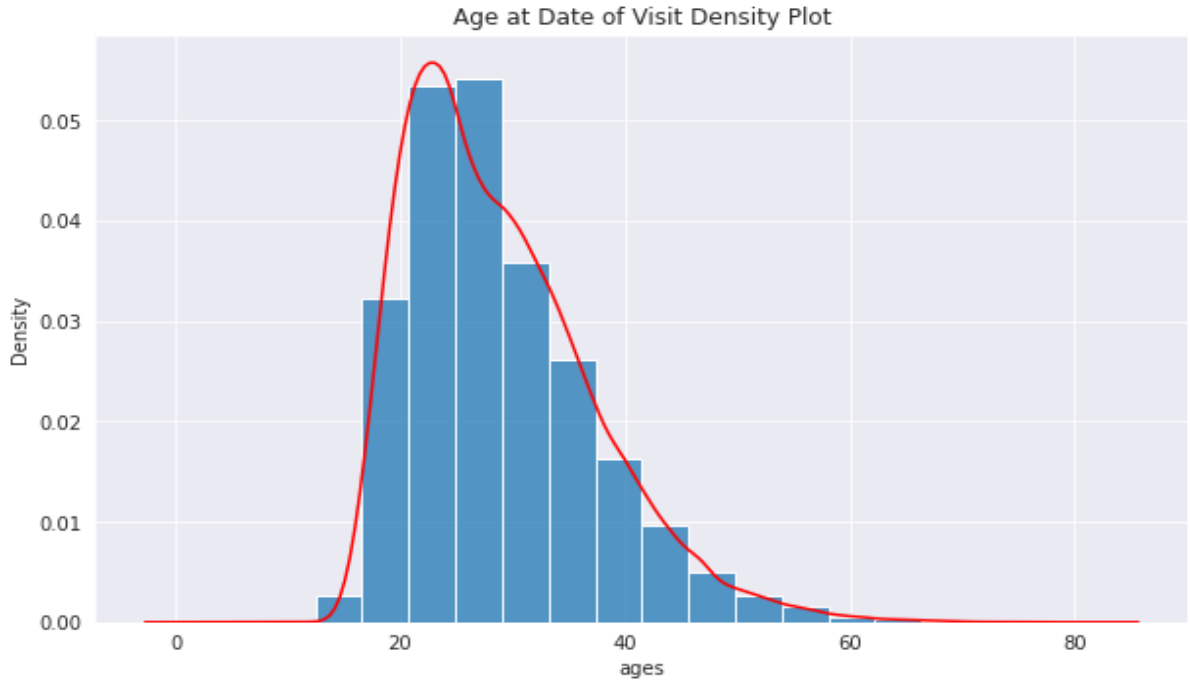


Figure 4.2 The distribution of age at first visit for clients registered at the Sisters clinics.

Figure 4.2 is a density graph that is normally distributed, depicting that most clients registered at the sisters clinics were between the ages of 30 and 35, followed by 25-30 with only a few clients falling in the age band of below 20 and over 60 age.

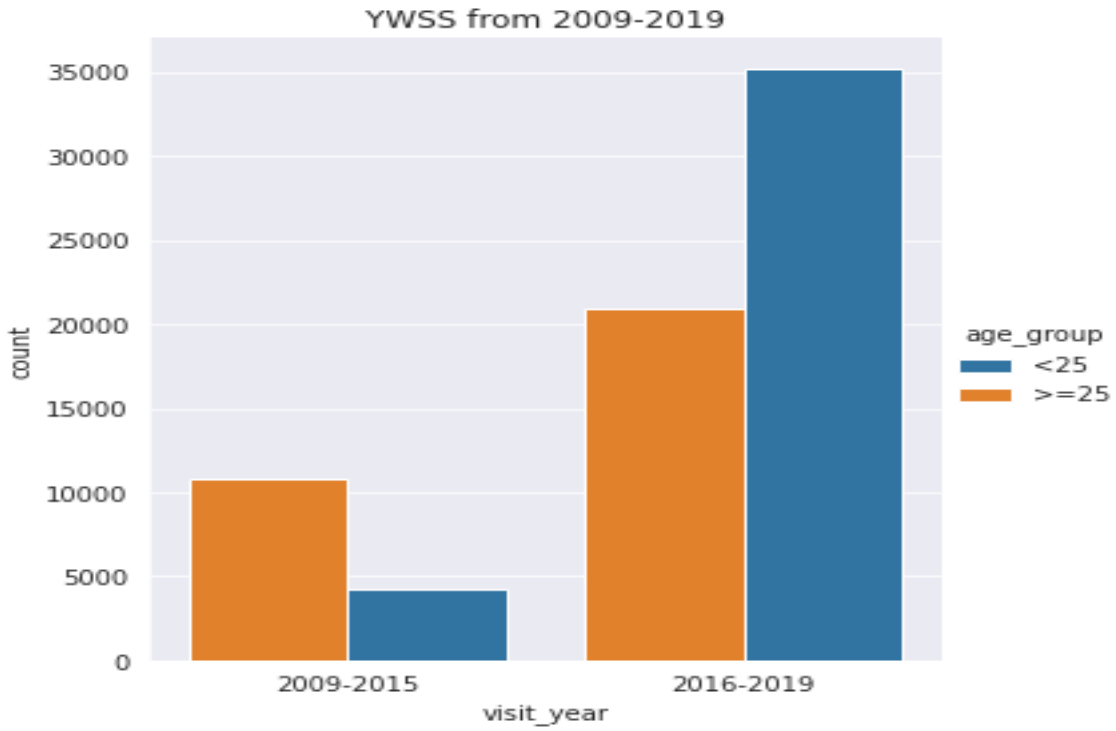


Figure 4.3 A comparison of age distribution at first visit for client registered between 2009-2015 and 2016-2019

Figure 4.3 describes age at first visit where it was noted that between 2009 and 2015, the proportion of young women selling sex (YWSS) who are aged less than 25 was less than 5000 (<25%) and there was an increase of the YWSS in noted between 2016 to 2019 to more 35000 . The graphs shows a shift between 2016 and 2019, where more YWSS were registered as compared to the older SW from the age of 25.

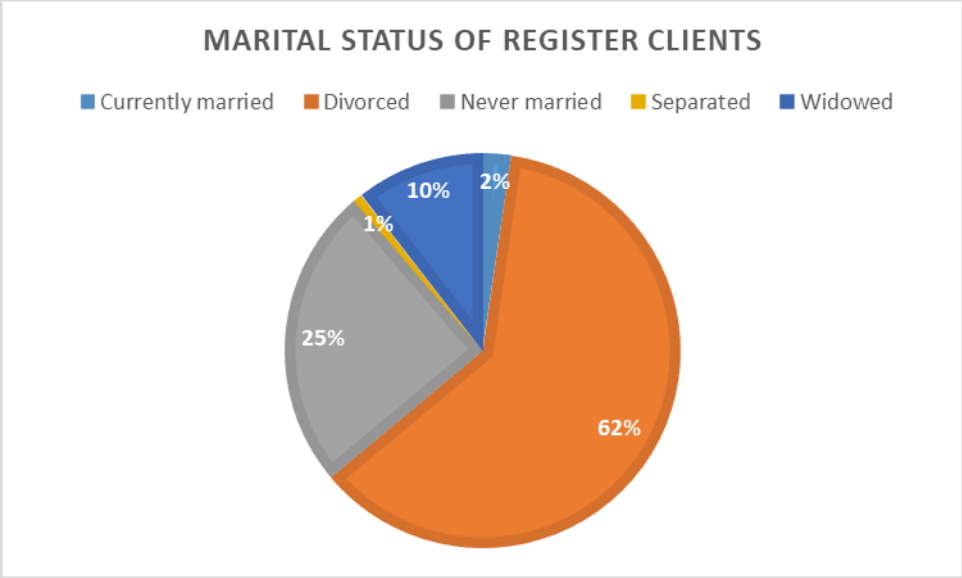


Figure 4.4 Marital status of sex workers at registration

Figure 4.4 describes the marital status of the sex workers from the demographic data which showed that 1% of the sex workers were separated, 2% reported as currently married, 10% widowed, 25% never married with 62% of the clients reporting that they were divorced at registration.

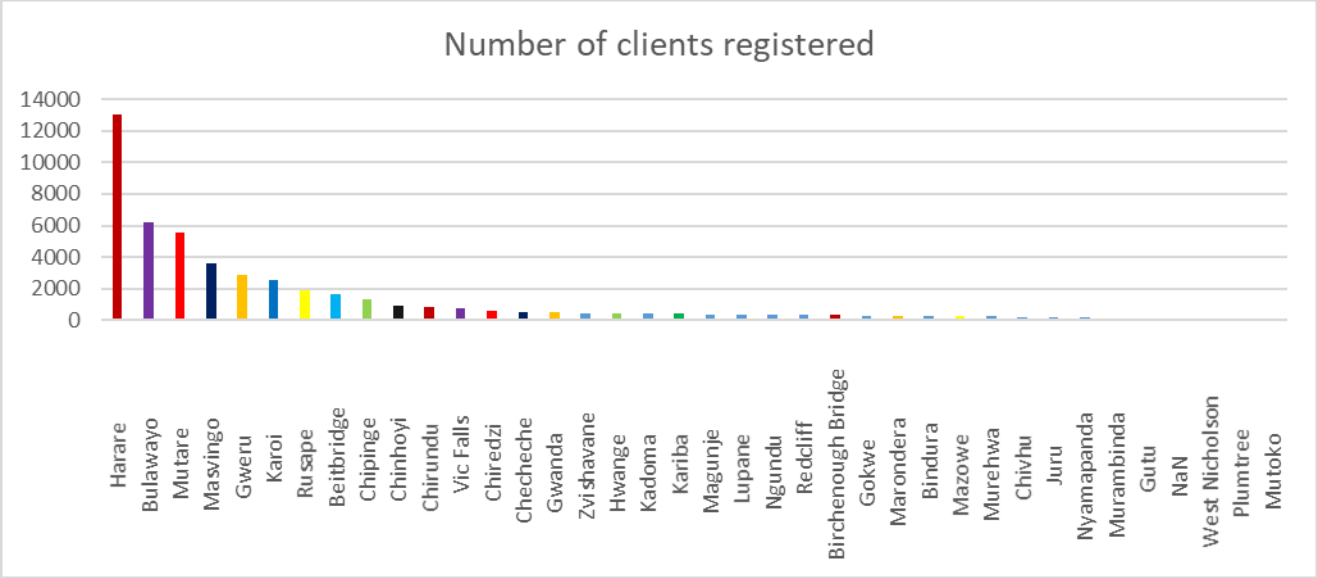


Figure 4.5 Number of clients registered at the Sisters clinics by site.

Figure 4.5 shows the number of clients registered in the link log data by site. From the above figure and supporting Table 4.1 below, Harare registered the highest number of clients with more than 13,000 sex workers registered, followed by Bulawayo with more than 6,200. The data from Figure 4.5 shows that overall static sites in Mutare, Masvingo, Gweru and Karoi enrolled more clients after Harare and Bulawayo with the lowest numbers being recorded in Plumtree, Mutoko and West Nicholson.

Table 4.1 Table showing number of clients registered by site

Site	Number of clients registered
Harare	13036
Bulawayo	6202
Mutare	5578
Masvingo	3589
Gweru	2889
Karoi	2579
Rusape	1861
Beitbridge	1616
Chipinge	1322
Chinhoyi	917
Chirundu	803
Vic Falls	796
Chiredzi	612
Checheche	552
Gwanda	494
Zvishavane	468
Hwange	462
Kadoma	419
Kariba	415
Magunje	395
Lupane	386
Ngundu	366
Redcliff	357
Birchenough Bridge	324
Gokwe	308
Marondera	290
Bindura	288
Mazowe	247
Murehwa	240
Chivhu	213
Juru	180
Nyamapanda	172
Murambinda	126
Gutu	118
NaN	74
West Nicholson	58
Plumtree	53
Mutoko	41

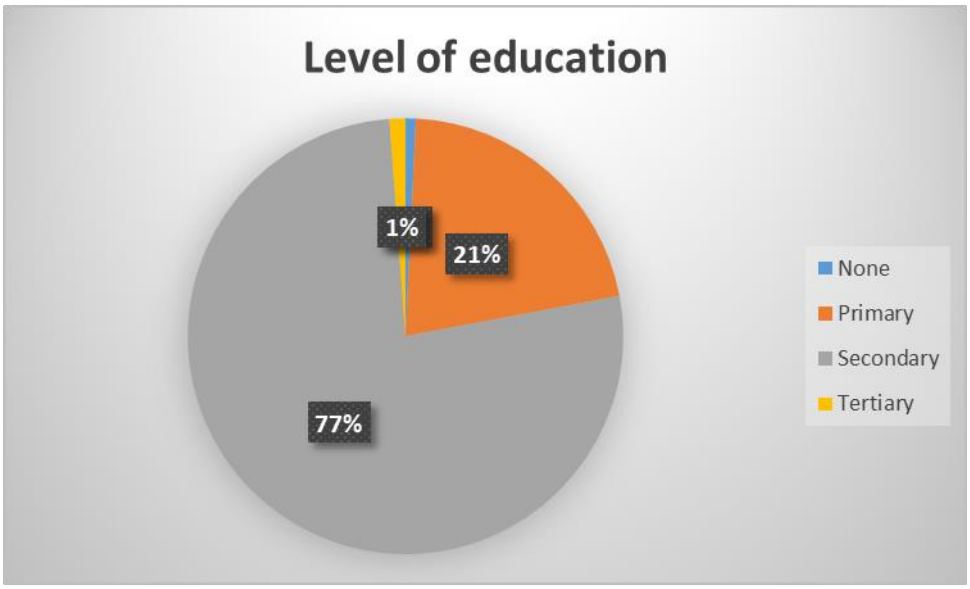


Figure 4.6 Level of education at registration for sex workers registered at the sisters' clinics

Figure 4.6 depicts level of education for sex workers at registration. The data showed that 77% of the sex workers who were registered at the sisters' clinics had gone up to secondary level education as highest level, with 21% indicating that they had gone up to primary level, less than 2% had tertiary education and 1% had no education.

4.3 Results of the record linkage process

The link log data set had 90797 records collected between 2009 and 2019 across the 31 clinics. The number of records in the link log were reduced to 48 846 records after data cleaning and pre-processing and the demographic data had 92 082 records before data cleaning and pre-processing which were reduced to 71 124. The 2 data sets were linked using `Idlabel` and produced 433 450 candidate links.

A compare object was created where exact method comparison was done on `birthday_date` and `visitdate` variables and the Jaro Winkler string method with parameter method was used on `Idlabel` and `UIC`. Where there were similarities between the variables, a score of 1 was computed and a summation of the scores was done between the 4 columns. Where the score was greater than 3, the links were identified as duplicate records. Out of the 433 450 -candidate links, 3 274 pairs were identified as duplicate records.

Table 4.2 Table showing sample matched pairs during record linkage

		uic	Idlabel	visitdate	date_of_birth
92028	90687	1	1	1	0
66304	9099	1	1	1	0
81658	32234	1	1	1	0
62841	3698	1	1	1	0
68326	12009	1	1	1	0
74845	21797	1	1	1	0
72751	18675	1	1	1	0

The identified duplicates were used as the labelled data to create a machine learning model to de-duplicate more records. To create a training data set, golden pairs were created where object were compared from candidate links. Support Vector machine was fit into the training data and were able to predict 3450 pairs as duplicate records.

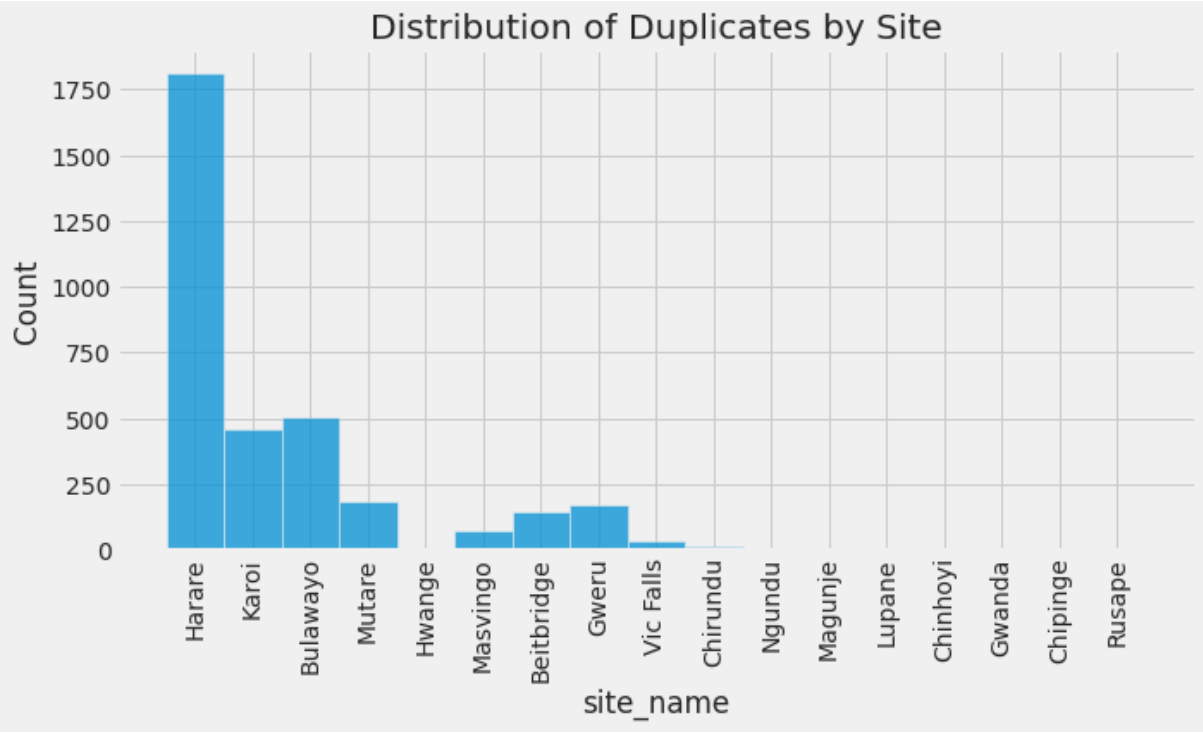


Figure 4.7 Duplicate records by site

Figure 4.7 above shows the distribution of the duplicates by site. From the above graph, more than half of the duplicates were in Harare 52.93% followed by Bulawayo at 14.68%.

Table 4.3 Table showing distribution of duplicates by site

Site	Proportion of duplicates
Harare	52.93
Bulawayo	14.68
Karoi	13.43
Mutare	5.43
Gweru	4.99
Beitbridge	4.26
Masvingo	2.1
Vic Falls	0.93
Chirundu	0.408
Ngundu	0.29
Hwange	0.17
Gwanda	0.14
Chinhoyi	0.058
Lupane	0.058
Rusape	0.029
Chipinge	0.029
Magunje	0.029

Table 4.3 shows the distribution of duplicates by site. Harare is topping the list with more than 50% of the duplicates being registered in Harare and the least being in the smaller mobile sites like Chipinge, Magunje and Rusape. The demographic data had 40 507 clients registered between 2017 and 2019 and this number was reduced to 36 654 after de-duplication. The duplicated data had less than 1% (23 out of 3450) sex workers who had responded that they had visited a sister's clinic before.

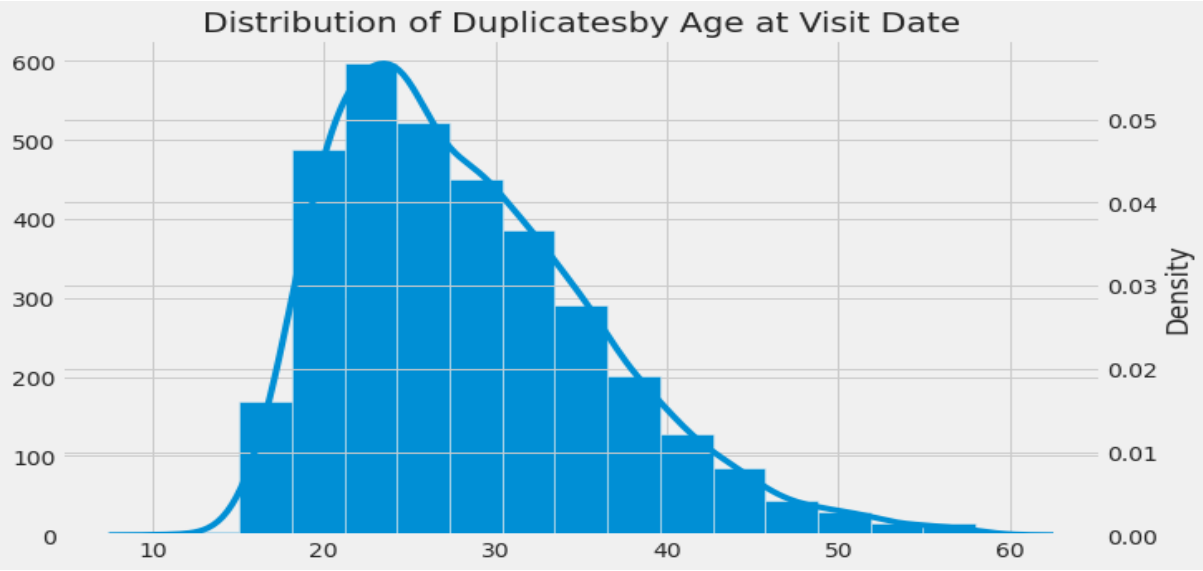


Figure 4.8 Duplicate records by age at first visit

The distribution of age at first visit for the duplicates is normally distributed and similar to the distribution of age in the whole data set with majority of clients ranging between 24-30 years of age at registration.

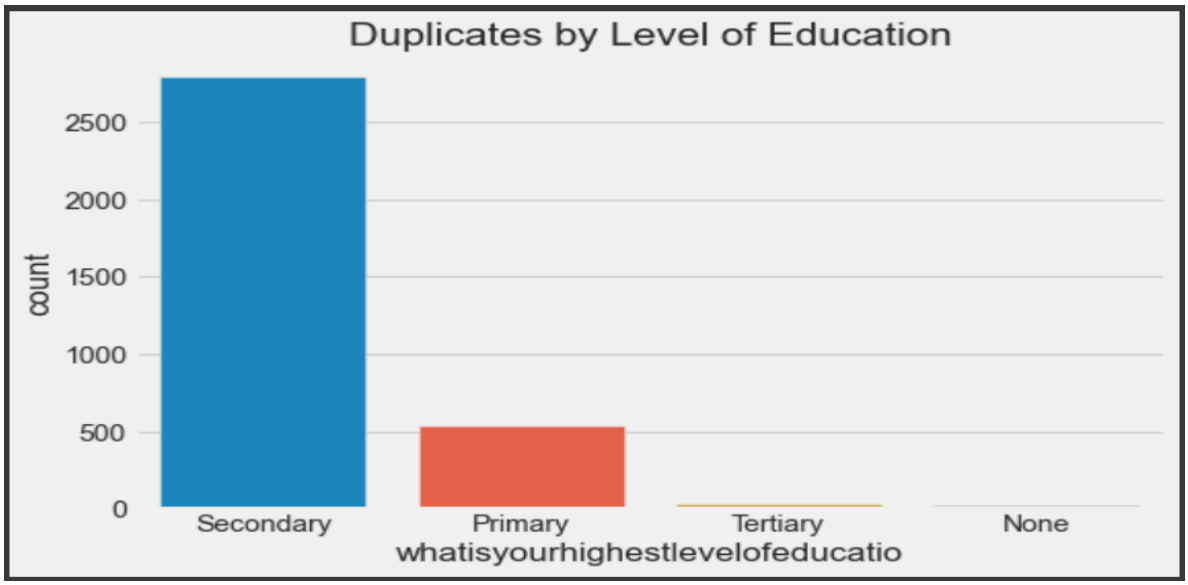


Figure 4.9 Level of education at registration for duplicate records

The data showed that 2798 of the clients from the duplicates had secondary education as their highest level of education, 536 had primary education and 29 had tertiary. 23 of the clients indicated that they had no education.

4.4 Evaluation of classification results

The two algorithms were evaluated and assessed for performance using the Precision, Accuracy, and Recall and F1 score metrics. Below are the results of the evaluation by method: -

SVM

The graph below shows the result of the confusion matrix. For the SVM algorithm, Precision was at 95,5%, Recall 1, Accuracy at 99,9% and an F score of 0.9778.

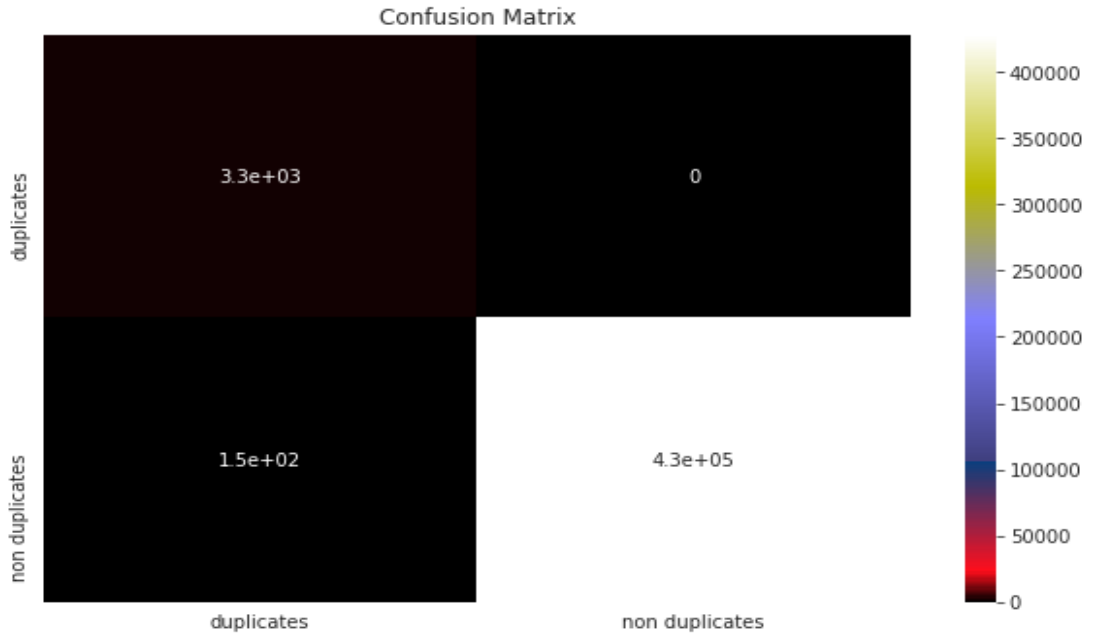


Figure 4.10 SVM Confusion matrix

NBC

The NBC algorithm was also assessed against the training data and results showed an Accuracy of 99.9%, Precision of 94.1%, Recall of 1and an F score of 0.9706.

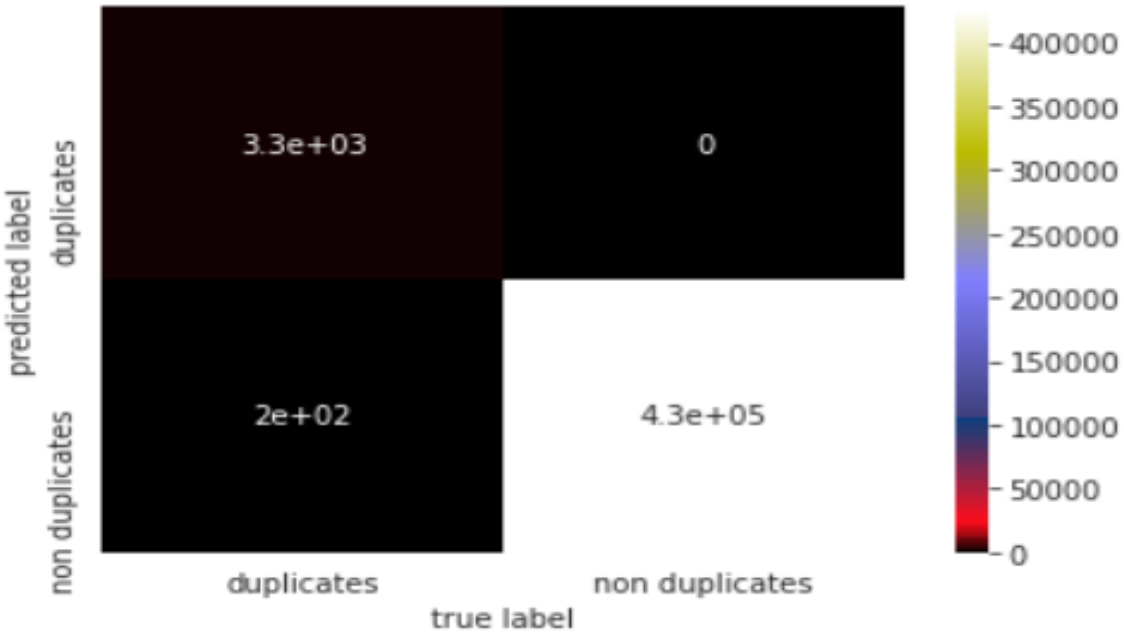


Figure 4.11 NBC Confusion matrix

Table 4.4 Table showing results of the ML algorithms

	Accuracy	**Precision**	**Recall**	**F1 score**
SVM	99.9%	95.6%	1	0.9778
NBC	99.9%	94.3%	1	0.9706

		Predicted SVM		Predicted NBC	
Actual	3450	3300 (TP)	150 (FP)	3200 (TP)	250 (FP)
	430000	430000 (TN)	0 (FN)	43000 (TN)	0 (FN)

SVM

$$Precision = \frac{True\ positive}{(True\ positive + False\ positive)}$$

$$= \frac{3300}{3300+150}$$

$$= \underline{95.7\%}$$

$$Recall = \frac{True\ positive}{(True\ positive + False\ negative)}$$

$$= \frac{3300}{3300+0}$$

$$\underline{=1}$$

$$Accuracy = \frac{True\ positive + True\ negative}{(True\ positive + False\ positive + False\ negative + True\ negative)}$$

$$= \frac{3300+430000}{3300+150+0+430000}$$

$$= \underline{99.9\%}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

$$= \frac{430000}{430000+150}$$

$$= \underline{99.9\%}$$

$$F\ score = \frac{2\ (Precision * Recall)}{(Precision + Recall)}$$

$$= \frac{2(95.7)}{95.7+1}$$

$$= \underline{0.9778}$$

NBC

$$Precision = \frac{True\ positive}{(True\ positive + False\ positive)}$$

$$= \frac{3300}{3300+200}$$

$$= \underline{94.3\%}$$

$$Recall = \frac{True\ positive}{(True\ positive + False\ negative)}$$

$$= \frac{3300}{3300+0}$$

$$= \underline{1}$$

$$Accuracy = \frac{True\ positive + True\ negative}{(True\ positive + False\ positive + False\ negative + True\ negative)}$$

$$= \frac{3300+430000}{3300+200+0+430000}$$

$$= \underline{99.9\%}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

$$= \frac{430000}{430000+200}$$

$$= \underline{99.9\%}$$

$$F\ score = \frac{2\ (Precision * Recall)}{(Precision + Recall)}$$

$$= \frac{2(94.3)}{94.3+1}$$

$$= \underline{0.9706}$$

This chapter presented the results of the record linkage and de-duplication process that was explained in Chapter 3 and gave descriptive results of the data. The next chapter discusses and interprets the results and makes deductions and inferences from the results.

5. DISCUSSION

5.1 Introduction

For any intervention program to be effective and efficient, there is need for accurate data to inform policies and allocation of resources. This study was motivated by the need to link and de-duplicate the data for sex workers registered in Zimbabwe to improve the quality and accuracy of the data for the sex work program. The chapter first gives descriptive statistics of the data sets that were used in the study, then considers the discussion of the findings identified in Chapters 4. The chapter concludes with recommendations for future studies and the limitations of the present study.

The aim of this study is to apply supervised machine learning approaches to match and de-duplicate records in the sex work registers at Sisters' clinics in Zimbabwe. This would enable accurate enumeration of unique individual sex workers that were registered at the clinics between 2017 and 2019. This study also seeks identify the best model for de-duplicating the data and how this study can contribute towards the use of program data from the sisters' clinics to correctly estimate the population size of sex workers in Zimbabwe.

5.2 Descriptive statistics of the data sets

Among the descriptive statistics that the study looked at was distribution of gender. The study showed that 98.8% of the total number of clients registered at the clinics were females, 0.8% were males, while 0.4% were transgender. The proportion of females in the program could be attributed to that when the sex work program started in 2009 it was only serving female sex workers until 2018 when the program expanded to serve sex workers in their diversity. The distribution of gender maybe be an indication of the proportion of male and transgender sex workers in Zimbabwe as compared to female sex workers. Moreover, it could imply poor uptake of health services by the male and transgender sex workers as compared to their female counter parts in the sex work community. Despite the efforts that have been made by CeSHHAR of engaging male and transgender sex workers as part of the outreach team for demand creation, we still witnessing low numbers with these type of sex workers. This suggests a need to investigate why there are few male and transgender sex workers visiting the sex work clinics.

A comparison of age showed that although there was an overall increase in the number of sex workers registered in both age groups, the proportion of young women selling sex (YWSS) aged between (16-24) was lower (26%) between 2009 to 2016 as compared to 2017 to 2020 (35%). The increase in number of YWSS who were registered at the sister's clinics might have

been influenced by the Determined Resilient Empowered AIDs Free Mentored Sister (DREAMS) intervention which targeted YWSS in the ages of 16-24 from 2016 in Mutare, Bulawayo, Gweru, Mazowe and Chipinge. The DREAMS intervention used a strategy of engaging peer educators who were within the 16–24-year age group to reach out to their peers and encourage the uptake of sexual reproductive health services among other interventions within the DREAMS services. Apart from this, the statistics could be an indication of an increase in the number of YWSS due to various reasons like the death of parents which tends to affect the girl child mostly. The young women end up taking the role of guardian for themselves and their sibling, with no guardian, no source of income and no education. The burden to provide for the family might be a push factor that forces the orphaned young women into sex work while some of the young women run away from home to escape poverty and end up in sex work. The overall increase in both age groups could be an indication that the sex work program was making strides in reaching sex workers which could also be attested by the increase in the number of clinics supporting the program between 2016 and 2019.

The reported marital status shows that 62% of the sex workers registered at the sisters' clinics were divorced, 25% never married, 10% widowed, 2% currently married and 1% of the sex workers were separated. These descriptive statistics suggests that most of the women have been negatively affected by harsh socio-economic factors that have triggered them to join sex work as a form of financial means. Furthermore, these findings could indicate that these are middle aged women, possibly single mothers raising children on their own who having lost financial support from their partners might have opted to enter into sex work to fend for themselves and their children. The high divorce rate could also be an indication that the women were sex workers who got into "short marriages" as they are referred to among the sex workers just for financial gain and got divorced when the partners were no longer able to meet their financial needs or due to gender based violence (GBV) which is common in these living arrangements as reported by the sex workers. The death of spouses as bread winners, might have led to financial insecurity and changed the lifestyle of the women and their children. This coupled with the psychological effects of the loss might have forced the women into sex work. There is therefore a need for intervention programmes to empower and support divorced and widowed women who do not have the capacity to fend for themselves with skills that can help them generate some income for their families.

The distribution of level of education at registration depicts that less than 2% of the sex workers had tertiary education which is a minimum requirement for most jobs in formal employment.

With 77% of the sex workers having secondary level education, 21% primary level, and 1% having no education at all, their chances of getting any formal employment are slim hence they are negatively affected social, financially and otherwise. The high unemployment rates in the country combined with the lack of education and capital to start income generating projects, can be triggers of sex work. These uneducated women with few opportunities end up having no options of survival and as such, they are forced to enter sex work to try and make a living.

An analysis of the link log data by site showed that more than 27% of the sex workers registered at the sisters' clinics were registered in Harare followed by Bulawayo at 13%. These high numbers reported for capital cities can possibly be attributed to the fact that Harare and Bulawayo are big cities in Zimbabwe. Consequently, there is dominant economic activity thus sex workers possibly move from the rural areas and smaller towns for sex related work in these big cities. Furthermore, an interesting observation is that there is a high hive of sexual work (mobile sites that were among the top fifteen in registrations) at border sites such as Beitbridge and Chirundu. This could be explained by the fact that these two towns are border towns where there is a hype of activity with truck drivers who are major clients of sex workers. In contrast although Plumtree is also a border town, it is among the bottom three sites in registration due to low volumes of sex work at the border site and this led to a closure of this clinic in 2016 as it was no longer economically feasible to have a team travel to Plumtree to offer services to few sex workers. The data implies that there is more trade between South Africa and Zambia as compared to Botswana, thus the low activity at the border post. Other sites had lower numbers registered as the clinics were only opened once per week and in some sites once per fortnight for shorter hours as compared to the static sites where the clinics are open Monday to Friday from 8am to 4.30pm.

The link log data set had 399 duplicated UICs before linking with the demographic data set. This suggests that the actual number of sex workers registered at the clinics could be lower than the reported figure. However, due to some clients changing their details at registration the assigned `UIC` and `Idlabel` which are unique identifiers in this data were also different. This made it difficult to identify the duplicates using traditional methods thus the need to employ machine learning techniques in this study.

From the above data one can conclude that poor socio-economic conditions, poverty and the harsh economy is one of the main causes that pushes women into sex work in a bid to provide for their families. The high unemployment rate coupled with low level of education means that most of the women who enter sex works have limited option for generating income as their

chances for securing a job are small and they end up resorting to sex work as a means of surviving. However, for us to better understand the reasons behind the increase in the number of sex workers, the sex work programme needs to consider routinely collecting data on reasons why women entered sex work and age at which they started sex work. This information will guide and inform the government to come up other intervention programmes to alleviate the number of women who end up in sex work. An increase in the number of sex worker might lead to a public health crisis where the country has an increase in incidence rate of HIV thus reversing the effort being made towards achieving epidemic control of HIV.

5.3 Record linkage

For the study to be able to fulfil its aim of record linkage and de-duplication, three research objectives were set, the first objective was to implement a record linkage framework to de-duplicate registers of the female sex workers registered at the Sisters with a voice clinic in Zimbabwe between 2017 and 2019 using supervised machine learning techniques. This objective was achieved through the record linkage process that was done using Python record linkage toolkits where matching was done for the demographic and link log datasets. Furthermore, this investigation managed to identify 3274 records as matches when the link log data set and the demographic data set were matched. Moreover, after the de-duplication this study managed to enumerate and ascertain the number of individual sex workers and find out the precise rate of repeat visits by sex workers registered at Sisters with a voice clinics in Zimbabwe. This implied that out of the total 40 507 clients who were registered in the demographic data between 2017 and 2019 less the 8% duplicate records, we got a more realistic reduced number of 37 223 individual sex workers registered after the de-duplication process.

The identified matches were used as labels and a training data set was created. SVM, NBC algorithms were fit into the models. The confusion matrix was used to evaluate the performance of the 2 algorithms to determine which one performs better. The algorithms were evaluated on Precision, Accuracy, and Recall and F score. Although from the results, SVM and NBC has similar Accuracy rates of 99.9% and a Recall of 1, SVM outperformed NBC with a Precision of 95,6% and an F score of 0.9778.

Validation analysis was done to measure the performance of the best optimum ML algorithm in matching as compared to the gold standard. The results of the evaluation suggest that the SVM was the best performing optimum ML algorithm as it had a higher Precision rate of 95.6% and low rate of false positives than NBC. The high evaluation results of the validation process could be attributed to the fact that the record linkage process was done using deterministic

approaches, the datasets where classical (not much overfitting or underfitting) and the ability of the SVM to handle non-linearity of the data.

An analysis of the duplicates by site showed that more than 50% of the duplicate records were in Harare. This could be explained by sex worker mobility where a sex worker could have first registered in another facility then re-registered in Harare or vice versa. As sex workers are not keen on being identified, there is likelihood that one might have re- registered at the same facility in Harare using a different name. Results showed that smaller towns had insignificant proportion of duplicate records with some sites having no duplicate records. This could be attributed to the facilities having few clients as such chances of being identified as a repeat client are higher as compared to high traffic facilities like the clinics in most static sites. Therefore future work can possibly focus on implementing biometric authentication techniques to reduce or avoid such possible duplications.

5.4 Limitations of the study

There are potential limitations to this study which could possibly have affected the results. The first limitation is to do with the sample size. Although the intention was to use all data from 2009 to June 2020, data collected prior to 2017 only had the date of birth as a common variable between the data sources. The missing variables meant that the study would not use most of the data collected before 2017 to link clients as there were limited variables to link the two data bases. This might have affected the rate of duplicate records as some clients might have registered for the first time or repeatedly before 2017. A second potential limitation is that in the data from 2017, there was missing data and some invalid data in the database for the variables of interest with some values being dropped during the data cleaning process where the values were invalid. The missing data might have limited the algorithms from performing optimally. Another limitation is the quality of the data being matched and de-duplicated, the identified unique identifier of `Idlabel` and `UIC` might not be unique after all since there was a likelihood of assigning one sex worker more than `Idlabel` or `UIC` should a sex work change all their details in re registration.

6. CONCLUSION AND RECOMMENDATIONS

The intent of this study was to match and de-duplicate records in the link log and demographic data sets for sex workers registered at Sisters' clinics in Zimbabwe. The aim of this study was fulfilled as results showed that the database had an 8% duplicate rate which could be used to adjust on previously reported data. The research objectives were to assess the performance of SVM and NBC in record linkage and de-duplication and to select an optimal algorithm in matching with the gold standard. The results presented in Chapter 4 show that SVM algorithm performs better than NBC in linking and matching data with an Accuracy of 95.6%.

Although the generalisation of the current results must be established by future research, the results of this study has provided enough evidence for us to conclude that SVM performs better than NBC in record linkage. With this in mind, I would recommend that the SVM model can be integrated and sit on an Application Programming Interface (web application linked to the databases) using the link log database. This will provide real-time linking and matching during registration. The new system could also implement the health level seven (HL7) Fast Health care interoperability (FHIR) standards which enables data sharing among health care systems. FHIR defines how healthcare information can be exchanged between different systems and enables fast, secure access of data to the authorised individuals. The FHIR standard would support real time data linkage and reduce likelihood of duplication. The other recommendation would be to have a biometric registration embedded on the linklog electronic system to generate the unique identifier code instead of using algorithms that are depended on the information provided by the sex worker. However, with biometrics being a sensitive subject especially among key populations, a feasibility and acceptability study will need to be done first before implementing biometrics for registration at the sister’s clinics.

Future investigations should take all the possible limitations into consideration. The performance of the supervised machine learning models needs to be examined with a bigger sample that has few missing information to assess the effects of the missing data on the results of this study. Considering the quality of the data and the limitation of the unique identifier highlighted, I would also recommend running the same study using unsupervised machine learning techniques and comparing the results with those obtained from this study. It would be useful to extend the current findings by implementing other record linkage techniques to try and link the data for all sex workers registered into the program since program inception (2009) to get more accurate data that can contribute to the size estimation of sex workers in Zimbabwe. In summary support vector machine learning is better than naïve bayes classifier in record

linkage. This study managed to do the record linkage and ascertain the number of repeat visits of sex workers. Evaluation of models was done and achieved high validation analysis scores. As such, the results of this study can prove beneficial in terms of health priorities, intervention programs, informing policy and practice and this can positively impact the mobilisation of resources to address issues around sex work in Zimbabwe and the global world.

7. REFERENCES

ANSOLABEHERE, S. H., EITAN D. 2017. ADGN: An Algorithm for Record Linkage Using Address, Date of Birth, Gender, and Name. *Statistics and Public Policy,* 4**,** 1-10.

AVERT. 2018. *Sex workers, HIV and AIDS* [Online]. Available: https://www.avert.org/printpdf/node/384 [Accessed 11 July 2019].

BIJU, S. M. 2008. Agile Software Development. *E-Learning and Digital Media,* 5**,** 97-102.

BRUIN, J. D. 2015. *Probabilistic record linkage with the Fellegi and Sunter framework Using probabilistic record linkage to link privacy preserved police and hospital road accident records.* Master of Science in Applied Mathematics, Delft University of Technology.

BRUIN, J. D. 2016. Available: https://recordlinkage.readthedocs.io/en/latest/about.html [Accessed 11 July 2019].

BRUIN, J. D. 2019a. *Record linkage* [Online]. Available: https://pypi.org/project/recordlinkage/ [Accessed 26 July 2019].

BRUIN, J. D. 2019b. Record Linkage Toolkit Documentation.

CESHHAR. 2019. *http://ceshhar.org/our-work/key-populations/* [Online]. Available: http://ceshhar.org/our-work/key-populations/ [Accessed 08 July 2019].

CHRISTEN, P. 2007. A Two-Step Classification Approach to Unsupervised Record Linkage. *6th Australasian Data Mining Conference.* Australia.

CHRISTEN, P. 2008. Automatic Record Linkage using Seeded Nearest Neighbour and Support Vector Machine Classification.

CHRISTEN, P. & GOISER, K. N.D. Quality and Complexity Measures for Data Linkage and Deduplication.

CHURCHES, T., CHRISTEN, P., LIM, K. & ZHU, J. X. 2002. Preparation of name and address data for record linkage using hidden Markov models. *BMC Medical Informatics and Decision Making,* 2**,** 9.

DANNELOV, J. 2018. Study on Record Linkage regarding Accuracy and Scalability

DUSETZINA, S., TYREE, S., MEYER, A., MEYER, A., GREEN, L. & CARPENTER, W. 2014a. Data for Health Services Research: A Framework and Instructional Guide. Agency for Healthcare Research and Quality: Agency for Healthcare Research and Quality.

DUSETZINA, S., TYREE, S., MEYER, A., MEYER. A, GREEN, L. & CARPENTER, W. 2014b. Linking data for health services research. *Agency for Healthcare Research and Quality*.

EUROPA N.D. Methodology Data linkage.

FELLEGI, I. P. & SUNTER, A. B. 1969. A theory for record linkage. *Journal of the American Statistical Association*.

FENG, C. 2016. Improve Record Linkage Using Active Learning Techniques.

GROOTHEEST, G. G. V., DE GROOT, M. C. H. M., ., VAN DER LAAN, D. J. J., SMIT, J. H. J., . & BAKKER, B. F. M. B. 2015. Record linkage for health studies: three demonstration projects. 71.

GURUSAMY, V. & KANNAN, S. 2014. PreprocessingTechniques forText Mining.

HAND, D. & CHRISTEN, P. 2018. A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing,* 28**,** 539-547.

HARRON, K., DIBBEN, C., BOYD, J., HJERN, A., AZIMAEE, M., BARRETO, M. L. & GOLDSTEIN, H. 2017. Challenges in administrative data linkage for research. *SAGE*.

HUANG, S., CAI, N., PACHECO, P. P., NARRANDES, S., WANG, Y. & XU, W. 2018. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics,* 15**,** 41-51.

KABUDULA, C., JOUBERT, J., TUOANE-NKHASI, M. & KAHN, K., RAO, C. 2014. Evaluation of record linkage of mortality data between a health and demographic surveillance system and national civil registration system in South Africa. *Population Health Metrics,* 12.

KABUDULA, C.W., Clark, B.D., Gómez-Olivé, F.X. et al. 2014. The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa. BMC Med Res Methodol 14, 71.

KELLEY, C. & MCKAY, M. 2011. *Coconut* [Online]. RTI International. Available: http://chrisekelley.github.io/coconut/ [Accessed 21 December 2019].

LAVRAKAS, P. J. 2008, *Encyclopedia of survey research methods*, vol. 0, Sage Publications, Inc., Thousand Oaks, CA, [Accessed 20 February 2021]

LUND RESEARCH 2012. Total population sampling available on https://dissertation.laerd.com/total-population-sampling.php [Accessed 20 February 2020].

MACHADO, C. J. 2004. A literature review of record linkage procedures focusing on infant health outcomes.

MERRIAM-WEBSTER. N.D. Available: https:/www.merriam-webster.com/dictionary/pseudonym [Accessed 24 January 2019].

MITCHELL, T. M. 2017. GENERATIVE AND DISCRIMINATIVE CLASSIFIERS: NAIVE BAYES AND LOGISTIC REGRESSION. *Machine Learning.* Second ed.: McGraw Hill.

NDORI-MHARADZE, T. 2018. Differentiated prevention and HIV testing for Female Sex. *IAS.*

OLIVEIRA, G. P., BIERRENBACH, A. L., CAMARGO, K. R. J., COELI, C. M. & PINHEIRO, R. S. 2016. Accuracy of probabilistic and deterministic record linkage: the case of tuberculosis. *Rev Saude Publica,* 50**,** 49.

OVERS, C. 2002. An analysis of HIV prevention programming to prevent HIV transmission during commercial sex in developing countries

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A. & VINCENT MICHEL 2011. Scikit-learn Machine Learning in Python.

RAMESH, B. & SATHIASEELAN, J. G. R. 2015. An Advanced Multi Class Instance Selection based Support Vector Machine for Text Classification. *Procedia Computer Science,* 57**,** 1124-1130.

RASCHKA, S. 2015. Python machine learning. Packt Publishing Ltd.

UNAIDS. 2011. Developing and Using Individual Identifiers for the Provision of Health Services including HIV 24–26 February 2009.

WINKLER, W. E. 2014. Matching and record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics,* 6**,** 313-325.

ZHU, Y. E. A. 2015. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *Journal of Biomedical Informatics,* 56**,** 80-86.

ZIMMERMAN, D. & KANG, K. ND. Clustering LInkage Functions in Record Linkage.

YANG, S. 2019. An Introduction to Naïve Bayes Classifier

Appendices

Appendix A Plagiarism declaration



PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I **Sithembile Musemburi** (Student number: **1179538)** am a studentregistered for the degree of Public Health Informatics in the Wits School of Public Health in the academic year 2021.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: _____ Date: 20 April 2021

Appendix B MRCZ Ethics Clearance

Telephone: 08644073772/791193
E-mail: mrcz@mrcz.org.zw
Website: http://www.mrcz.org.zw

Medical Research Council of Zimbabwe
Josiah Tongogara / Mazowe Street
P. O. Box CV 573
Causeway
Harare

APPROVAL

REF: MRCZ/E/266 **5 February, 2020**

Prof. F. Cowan
Dr F. Machingura
CESHHAR
9 Monmouth Road
Avondale West
Harare

RE: Secondary analyses of anonymised programme data collected through the Sisters with a Voice Programme by CeSHHAR Zimbabwe on behalf of MoHCC and NAC.

Thank you for the application for review of Research Activity that you submitted to the Medical Research Council of Zimbabwe (MRCZ). Please be advised that the Medical Research Council of Zimbabwe has **reviewed** and **approved** your application to conduct the above titled study.

Thisapproval is based on the review and approval of the following documents that were submitted to MRCZ for review: -

1. Study protocol

- **APPROVAL NUMBER** : **MRCZ/E/266**

This number should be used on all correspondence, consent forms and documents as appropriate.

- **TYPE OF MEETING** : **EXEMPT**
- **APPROVAL DATE** : **5 February, 2020**
- **EXPIRATION DATE** : **4 February, 2021**

After this date, this project may only continue upon renewal. For purposes of renewal, a progress report on a standard form obtainable from the MRCZ offices should be submitted three months before the expiration date for continuing review.

- **SERIOUS ADVERSE EVENT REPORTING**: All serious problems having to do with subject safety must be reported to the Institutional Ethical Review Committee (IERC) as well as the MRCZ within 3 working days using standard forms obtainable from the MRCZ Offices or website.
- **MODIFICATIONS**: Prior MRCZ and IERC approval using standard forms obtainable from the MRCZ Offices is required before implementing any changes in the Protocol (including changes in the consent documents).
- **TERMINATION OF STUDY**: On termination of a study, a report has to be submitted to the MRCZ using standard forms obtainable from the MRCZ Offices or website.
- **QUESTIONS**: Please contact the MRCZ on Telephone No. (0242) 791193, 0864407377203 or by e-mail on mrcz@mrcz.org.zw

Other

- Please be reminded to send in copies of your research results for our records as well as for Health Research Database.
- You're also encouraged to submit electronic copies of your publications in peer-reviewed journals that may emanate from this study.
- In addition to this approval, all clinical trials involving drugs, devices and biologics (including other studies focusing on registered drugs) require approval of Medicines Control Authority of Zimbabwe (MCAZ) before commencement

Yours faithfully

MRCZ SECRETARIAT
FOR CHAIRPERSON
MEDICAL RESEARCH COUNCIL OF ZIMBABWE

MEDICAL RESEARCH COUNCIL OF ZIMBABWE
2020 -02- 05
APPROVED

PROMOTING THE ETHICAL CONDUCT OF HEALTH RESEARCH

Appendix C List of variables

Idlabel

Visitdate

Date of birth

Age

Gender

Marital status

Education level

UIC

Site

Appendix D Ethical clearance certificate -Wits

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Mrs Sithembile Musemburi

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M200841

NAME: (Principal Investigator)	Mrs Sithembile Musemburi
DEPARTMENT:	School of Public Health Center for Sexual Health HIV and Aids Research
PROJECT TITLE:	Record linkage to de-duplicate sex-worker registers at the sex-work clinics in Zimbabwe using supervised learning techniques
DATE CONSIDERED:	28/08/2020
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Mr Michael Mapundu
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	31/08/2020

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions.Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. I **agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **August** and will therefore be due in the month of **August** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix E Approval letter to access CeSHHAR data

The Centre for Sexual Health and HIV AIDS
Research Zimbabwe (CeSHHAR Zimbabwe)
4 Bath Road, Belgravia
Harare, Zimbabwe

Phone +263 (0) 4 333393

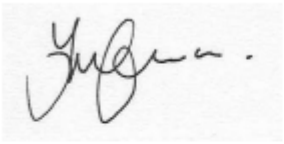
4th June 2020

To whom it may concern

Re: Permission to use the programme data from the Sisters with a Voice programme for female sex workers for a project entitled **"Record linkage to de-duplicate sex-worker registers at the sex-work clinics in Zimbabwe using supervised learning techniques".**

This letter provides permission for Sithembile Musemburi to use programme data from the Sisters with a Voice programme for female sex workers for her MSc Epidemiology – Public Health Informatics dissertation project at University of Witwatersrand. The data have been collected from female sex workers when they attend the programme for clinical services. The project is important and will provide useful information for the programme.

Yours faithfully



Frances M Cowan
Executive director, CeSHHAR Zimbabwe
Professor of Global Health, Liverpool School of Tropical Medicine.

Appendix F Originality report

