UNIVERSITY OF GHANA
COLLEGE OF HEALTH SCIENCES
SCHOOL OF PUBLIC HEALTH

PERFORMANCE COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTING HIV STATUS AMONG FEMALE SEX WORKERS IN GHANA

DENNIS ADJEI ANNANG

DEPARTMENT OF BIOSTATISTICS

OCTOBER 2018

UNIVERSITY OF GHANA
COLLEGE OF HEALTH SCIENCES
SCHOOL OF PUBLIC HEALTH

PERFORMANCE COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTING HIV STATUS AMONG FEMALE SEX WORKERS IN GHANA

BY

DENNIS ADJEI ANNANG
(10242975)

A PRACTICUM REPORT SUBMITTED TO THE UNIVERSITY OF GHANA SCHOOL OF PUBLIC HEALTH IN PARTIAL FULFILMENT OF THE AWARD OF DEGREE OF MSc HEALTH INFORMATICS

DEPARTMENT OF BIOSTATISTICS

OCTOBER 2018

DECLARATION

I hereby declare that, except for other people's works, which have been duly acknowledged, this work is the result of my own research. This project work, either in whole or in part has not been presented elsewhere for another degree.

Mr. Dennis Adjei Annang	Dr. Samuel K. K. Dery
(Student)	(Supervisor)
……	……..
Signature	Signature
…….	……
Date	Date

ABSTRACT

Introduction: The Human Immunodeficiency Virus (HIV) and the Acquired Immunodeficiency Syndrome (AIDS) disease still remains a global public health issue. An intriguing observation is the increasing rate of the infection among Female Sex Workers (FSW). HIV Testing Services (HTS) is an essential entry point for any HIV intervention service for FSW and studies like the Integrated Bio-Behavioral Surveillance Surveys (IBBSS) anonymously link the HIV test results of FSW with their respective socio-demographic and behavioral characteristics. These studies report summaries of HIV status of FSW with their respective socio-demographic and behavioral variables using traditional statistical methods. This approach limits information required to scale up knowledge on HIV status among FSW in Ghana hence innovative solutions like data mining is needed to explore approaches to the prediction of HIV from available information.

Objective: The purpose of this paper is to develop a data mining solution that predicts the HIV status of FSW using socio–demographic and behavioral characteristics.

Methodology: The approach adapted was the CRISP–DM which followed six main steps: business understanding, data understanding, data preparation, modeling, evaluation and deployment. Ghana's 2015 FSW IBBSS data set was used for the study. Microsoft Excel was used for data preparation and WEKA 3.6.9 was used as data mining tool to implement experimentations using 5 algorithms: Random Tree, J48, Naïve Bayes, Logistic regression and Neural Network.

Results: The target dataset contained 3,092 female sex worker participants. Out of which 2,491 (80.56 %) of the FSW were roamers while the remaining 601 (19.44%) were seaters. The age ranged from 16 to 64 years old. The study showed that Random Tree classifier out of the five classifiers was the best classifier to predict HIV status with an accuracy of 98.9%. Age, highest educational level, marital status, and average income from sex work, sex work experience, relationship with most recent sexual partner, HIV prevention knowledge, number of sex partners in a week, frequency of condom use among paying partners, had HIV test before, condom use by paying client, religion, had anal sex, drug usage, drank alcohol before sex and FSW type were found to be predictors of HIV status of FSW. The association rule extracted showed that there is direct relationship between having anal sex and the HIV status of a FSW.

Conclusion: The results from the study proved that data mining can be used to extract relevant information for HIV prediction for FSW and that socio demographic and behavioral attributes are sufficient enough to predict HIV status of a FSW.

DEDICATION

This project work is dedicated to the Almighty God who has given me life, good health and wisdom to enable me undertake this project work. Also to my lovely wife Eunice Annang, my mother Agnes Annan and my brothers Derrick Annang and Daniel Annang for their unflinching support and inspiration.

ACKNOWLEDGEMENT

I acknowledge the invaluable contributions of my supervisor, Dr. Samuel K. K. Dery to this project work. Thanks and appreciation also goes to the entire lecturers in the Biostatistics Department of the School of Public Health, University of Ghana whose support provided the technical guidance to the success of this practicum.

Many thanks go to Mr. Kyeremeh Atuahene, Director – General of Ghana AIDS Commission (GAC), Mrs. Jewel Lamptey, Research Coordinator of GAC, Miss. Cynthia Asante, Data Manager of GAC and Mr. Isaiah Doe Kwao, Data Quality Assurance of GAC.

I am grateful to my family for all the support while I was in school. Special thanks go to the management of the Ghana AIDS Commission for granting me this opportunity to undertake this course.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AIDS	-	Acquired Immune Deficiency Syndrome
ART	-	Antiretroviral Therapy
CDC	-	Centre for Disease Control
CRISP-DM	-	Cross Industry Standard Process for Data Mining
FSW	-	Female Sex workers
GAC	-	Ghana AIDS Commission
HIV	-	Human Immunodeficiency Virus
HTS	-	HIV Testing Services
IBBSS	-	Integrated Behavioural Bio Surveillance Study
KDD	-	Knowledge Discovery in Database
KP	-	Key population
MSM	-	Men who have Sex with Men
Non-PP	-	Non Paying Partners
NSP	-	National Strategic Plan
PLHIV	-	People Living With HIV
PWID	-	People Who Inject Drugs
SEMMA	-	Sample Explore Modify Model Assess
SMOTE	-	Synthetic Minority Over-sampling Technique
STI	-	Sexually Transmitted Infection
WHO	-	World Health Organisation
UNAIDS	-	Joint United Nations Programme on HIV/AIDS

CHAPTER 1

INTRODUCTION

1.1. Study Background

The Human Immunodeficiency Virus (HIV) and the Acquired Immunodeficiency Syndrome (AIDS) disease still remain a global public health issue and a key hurdle for any development agenda. A higher burden of both the infection and the disease has an adverse effect on both political and socio-economic sectors for any country. The disease has contributed significantly to about 35.0 million deaths from AIDS related illnesses as at 2016 with sub-Saharan Africa receiving the greatest impact of these deaths.

HIV attacks the immune systems specifically the CD4 cells designed to fight off certain kinds of infections including tuberculosis, cancers or involuntary weight loss (CDC, 2018). An individual is said to have the HIV infection if he or she tests positive and is classified as having AIDS when his or her immune system becomes weak and infections take over the body system (U.S. Department of Veterans Affairs, 2018). There is currently no cure for the disease and it is transmitted through various means including from mother to child during the process of conceiving, delivery or breastfeeding, unprotected sexual intercourse, sharing infected needles and contaminated blood transfusions (UNAIDS, 2010) .

Since its advent, about 76 million individuals, made up of children and adults, have been infected with the virus. Globally, it is estimated that about 37 million people, made up of about 2.1 million children, were living with HIV as at the end of 2016. Sub-Saharan Africa countries including Ghana, are the worst hit with a vast majority of People Living with HIV (PLHIV) which accounts for about two-thirds (69%) of PLHIV population worldwide (Ogaji, A.S, & I, 2013). The global HIV prevalence among adults aged 15–49 is estimated at 0.8%.

The UNAIDS (2017) estimates global new HIV infections at 1.8 million in 2016 with sub-Saharan Africa contributing about 44%. This represents a sharp decline from the 2015 figure of about 2 million. New infections among children below 15 years is estimated at about 160,000 with most infections occurring as mother-to-child transmission during pregnancy, childbirth or breastfeeding. It is also estimated that new infections among the general population has significantly declined by 11% and 47% from 2010 to 2016 for adults and children respectively (UNAIDS, 2017).

Although there is a global reduction in new infections, research has shown that more is required to improve HIV knowledge and testing among the various population groups. An intriguing observation is the increasing rate of HIV infection among key population subgroup. Key population is a sub population made up of groups of individuals said to be most at risk or vulnerable to getting the HIV infection. This is made up made up of Female Sex Workers (FSW), Men who have Sex with Men (MSM), transgender and People Who Inject Drugs (PWID) (WHO, 2013). They account for about 15% to 25% of all new infections globally (Prüss-Ustün et al., 2013). This requires interventions to target high risk populations especially in high HIV burden countries to control the spread of the disease. This is because in many countries about 60% of all new infections is attributed to the risky behaviours of this group. (Fettig, Swaminathan, & Murrill, 2014).

Globally, deaths related to AIDS has dropped significantly by 48% since 2005 and it is a sharp contrast for the age group 10-19 years who have witnessed an increase in death by 50% (Kabiri, 2016). The introduction of antiretroviral therapy has significantly reduced deaths related to HIV and AIDS as 54% of PLHIV across the world had access to treatment in 2016 (UNAIDS, 2017). This achievements has in effect dropped HIV and AIDS from the global top 10 causes of death (WHO, 2018). Research has shown that identifying and focusing on the major sources

of new HIV infections are pivotal to any prevention efforts as these tend to vary for many countries. Fettig, Swaminathan, Murrill, & Kaplan (2014) argue that new infections and modes of transmission for HIV are mainly due to sex work, needle sharing and among stable couples. This is the case for most countries in sub-Saharan Africa where unprotected heterosexual intercourse has been identified as the main HIV transmission mode. (UNAIDS, 2013). Consequently, unique prevention guides are required to address specific vulnerability patterns for all countries and emphasis placed on individuals and sub populations engaged in high risk activities.

A prominent sub population of interest to new HIV infections and modes of transmission is the key population group. HIV epidemic among these groups are defined as a concentrated epidemic since HIV spreads rapidly among these sub groups and they have prevalence greater than 5% (UNICEF, 2008).Female Sex Workers (FSW) are considered as a highly vulnerable to HIV infection due to the nature of their work which is summarized in the relationship that exist between the risks associated with HIV transmission and the number of sexual clients of FSWs. (Fettig, Swanminathan, Murrill, 2014). Sex work is defined as "the provision of sexual services in exchange for money, goods, or other benefits" (Prüss-Ustün et al.,2013). This definition covers all women selling sex for cash and for in-kind benefits.

FSWs are highly stigmatized population and because their trade is outlawed in Ghana, they are reluctant to disclose their profession in most cases making them hard to reach. According to Boily et al. (2009) unsafe sex work contributes to about 15% of HIV infections among adult female population, and sub-Saharan Africa has the highest fraction. Available research also suggests that there is a high prevalence among women engaged in the sex trade against women who do not and also significantly higher prevalence in men who patronize the services of FSWs as compared to men who do not (Fettig et al., 2014).

In general, FSWs contribute to the spread of the HIV either within the same population or other key population group and subsequently these infected partners of FSWs then infect their respective spouses. Low and middle income countries have an estimated HIV prevalence of 12% for FSWs with an odds ratio of 13.5 as related to women in their reproductive age (Baral et al., 2012). It is further estimated that, globally, about 106,000 AIDS related deaths can be attributed to FSW with about 98,000 of these occurring in Africa (Prüss-Ustün et al., 2013). Other risky behaviours of FSWs include injecting drugs hence making them more vulnerable to HIV transmission. Consequently, there is the need for a composite HIV intervention package which safeguards the life of FSWs and their clients and partners.

HIV Testing Services (HTS) can be described as the doorway into HIV prevention and treatment service. Globally, HTS is recognised as one of the interventions necessary for realising the UNAIDS flagship 90-90-90 targets toward ending AIDS by 2030 in line with the Sustainable Development Goals. HTS plays an important role in encouraging safer sex practises especially for individuals who test negative (Sherr, Lopman, Kakowa, & Dube, 2007). Research has also established that getting people to access HTS as early as possible is very effective in ensuring low risky behaviours but this effectiveness is highly dependent on availability and accessibility of HTS by various target populations (Foster et al., 2017). It is therefore essential to encourage uptake of HTS by both the general and key populations because high HIV infection rates results in high and increased rates in morbidity, mortality and transmission (Foster et al., 2017).

High HIV prevalence among FSWs can be attributed to the risky nature of sex work and to their lack of knowledge about their HIV status thus making it very difficult to plan adequately for them. A study by Elhadi et al., (2013) concluded that HIV programs and interventions for FSW must include and emphasis on HIV Testing Services. Though FSWs are continuously

been engaged to increase their knowledge in HIV prevention approaches and HTS, only a handful have access to these services. It is reported that about 64% of FSWs globally actually are aware of their status and 43% of them are being treated (AIDS MAP, 2018). This requires an integrated approach to increase uptake of HIV testing among FSWs and their enrolment into care. The key barriers to the low uptake of HTS by FSWs are stigma and discrimination, perceived low risk, cultural beliefs and the implications of the test results.

Behavioural Surveillance Surveys (BSS) serve as one of the information guides to design relevant HIV prevention interventions for key populations who may be hard to reach due to the nature of their work (Brown & Mills, 2000) . Information from these surveys shape how strategic documents are developed to manage the national response to HIV for many countries. Surveys like this provide appropriate methods to track risky behaviours of key population who have a high probability of being infected with HIV and further infecting others (Ghana AIDS Commission, 2015) . An increasing HIV prevalence is an indication that the HIV prevention programs are failing.

A critical benefit of BSS is its ability to anonymously link the HIV test results of a respondent and in this case a member of the key population with their respective socio-demographic, and behavioral characteristics (Global Network of Sex Work Projects, 2015). Most BSS studies for FSWs report the total HIV positive and negative FSWs in relation to socio-demographic and behavioural characteristics. As a result, voluminous data are collected on HTS among this sub population where HIV program based information are churned through the use of traditional statistical methods. A limitation to these traditional statistical methods is that it has less capacity to discover new and unanticipated patterns in the data set (Nicole & Tim, 2006).

It is therefore essential for policy makers and programmers to devise other ways of using available data to scale up knowledge of HIV status among FSWs since testing is the only means of knowing one's HIV status. To achieve this, there is the need to employ innovative solutions which include the ability to predict HIV status of FSWs from available information which includes information from Behavioural Surveillance Surveys. This approach involves the application of data mining techniques which requires the transformation of data set to suit the context.

Data mining is the process of recognising hidden, valuable and intelligent relationships which is not inherent in huge data sets or databases (Zewdu, 1998). These relationship are often termed as models or patterns. The current technology drive across most disciplines and area of study including health, engineering, education, business and many others makes enormous data available in digital form (Taniar & Chen, 2011). This requires the application of sophisticated computational theories to ensure competitive advantage and sustainable growth (Padhy, 2012). Therefore data mining is an essential tool for transforming huge data from its raw state to a state that aides the decision making process and can be helpful in solving problems (Padhy, 2012).

1.2. Problem Statement

The HIV and AIDS epidemic in Ghana can be described as a generalized epidemic with a prevalence more than 1% in the general population (Ghana AIDS Commission, 2016). Ghana has experienced an average decline in HIV prevalence from 2.3% in 2000 to 1.67% in 2017. It is estimated that 313,063 people were living with HIV in Ghana in 2017 of which 9% were children. It is further estimated that there were 15,694 AIDS related deaths in 2016 up from 15,116 in 2016. New HIV infections for 2017 is estimated at 19,101, and AIDS orphans are set to hit an all-time high of 186,059 in 2017.

The epidemic among key populations in Ghana is described as concentrated epidemic with a far higher HIV prevalence compared to the general population. Ghana's National Strategic Plan for 2011 – 2015 and 2016-2020 classify FSWs as a higher risk KP group responsible for driving most new HIV infections. To this end, the Ghana AIDS Commission initiated a series of activities to help HIV programming for FSW. This included tow national Integrated Bio-Behavioral Surveillance Survey (IBBSS) of FSWs which was conducted in 2011 and 2015, and recorded HIV prevalence among FSW as 11.1% (Seaters 21.4%; Roamers 6.6%) and 7% (Seaters 13.2%; Roamers 5.4%). This provided first-hand information on data among FSWs in Ghana. The two studies estimated the size of FSWs in Ghana to be about 52,000 and described various interactions including HIV testing, comprehensive HIV knowledge, perceived risk and consistency of condoms use, multiple non-paying partners, alcohol and drug use, and other vulnerabilities among FSW. It also noted a near universal condom use by FSWs with paying clients, often initiated by the FSWs themselves. Conversely, almost 40% of FSWs stated that they never used condoms with their non-paying partners (Non-PPs), suggesting a high potential for sexually transmitted infection (STI) and HIV transmission via these Non-PPs.

In line with the National HIV and AIDS Strategic Plan 2011-2015, the Ghana AIDS Commission (GAC) developed the National Key Population strategy to guide the implementation of comprehensive HIV prevention interventions which cover HIV Testing Services for FSW to know their status, treatment and care to KPs throughout the country. The national HIV and AIDS, STI policy also gives direction for on how key populations including FSWs can and should have access to information on HIV prevention and mitigation services. It also gives guidance on the need for the protection of human rights of FSWs (GAC, 2013). Though there have been numerous efforts in implementing the HIV prevention activities the 10 regions of Ghana, testing for HIV still remains a big issue among FSWs as some fail to avail themselves for the test. Though there seems to be improvement in the uptake of HIV testing

among FSWs as a result of approaches like Peer Education and outreaches, it is still an issue of great concern. The IBBSS is the only survey that is able to link anonymously, the results of the HIV testing to both socio-demographic and behavioral characteristics of FSW. The 2015 IBBSS provides an overview for understanding HIV dynamics among FSWs in Ghana since it captures behavioural risk and stigma factors and socio-demographic factors and generate valuable knowledge associated with HIV status of FSW.

Summaries of reported number of HIV negative and positive clients with their respective socio-demographic and behavioral variables can be obtained from the 2015 IBBSS using traditional statistical methods approach. These methods have a limited ability to unearth new relationships that exit and are hidden in the data sets. In view of these identified hindrances it is imperative to identify a data mining solution and an algorithm that has the highest predictive powers to determine the HIV status of an FSW based on information on demographic and behavioral characteristics.

1.3. General Objective

To investigate the possibility of using a data mining predictive solution to determine the HIV status of female sex workers in Ghana using information on their socio-demographic and behavioural characteristics.

1.3.1. Specific Objectives

1. To assess the main determinants of HIV Testing outcome among FSWs in Ghana
2. To determine the best data mining methodology for identifying the HIV Testing outcome among FSWs in Ghana
3. To compare the accuracy and precision of various data mining algorithms ability to determine and predict the HIV testing status of FSWs in Ghana

1.4. Research Questions

1. What are the main determinants of HIV Testing outcome among FSWs in Ghana
2. What is the appropriate data mining methodology for predicting the HIV Testing outcome among FSW in Ghana
3. What data mining technique has the best predictive power to develop the HIV testing prediction model for FSW in Ghana

1.5. Justification for the Study

The data set for the 2015 IBBSS study for female sex workers in Ghana contains enormous data about characteristics of FSW and HIV testing. Predictive patterns that exist in this data set cannot be revealed using the standard and known statistical approaches or software. It is therefore of interest to researchers, policy makers and HIV program managers to have the capability of transforming data from surveys like IBBSS into a tool for predicting the HIV status of FSWs. This will help translate discoveries in the academic world into useful findings applicable in a real world setting. As the Ghana AIDS Commission continues to develop interventions to increase uptake of HIV testing services among FSWs, predictive solutions like this will complement effort on effective HIV programming and planning. This is because, if the HIV status of female sex workers can be accurately predicted by a certain percentage it would also help address concerns associated with HIV Testing among FSW.

CHAPTER 2

LITERATURE REVIEW

2.1. Overview of Data Mining

The introduction and integration of information communication technologies in most business processes have made huge amount of data in digital format available(Cenzer & Lee, 2000). This requires the automation and analysis of these large data sets to discover knowledge. Data mining can therefore be explained as the process of extracting interesting patterns which can be useful, non-trivial or implicit from huge data sets (Han, 2007). The main objective of data mining is to gather useful information from various data sources and investigating unexpected note-worthy patterns or valuable structures in a huge dataset. Data mining applications tend to support data analysis and decision support process which is implemented in market analysis, risk analysis, fraud detection and text mining(Hand, 2007).

Data mining is generally composed of many disciplines like database, machine learning and statistics. The integration of these disciplines provides the requisite methods and techniques, tools and expertise which facilitates the discovery of useful knowledge from large data set. This is illustrated in the figure 2.1below.

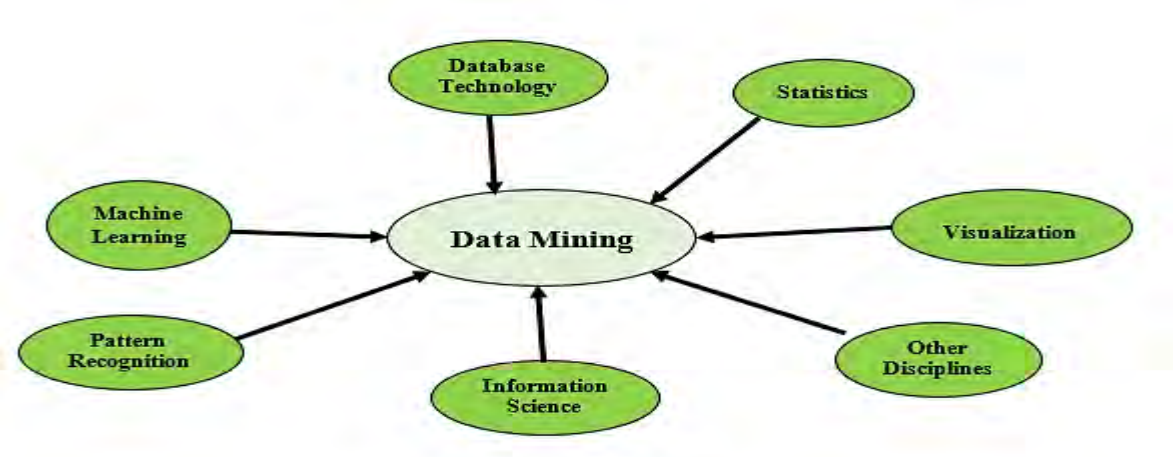

Figure 0.1: Associated fields of Data mining
Source: Han (2007)

Data mining processes for every task involves identifying and choosing the data of interest, data preprocessing, transforming the data, implementing the data mining technique to extract useful relationships and finally evaluating the discovered knowledge (Bramer, 2003). It is a critical phase in the knowledge discovery process because different algorithms are used to search through the target data to identify relationships as shown in figure 2.2.

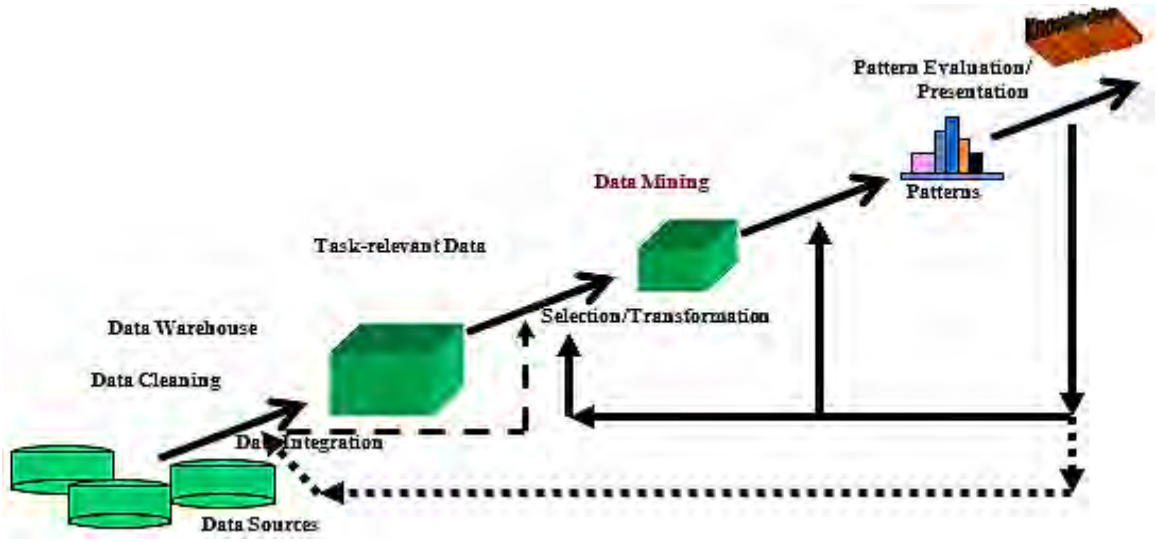

Figure 0.2: KDD Diagram
Source: Han (2007)

These relationships and summaries often termed as models or patterns are fitted to determine interesting patterns (Nicole & Tim, 2006). Research shows that there are various approaches to data mining and these approaches differ in how the models are built or finding patterns (Hand, 2007). These approaches are grouped under the following:

1. Exploratory Data Analysis (EDA) for exploring the data set
2. Descriptive Modelling for describing the process involved in generating the data
3. Predictive Modelling for predicting a variable from known independent variables.
4. Discovering Patterns and Rules: for discovering association and relationship among attributes
5. Retrieval by content for finding a similar pattern in the data set

(Nicole & Tim, 2006)

Data mining is also synonymous to terms like knowledge extraction, data pattern processing, information discovery or Knowledge Discovery in Database (KKD) which also finds useful patterns in any given data set (Han, 2007). In most cases, data mining is viewed as an important and integral component of KDD (Han, 2007). This involves choosing the data of interest, data preprocessing and transformation, extracting the useful patterns and relationships and then evaluating the discovery.

Methodologies or processes in data mining are made up of Fayyad et. al Knowledge Discovery in Database (KDD), Sample, Explore, Modify, Model, and Assess (SEMMA) and CRISP-DM (Padhy, Mishra, & Panigrahi, 2012). Each method has its own set of concepts, theories and practices and performs functions which include Exploratory Data Analysis (EDA), descriptive and predictive modelling and discovering Patterns and Rules to build models. These functions are based on algorithms to discover knowledge. For instance, to carry out a predictive analysis under clustering requires using algorithms such a decision tree.

Other available algorithms include neural networks, and hybridizations among others. It is important to note that choosing an algorithm is highly influenced by the data mining task at hand since specific algorithms are better suited for specific task (Hand, 2007). Predictive modeling as a function of data mining allows the ability to predict and determine an unknown value of an outcome variable (target variable) based on the values of independent variables. The model is made up of independent variables which are called predictors which have a high probability of influencing the outcome variable.

2.2.1 Data Mining Process/Methodologies

In data mining research there are three main process or methodology, namely:

a. Fayyad et al. Knowledge Discovery in Database (KDD)

According to Fayyad, Piatetsky-Shapiro & Smyth (1996) this process comprises several steps from data preparation to pattern searching, to evaluating the discovered knowledge and finally to modification. This process also consist of multiple iterations during the process and includes sophisticated computations to arrive at the discovered pattern or knowledge. The new knowledge must also be valid for any given data set with some degree of certainty and must be novel and potentially useful and understandable. This approach is an academic-based approach and consist of nine main steps.(Krzysztof, Witold, Roman, & Lukasz, 2007). These steps include understanding and developing the application domain, creating a target set, data cleaning and pre-processing, data reduction and projection, choosing the data mining algorithm, data mining, interpreting mined patterns and consolidating discovered knowledge (Fayyad et al., 1996).Figure 3 summarizes the various steps

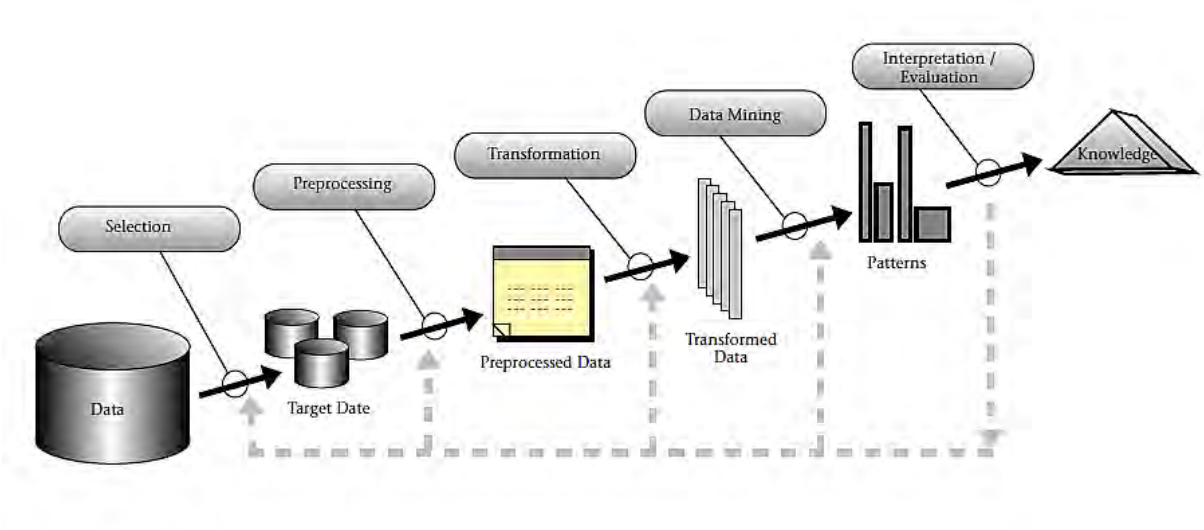

Figure 0.3: Data Mining and its Associated Fields
Source: Fayyad et al. (1996)

b. Sample, Explore, Modify, Model, and Assess (SEMMA)

This methodology is considered as one of the processes or methodologies for undertaking a project in data mining and operationalized by the SAS Institute. SEMMA makes it easy for the application of exploratory statistical and visualization techniques. This is done through the selection and transformation of the variables to be used for prediction. This to ensure that these variables in the prediction model have the ability to predict the outcome taking into consideration the accuracy of model. The SAS Institute in developing the SEMMA process came up with 5 stages and this included extracting a sample from the data set, exploring the data set for unanticipated trends, modification through variable transformation which focus on model construction, creation of the model where user focus on a combination of variables that are in the best position to predict the preferred outcome. This is followed by evaluating how the findings from the SEMMA process is useful or reliable. The SEMMA process make it possible for new questions and ideas to be investigated out of the results from each stage. This is done by returning to data and refining it to meet the needs of the new questions.

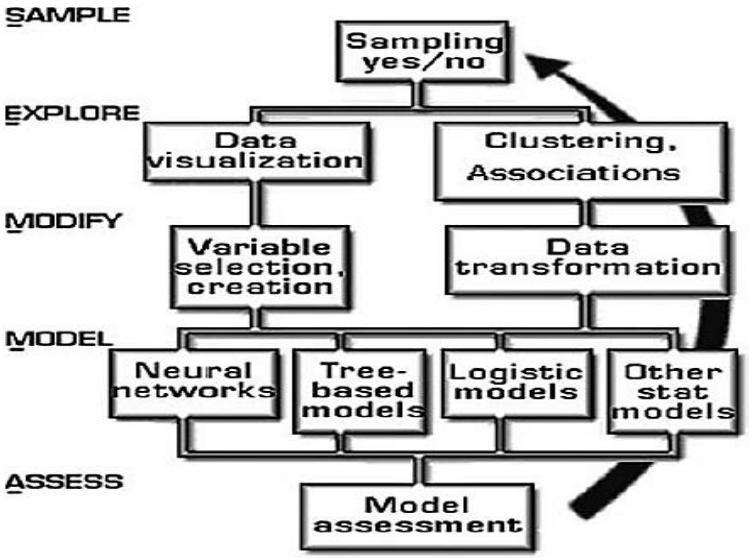

Figure 0.4: SEMMA Cycle
Source: Delen & Olson (2008)

c. Cross-Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM can be described as a non-proprietary model but mostly preferred by industry experts since it focuses more on business and industry. It follows an iterative process like the other methodologies and very appropriate for people new to the data mining field since it follows an easy-to-read steps (Jair et al., 2017). It follows a six stage cycle from understanding the business to deployment of the system (A. Azevedo & Santos, 2014). Figure 5. Summaries the various steps in the CRISP-DM

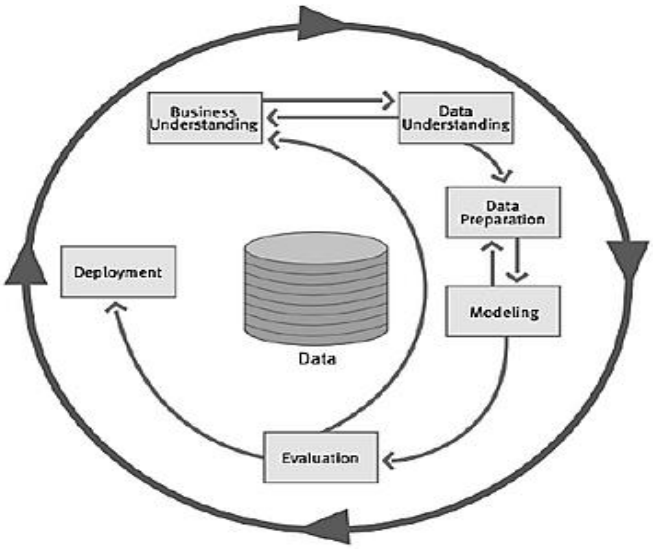

Figure 0.5: CRISP – DM Cycle
Source: Azevedo & Santos(2014)

2.2.2 Comparing the three Methodologies

All three methodologies or processes have their uniqueness making them suitable for a specific data mining process. Fayyad et al. Knowledge Discovery in Database (KDD) employs a methodology that extracts knowledge from a database through the pre-processing of the database. SEMMA is seen as an easy- to-understand process which allows for development and maintenance of data mining projects. CRISP –DM has all its stages documented and organized hence easy to understand the flow of the process and easy to revise it (A. Azevedo & Santos, 2014)

A study by Azevedo & Santos, (2014) suggested that Fayyad et al. KDD and SEMMA are similar in nature since the first five stages of SEMMA can be linked to Fayyad et al. KDD. The same study compared the Fayyad et al. KDD with CRISP-DM and concluded that these two processes are not similar as in the first instance. Another comparison study by Umair & Haseeb (2014) found that some or most of the steps in each of the models can be found in other models. An instance is the data transformation stage in Fayel which is similar to data preparation in CRISP-DM and modify in SEMMA (Umair & Haseeb, 2014) concluded that most data mining researchers follow KDD because of its accurcay and completeness and that CRISP-DM and SEMMA are mostly industry orriented. The study concluded that CRISP-DM is more accurate and complete than SEMMA.

Also Jair et al. (2017) in their study to compare CRISP-DM and SEMMA concluded that CRISP-DM is a more appropriate and easy-to-use method as compared to SEMMA since it allows for a better understanding of the data mining project at hand which guarantees success. Also there is available documentation for each case study and CRISP – DM dedicates a whole section for data transformation whiles SEMMA is designed for SAS Enterprise Miner hence all available documentation focuses on this tool. The above suggest that industry experts prefer CRISP-DM for its reliability, accuracy and completeness.

2.2.3 Data Mining Algorithms and Techniques

Srihari, (2000) defines algorithms in data mining as a step-by-step procedure where all the procedures are well defined to receive data as an input into the system and generate output in the form of models and patterns. Nicole & Tim, (2006) also suggests that algorithms in data mining generally are made of three sections which are the model, the preference criterion and the algorithm search. Nicole & Tim (2006) described the model component to be responsible for the function of the algorithm which explains the specific task for the data mining process.

Also the preference criterion is explained as deciding the best-fit model for the dataset on which the data mining task will be performed. The process for finding the models and parameters following an algorithm is termed the search algorithm. Srihari (2000) also classifies the components of data as the task, structure, score function, search method and data management technique.

It is critical to know the type of data mining task to undertake for any data mining project since this help to understand what data mining algorithm to apply. Aleksey (2017) classifies the data mining tasks into two groups which is descriptive and predictive. Predictive data mining tasks involves creating a model that can be used to predict a behavior pattern in a system which has been analyzed which previously did not exist whereas descriptive data mining tasks search for hidden patterns in any given data set (Padhy, 2012). Figure 2.6 summaries the description of the task and what kind of analysis can be performed under each data mining task. Also key analysis that can be done include association analysis where dependencies in the data sets are done (Aleksey, 2017)

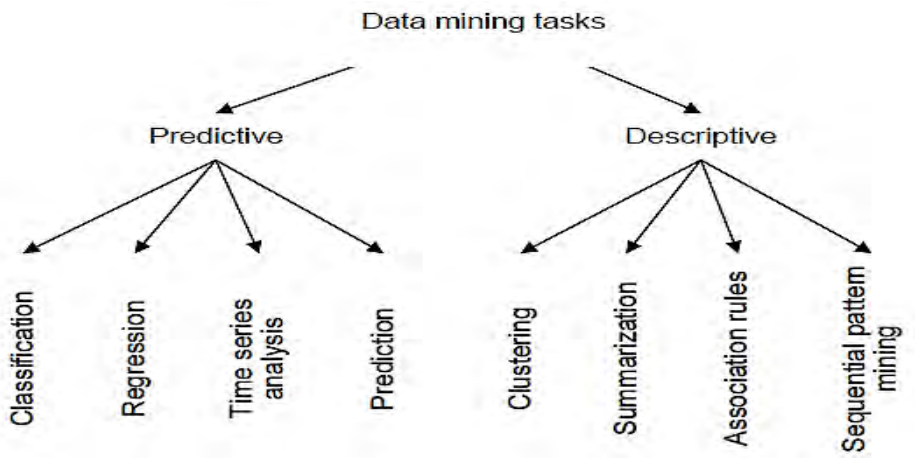

Figure 0.6: Data Mining Task Classification
Source: Aleksey, (2017)

Various data mining algorithms are being applied to various domains or field of study but after a series of comparative review of articles, five popular algorithms were identified for predictive data mining. These are

a. J48 and Random Tree (Decision Trees) Algorithm. Decision trees as described by Sundaram (2012) is very useful for prediction and classification. It is also seen as a classifier such that the decision tree generation is based on the inputted attribute (Patel & Patel, 2016). The decision tree has a tree-like configuration and begins with the roots which in this case are attributes and follows through to branches made up of different attributes. The decision tree ends with the leaves which are the nodes and each leaf represent a class. Decision trees follow a set of rules which leads to a value and it is a reasonable model since it uses explicit rules for decision making. This makes Decision trees appropriate for classification which is a directed data mining. There are different types of decision tree algorithms in data mining this include J48 and random tree. These two models together are as a result of implementing the C4.5 decision tree learner which produces models. Identify from decision trees is done by following each path from the root to the node which has a rule. (Patel & Patel, 2016). These rules are made readable and understandable through transforming the node into an "if - then" classification rules. The testing component of the path is defined by the "if" and the "then" specifies final classification.

b. Logistics Algorithm. Logistic regression algorithm in data mining is a predictive algorithm which follows the Ordinary Least Square regression approach for prediction (Bramer, 2003). It is suitable for predicting outcome attributes which are discrete in nature as against linear regression which will not be able to model attributes or variables of this nature. This in effects helps to predict the odds of an event building models around the event instead of the point estimates.

c. Neural Networks as described by Ashraf, Ahmad, & Ashraf (2018) is seen to be likened to the neurons in the human brain and follows the reasoning approach which is made up of masses of neurons which are connected through synapses. Therefore a neural network as applied to data mining task can be described as an input and output connected together as units with a weight associated with the connection. The algorithm is seen to be a simple model since the connections are similar to the interconnection in the brain. This algorithm is used for classification and prediction after learning using a training set which aides to developing universal patterns from the data set. This can be practicalized in instances of undirected data mining and time-series prediction. Advantages of this algorithm includes being able to tolerate noise in the data set and being able to classify unseen patterns in the data set; this makes neural networks best suitable for predictive data models (Sundaram, 2012).

d. Naives Bayes algorithm is mainly a supervised learning approach used for classification (Padhy et al., 2012). This classifier with the help of probability distribution classifiers all given data inputs into classes. The attributes are independent of the class and this satisfy's the Naïve component and bayes section is due to the working out the correct probability for a given class for particular attributes. Parameters under this section are estimated by the maximum likelihood (Berka & Rauch, 2010). Advantages of this algorithm is that it is very easy to implement and gives accurate results for predicting problems. Also it requires less amount of data and has a great classification performance. Naïve Bayes has a zero problem if there is no occurrence of the class label and attribute label

Data mining techniques as suggested by Padhy (2012) is divided into two which are the unique table approach mining and structured data mining. In the unique table approach technique,

characteristics of each individual in a central database are termed as attributes which are represented in a vector format. The central database is in the form of a table with the rows or tuples matching the individuals and the columns matching the attributes. The second technique which is the Structured Data Mining is noted to be able to handle complex data where the table properties including rows or columns are not related to each other.

2.2 Human Immunodeficiency Virus and Acquired Immune Deficiency Syndrome (HIV and AIDS)

HIV attacks the immune system and makes it incapable of functioning properly. The immune system which is made up of CD4 cells is in charge of the overall organization of the human immune system reaction to infections (Sacramento County Department of Health Unit, 2002). Through the attacks of the virus there is a progressive depletion of the immune system of the cells hence given rise to an immune deficiency. An immune deficiency prones one to opportunistic infections and cancers.

2.2.1 HIV Mode of operation

HIV first enters the CD4 cells and then replicates itself by copying the cell's DNA thus making it difficult to identify and as a result causes harm to the immune system (Nicole & Tim, 2006). New CD4 cells are then destroyed by the replicated HIV. Usually the initial stages of the infection has a lot of cells being infected; this makes the body to contain a high amount of HIV particles. At this stage, called the window period, it is difficult to detect an individual's HIV status in this case through tests since there are insufficient antibodies formed. If left undetected and untreated, the infection progresses to cause immunodeficiency which then results in cancers and AIDS related deaths.

2.2.2 HIV transmission

HIV can be transmitted or contracted through specific activities. The commonest means are through sexual contact with an infected person, sharing needle or syringe with an infected person and MTCT. Only certain body fluids, namely, blood, semen (*cum*), pre-seminal fluid (*pre-cum*), rectal fluids, vaginal fluids, and breast milk, from a person who has HIV can transmit HIV. These fluids must come in contact with a mucous membrane or damaged tissue or be directly injected into the bloodstream (from a needle or syringe) for transmission to occur. Mucous membranes are found inside the rectum, vagina, penis, and mouth. (CDC 2018). Sexual contact can be described as the foremost transmission mode. (Nicole & Tim, 2006) suggested that the presence of STDs sexual transmission of HIV is influences the natural chances of transmission.

2.2.3 Diagnosing and treating HIV

HIV diagnosis can be done in two ways: Clinical Diagnosis and laboratory diagnosis. The clinical diagnosis is able to detect HIV in its early stages In the case of laboratory diagnosis, the HIV antigen is detected when the amount of HIV particles in the body are high and also an easier way of HIV detection is the presences of antibodies. There are two tests that can confirm HIV presence in the blood which are Enzyme Linked ImmunoSorbent Assay (ELISA) Test, and Western blot assay (Nicole & Tim, 2006). To measure the test accuracy, sensitivity and specificity are used and high rates from each measure is interpreted. It is of importance to be mindful of how to apply a combination of the test. Currently, there is no cure for HIV. Available treatment, through the administration of antiretroviral drugs, only impedes or stops the virus from duplicating itself leading to viral load suppression.

2.3 Female Sex Workers and HIV

Female sex workers as defined by Overs (2002) is the exchange of sexual services for the monetary or in kind benefits by women who consider the service as source of income.. A study by Vandepitte et al. (2006) concluded that in the scope of HIV, FSW are very vulnerable and susceptible to the infection. This is a similar case for FSWs in Ghana where HIV prevalence is concentrated among this group. Also HIV and AIDS in Ghana is generally being fuelled by commercial sex which characterise the work of female sex workers.

Vandepitte et al., (2006) attributed motivation for sex work among majority of women to a combination reasons: debt alleviation, wealth acquisition and survival. Ghana's National Strategic Plan (2016 -2020) classifies FSW in Ghana into four main groups which are Seaters, Roamers, Clandestine Groups and those who do not self-identify as FSW although they are engaged in sex work. Wariki et al., (2012) also classifies FSW into those who portray themselves as sex workers and earn a living from selling sex where the sex work of these women is not the first source of income (this group is made up of waitresses, hairdressers, and tailors, massage girls, street vendors, or beer promotion girls). The meeting place for the sex is often determined by both the social status and demographic characteristics of a sex worker. This is typical case for Ghana where FSW with low socio economic status work in bars, on the street whiles those with high socio economic status work with escort agencies.

A study by Prüss-Ustün et al., (2013) concluded that 15% to 25% of all new infections in West Africa occur among FSWs and their clients. The same study also suggested that 20% to 30% of new infections occurred in low-risk and partners of the people with high risk behavior. The prevalence of HIV among FSW is estimated at 11% after dropping from 25% in 2009 and contributing about 2.4% of new infections in Ghana. This figure is higher than the prevalence in the general population. The Ghana AIDS Commission approximates that for new HIV

infections, there are 6.5% among clients of sex workers and 22.2% among partners of clients of sex workers (GAC, 2010). Ruxrungtham, Brown, & Phanuphak (2004) in their study to understand the spread of HIV in Asia attributed high HIV prevalence to high patronage of commercial sex. Also a study conducted by Gaffey et al., (2011) in India reveals that about 4% of Indian men visited a FSW. This gives a clear indication of the impact of sex work and female sex workers contributing to the spread of HIV. This has largely influenced how interventions targeted at FSW and their clients have been conducted to include behavioral change interventions like condom use and health education for FSW (Wariki et al., 2012).

2.3.1. Social and risky behaviour among FSW

An increase in HIV infection among FSW can be linked to socio-demographic factors such as age, marital status, poverty rates and social status (Blanchard et al., 2005). Ghana's National MARP (now KP) Strategy classifies risk behaviour among FSW to include unprotected sex with clients, untreated STIs, multiple sexual partners and alcohol and drug use. Additionally certain risk factors make FSWs very vulnerable to the infection. These include harassment and abuse of human rights by the police, sexual partners and public officers. (Muula & Twizelimana, 2015)

A study by Ramesh et al., (2008) showed that the current marital status of FSW could help determine the HIV status of an FSW. This claim is supported in the study by Gangakhedkar et al.(1997) which found that the spread of HIV among FSW who were married were likely being infected by their partners. Also a study by Scorgie et al. (2012) mentioned that the place of work is directly linked to the social status of FSWs in sub-Saharan Africa and how much they are paid. The same study also found that condom use among FSW was based on monetary considerations. Whereas professional FSW strictly adhered to 'no condom, no sex' policy, the unprofessional ones give in due monetary allurement. A study conducted in Ghana by Adu-

Oppong, Grimes, Ross, Risser, & Kessie (2007) found out that the level of condom education among FSW was low.

Social factors like poverty make FSW vulnerable to HIV infection since these women are forced to indulge in risky behaviours to earn a living. Another factor of key concern is stigma discrimination from family circle or the society at large and issue of gender-based violence which seems to be on the rise among FSW. The FSW population is hard-to-reach hence the difficulty in understanding the context in relation to risk factors. It is proposed that having a high risk perception of HIV can indirectly affect how one behaves hence the need for a comprehensive knowledge of HIV for FSW (Muula & Twizelimana, 2015). It is also observed that knowledge of HIV among FSWs can help in planning, designing and implementing intervention to address the needs of this population.

Some studies conducted in sub-Saharan Africa noted that FSW have reported having unprotected sex with clients from high class in society and considered them as disease free. Some studies in Ghana (Ankomah et al., 2011) have also shown that FSW did not consider their trade as risky and consider the general population as rather at risk of the HIV infection. This is because they have access to HIV information on preventive measures like condoms which is not the case for most women who have not heard about it. Even though FSW are aware of their risk to HIV, issues of comprehensive knowledge about the disease come to play since a study conducted by (Aklilu et al., 2001) mentioned that half of the respondent were of the view that healthy looking persons cannot have the HIV infection.

Alcohol use is a common practise among FSWs and their clients. It is seen as a catalyst for sex and FSW will normally report having unprotected sex due to alcohol intake although some have reported it as boost or motivation to negotiate sex with sexual clients (Scorgie et al., 2012). This can be seen as very risky behaviour since been intoxicated with alcohol affects a person's

judgement and decision making which could undermine the proper use of condom during sexual intercourse (Chersich et al., 2007; Chersich et al., 2009. A study conducted in China concluded that about 30% of participants reported having taking alcohol before sex and reported a limited use of condom. Similar findings were obtained in an Indian study conducted within the same year (Samet et al., 2010).

2.3.2. FSW HIV Testing behaviour

Testing for HIV has been seen to have a correlation with HIV prevention since available evidence shows that individuals testing negative for HIV are more likely to take the necessary steps to remain negative (Keetile, 2014; Udoh, 2013). Also, according to WHO (2011) heterosexual couples who are HIV positive are more likely to have a different attitude towards their sexual behaviour than people who are negative or those who do not know their status. It is difficult to extend this observation to FSW who test positive since there is little or no evidence to support this claim. Research has shown that HIV Testing among FSW have over the period been a difficult thing to do because of the nature of their work and since they indulge in risky behaviours like having unprotected sex hence a high probability of testing positive to HIV. (Aunon et al., 2015). This is one of the few barriers deterring FSW from getting tested to know their status although they are familiar and know places to access HIV Testing Services.

It is important to be aware of structural barriers to HIV testing among FSW as it is has been reported by FSWs that HTS services are expensive. Also it issues of confidentiality have been reported where FSWs do not trust health workers to guarantee the safety of their test results. In addition, most FSW have accepted the view that by the nature of their work, they have contracted the disease hence there is no need for the test. This in some cases has resulted in FSW threatening to commit suicide in the case they test positive (Aunon et al., 2015). A number of studies have shown that for FSWs who undertook HTS, only about 60% of those testing

positive did not return to receive their results. This the researchers have attributed to high fluid nature of FSW who are in constant motion and will not want to bear transport cost and time for these return visits. (Chersich et al., 2013). However, this situation vary for some HIV programs in some parts of Africa where social support programs are available for FSW who have registered under such programs and are required to visit their clinic regularly (Steen et al, 2000). This prevails in countries with significantly low HIV prevalence and there is an effort estimate the prevalence by offering free voluntary HTS to participants.

Sex work is outlawed by the laws of most countries. Such laws are antithetical to public health practise which seeks to make health services available to everyone regardless of your occupation or gender (Chersich et al., 2013; Scorgie et al., 2012). The resultant barrier created by such laws makes it difficult to reach FSWs with prevention interventions. According to WHO (2011) one third of FSW in Africa actually get access to adequate HIV prevention services with less numbers getting treatment, care and psychosocial support.

2.4 Related Works

Data mining as a field of study has contributed tremendously to the acquisition of the knowledge. This is seen the application of various data mining techniques in the acquisition and extraction of knowledge as shown below.

2.4.1 Data Mining application in health and medical practise

The business processes for organizations in the health care sector involve capturing, storing and generating various kind of data. Most of the data gathered are in connection with patient information, research, clinical duties and trials, administrative duties and resources. The use of traditional statistical methods or techniques on health data limits knowledge derived to only operation information since health data in recent times have become voluminous and complex.

Therefore using data mining techniques ensure that a more explorative and interactive approach is applied to health data to discover knowledge in care in terms of associations, sequential patterns, classifications, predictions, and symbolic rules. This brings to bear an in-depth analysis of health care data to support improvement in the health care industry. Medical practise is currently expanding and it is the requirement of all clinicians to be empowered to protect patients form diseases and side effects. Tools in data mining provide the necessary empowerment for clinicians to expand their knowledge base since these techniques can be applied to every aspect of the health care sector.

2.4.2 Data Mining application to HIV Testing

In recent data mining research, various algorithms have been applied to data set from various sources to get information on HIV. Various data mining algorithms have been used to establish relationships and discover new knowledge in HIV. A study conducted by Hailu (2015) investigated the prediction power of various data mining algorithms. The study, which used J48, random tree, Naive Bayes, Neural network, and logistic regression algorithms made it possible in determining if an individual has ever tested for HIV. It found that out of the four methods, random tree had the highest accuracy of 96% and employed the CRISP-DM methodology to data mining. The study concluded that data mining techniques are essential for utilization of HTS.

In the study to predict the HIV status of individuals in Addis Ababa, (Zewdu, 1998) found that J48 classifier was the best algorithm to predict the HIV status of an individual with an accuracy of 93.95% out of the 11 algorithms used for the study where algorithms like Naïve Bayes performed badly. The study incorporated the CRISP –DM methodologies and concluded that it is possible to predict the HIV status of an individual knowing socio-demographic and behavioural characteristics. This also disproved a study by (Nicole & Tim, 2006) which

suggested that using demographic and behavioural data it is not enough to predict the HIV status of and individual.

Leke-betechuoh, Marwala, Tim, & Lagazio (2006) in their study to predict HIV status from demographic data demonstrated how data mining algorithm neural network can be used to classify the HIV status of individuals using their characteristics such as demographic data and socio-economic status. The study used demographic data such as age, education and race found an accuracy of 84.24% for the design. This support the claim that an individual's HIV status can be determined from demographic characteristics and concluded that the neural networks were more accurate. A similar study conducted by Haile Mariam (2015) to predict the CD4 status of patient on ART used the CRISP –DM methodology and had accuracy of 88.79% for the J48 decision tree. It concluded that J48 decision tree was a better model than PART rule algorithm and that ART status, marital status are significant to determine the CD4 status of a PLHIV.

Database for information system are growing in volumes and this requires automated data processing systems. It is clear that using classification and predictive models in health care is becoming a common task. It is therefore necessary and promising to apply data mining techniques to predict whether an individual HIV status from available data.

CHAPTER 3

METHODOLOGY

3.1.Study Area

This study was based on the data from Ghana's 2015 FSW Integrated Bio – Behavioral Surveillance Survey (IBBSS) which was carried out across the ten regions of Ghana among FSW.

3.1.1. Data Source

The 2015 IBBSS is the most current data set which contains HIV testing results for FSWs in Ghana. The study has national representation of FSWs and was conducted by Ghana AIDS Commission in collaboration with FHI 360. Noguchi Memorial Institute for Medical Research (NMIMR) was in charge of testing the FSW for HIV. The 2015 IBBSS for FSW is the second national IBBS conducted for FSWs in Ghana.

The study collected information on bio-behavioral characteristics of FSWs and covered topics on socio demographic sex work, risk perception, sexual behavior, sexual partner, condom and lubricant use with clients husband and boyfriends, STIs, knowledge opinions and attitudes towards HIV and AIDS, alcohol and narcotic drug use, stigma and discrimination and exposure to the interventions and self-esteem. The 2015 IBBSS for FSW (HIV testing component whether negative or positive) served as the data source for this study to predict the HIV status of FSW using data mining techniques. The 2015 IBBSS consisted of both biological and behavioral components and consisted of 3,865 instances (only this number of FSW participated in the biological test out of 4,751 FSW who took part in the study) and 318 attributes/variables.

3.2.Basic Data Mining Methodology

The data mining methodology employed in this project is critically dependent on first identifying the proposed objective or specialty whose tools, when applied, provide the most meaningful and desirable outcomes or useful results. The data mining methodology employed for this study was the Cross Industry Standard Process for Data Mining (CRISP-DM). The methodology is best summarized as shown in figure 2.5 and explained as follows:

3.2.1. Business Understanding

The understanding of the business process required reviewing all critical and relevant documents related to HTS among FSWs in Ghana to have a better understanding of the scope of HTS among FSWs. The approach adapted was a desk review of these relevant documents: NSP 2016 – 2020, Standard Operating Procedures for Implementing HIV Programmes among Key Populations, National HIV and AIDS STI Policy and National HIV Treat All Policy. These documents provide direction and guidance on HIV testing among FSW in Ghana and hence the review was critical to understand what the Ghana Government through the Ghana AIDS Commission is doing in terms of HIV testing among FSW.

The review identified gaps in HTS provision for FSW in Ghana and hence the research problem for this study was constructed. In effect the aim and objectives of the Ghana AIDS Commission have been identified and drawn to this data mining project. In this case five popular data mining algorithms were used in predicting the HIV status (negative or positive) of female sex workers in Ghana.

3.2.2. Data Understanding

The data understanding stage requires been familiar with the initial data set, selecting and identifying quality issues with the data set. To meet this requirement, the task undertaken under this phase included checking if all questions in the questionnaire were covered in the data set, following through questions where respondents skipped certain questions, and summaries of variables.

3.2.3. Data Preparation

The final data set used for this study was constructed from the initial raw data from the 2015 IBBSS through attribute selection, data cleaning and transformation. The final target data set was made up of 21 attributes and 3,092 instances.

3.3.3.1. Attribute Selection

Any data mining task requires the selection of appropriate and relevant attributes which addresses the data mining project objectives. The presence of irrelevant attributes tend to affect the efficiency of data mining algorithms and in most cases create poor classifiers. In view of this, removing irrelevant attributes from the data set was important since it increased accuracy, simplified the results and saved on computational time. Irrelevant attributes (particular to this data mining project) which literature has suggested not to be relevant to this study were identified and deleted (column deletion) from the initial 2015 IBBSS data set. The selected attributes were based on relevance to the subject matter (in this case HIV testing among FSW) and general facts from reviewing literature. An initial data set of 4,751 instances and 318 attributes was reduced to 22 attributes and 3,092 instances. This was prepared using STATA version 15 (IC) and MS Excel before importing into WEKA 3.6.9. The attributes of interest to the outcome variable (in this case was the HIV testing results for FSW) was arrived at using

the select attributes option in WEKA using the "*CFsSubsetEval*" as the attribute evaluator and "*BestFirst*" as the search method.

3.3.3.2. Data Cleaning and Missing values

Data cleaning is required to be done for any dataset as it helps with data analysis. This usually involves completing missing values or removing attributes that has incomplete data to remove noise in the data. Therefore the approach adapted in this study was removing and deleting records with missing values in the data set by either row or column deletion and this did not affect the entire data set as this was in order to improve the quality and accuracy of the model. In all 773 instances ore records were deleted to form the final data set of 3,092 instances.

3.3.3.3. Data Transformation and Reduction

This phase requires transforming variables into forms that can be handled by the data mining tools by adapting data reduction techniques like data discretization. Consequently, Knowledge in HIV and AIDS prevention was built from responses to 7 questions and classified as low knowledge (mark $\leq$ 1), average knowledge (mark 2-4), or high knowledge (mark 5-7) knowledge. Nine questions were used to generate the "drug usage" attribute and was categorized as "Used 5 or more drugs before" (score 5-9) and "Used less than 5 drugs before" (score $\leq$ 4). The source of HIV information attribute was also generated using 9 questions and was categorized as "Have at least one source Information on HIV" (score $\geq$1), and "No source of information HIV" (score < 1). Though a chi square test carried out showed that there is no statistical association between the outcome variable and these indexes, they were maintained since literature shows their importance to predicting HIV status. Attributes such as "age-group", "marital status", "years of sex work", "relationship with last sexual partner" and "income from sex work" having multiple distinct values were discretized into a more general concept. Most of the attributes used for this study are not in their original state since these

attribute have been transformed to make it easy to understand and give better interpretation to the results. The attribute of interest was considered to be unbalanced with 93% testing negative to HIV and 7% testing positive hence bias to evaluate the classifier method since well-balanced dataset is very important for creating a good prediction model. This requires a balance for these two classes, hence the Synthetic Minority Over-sampling Technique (SMOTE) and randomizing techniques in WEKA 3.6.9 pre-processing option was used to balance the data.

3.2.4. Modeling

This phase requires selecting modelling techniques with their respective parameters to obtain optimal values. Additionally, the task under this phase forms part of data mining task in some descriptive and predictive process. The descriptive task employed for this study covered association and classification for the prediction task.

3.2.4.1. Classification

Classification as a learning function links a data item into one of several predefined classes and this study therefore classified FSW as tested negative or positive for HIV. The study employed the use of 5 *classify* functions in WEKA 3.9.6 which included: J48, Random Tree, Logistics, Neural Networks and Naives Bayes.

3.3.4.2. Association Rules

Association rules as part of descriptive data mining task were employed in this study to uncover patterns in the data set. The *Apriori* function under *associate* in WEKA was used to examine which instances of testing negative or positive to HIV by FSW frequently occurred in a database. Therefore the association rules used for this study was configured to have a minimum support of 95% and with 90% of confidence level with the selected attributes as inputs.

3.3.4.3.Methods of Training and Testing

Witten and Frank (2005) stipulate that classifiers mostly depend on training before it can be used. This means that exposing classifiers during the training phase make them reliable and since they are trained we can confidently use them. Additionally, assessing the error rate of a classifier on a new data set using stratified 10-fold cross-validation is also suggested to be one of the ways to predict the performance of a classifier. In this study the method of training and testing was achieved using WEKA 3.6.9 to test and train the classifiers. The *training* test option in WEKA 3.6.9 was used to train the 3,092 instances for all five classifiers. This was followed by using the cross validation test parameter to divide the instances randomly into 10 parts. After this process the *supply* test parameter in WEKA was used to train the instances of FSW HIV status to determine if the FSW HIV results will be negative or positive.

3.3.5. Evaluation

The evaluation stage of the CRISP-DM involves assessing the various models that have been obtained. The evaluation is done by assessing the accuracy, the processes involved to construct the models and the confusion matrix of the models. Additionally the complex nature of the models are evaluated in terms of the number of leaves and trees. This is in an effort to meet the business objectives of the research. Table 3.1 summarizes the confusion matrix used in this study for classifying the HIV status.

Table 0.1: Confusion Matrix Template

		Predicted HIV status		
		Negative (N)	Positive (P)	Total
Actual HIV status	Negative (N)	True Negative (TN)	False Negative (FN)	TN + FN
	Positive (P)	False Positive (FP)	True Positive (TP)	FP + TP
	Total	TN + FP	FN + TP	TN + FP + FN + TP

Note: TN: The number of FSW testing Negative for HIV that are classified as negative. FN: The number of FSW testing negative for HIV that are classified as positive. FP: The number of FSW testing negative for HIV that are classified as positive. TP: The number of FSW testing positive for HIV that are classified as positive. TN + FP: The total number of FSW predicted to test negative for HIV. FN + TP: The total number of FSW predicted to test positive for HIV. TN + FN: The total number of FSW that tested negative for HIV. FP + TP: The total number of FSW that tested positive for HIV. TN + FP + FN + TP: The total number of FSW.

The measurement of the performance of the models used for the study are given as follows:

- **Correctly Classified Instances (Accuracy)** = (TN + TP) / (TN + FN + FP + TP)
- **Sensitivity** = TP/TP+FN
- **Specificity** = TN/TN+FP
- **Receiver Operating Characteristic (ROC) curves**

3.3.5.1. Data mining experiments

The five algorithms (random tree J48 Bayes, Naïve Bayes, logistic regression and neural network) used 3,092 instances in developing the model for predicting the HIV status of FSW. These experiments were done in WEKA using different evaluation and testing options. The training test option achieved the best accuracy for this study hence the training option was used in evaluating the performances of the models. The study used data mining techniques and WEKA 3.6.9 to address the research objectives.

3.3.6. Deployment

Currently, numerous data mining software are available which are being employed both in academia and industry. These may either be proprietary (SAS Miner, Intelligent Miner by IBM, CLEMENTINE by SPSS) or open-source (WEKA, and R). In this research, WEKA version

3.6.9 was used as the data mining tool. Microsoft Excel and STATA version 15 was used for the data preparation before transferring to WEKA. The findings and results from the study are presented in this project work report in chapter 4. A review of the results and recommendations are presented in Chapter 5 and 6 will be beneficial to both HIV program managers and stakeholders.

3.4. Ethical Consideration

Approval was obtained from the management of the Ghana AIDS Commission to use the IBBSS data. This research involved the secondary use of the 2015 Integrated Bio-Behavioral Surveillance Survey (IBBSS) for FSW in Ghana. Ethical approval was given for the IBBSS study to be undertaken in 2015.

CHAPTER 4

RESULTS

4.1.Demographic characteristics of respondents

The target dataset contained 3,092 female sex workers as study participants. Out of which 2,491 (80.56 %) of the FSW were roamers while the remaining 601 (19.44%) were seaters. The age ranged from 16 to 64 years old with modal age of 20 – 24 years old. Vast majority of FSWs were Ghanaians (2,814, 91%) whilst only 9% (278) were foreigners (mostly Nigerians). Majority of the FSW are from the Akan origin (50.78%, 1,570) while the rest were from other tribes including Ewe, Ga/Adangme and Mole-Dagbon, Gruama and Hausa.

Out of the total number of FSW, 2,426 (78.46%) were single whilst 139 (4.43%) were currently married. The rest were widowed, separated, divorced or cohabiting with their partners. Among the study participants, 2,700 (87.32%) tested negative to HIV whilst the rest 392 (12.68%) tested positive to HIV.

Majority of the participants had completed JSS (1,292, 41.79%) followed by Secondary/SSS/Vocation with 683 (22.09%). On religious affiliation, 2,466 (79.750 %) of participants said they were Christians, 451 (14.59%) were Muslims, 164 (5.30%) said they were affiliated to no religion and 11 (0.36%) said they were practitioners of traditional religion.

Table 0.1: Socio-Demographic Characteristics of Study Participants

Characteristic	Frequency (n=3,092)	Percentage (%)
Age (years)		
15-19	423	13.68
20-24	1,081	34.96
25-29	875	28.30
30-34	389	12.58
35+	324	10.48
Ethnicity		
Akan	1,570	50.78
Dagbani/Mole	253	8.18
Ewe	396	12.81
Ga Adangbe	240	7.76
Grusi	258	8.34
Gruma	20	0.65
Hausa	77	2.49
Foreign	278	8.99
Marital Status		
Cohabiting	20	0.65
Currently married	139	4.50
Divorced	276	8.93
Separated	137	4.43
Single	2,426	78.46
Widowed	94	3.04
Level of Education		
Higher	83	2.68
JSS	1,292	41.79
Middle	69	2.23
No school	387	12.52
Primary	578	18.69
Secondary/SSS/Vocation	683	22.09
Religious Affiliation		
Christian	2466	79.75
Moslem	451	14.59
No Religion	164	5.30
Traditional	11	0.36
FSW Type		
Roamer	2,491	80.56
Seater	601	19.44
Had Anal sex		
Yes	145	4.69
No	2,947	95.31

4.2.Predictors of HIV Status

The *wrapper* method in WEKA 3.6.9 showed that 16 attributes were most effective for determining the HIV status of a female sex worker. Age group, sex work experience, relationship with most recent sexual partner, HIV prevention knowledge, highest educational level, Number of sex partners in a week, Frequency of condom use among paying partners, Had HIV test before, Marital Status, Condom use by paying client, Religion, Had anal sex, Average income from sex work, Drug usage, Drank alcohol before sex, FSW type were found to be predictors for HIV status of FSW. Table 4.2 below shows the selected variables and the contribution of each variable to predicting HIV status of FSW using the *information gain* attribute evaluator algorithm in WEKA 3.69.

Table 0.2: Likely attributes for predicting the HIV status of FSW

Rank	Attribute	Attribute Contribution	Data type	Categories	Missing values (%)
1	Age group	0.042699	Text	5	0%
2	Sex work experience	0.038818	Text	4	0%
3	Relationship with most recent sexual partner	0.02972	Text	4	0%
4	HIV prevention knowledge	0.26303	Text	3	0%
5	Highest educational level	0.026154	Text	6	0%
6	Number of sex partners in a week	0.016114	Text	2	0%
7	Frequency of condom use among paying partners	0.013456	Text	4	0%
8	Had HIV test before	0.01156	Text	2	0%
9	Marital Status	0.006557	Text	6	0%
10	Condom use by paying client	0.004897	Text	2	0%
11	Religion	0.00434	Text	4	0%
12	Had anal sex	0.002009	Text	2	0%
13	Average income from sex work	0.001882	Text	4	0%
14	Drug usage	0.000761	Text	2	0%
15	Drank alcohol before sex	0.000469	Text	2	0%
16	FSW type	0.00017	Text	2	0%

4.3.Classification

The tables (table 4.3 – 4.7) below shows the results for testing each of the five algorithms with the four respective testing options. After the four parameters were used, the *training* test parameter was selected since it achieved the best classification accuracy.

Table 0.3: Results from J48 for the various testing parameters

Test parameters	TP	Precision	Recall	ROC	Class
Training	0.993	0.933	0.993	0.789	Negative
	0.508	0.917	0.508	0.739	Positive
Cross validation	0.988	0.932	0.988	0.76	Negative
	0.503	0.857	0.503	0.76	Positive
Percentile	0.986	0.928	0.986	0.763	Negative
	0.478	0.831	0.478	0.763	Positive
Supply	0.993	0.933	0.993	0.579	Negative
	0.015	0.143	0.015	0.0579	Positive

Table 0.4: Results from Random Tree for the various testing parameters

Test parameters	TP	Precision	Recall	ROC	Class
Training	0.996	0.991	0.996	0.999	Negative
	0.939	0.974	0.939	0.999	Positive
Cross validation	0.934	0.933	0.934	0.737	Negative
	0.541	0.542	0.541	0.737	Positive
Percentile	0.936	0.931	0.936	0.74	Negative
	0.522	0.543	0.522	0.74	Positive
Supply	0.996	0.993	0.996	0.999	Negative
	0.898	0.946	0.898	0.999	Positive

Table 0.5: Results from Naïve Bayes for the various testing parameters

Test parameters	TP	Precision	Recall	ROC	Class
Training	0.954	0.932	0.954	0.799	Negative
	0.523	0.625	0.523	0.799	Positive
Cross validation	0.953	0.931	0.953	0.789	Negative
	0.515	0.616	0.515	0.789	Positive
Percentile	0.96	0.93	0.96	0.788	Negative
	0.507	0.648	0.507	0.788	Positive
Supply	0.954	0.935	0.954	0.606	Negative
	0.087	0.121	0.087	0.606	Positive

Table 0.6: Results from Neural Network for the various testing parameters

Test parameters	TP	Precision	Recall	ROC	Class
Training	0.996	0.975	0.996	0.903	Negative
	0.824	0.967	0.824	0.903	Positive
Cross validation	0.96	0.933	0.96	0.784	Negative
	0.523	0.655	0.523	0.784	Positive
Percentile	0.952	0.932	0.952	0.795	Negative
	0.522	0.614	0.522	0.795	Positive
Supply	0.996	0.976	0.996	0.806	Negative
	0.668	0.923	0.668	0.806	Positive

Table 0.7: Results from Logistic Regression for the various testing parameters

Test parameters	TP	Precision	Recall	ROC	Class
Training	0.993	0.913	0.993	0.809	Negative
	0.349	0.873	0.349	0.809	Positive
Cross validation	0.99	0.931	0.99	0.788	Negative
	0.349	0.84	0.349	0.788	Positive
Percentile	0.993	0.909	0.993	0.8	Negative
	0.321	0.878	0.321	0.8	Positive
Supply	0.993	0.974	0.993	0.627	Negative
	0.026	0.007	0.026	0.627	Positive

The performance of the predictive models used in this study were measured based on accuracy, sensitivity and specificity as summarized in table 4.8 and 4.9. The random tree algorithm found accuracy, sensitivity, and specificity to be 98.9%, 99.1% and 97.4% respectively. Also the decision tree of J48 algorithm showed an accuracy of 93.18%, followed by sensitivity of 93.3% and then a specificity value of 91.7%. Additionally using Naïve Bayes algorithm found an accuracy to be 89.97%, sensitivity to be 93.2% and specificity to be 62.5%. However, using neural networks model an accuracy of 97.41%, a sensitivity of 97.5% and specificity of 96.7% were identified. Furthermore, logistic regression achieved an accuracy of 91.12%, a sensitivity of 91.3% and a specificity of 87.3%.

Table 0.8: Comparison of the five algorithms

Classification Technique	Class: HIV status	Precision	Recall	ROC	Accuracy (%)
J48	Negative	0.933	0.993	0.789	93.18
	Positive	0.917	0.508	0.789	
Random tree	Negative	0.991	0.996	0.999	98.90
	Positive	0.974	0.939	0.999	
Naïve Bayes	Negative	0.932	0.954	0.799	89.97
	Positive	0.625	0.523	0.799	
Neural Network	Negative	0.975	0.996	0.903	97.41
	Positive	0.967	0.824	0.903	
Logistic	Negative	0.913	0.993	0.809	91.12
	Positive	0.873	0.349	0.809	

It is clear from the table that the random tree algorithm is the best model predictor among the five algorithms with an accuracy of 98.9%. This means that the random tree algorithm has a high accuracy for predicting the HIV status of a female sex worker based on the 16 attributes selected.

The complete confusion matrix is presented in Table 4.9 which gives a matrix illustration of the results from the classification. Additionally, it summarizes the sensitivity and specificity percentages for each algorithm. The random tree algorithm had the highest sensitivity that is predicting positive FSW who are actually positive and specificity (predicting negative FSW who are actually negative). The random tree and naïve Bayes had the least computational time.

Table 0.9: Summary performance of the five algorithms

Testing criteria	J48		Radom tree		Naïve Bayes		Neural network		Logistic regression	
	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive
Confusion matrix	2,682	18	2,690	10	2,577	123	2,689	11	2,680	20
	193	199	24	368	187	205	69	323	255	137
Accuracy (%)	93.18		98.9		89.97		97.41		91.12	
Sensitivity (%)	93.3		99.1		93.2		97.5		91.3	
Specificity (%)	91.7		97.4		62.5		96.7		87.3	
Area under the ROC (%)	78.9		99.9		79.9		90.3		80.9	
Computation time (seconds)	0.02		0.01		0		35.53		0.46	

4.4.The Receiver Operating Characteristic (ROC) Curve Results

Figure 4.2 to 4.6 below shows the area under the ROC curve for the instances of HIV status of FSW was produced from each of the 5 algorithms or models. The vertical axis (Y-axis) of ROC curve represents the true tested rate. The curves are represented with true tested rate on the vertical axis and false tested rate on the horizontal axis. The HIV status class value (positive) gives the ROC accuracy for the J48, random tree, Naïve Bayes, neutral network and logistic as 78.9%, 99.9%, 79.9%, 90.3 and 80.9% respectively. The ROC are all closer to 1 as shown in the figures below.

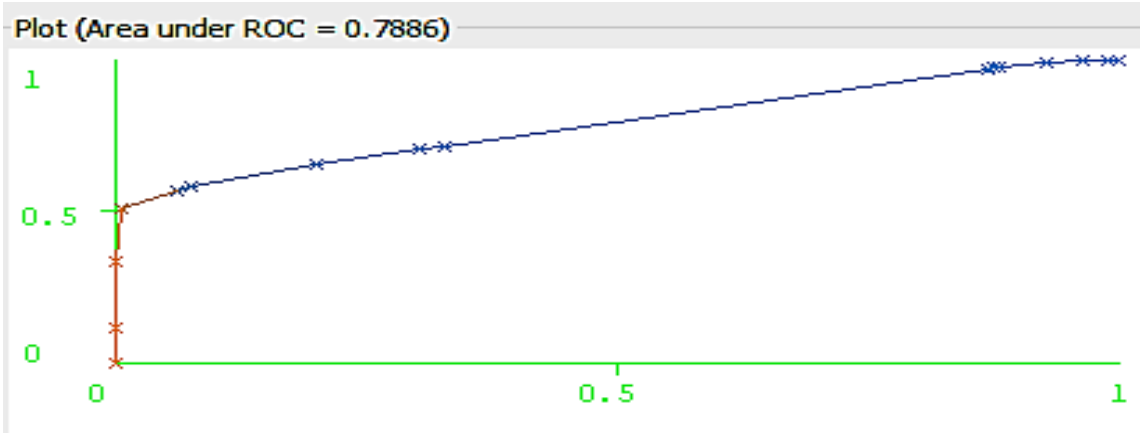

Figure 0.1: The ROC curve analysis – J48

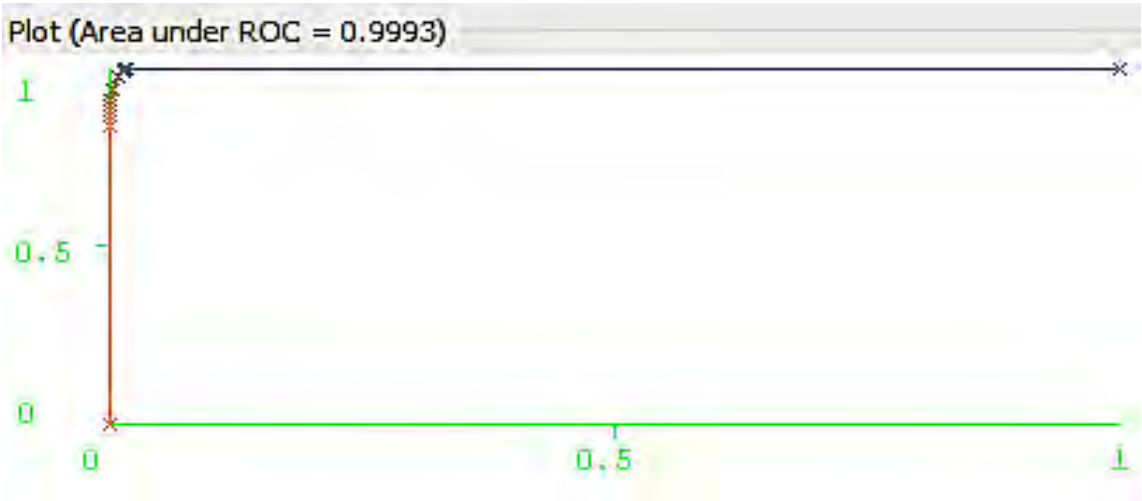

Figure 0.2: The ROC curve analysis – Random tree

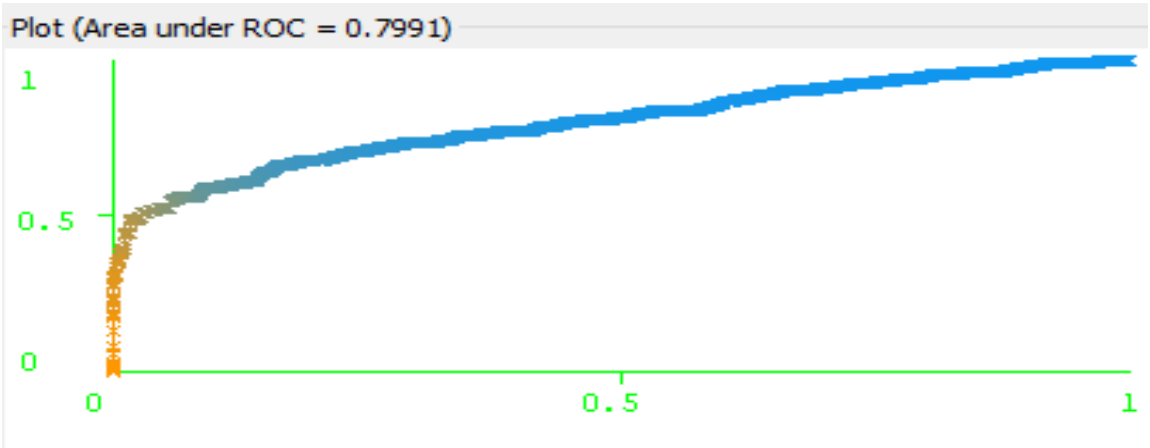

Figure 0.3: The ROC curve analysis – Naïve Bayes

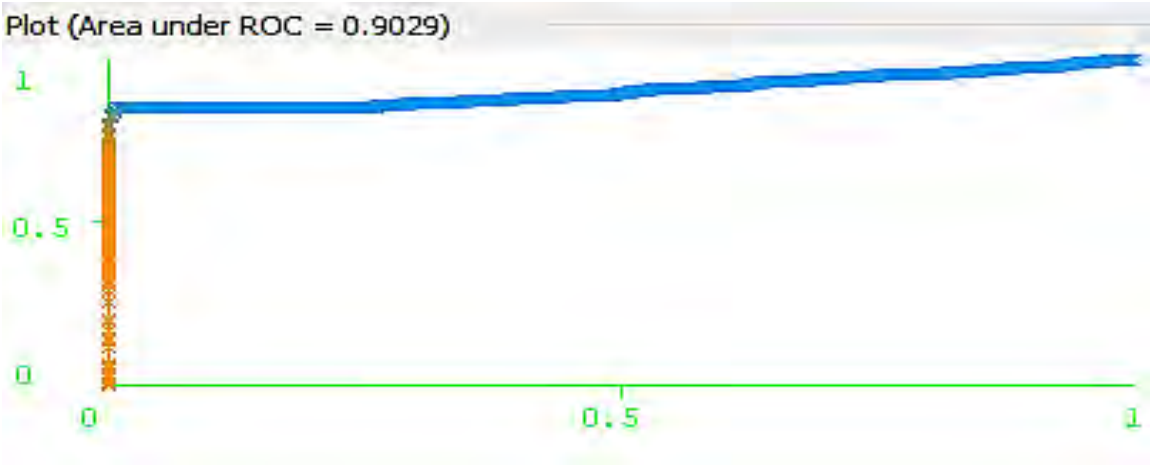

Figure 0.4: The ROC curve analysis – Neural network

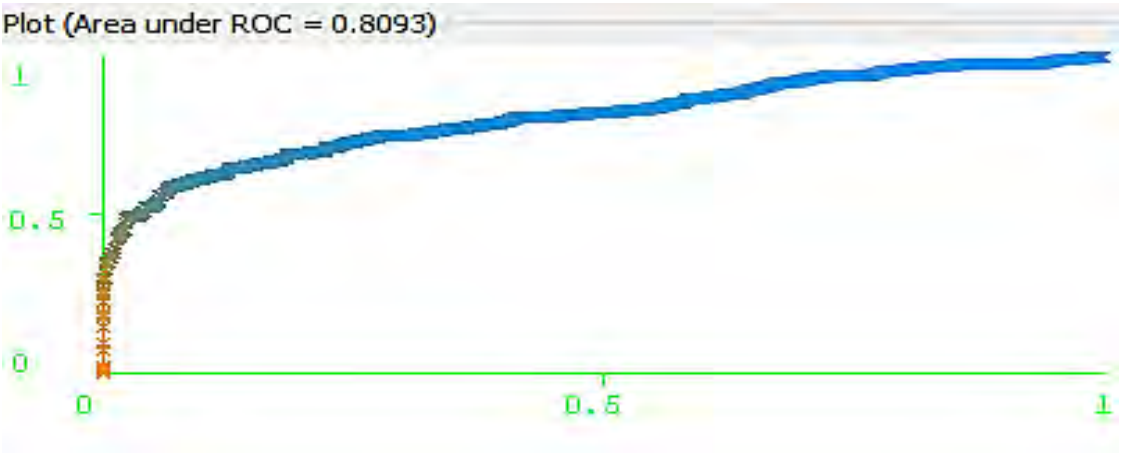

Figure 0.5: The ROC curve analysis – Logistic

This gives an indication that using the 16 attributes as input, the classifiers are better used as compared to a random model to determine the HIV status of an FSW (whether negative and positive) since the ROC curve scores for the five algorithms are above 50%. The curves also show sharp move up from 0, meaning that there is the presence of more truly testing positive than falsely testing positive rates and follows the flattened horizontal as it encounters less truly testing positive and more falsely testing positive rates. Furthermore, the random tree and neural network algorithms will be most preferred since it achieved a better accuracy than the other classifiers. Figure 4.2 to 4.6 shows the performances of each algorithm.

4.5. Behavioral Characteristics

The behavioral characteristics of most FSW make them highly probable to testing positive for HIV. These behavioral characteristics include drinking before sexual contact and sometimes less frequency in the use of condoms with clients. It is therefore important for policy makers and any HIV intervention programs to assess how these characteristics can lead to classify the HIV status of a FSW. This study tried to investigate behavioral characteristics that can help classify the HIV status of a FSW through the *wrapper* method in WEKA by limiting the attributes to only behavioral attitudes.

Table 0.10: Behavioral attributes predicting the model

Rank	Attribute	Attribute Contribution
1	Sex work experience	0.038818
2	Relationship with most recent sexual partner	0.02972
3	HIV prevention knowledge	0.26303
4	Number of sex partners in a week	0.016114
5	Frequency of condom use among paying partners	0.013456
6	Had HIV test before	0.01156
7	Had anal sex	0.002009
8	FSW type	0.00017

The results shows that 8 behavioral attributes can classify the HIV status of a female sex worker as shown in table 4.10 above.

4.5.1. Behavioral association rules

The *apriori* algorithm in WEKA was used in implementing the association rule to identify relationships and frequent patterns that exist among the selected behavioral attributes and the outcome attribute. The 10 best rules are shown in figure 4.7.

```
Best rules found:

 1. payingClient_often_condom=Every time 2474 ==> had_analsex=No 2383    conf:(0.96)
 2. payingClient_often_condom=Every time had_HIVtest=Yes 1886 ==> had_analsex=No 1816    conf:(0.96)
 3. payingClient_often_condom=Every time HIV_status=Negative 2110 ==> had_analsex=No 2025    conf:(0.96)
 4. FSW_Type=Roamer payingClient_often_condom=Every time 1920 ==> had_analsex=No 1842    conf:(0.96)
 5. had_HIVtest=Yes 2233 ==> had_analsex=No 2135    conf:(0.96)
 6. num_sexPartner_pastweek=Greater than 5 1939 ==> had_analsex=No 1850    conf:(0.95)
 7. had_HIVtest=Yes HIV_status=Negative 1895 ==> had_analsex=No 1803    conf:(0.95)
 8. HIV_status=Negative 2700 ==> had_analsex=No 2563    conf:(0.95)
 9. FSW_Type=Roamer 2491 ==> had_analsex=No 2363    conf:(0.95)
10. FSW_Type=Roamer HIV_status=Negative 2169 ==> had_analsex=No 2049    conf:(0.94)
```

Figure 0.6: Apriori 10 Best Rules

The association rule extracted some interesting patterns which will be helpful in integrating into any HIV and epidemiological programs. Some of the rules state that FSW who have tested for HIV before and whose test result showed HIV negative are more likely not indulge in anal sex (rule 5). Also for FSW who test negative to HIV are more likely not indulge in anal sex (rule 7). This association rule also demonstrated there is relationship between having anal sex and the HIV status of a FSW.

CHAPTER 5

DISCUSSION

5.1. Determinants of HIV status

Globally, it has been suggested that having universal access to HIV prevention, treatment and care serves as the guiding principle for any HIV intervention, and HIV Testing Services is considered a key and critical intervention and entry point for other interventions. It is beneficial to know one's HIV status since it allows for the adoption of healthier sexual lifestyle to maintain one's status if test results show negative. On the other hand, testing positive will ensure the individual receives counseling for healthy living and access to life saving antiretroviral therapy. This in effect shows that knowing your status helps to reduce incidence and prevalence of HIV.

It has been established by research that, Female Sex Workers are considered to be generally at a higher risk of contracting HIV as compared to other groups. Consequently, it is crucial for FSWs to know their HIV status to help them make well informed choices about healthy lifestyle. This study identified certain socio demographic and behavioral characteristics as the determinants of HIV status of FSWs in Ghana. The socio demographic characteristics include age, highest educational level, marital status, and average income from sex work. This is similar to the findings from a study in Brazil which found low educational level, street based FSW (roamers) and sex work characteristics as associated with HIV infection (Szwarcwald et al., 2018) Also the first IBBSS for Mozambique concluded that HIV status is highly influenced by age, age started sex work, educational level, and having had a genital ulcer (Augusto & Young, 2015). This is similar to the finding from this study.

Furthermore the behavioral characteristics of FSW which are determinants of HIV status identified from this study include sex work experience, relationship with most recent sexual

partner, HIV prevention knowledge, number of sex partners in a week, frequency of condom use among paying partners, had HIV test before, condom use by paying client, had anal sex, drug usage, drank alcohol before sex and FSW type. A study by Medhi et al.(2012) also found similar characteristics to be the determinants of HIV. This is also collaborated by the findings from a study conducted by Augusto & Young (2015) who found some of the variables in this study to be significantly associated with HIV. These characteristics as identified by this study shows how important and critical these attributes and characteristics are to determining the HIV status of FSWs.

5.1.1. Socio demographic characteristics

Based on the finding of this study, it is clear that certain socio demographic characteristics are important since they can influence the HIV status of an individual. This is corroborated by a study conducted by Mirandola et al., (2016) which found that socio demographic characteristics like age and educational status have a direct impact on testing seeking behavior of men who have sex with men (MSM). In another study conducted in Nigeria by Fagbamigbe, Adebayo, & Idemudia (2016) it was realized that marital status was highly associated with HIV infection among women. The study concluded that HIV is more prevalent among formerly married women compared with the never married and currently married women in Nigeria.

It is also significant to note that another study conducted in Nepal using results from the FSW IBBSS investigated the most important determinants of HIV using logistic regression models (Kakchapati, Singh, Rawal, & Lim, 2017). The study found that age, educational status and sex work experience and street based FSW were significant to determine the presence of HIV in female sex workers. This is similar to the findings of this study which associates the presence of HIV in FSW to be from these same characteristics. Another major determinant of HIV

among FSW as shown by this study is religion, and this is corroborated by the Oliveira, Faria, Gaio, & Reis (2017) in their study titled "Data Mining in HIV-AIDS Surveillance System".

Due to varying socio demographic characteristics of FSW especially in Ghana, urgent efforts are required to target FSW who have different socio demographic backgrounds (Laar, Sutherland, Ankomah, Asampong, & Dako-Gyeke, 2014). This should be part of the routine HIV prevention programme targeted at FSW to ensure that they are well armed to live healthier lifestyles.

5.1.2. Behavioural characteristics

Female Sex Workers (FSW) by their nature are associated with high risk of contracting the HIV infection because of the nature of their work. This study shows that it is possible to predict the HIV status of a female sex worker in Ghana knowing at least 8 behavioural characteristics (Sex work experience, Relationship with most recent sexual partner, Number of sex partners in a week, Frequency of condom use among paying partners, Had HIV test before, Had anal sex, FSW type and HIV prevention knowledge) and a combination of socio demographic and behavioural characteristics.

A study by Vuylsteke et al., (2003) established that there is an association between condom use by FSW and HIV status and reported a high condom use among FSW in Abidjan. This claim is also supported by a study by Richter, Chersich, Temmerman, & Luchters (2013) which concluded that behaviour characteristics like condom use during anal sex and alcohol and drug consumption can influence the chance of female sex work testing positive to HIV. The presence of attributes like relationship with last sexual partner and frequency of condom use as behavioural determinants of HIV among FSW is explained by the association established in a study conducted by Richter et al. (2013). The study found that FSWs having a romantic partner

was significantly associated with number of sexual partners they will have and condom use among these sexual partners.

These behavioural characteristics are shaped by both societal and cultural foundation and can only be addressed by HIV prevention and educational programmes (Ghana AIDS Commission, 2015). This requires consistent monitoring of behavioural characteristics through regular IBBSS studies and using the information from these studies to inform HIV programming.

5.2. HIV status prediction using socio demographic and behavioural characteristics

The purpose of this study was to investigate the possibility of predicting the HIV status of FSW given certain socio demographic and behavioral characteristics about them. It found 16 attributes to be significant to the determination of HIV status. This objective was achieved using the decision tree classifier (random tree algorithm) which showed an accuracy of 98.9% in predicting the HIV status of a female sex worker. This gives an indication that it is possible to predict HIV status of FSW using certain characteristics. This conclusion contradicts the findings from a similar study conducted by (Nicole & Tim, 2006) which concluded that demographic data is not enough to accurately predict the HIV status of an individual and hence inadequate for medical classification. The study found an accuracy of 62% for the prediction of HIV status of an individual and accuracy value can attributed to weaknesses identified in the study. The weaknesses data mining algorithm used, sample size too small, imbalance nature of the outcome variable and how missing values were handled.

A study by Zewdu (1998) also supports the findings of this research and disproves the finding by Nicole & Tim (2006). Zewdu (1998) in his study found out that it is possible to classify the HIV status of individual given certain socio demographic, behavioral and clinical results. The study compared the accuracy of 11 algorithms (did not include random tree, neural networks

and logistic algorithms) and found that decision tree (Pruned J48 algorithm) was the best classifier with an overall accuracy of 93.95% in predicting the HIV status of an individual. Hailu (2015) in his study to compare the prediction power of data mining techniques concluded that it is possible to determine whether an individual will test for HIV or not and found the decision tree (J48 algorithm) as the best classifier using the demographic characteristics. This support the claim found in this study that decision trees are better and faster classifiers.

5.3. Performance of study algorithms in predicting HIV

The purpose of this objective was to examine five popular data mining algorithms that can be used to predict the HIV status of an FSW in Ghana using the 2015 IBBSS. These algorithms were made up of Decision tree (J48 and random tree), Naïve Bayes, Neural network and logistic regression. Evaluating the results from the experiment, the accuracy, sensitivity, specificity and the area under the curve from the experiments were considered. This study discovered that the random tree algorithm is the best with an accuracy of 98.9%, the next best was neural networks with an accuracy of 97.41%, then J48 (93.18%) and logistic regression (91.12%) in that order. Naïve Bayes was the last with an accuracy of 89.97%. This gives a clear indication that random tree, neural network, J48 and logistic regression and Naïve Bayes classifiers are able to predict if a female sex worker will test negative or positive for HIV.

Kamruzzaman & Jehad Sarkar (2011), suggests that neural networks are best at identifying relationships that exist in data sets and are deployed in database technology. Neural Networks are used to identify complex relationships from inputted data. (A. I. R. L. Azevedo & Santos, 2008). This is a sharp contrast to the finding of this study which showed that random tree is the best predictor of HIV status for FSW. The study conducted by Nicole & Tim, (2006) using neural networks to predict HIV status had an accuracy of 61% and study revealed that there was no complex relationship among the attributes hence the low prediction rate. Also the neural

network proved to be a bad predictor because the data set used for the study had a small sample size, besides it was difficult to update and interpret the rules generated by neural networks. Additionally, there was an imbalance in the number for the two classes (negative and positive HIV status) hence creating a situation of over-fitting which reduces the accuracy on the test data. In view of this weakness the study was corrected in this study which gave an accuracy of 97% using neural networks. Additionally, Logistic regression is best suitable for binary classification of the outcome variable which is the case for this study. The logistic regression measures the relationship between dependent variable and independent variables by estimating probabilities using the underlying logistic function and then transforming the probabilities into binary values in order to actually make a prediction. This is a common predictive tool used in many sub disciplines under medical research. The results from this study though demonstrated that logistic regression could be used to predict HIV status with an accuracy of 91.9%. This was not high enough and not the best classifier as compared to decision tree. A study by Idowu, Aladekomo, Agbelusi, Alaba, & Balogun (2017), in predicting the pediatric HIV and AIDS survival in Nigeria using Naïve Bayes' Approach found prediction accuracy of 81.02% .

Though these mentioned algorithms are powerful predicting classifiers, the random tree stands tall amongst them all for this study. This is because the algorithm under the decision tree creates an attractive environment for data mining task. It is also supported by the fact that the all the attributes used for this study were transformed into categories and hence making the resultant classification model easy to understand by users. Also, the complexity of the tree constructed by the algorithm has an essential effect on its accuracy (Breiman, 2017). This is supported by the claim by Breiman (2017), that for most decision trees algorithm the tress constructed are relatively fast and accuracies from the decision trees are better as compared to classifiers. This evidence is validated in this study where the computational time for random tree and J48 was

less than 0.02 seconds. This largely makes the attributes for this study more convenient for random tree hence providing the best predictive power.

In addition the association rule mining was also performed for the behavioral attributes by implementing the apriori algorithm to identify frequent patterns with the selected predictors and HIV status. This has shown that FSW who have tested for HIV before and had results showing negative were more likely not to indulge in anal sex. The association rule extracted from this study indicated that because an FSW tested negative to HIV for previous test, they are more likely to make healthier life choices which include not indulging in risky behaviors like having anal sex. It is also clear FSW who tested negative are also trying to avoid complications with risky behavior which in effect has other complications which are or not associated with HIV. In addition, this indicates that HIV status and having anal sex are interlinked and have many common objectives (Kumar, 2014). This is consistent with the findings of a research conducted by Bancroft, Carnes, & Janssen (2005) which found that anal sex can be classified under sexual risk variables for determining HIV status. The authors also found that anal sex with last clients was frequent among FSW and higher among professional FSW than among nonprofessionals. They further concluded that FSW HIV positivity was significantly associated with anal sex with last clients.

5.4. Limitations of the Study

This study has some limitations which included information provided by FSW for variables like “Had HIV test before” and “Condom use by paying client” was mostly based on self-reported responses. In effect the study result can be affected by recall bias if the case of testing and condom use did not occur recently, and also by FSW not preferring to give their desirable answers. Another limitation of this study is that comprehensive knowledge on HIV could not

be calculated since the data set collected from the Ghana AIDS Commission did include variables about common misconception about HIV and AIDS.

The other major limitation of this study is that the primary data source (2015 FSW IBBSS) is a cross-sectional survey and does require following on FSW to monitor changes in behavior over time. Also, attributes selected were based on business process analysis and facts from literature. Additionally, data mining requires that the extracted knowledge be exploited. This means the knowledge discovered in this study must be evaluated by health experts to determine if they are logical, actionable and novel to fuel new biological and clinical research directions.

CHAPTER 6

CONCLUSION AND RECOMMENDATION

6.1. Conclusion

This study employed the use of supervised learning as a classification technique to classify the HIV status of a female sex worker. The study showed that the predictive models: random tree, neural networks, J48 algorithms, logistic regression and naïve bayes were able to predict whether a female sex worker will test negative or positive for HIV given certain socio-demographic and behavioral factors as inputs with an accuracy of 98.9%, 97.41%, 93.18%, 91.12% and 89.97% respectively. The socio-demographic and behavioral factors considered include age, highest educational level, marital status, average income from sex work, sex work experience, relationship with most recent sexual partner, HIV prevention knowledge, number of sex partners in a week, frequency of condom use among paying partners, had HIV test before, condom use by paying client, religion, had anal sex, drug usage, drank alcohol before sex and FSW type. The findings of this study has shown that demographic and behavioral data are sufficient to predict HIV status of an FSW.

Furthermore, the association rule extracted for behaviour characteristics from this study indicated that female sex workers who test negative for HIV were more likely not indulging in anal sex than female sex workers who test positive for HIV. In conclusion, the results obtained from this study shows that data mining techniques are relevant and critical to extracting relevant knowledge and information for the effective HIV prevention intervention targeted at FSW and key populations in Ghana. This will lead to exploring more on the application of data mining approaches in the health sector of Ghana.

6.2. Recommendation

Based on the findings of the study the following are recommended for consideration by the Ghana AIDS Commission:

1. There is the need to continue, sustain and intensify all HIV prevention interventions which focuses on the educating FSW on how to prevent HIV. This is because behavioural characteristics like condom usage, indulge in anal sex, and the number of sex partners in week are associated with the HIV status of a female sex worker.
2. An aggressive approach in reaching FSW with HIV prevention intervention since there is a low knowledge among about HIV intervention among FSW. This is to be followed up with HIV Testing Services which will help FSW to know their status

6.3. Suggestions for Future Research

Future research should apply other feature selection algorithms in WEKA for the purposes of comparison since attributes selected for this study was based on business process analyses and facts from literatures because data set in this domain is usually too imbalanced for conventional algorithms to perform equally and correctly on the two classes. Additionally future research in this area should also focus on using the IBBSS data for MSM and compare the results with results from this study.

REFERENCES

Adu-Oppong, A., Grimes, R. M., Ross, M. W., Risser, J., & Kessie, G. (2007). Social and Behavioral Determinants of Consistent Condom Use Among Female Commercial Sex Workers in Ghana. *AIDS Education and Prevention*, *19*(2), 160–172. https://doi.org/10.1521/aeap.2007.19.2.160

AIDS MAP. (2018). HIV & AIDS Information :: High uptake of HIV self-testing by female sex workers in African countries. Retrieved June 14, 2018, from http://www.aidsmap.com/page/3159057/

Aklilu, M., Messele, T., Tsegaye, A., Biru, T., Mariam, D. H., van Benthem, B., … Fontanet, A. (2001). Factors associated with HIV-1 infection among sex workers of Addis Ababa, Ethiopia. *AIDS (London, England)*, *15*(1), 87–96. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11192872

Aleksey, M. (2017). CLASSIFICATION OF TASKS OF DATA MINING, *3*(3).

Ankomah, A., Omoregie, G., Akinyemi, Z., Anyanti, J., Ladipo, O., & Adebayo, S. (2011). HIV-related risk perception among female sex workers in Nigeria. *HIV/AIDS - Research and Palliative Care*, *3*, 93. https://doi.org/10.2147/HIV.S23081

Ashraf, N., Ahmad, W., & Ashraf, R. (2018). A Comparative Study of Data Mining Algorithms for High Detection Rate in Intrusion Detection System, *2*(1), 49–57.

Augusto, Â. do R., & Young, P. (2015). High burden of HIV infection and risk behaviors among female sex workers in three main Urban areas of Mozambique. *AIDS and Behavior*, *20*(4), 799–810. https://doi.org/10.1007/s10461-015-1140-9.High

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS - DM*. Retrieved from http://recipp.ipp.pt/handle/10400.22/136

Azevedo, A., & Santos, M. F. (2014). KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW, (June).

Bancroft, J., Carnes, L., & Janssen, E. (2005). Unprotected Anal Intercourse in HIV-Positive and HIV-Negative Gay Men: The Relevance of Sexual Arousability, Mood, Sensation Seeking, and Erectile Problems. *Archives of Sexual Behavior*, *34*(3), 299–305. https://doi.org/10.1007/s10508-005-3118-6

Baral, S., Beyrer, C., Muessig, K., Poteat, T., Wirtz, A. L., Decker, M. R., … Kerrigan, D. (2012). Burden of HIV among female sex workers in low-income and middle-income countries: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, *12*(7), 538–549. https://doi.org/10.1016/S1473-3099(12)70066-X

Berka, P., & Rauch, J. (2010). Machine Learning and Association Rules. *Proceedings of COMPSTAT'2010*, (2010). Retrieved from https://www.rocq.inria.fr/axis/COMPSTAT2010/TU_Berka-Rauch_paper.pdf

Blanchard, J. F., O'Neil, J., Ramesh, B. M., Bhattacharjee, P., Orchard, T., & Moses, S. (2005). Understanding the Social and Cultural Contexts of Female Sex Workers in Karnataka, India: Implications for Prevention of HIV Infection. *The Journal of Infectious Diseases*, *191*(s1), S139–S146. https://doi.org/10.1086/425273

Boily, M.-C., Baggaley, R. F., Wang, L., Masse, B., White, R. G., Hayes, R. J., & Alary, M.

(2009). Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. *The Lancet Infectious Diseases*, *9*(2), 118–129. https://doi.org/10.1016/S1473-3099(09)70021-0

Bramer, M. (2003). *Principles of Data Mining* (Vol. 2001).

Breiman, L. (2017). *Classification and Regression Trees*. Routledge. https://doi.org/10.1201/9781315139470

Brown, T., & Mills, S. (2000). Behavioral urveillance Surveys. *Health San Francisco*, *81*(1), 33–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19738187

CDC. (2018). About HIV/AIDS | HIV Basics | HIV/AIDS | CDC. Retrieved May 29, 2018, from https://www.cdc.gov/hiv/basics/whatishiv.html

Cenzer, I. S., & Lee, S. J. (2000). Converting Electronic Medical Records Data into Practical Analysis Dataset. Retrieved from https://www.lexjansen.com/wuss/2013/120_Paper.pdf

Delen, D., & Olson, D. L. (2008). 2 Data Mining Process. *Advanced Data Mining Techniques*, 9–35.

Elhadi, M., Elbadawi, A., Abdelrahman, S., Mohammed, I., Bozicevic, I., Hassan, E. A., … Setayesh, H. (2013). Integrated bio-behavioural HIV surveillance surveys among female sex workers in Sudan, 2011-2012. *Sexually Transmitted Infections*, *89 Suppl 3*(Suppl 3), iii17-22. https://doi.org/10.1136/sextrans-2013-051097

Fagbamigbe, A. F., Adebayo, S. B., & Idemudia, E. (2016). Marital status and HIV prevalence among women in Nigeria: Ingredients for evidence-based programming. *International Journal of Infectious Diseases*, *48*, 57–63. https://doi.org/10.1016/J.IJID.2016.05.002

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *Social Science Research*, *17*, 37. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=14728591308876110427related:W1p3j8J3ZswJ%5Cnpapers2://publication/uuid/ABD8C83F-C4AA-4694-8369-5651091CBA9E

Fettig, J., Swaminathan, M., & Murrill, C. (2014). Global epidemiology of HIV. *Infectious Disease Clinics of North America*, *28*(3), 323–37. https://doi.org/10.1016/j.idc.2014.05.001

Foster, C., Pace, M., Kaye, S., Hopkins, E., Jones, M., Robinson, N., … Frater, J. (2017). Early antiretroviral therapy reduces HIV DNA following perinatal HIV infection. *AIDS*, *31*(13), 1847–1851. https://doi.org/10.1097/QAD.0000000000001565

Gaffey, M. F., Venkatesh, S., Dhingra, N., Khera, A., Kumar, R., Arora, P., … Jha, P. (2011). Male Use of Female Sex Work in India: A Nationally Representative Behavioural Survey. *PLoS ONE*, *6*(7), e22704. https://doi.org/10.1371/journal.pone.0022704

Gangakhedkar, R. R., Bentley, M. E., Divekar, A. D., Gadkari, D., Mehendale, S. M., Shepherd, M. E., … Quinn, T. C. (1997). Spread of HIV Infection in Married Monogamous Women in India. *JAMA: The Journal of the American Medical Association*, *278*(23), 2090. https://doi.org/10.1001/jama.1997.03550230066039

Ghana AIDS Commission. (2015). Bio-Behavioral Survey Among Female Sex Workers and

Their Non-Paying Partners in Ghana. Ibbss. *[2015 Bio-Behavioral Survey among FSW and Their Non-Paying Partners in Ghana. FIRST DRAFT REPORT: JUNE 21ST 2016*. Retrieved from www.ghananaids.gov.gh

Ghana AIDS Commission. (2016). *National HIV and AIDS Strategic Plan 2016-2020*.

Global Network of Sex Work Projects. (2015). Mapping and Population Size Estimates of Sex Workers: Proceed with Extreme Caution.

Haile Mariam, T. (2015). Application of Data Mining Techniques for Predicting CD4 Status of Patients on ART in Jimma and Bonga Hospitals, Ethiopia. *Journal of Health & Medical Informatics*, *6*(6). https://doi.org/10.4172/2157-7420.1000208

Hailu, T. G. (2015). Comparing Data Mining Techniques in HIV Testing Prediction, (May), 153–180.

Han, J. (2007). Data Mining: Concepts and Techniques. *University of Illinois at Urbana-Champaign*, 1–136. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Hand, D. J. (2007). Principles of Data Mining. *Drug Safety*, *30*(7), 621–622. https://doi.org/10.2165/00002018-200730070-00010

Idowu, P. A., Aladekomo, T. A., Agbelusi, O., Alaba, O. B., & Balogun, J. A. (2017). *International journal of child health and human development : IJCHD. International Journal of Child Health and Human Development* (Vol. 10). Nova Science Publishers, Inc. Retrieved from https://www.questia.com/library/journal/1P4-1936442427/prediction-of-pediatric-hiv-aids-survival-in-nigeria

Jair, H., Palacios, G., Andrés, R., Toledo, J., Albeiro, G., Pantoja, H., & Martínez, Á. A. (2017). A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change, *2*(3), 598–604.

Kabiri, M. (2016). Factors Influencing Uptake of Hiv Testing and Counselling Among the Youth in Kintampo South District. Retrieved from http://ugspace.ug.edu.gh/handle/123456789/21668

Kakchapati, S., Singh, D. R., Rawal, B. B., & Lim, A. (2017). Sexual risk behaviors, HIV, and syphilis among female sex workers in Nepal. *HIV/AIDS - Research and Palliative Care*, *Volume 9*, 9–18. https://doi.org/10.2147/HIV.S123928

Kamruzzaman, S. M., & Jehad Sarkar, A. M. (2011). A new data mining scheme using artificial neural networks. *Sensors (Basel, Switzerland)*, *11*(5), 4622–47. https://doi.org/10.3390/s110504622

Krzysztof, J. C., Witold, P., Roman, W. S., & Lukasz, A. K. (2007). *Data Mining: A Knowledge Discovery Approach - Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz Andrzej Kurgan - Google Books*. Springer. Retrieved from https://books.google.com.gh/books?id=YvTxwaLJJ2kC&pg=PA9&lpg=PA9&dq=logical,+cohesive,+well-thought-out+structure+and++approach+that+can+be+presented+to+decision-makers+who+may+have+difficulty+understanding+the+need,+value,+and+mechanics+behind+a+KDP&sour

Kumar, S. G. (2014). Control of HIV: Role of female sex workers in risk of HIV transmission. *Indian Journal of Sexually Transmitted Diseases and AIDS*, *35*(2), 165–6.

https://doi.org/10.4103/0253-7184.142421

Laar, A., Sutherland, E., Ankomah, A., Asampong, E., & Dako-Gyeke, P. (2014). A performance evaluation of the National HIV Prevention Program for FSW and MSM in Ghana. TT -, (July). Retrieved from http://www.cpc.unc.edu/measure/publications/tr-14-97

Leke-betechuoh, B., Marwala, T., Tim, T., & Lagazio, M. (2006). Prediction of HIV Status from Demographic Data Using Neural Networks, 2339–2344.

Medhi, G. K., Mahanta, J., Paranjape, R. S., Adhikary, R., Laskar, N., & Ngully, P. (2012). Factors associated with HIV among female sex workers in a high HIV prevalent state of India. *AIDS Care*, *24*(3), 369–376. https://doi.org/10.1080/09540121.2011.608787

Mirandola, M., Gios, L., Joanna Davis, R., Furegato, M., Breveglieri, M., Folch, C., … Stehlíková, D. (2016). Socio-demographic factors predicting HIV test seeking behaviour among MSM in 6 EU cities. *The European Journal of Public Health*, *27*(2), ckw144. https://doi.org/10.1093/eurpub/ckw144

Muula, A., & Twizelimana, D. (2015). HIV and AIDS risk perception among sex workers in semi-urban Blantyre, Malawi. *Tanzania Journal of Health Research*, *17*(3), 1–7. https://doi.org/10.4314/thrb.v17i3.

Nicole, T., & Tim, H. (2006). Predicting HIV Status Using Neural Networks and Demographic Factors, (April).

Ogaji, D. ., A.S, O., & I, I. (2013). & PRIMARY HEALTH CARE Awareness , willingness and use of Voluntary HIV testing and counseling services, *25*(2), 36–44.

Oliveira, A., Faria, B. M., Gaio, A. R., & Reis, L. P. (2017). Data Mining in HIV-AIDS Surveillance System. *Journal of Medical Systems*, *41*(4), 51. https://doi.org/10.1007/s10916-017-0697-4

Overs, C. (2002). Sex Workers Part of The Solution : An Analysis of HIV Prevention Programming to Prevent HIV Transmission During Commercial Sex in Developing Countries. *Who*.

Padhy, N. (2012). Multi Relational Data Mining Approaches : A Data Mining Technique, *57*(17).

Padhy, N., Mishra, D. P., & Panigrahi, R. (2012). The Survey of Data Mining Applications And Feature Scope. https://doi.org/10.5121/ijcseit.2012.2303

Patel, H., & Patel, D. (2016). A Comparative Study on Various Data Mining Algorithms with Special Reference to Crop Yield Prediction, *9*(June). https://doi.org/10.17485/ijst/2016/v9i22/92713

Prüss-Ustün, A., Wolf, J., Driscoll, T., Degenhardt, L., Neira, M., & Calleja, J. M. G. (2013). HIV Due to Female Sex Work: Regional and Global Estimates. *PLoS ONE*, *8*(5), e63476. https://doi.org/10.1371/journal.pone.0063476

Ramesh, B. M., Moses, S., Washington, R., Isac, S., Mohapatra, B., Mahagaonkar, S. B., … Blanchard, J. F. (2008). Determinants of HIV prevalence among female sex workers in four south Indian states: analysis of cross-sectional surveys in twenty-three districts. *AIDS*, *22*(Suppl 5), S35–S44. https://doi.org/10.1097/01.aids.0000343762.54831.5c

Richter, M. L., Chersich, M., Temmerman, M., & Luchters, S. (2013). Characteristics, sexual behaviour and risk factors of female, male and transgender sex workers in South Africa. *South African Medical Journal = Suid-Afrikaanse Tydskrif Vir Geneeskunde*, *103*(4), 246–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23547701

Ruxrungtham, K., Brown, T., & Phanuphak, P. (2004). HIV/AIDS in Asia. *The Lancet*, *364*(9428), 69–82. https://doi.org/10.1016/S0140-6736(04)16593-8

Sacramento County Department of Health Unit. (2002). What is AIDS? *Medical Care*, *916*, 874–7720. Retrieved from http://www.dhhs.saccounty.net/PUB/Documents/AZ-Health-Info/PUB_AIDSBrochure.pdf

Scorgie, F., Chersich, M. F., Ntaganira, I., Gerbase, A., Lule, F., & Lo, Y.-R. (2012). Socio-Demographic Characteristics and Behavioral Risk Factors of Female Sex Workers in Sub-Saharan Africa: A Systematic Review. *AIDS and Behavior*, *16*(4), 920–933. https://doi.org/10.1007/s10461-011-9985-z

Srihari, S. (2000). A Systematic Overview of Data Mining Algorithms.

Sundaram, V. (2012). C OMPARATIVE S TUDY OF D ATA M INING A LGORITHMS FOR, *4*(2), 173–178.

Szwarcwald, C. L., Damacena, G. N., de Souza-Júnior, P. R. B., Guimarães, M. D. C., de Almeida, W. da S., de Souza Ferreira, A. P., … Brazilian FSW Group. (2018). Factors associated with HIV infection among female sex workers in Brazil. *Medicine*, *97*(1S Suppl 1), S54–S61. https://doi.org/10.1097/MD.0000000000009013

Taniar, D., & Chen, L. (Eds.). (2011). *Integrations of Data Warehousing, Data Mining and Database Technologies*. IGI Global. https://doi.org/10.4018/978-1-60960-537-7

U.S. Department of Veterans Affairs. (2018). What is AIDS? - HIV/AIDS. Retrieved June 14, 2018, from https://www.hiv.va.gov/patient/basics/what-is-AIDS.asp

Umair, S., & Haseeb, Q. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA), *12*(1), 217–222.

UNAIDS. (2010). New HIV Infections by mode of transmission in West Africa: A Multi - Country Analysis, 27. Retrieved from http://www.unaids.org/sites/default/files/en/media/unaids/contentassets/documents/countryreport/2010/201003_MOT_West_Africa_en.pdf

UNAIDS. (2013). *GLOBAL REPORT: UNAIDS report on the global AIDS epidemic 2013*. *Unaids*. https://doi.org/JC2502/1/E

UNAIDS. (2017). Fact sheet - Latest global and regional statistics on the status of the AIDS epidemic. *Unaids*, (June), 8. https://doi.org/2017

UNICEF. (2008). Children and HIV and AIDS - How widespread is the HIV/AIDS epidemic? Retrieved June 13, 2018, from https://www.unicef.org/aids/index_epidemic.html

Vandepitte, J., Lyerla, R., Dallabetta, G., Crabbé, F., Alary, M., & Buvé, A. (2006). Estimates of the number of female sex workers in different regions of the world. *Sexually Transmitted Infections*, *82 Suppl 3*(suppl 3), iii18-25. https://doi.org/10.1136/sti.2006.020081

Vuylsteke, B. L., Ghys, P. D., Traoré, M., Konan, Y., Mah-Bi, G., Maurice, C., … Laga, M. (2003). HIV prevalence and risk behavior among clients of female sex workers in Abidjan, Côte d'Ivoire. *AIDS (London, England)*, *17*(11), 1691–4. https://doi.org/10.1097/01.aids.0000060419.84040.61

Wariki, W. M., Ota, E., Mori, R., Koyanagi, A., Hori, N., & Shibuya, K. (2012). Behavioral interventions to reduce the transmission of HIV infection among sex workers and their clients in low- and middle-income countries. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD005272.pub3

WHO. (2018). The top 10 causes of death. Retrieved June 1, 2018, from http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

World Health Organization (WHO). (2013). WHO | Definition of key terms. *WHO*. Retrieved from http://www.who.int/hiv/pub/guidelines/arv2013/intro/keyterms/en/

Zewdu, T. (1998). Prediction of HIV Status in Addis Ababa using Data Mining Technology.