Intelligent Information Management, 2015, 7, 153-180
Published Online May 2015 in SciRes. http://www.scirp.org/journal/iim
http://dx.doi.org/10.4236/iim.2015.73014

Comparing Data Mining Techniques in HIV Testing Prediction

Tesfay Gidey Hailu

School of Interdisciplinary, Department of Statistics, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia
Email: tesfaygdey@yahoo.com

Received 28 February 2015; accepted 25 May 2015; published 28 May 2015

Abstract

Introduction: The present work compared the prediction power of the different data mining techniques used to develop the HIV testing prediction model. Four popular data mining algorithms (Decision tree, Naive Bayes, Neural network, logistic regression) were used to build the model that predicts whether an individual was being tested for HIV among adults in Ethiopia using EDHS 2011. The final experimentation results indicated that the decision tree (random tree algorithm) performed the best with accuracy of 96%, the decision tree induction method (J48) came out to be the second best with a classification accuracy of 79%, followed by neural network (78%). Logistic regression has also achieved the least classification accuracy of 74%. Objectives: The objective of this study is to compare the prediction power of the different data mining techniques used to develop the HIV testing prediction model. Methods: Cross-Industry Standard Process for Data Mining (CRISP-DM) was used to predict the model for HIV testing and explore association rules between HIV testing and the selected attributes. Data preprocessing was performed and missing values for the categorical variable were replaced by the modal value of the variable. Different data mining techniques were used to build the predictive model. Results: The target dataset contained 30,625 study participants. Out of which 16,515 (54%) participants were women while the rest 14,110 (46%) were men. The age of the participants in the dataset ranged from 15 to 59 years old with modal age of 15 - 19 years old. Among the study participants, 17,719 (58%) have never been tested for HIV while the rest 12,906 (42%) had been tested. Residence, educational level, wealth index, HIV related stigma, knowledge related to HIV, region, age group, risky sexual behaviour attributes, knowledge about where to test for HIV and knowledge on family planning through mass media were found to be predictors for HIV testing. Conclusion and Recommendation: The results obtained from this research reveal that data mining is crucial in extracting relevant information for the effective utilization of HIV testing services which has clinical, community and public health importance at all levels. It is vital to apply different data mining techniques for the same settings and compare the model performances (based on accuracy, sensitivity, and specificity) with each other. Furthermore, this study would also invite interested researchers to explore more

How to cite this paper: Hailu, T.G. (2015) Comparing Data Mining Techniques in HIV Testing Prediction. *Intelligent Information Management*, **7**, 153-180. http://dx.doi.org/10.4236/iim.2015.73014

on the application of data mining techniques in healthcare industry or else in related and similar settings for the future.

Keywords

Data Mining, Comparison, Predictive Modeling, HIV Testing, Ethiopia

1. Background

According to Ministry of Health Report (2006), the adult prevalence of HIV infection in Ethiopia was estimated at 2.1% where most of the burden occurred among the young age group [1]. Although HIV voluntary counseling and testing has been carried out in different places with less cost and a gateway to most HIV related services including provision of antiretroviral drugs [2] [3], many people (in most sub-Saharan African countries) still do not know their HIV status [4]. Some treatment programmers have also reported that high early mortality in patients receiving antiretroviral therapy because of late HIV testing [5].

Different studies have also demonstrated that early HIV testing is effective in omitting sexual risk behavior in people attending counseling and testing centers, but few studies have specifically looked at young people [6]. Therefore, even lately detection of HIV infection is beneficial for both individuals and society since it is associated with increased morbidity, mortality, and probability of transmission [7]. By linking prevention and care, HIV testing can help reduce this burden. Through early diagnosis and treatment, HIV testing leads to improved clinical outcomes [8]. It also reduces the risk of transmission since there is growing evidence that compliance with antiretroviral treatment causes individuals to be less infectious [9] and data suggest that many people reduce their sexual risk behavior after testing positive for HIV [10] [11]. Hence, to meet these benefits, it is important to promote HIV testing among adults.

Despite the various efforts made to implement HIV prevention activities [12], HIV testing is a critical issue among adults in Ethiopia though there is a good progress compared to the reports in EDHS 2005. The 2011 EDHS is the second EDHS survey to anonymously link HIV testing results with demographic, socioeconomic, and behavioral characteristics of survey respondents. As a result, the EDHS 2011 has collected huge amount of data on HIV testing services. Therefore, traditional statistical techniques were used to derive some operational information from the data but had limited capacity to discover a novel idea from huge databases. There is huge amount of information within the healthcare systems. Today, the major challenge in healthcare industry is the provision of quality services at affordable costs [13]. In records of hospitals there is a large amount of unanalyzed data that can be turned into useful information. However, there is a lack of useful analysis tools to realize hidden relationships and trends in data. In this perspective, data mining techniques can be used to automatically infer HIV testing results from descriptions of successfully tested people and can help health professionals to make the HIV testing process more objective and reliable. The main objective of this research is to develop a novel technique to categorize into two categories whether an individual was being tested for HIV or not through data mining techniques.

Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns. One of the underlying concerns of health service providers is the scaling up of the knowledge of HIV status. Beside, testing for HIV is the only means to determine HIV status. Therefore, predictive modeling is one of the data mining tasks that allow predicting the unknown value of a variable of interest (being tested for HIV) given known values of other variables. A predictive model is made up of a number of predictors, which are factors that are likely to influence future behavior or results. Thus, the goal of predictive modeling is to estimate a mapping that can predict the value of being tested for HIV given an input vector of measured values of predictors and a set of estimated parameters for the model. Different studies that have used traditional statistics to predict whether an individual was being tested for HIV had shown relatively limited accuracy [14] [15]. Moreover, traditional statistical techniques alone are not enough (have limited capacity) to discover new and unanticipated patterns and relationship that are hidden in conventional databases (EDHS) [14]. Furthermore, as to the knowledge of the researcher, no previous studies have been made to predict HIV testing among adults at a country level using data mining techniques. It is therefore postulated that the incorporation of different data mining algorithms may improve the accuracy of this prediction. The objective of this study hence is to compare the prediction power of the different data mining techniques used to develop the HIV testing prediction model.

Health programs are unable to provide appropriate HIV/AIDS care, treatment, counseling and support to individuals or families without knowing who is infected and knowledge about HIV status is determined only through HIV testing. This implies that identifying the "best predictive model" for HIV testing is critical for the effective utilization of the service which has clinical importance at individual level and public health justification at societal level. The findings of this research might be helpful for both health programmes and researchers. For the health programmes, the findings of this study may help them to design a special intervention program or improve the existing one. As a result, the population might be benefited if the service is improved or new program is designed and implemented based on the output of this study. For researchers, the study can contribute on how the application of data mining is helpful in predicting HIV testing and identifies determinants of HIV testing through different algorithms. Hence, it can also invite interested researchers to explore more in related and similar areas for the future. Similarly, the benefits of HIV testing can be seen at the individual, community and population levels.

1.1. Over View on Data Mining

Data mining is a process which finds useful patterns from large amount of data. Data mining techniques provide people with new power to research and to manipulate the existing large volume of data. Across a wide variety of fields, data are being collected and accumulated at a dramatic speed. Hence, there is an urgent need for a new generation of computational theories and tools to assist decision makers in extracting useful information (knowledge) from the rapidly growing volumes of digital data [16]. Now a day data mining technology is widely used in every field (for example, in health, agriculture, education, business, etc.) for extracting hidden knowledge which is crucial for competitive advantage and sustainable growth. In this study, it has been attempted to review and incorporate a range of standards and steps used in data mining methodologies, various concepts, theories and practices of data mining in relation with practical health care problems of HIV testing.

The finding useful patterns in data has been given a variety of names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing [17]. Many people treat data mining as a synonym for another popularly used term Knowledge discovery in database (KDD). Others view data mining as an essential step in the process of KDD. One of the aims of data mining can be seen as the analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful. This relationships and summaries derived through data mining are often referred to as models or patterns.

The data mining process basically involves selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures will be done [18]. There are many achievements of application from data mining techniques to various areas such as engineering, marketing, financial, medical and so on. In the data mining literature, various "general frameworks" have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements.

1.2. Methodologies of Data Mining

There are three popular methodologies applied in data mining research: KDD (Knowledge Discovery in Database), SEMMA (Sample, Explore, Modify, Model, and Assess) and CRISP-DM [16] [19]. CRISP-DM was used for this particular study.

1.3. Knowledge Discovery in Database (KDD)

Data Mining or "the efficient discovery of valuable, on-obvious information from a large collection of data" [20] has a goal to discover knowledge out of data and present it in a form that is easily comprehensible to humans. Knowledge detection in databases is precise process consisting of a number of distinct steps [21]. A formal definition of Knowledge discovery in databases is given as: "Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data". Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions [22]. Hence, different methodologies of data mining research attempt to shape the activities the researcher performs in a typical data mining process [23].

The five basic steps of KDD [18] are presented in **Figure 1** below.

Therefore, the KDD process can be defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [17].

1.3.1. SEMMA

The SEMMA process was developed by the SAS Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model and Assess and refers to the process of conducting a data mining project. The SAS Institute considers a cycle with 5 stages for the process. By assessing the results gained from each stage of the SEMMA process, one can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data.

1.3.2. CRISP-DM

CRISP-DM is a standard process for data mining which is a non-proprietary model. It is an application or industry neutral data mining methodology which mainly focuses on business issues as well as technical analysis [24]. CRISP-DM is the most suitable for novice data miners due to the easy to read documentation and intuitive, industry-applications-focused description [21]. It begins from understanding the business and ends with the deployment of the system. The CRISP-DM process was developed by the means of the effort of a consortium initially composed with Daimler Chryrler, SPSS and NCR. It consists on a cycle that comprises six stages [23] (**Figure 2**).

1.4. Comparisons of the Methodologies

Recently, the researcher efforts have been focused on proposing new models, rather than improving design of a single model or proposing a generic unifying model. Despite the fact that most models have been developed in isolation, a significant progress has been made these days. Most of the models provide generic and appropriate descriptions that are not tied specifically to academic or industrial needs, but rather provide a model that is independent of a particular tool, vendor, or application [24].

It has been summarized the association of the three popular methodologies as seen in **Table 1** below and most of the steps to be followed in all methods seem somehow similar [19]. Though KDD and SEMMA do not have a step before selection and sample steps respectively, obviously understanding the domain is often a requirement to perform selection or sampling. All methodologies contain the following set of main tasks in one form or another [23]. Typically, these steps are performed iteratively and not necessarily in the presented linear order.

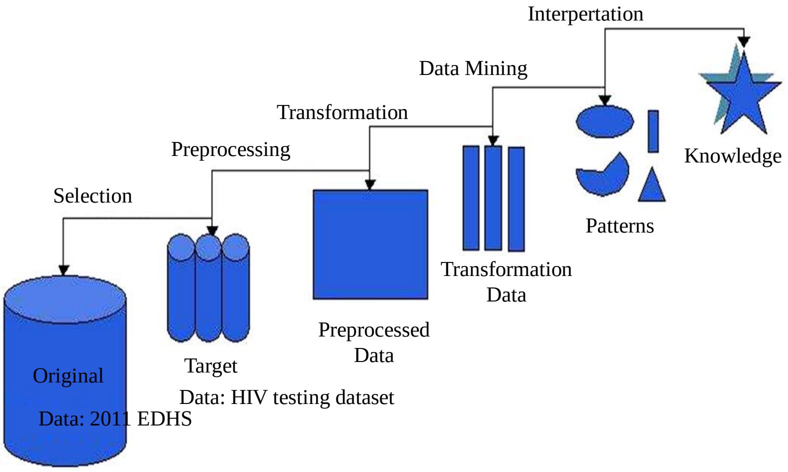


Figure 1. Overview of the steps constituting the KDD process.

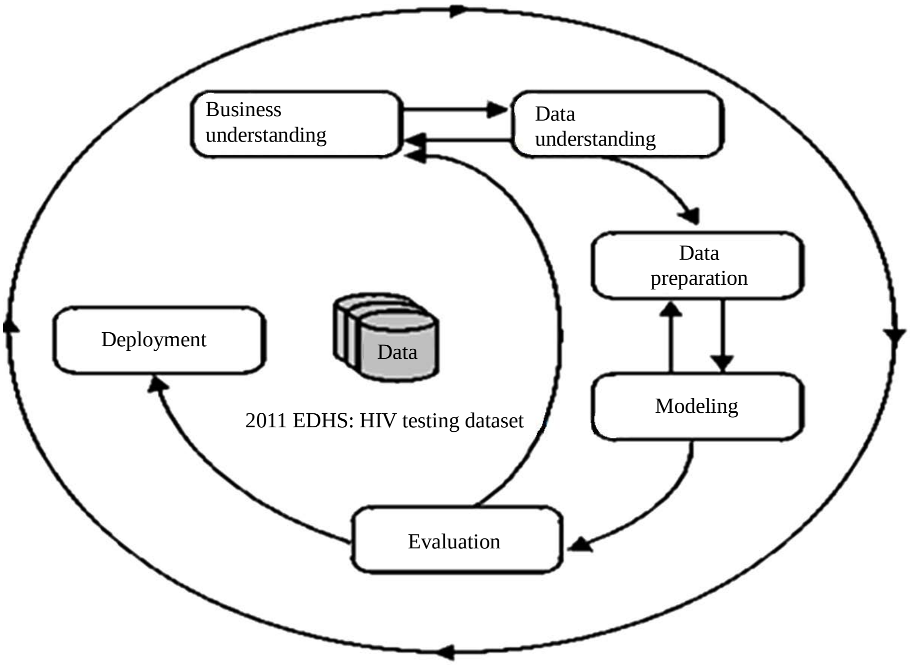


Figure 2. The CRISP-DM process model by Refaat, M. (2007) [23].

Table 1. Comparison of KDD, SEMMA and CRISP-DM methodologies.

KDD	SEMMA	CRISP-DM
Pre KDD	-----	Business understanding
Selection	Sample	Data understanding
Preprocessing	Explore	Data understanding
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	-------	Deployment

1.5. Related Works

This section presented a number of researches that demonstrate the application of data mining in health care domain in general and HIV/AIDS control and prevention program in particular are discussed.

1.5.1. Application of Data Mining in Health Care Industry

The generation and collection of various kinds of data relating to clinical practices, clinical trials, patient information, resource administration, policies and research are involved with in the workflow of health care organizations. Traditionally, statistical techniques are used to derive some operational information from the data. However, these huge amounts of data generated by health care transactions are too complex and voluminous to be processed and analyzed by traditional methods. Therefore, a new method, called data mining, provides the opportunity to derive, in an exploratory and interactive manner, valuable health care knowledge in terms of associations, sequential patterns, classifications, predictions and symbolic rules. Such inductively-derived health care knowledge can provide strategic insights into the practice delivery of health care [14].

As the medical field expands, it is the duty of each physician to evaluate and protect each patient from diseases, side effects and medical disasters. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers to detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services. Thus, physicians are now armed with data mining as a tool for expanding their knowledge base [16].

Data mining provides the methodology and technology to transform these mass of data into useful information for decision making [14]. In addition, the availability of huge amounts of medical data leads to the need for powerful data analysis tools to extract useful knowledge. Researchers have long been concerned with applying statistical and data mining tools to improve data analysis on large data sets. Disease diagnosis is one of the applications where data mining tools are providing successful results. Several researchers are using statistical and data mining tools to help health care professionals in the diagnosis of heart disease [22] [25].

1.5.2. Application of Data Mining in HIV Testing

Since the development of the neural networks (NNs), they have received considerable attention and have been applied to a variety of problems in classification and prediction. NNs have been applied successfully for development of nonparametric statistical models and more reliable outcome research has been explored in the area of pattern classification and pattern prediction [25].

Neural networks are known to be able to identify relationships even when some of the input data are very complex, ill-defined and ill-structured. One of the advantages of an NN is that it can discriminate linearly inseparable data. Recent study on neural networks made a substantial contribution to the HIV/AIDS care and prevention planning area by comparing the impact of various NN methodologies on the classification of HIV/AIDS-related persons [26]. Artificial neural networks (ANNs) also have been used to classify and predict an individual tested for HIV [26].

A study conducted by Brain *et al*. showed that Neural networks were used as pattern recognition tools in data mining to classify HIV status of individuals based on demographic and socio-economic characteristics [27]. The data under that study consists of seroprevalence survey information and contains variables such as age, education, location, race, parity and gravidity. The radial basis function (RBF) neural network architecture was used for that study since as preliminary design showed this architecture was the most optimal. Moreover, the Bayesian method of training used was approximated with the evidence framework. The design of classifiers involves the assessment of classification performance, and an accuracy of 84.24% was obtained in that design. This implies that the HIV status of an individual can be predicted using demographic data with 84.24% accuracy [27].

A study was conducted using the application of data mining technology to identify determinants of HIV infection and has noticed a new insight about risk feeling of the clients and HIV testing. According to the researcher previously it’s known that the clients whose reason for testing is plan for future are associated with HIV-negative class. This truth has also been verified with experiment too. However, the experiment discloses that people whose reason for testing is having risk, suspect or symptoms is also associated to HIV-negative result with promising evidence [28]. This could imply that an individual who has risky perception of oneself has a better chance to be tested for HIV.

A study was also performed to investigate the applicability of data mining on utilization of HIV testing the case of CDC [29]. CRISP-DM methodology was employed to develop a model and it has also used clustering and classification data mining techniques. K-means and Expectation maximization (EM) algorithms of clustering techniques are used to define group of similar VCT client and to see how these grouped affect the classification outcome. From the two clustering algorithms EM is selected and the cluster indices created using this algorithm is then used as class for the classification purpose. The study were used decision tree (J48), random tree and multi-layer perception classifiers to predict the level of risks of clients as high risk or low risk based on the clustered indices [30]. The performance of the model indicated that decision tree had shown better performance and appropriate to the domain. This is due to the fact that decision tree algorithm has a simple feature which can be easily understood by non-technical staff. Above all the researcher recommended further research using large dataset and other data mining techniques to boost the performance of the model [30].

Another study was also conducted to determine the status of clients being HIV positive and negative for data collected from Addis Ababa Administration HCT users. The researcher used decision tree (J48 algorithms) to predict client’s status. Association rule of Apriori algorithms was used to see interesting association among the selected attributes within the database. The study used CRISP-DM methodology to develop the predictive model. The result of the study had indicated that 81.8% performance was achieved to associate the attribute with HIV status [31]. The researcher has also recommended that in order to scale up HIV testing the implementation of the same research problem with other data mining techniques by increasing the number of dataset, number of variables could provide better performance.

Despite the numerous applications of ANNs to classification in medicine, very little attention has been made to the HIV/AIDS prevention and planning [32]. ANNs have been used to classify and predict the symptomatic status of HIV/AIDS patients using data from a publicly available AIDS Cost and Services Utilization Survey performed in the USA multilayer perceptron with 15 linear inputs and 3 hidden logistic nodes and one output was trained using 200 epochs with a learning rate of 0.1 and momentum of 0.1. 1026 cases were used for training and 667 HIV cases were used for testing. The inputs are: sex, race, exposure rate (homosexual, IV drug user, heterosexual), medical records (total number of patient admission, total number of inpatient nights, total number of ambulatory visits, total number of emergency room visits, total number of hospital clinic visits, total number of private physician visits). The best accuracy obtained was 88% [32].

Another study was also performed to predict the functional health status of HIV/AIDS patients and a binary outcome defined as well or not well, using neural networks [33]. The study was used medical care access, such as number of emergency room visits and inpatient nights as inputs. Most other applications of neural networks in HIV/AIDS research are in bioinformatics pertaining to modeling of the virus on a molecular level, such as the prediction of HIV-1 Protease Cleavage Sites [33].

Moreover, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, medical and scientific discovery [34]. Although humans have been manually extracting patterns from data for centuries, the increasing volume of data in modern times has called for more automated approaches.

As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms (1950s), decision trees (1960s) and support vector machines (1980s). Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. A common task in medicine is thus classification using predictive models. It is therefore applying data mining techniques on predicting whether an individual is being tested for HIV may provide a promising result.

2. Methods

2.1. Data Mining Approach

In this study, Cross Industry standard process for data mining (CRISP-DM) methodology was used on HIV testing dataset to explore the application of data mining. CRISP-DM is preferred to knowledge discovery in database (KDD) and SEMMA because it has additional features on understanding the business perspective and its deployment.

CRISP-DM is a standard process for data mining which is a non-proprietary model. It is an application or industry neutral data mining methodology mainly focuses on business issues as well as technical analysis [32]. CRISP-DM is the most suitable for novice data miners due to the easy to read documentation and intuitive, industry-applications-focused description [33]. It begins from understanding the business and ends with the deployment of the system. The CRISP-DM process was developed by the means of the effort of a consortium initially composed with Daimler Chryrler, SPSS and NCR. It consists on a cycle that comprises six stages (**Figure 2**).

2.1.1. Business Understanding

A critical review of related documents associated with HIV testing utilization among adults in Ethiopia was done. Hence, it has been attempted to understand what the Ethiopia’s government need to accomplish on HIV testing. HIV testing is a key strategic entry point to prevention, treatment, care and support services. This is critically important for individuals and couples to learn about their HIV status and make informed decisions about their future. The current rapid development in counseling and testing has necessitated reviewing and updating the guidelines for HIV counseling and testing, especially in the area of policy and implementation. Improved availability of antiretroviral medications and better treatment of opportunistic infections have created the opportunity to expand provider-initiated testing and counseling in health facilities thereby increasing access.

Moreover, the success of the first phase of the Millennium AIDS Campaign (MAC I) Ethiopia demonstrated the potential to provide extensive counseling and testing services throughout the country. And the government effort is to scale up the numbers of sites and people tested, using existing and new cadres of community counselors, who both counsel and conduct testing. After understanding the gap on HIV testing utilization, the research problem of this study was formulated. The business objectives of the Ethiopian government on HIV testing services were identified and tried to map to data mining problem. Therefore, a number of well-known data mining algorithms have been used to predict whether an individual is being tested or not for HIV.

2.1.2. Data Understanding

At this phase, attributes are selected for data mining purpose and have been familiar with the nature of the data, data quality problems are identified, first insights into the data is performed or interesting subsets is detected to form hypotheses for hidden information. The basic tasks before data mining such as data cleaning, attribute selection and transformation are performed.

2.1.3. Data Preparation

A final target data set especially on HIV testing with "17 attributes" and 30,625 instances/records have been constructed from the initial raw data basically from the 2011 EDHS.

2.1.4. Modeling

Various modeling techniques of classification are applied and their parameters have been calibrated to optimal values.

2.1.5. Evaluation

At this stage the model (or models) obtained have been thoroughly evaluated based on their accuracy and the steps executed to construct the model have been reviewed to be certain if properly achieves the business objectives.

2.1.6. Deployment

The final report of this study will be reviewed and deployed to the concerned body. The life cycle of a data mining project consists of six phases, as shown in **Figure 2**. The sequence of the phases is not rigid. Moving back and forth between different phases is always required [23]. In addition, data mining tools, techniques and expertise in WEKA 3.7.7 are used as means to address the stated research problem. Other intermediate data processing tools such as SPSS version 20 and Microsoft Excel 2010 were used to transfer and preprocess the data. In order to come up with reliable and stable model standard logical steps like in CRISP-DM has also been followed. The research design is represented diagrammatically as shown in **Figure 3** below.

3. Methods of Data Understanding and Data Preparation

3.1. Study Area

This study was conducted in Ethiopia.

3.2. Data Source

This study was based on data from the 2011 EDHS, the most recent national dataset on HIV testing that is available (as of January 2012). The 2011 EDHS included a nationally representative sample of women (aged 15 - 49 years old) and men (aged 15 - 59 years old) from all eleven administrative regions in the country. The 2011 Ethiopia Demographic and Health Survey (EDHS) were conducted by the Central Statistical Agency (CSA) under the auspices of the Ministry of Health. The Ethiopian Health and Nutrition Research Institute (EHNRI) were responsible for the testing of HIV from the dried blood samples (DBS). This is the third Demographic and Health Survey (DHS) conducted in Ethiopia, under the worldwide MEASURE DHS project, a USAID-funded project providing support and technical assistance in the implementation of population and health surveys in countries worldwide. The three EDHS surveys have been conducted at five-year intervals since 2000, and the 2011 EDHS is the second survey presenting results on HIV and anemia prevalence.

The 2011 Ethiopia DHS survey collected information on the population and health situation, covering topics

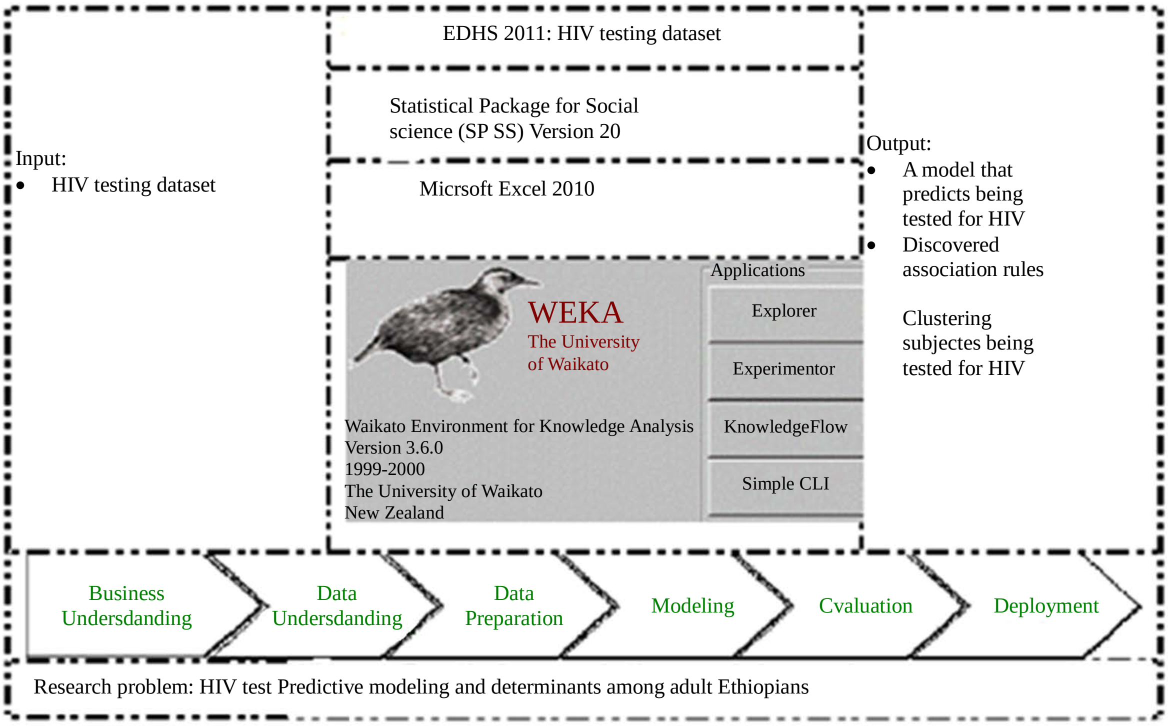

EDHS 2011: HIV testing dataset
Statistical Package for Social science (SP SS) Version 20
Micrsoft Excel 2010
Input:
• HIV testing dataset
Output:
• A model that predicts being tested for HIV
• Discovered association rules
Clustering subjectes being tested for HIV
WEKA
The University of Waikato
Waikato Environment for Knowledge Analysis
Version 3.6.0
1999-2000
The University of Waikato
New Zealand
Applications
Explorer
Experimentor
KnowledgeFlow
Simple CLI
Business Undersdanding
Data Undersdanding
Data Preparation
Modeling
Cvaluation
Deployment
Research problem: HIV test Predictive modeling and determinants among adult Ethiopians

Figure 3. Diagrammatic overview of the overall research design and methodology.

on family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, women's empowerment, and knowledge of HIV/AIDS were provided for the nine regional states and two city administrations. In addition, this report also provides data by urban and rural residence at the country level. In line with this, this study used the 2011 EDHS as a source of data especially on HIV testing (ever been tested for HIIV) to predict whether an individual was being tested for HIV among adults in Ethiopia using data mining technology. The 2011 EDHS dataset consists two important datasets which are relevant for this particular study. Generally, there are 3807 attributes/variables and 16,515 instances/records for women, 547 attribute and 14,110 instances for men.

3.3. Attribute Selection

The importance of selecting relevant variables (attributes) in any data mining task, as it has been noted that the abundance of potential features constituted a serious obstacle to the efficiency of most learning algorithms [35]. Popular methods such as k-nearest neighbor, C4.5, and back propagation are slowed down by the presence of many features, especially if most of these features are redundant and irrelevant to the learning task. The authors further stated that some algorithms might be confused by irrelevant or nosily attributes and construct poor classifiers. Therefore, eliminating some attributes, which are assumed to be irrelevant to build the model, can increase the accuracy of the classifier, save the computational time, and simplify results obtained.

Some of the data or attributes in the initial dataset were not relevant (particular to this study) to the data mining goal and were ignored. The attributes were selected from the subject matter expert and general facts from different literature reviews by considering the impact of these attributes on HIV testing. The initial dataset was from the EDHS 2011 with 16,515 instances/records and 3807 attributes for women and 14,110 instances and 547 attributes for men. Data integration technique has been used to merge these two separated databases. Hence, a new target dataset especially on HIV testing with 17 attributes and 30,625 records has been prepared in MS Excel and imported to WEKA 3.7.7 for analysis purpose. The dependent variable is a binary outcome of people who ever been tested for HIV. The attributes used for predicting the model have been ranked using information gain attribute evaluator algorithm in WEKA 3.7.7. This algorithm used to select the attributes based on their contribution to the model and the list of attributes used for prediction are annexed (**Table 2**).

Table 2. List of possible attributes for predicting the model and the dependent variable (being tested for HIV).

Rank	Attributes name	Contribution of each attribute to the model	Data type	Distinct values	% of missing values
1	Know place where to get test	0.2156	Text	2	0%
2	Wealth index	0.0858	Text	5	0%
3	Highest-level of education	0.0691	Text	4	0%
4	HIV related stigma	0.0667	Text	2	3%
5	Place of residence	0.0600	Text	2	0%
6	Heard family planning in mass media	0.0594	Text	2	0%
7	Region	0.0546	Text	11	3%
8	Ethnicity	0.0485	Numeric	57	0%
9	HIV related knowledge	0.0447	Text	4	3%
10	Frequency reading to newspapers	0.0171	Text	2	0%
11	Age group	0.0149	Text	10	0%
12	Religion	0.0138	Text	5	0%
13	Relationship with most recent sexual partner	0.0122	Text	4	15%
14	Risky sexual behaviors	0.0052	Text	3	3%
15	Marital status	0.0011	Text	4	3%
16	Sex	0.0000	Text	2	0%

N.B: The attributes/variables are ranked based on their information gain to the model [from high (1) to low (16)].

3.4. Data Cleaning

Data cleaning is a process which fills in missing values, removes noise and corrects data inconsistency. Usually, real world database contains incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process [33]. As a result, in order to improve the quality of data and performance of the model (accuracy and efficiency) data cleaning has become a must. This technique involves removing the records that had incomplete, noise (invalid) data and filling missing values under each column. Removing of such records was done as the records with this nature were few and their removal did not affect the entire dataset.

3.5. Missing Value Handling

Missing values refer to the values for one or more attributes in a data that do not exist. Data are rarely complete in real world application. When the dataset is small or the number of missing fields is large, not all records with a missing field can be deleted from the sample. Moreover, the fact that a value is missing may be significant itself. A widely applied approach is used to calculate a substitute value for missing fields, for example, the median or mean of a variable [36]. Accordingly, the investigator has analyzed the HIV testing dataset and identified missing values and took measures to solve the problem of missing. The missing values for numeric data type and normally distributed data were substituted by the mean of the variable. For the categorical variable, the missing values were replaced by the modal value of the variable [36].

The missing values were clearly identified and calculated the percentage of the value against total records and annexed in (**Table 2**). As a result, the entire selected attributes illustrated have less than five percent missing values (5%) except for condom used during last sex with most recent partner, respondents type of earning, can get condom, ever been married and can get female condom attributes which of them accounts for 37%, 29%, 44%, 59% and 85% respectively and discarded from analysis. Therefore, based on the above principles the investigator handled the missing values and WEKA preprocessing techniques such as replace missing value has been used. WEKA fills using the most frequent (modal) value methods which is the same as the above principle.

Thus, the dataset has been cleaned and prepared in such ways that are suitable for data mining tools.

3.6. Data Transformation and Reduction

Data mining often requires data integration or the merging of data from multiple data sources [34]. The dataset for this study which were used for the subsequent predictive model building are prepared and derived from one source, HIV testing from EDHS 2011.However, the datasets for men and women were prepared separately hence these databases have been integrated into one database. In data transformation, the collected data were transformed into forms which are appropriate for data mining tools. The process of data transformation included attribute construction, where new attributes were constructed and added from the given set of attribute to help the mining process [37].

In order to make the analysis procedures manageable and cost- effective the data needed to be reduced. Data reduction techniques include data discretization which is one of data transformation methods used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals [36] [37]. In this research some attributes were discretized to reduce the unlike values of the attribute to obtain knowledge (pattern) and to make the dataset suitable for data mining tools.

Accordingly, HIV/AIDS-related knowledge index was built from the answers to eight questions (**Table 3**): three questions on knowledge of HIV prevention and five on misconceptions about modes of HIV transmission.

Table 3. Distribution of attributes for HIV related knowledge, Ethiopia, 2013.

S. no	Attributes name	Category	n	%
1	Know place where to get test	Yes	23,598	77.05
		No	7,027	22.95
2	Using Condom during Sex	Yes	21,040	68.70
		No	4,960	16.20
		Don’t know	4,625	15.10
3	Don’t have sex at all?	Yes	24,116	78.75
		No	4,714	15.39
		Don’t know	1,795	5.86
4	Do you have one sex Partner only?	Yes	20,420	66.68
		No	6,836	22.32
		Don’t know	3,369	11.00
5	Can get HIV by sharing food with AIDS person	Yes	3,856	12.59
		No	25,170	82.19
		Don’t know	1,599	5.22
6	Healthy looking Person can have HIV	Yes	21,495	70.19
		No	5,831	19.04
		Don’t know	3,299	10.77
7	Can get HIV by super natural?	Yes	6,325	20.65
		No	22,454	73.32
		Don’t know	1,846	6.03
8	Can get HIV from mosquito?	Yes	7,589	24.78
		No	17,962	58.6
		Don’t know	5,074	16.57

It was categorized as low (score ≤ 4), high (score 5 - 6), or comprehensive (score >= 7) knowledge. In order to assess their risky sexual behaviour, the study used five questions related to their sexual behaviours and categorized as “No risk” (score 0), “Some risk” (score 1) and “High risk” (score ≥ 2) (**Table 4**). Five questions that reflected negative attitudes towards to people living with HIV/AIDS were used to create a stigma index. This index was categorized as “No stigma” (score 5), “Low stigma” (score 4), “Moderate stigma” (score 2 - 3), and “High stigma” (score ≤ 1) and the indexes were annexed (**Table 5**). A chi-square test was used to test the statistical association of knowledge, stigma, and risky sexual behaviors related attributes with ever being tested for HIV (**Table 6**).

In addition, multiple distinct values of some attributes such as “age”, “ethnicity” and “type of earning” are discretized with explicit data grouping technique. That means detailed concepts were grouped into the required level of more general concept. Almost all the selected attributes have been transformed from their original state in such a way that could be easily understandable and interpretable.

An attribute religion had six distinct values (Orthodox, Traditional, Protestant, Catholic, Muslim, and others) and later transformed into three distinct values as Christian (Orthodox, Protestant and Catholic), Muslim and others. In addition to that, ethnicity was originally with 57 distinct values but it has been converted into ten distinct categories as: Tigrean, Affar, Amara, Gurage, Somalie, Sidama, Nuwer, Welaiyta, Oromo and others. The percentage of HIV test class data size consists about 57.86% was not tested and 42.14% tested for HIV. This class size was considered to be unbalanced data which might be a bias to evaluate the classifier method. This indicates that there is a need to balance these two classes hence an equal amount of both HIV tested and not tested was taken randomly using WEKA 3.7.7 pre-processing option.

3.7. Data Mining Tasks

Data mining functionalities were used to specify the kind of patterns to be found in data. Accordingly, both data mining tasks (descriptive and predictive) were employed. Descriptive mining tasks used to characterize the general properties of the data in the database [17]. Basically, the goal of descriptive data mining is to gain an understanding

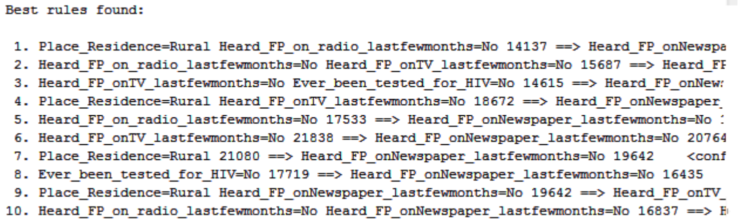
Table 4. Distribution of attributes on HIV related risk behaviors, Ethiopia, 2013.

S. no	Attributes name	Category	n	%
1	Had any STIs in last 12 months	Yes	172	0.56
		No	30,442	99.40
		Don’t know	11	0.04
2	Had genital ulcer in last 12 months	Yes	264	0.86
		No	30,181	98.55
		Don’t know	180	0.59
3	Condom use	Yes	24,116	78.75
		No	4,714	15.39
		Don’t know	1,795	5.86
4	Condom used last sex with most recent partner	Yes	1,132	3.70
		No	29,493	96.30
5	Chewing tobacco	Yes	536	1.75
		No	30,089	98.25
6	Uses Shisha	Yes	159	0.52
		No	30,466	99.48
7	Recent sexual activity	Active	14,279	46.63
		Not active	16,346	53.37
9	Smoking cigarettes	Yes	1,724	5.63
		No	28,901	94.37

Table 5. Distribution of attributes for HIV related stigma indicator.

S.No	Attributes name	Category	n	%
1	Would buy vegetables from vendor with AIDS	Yes	13,681	44.67
		No	16,944	55.33
2	Female teacher with HIV should continue to teach	Yes	20,234	66.07
		No	8,683	28.35
		Don't know	1,708	5.58
3	Ever heard of AIDS	Yes	29,812	97.35
		No	813	2.65
4	Willing to care for relatives with HIV	Yes	26,809	87.54
		No	3,489	11.39
		Don't know	327	1.07
5	Would want HIV infection in family remain secret	Yes	11,347	37.05
		No	18,475	60.33
		Don't know	803	2.62

A. Association rule mining.



```
Best rules found:

 1. Place_Residence=Rural Heard_FP_on_radio_lastfewmonths=No 14137 ==> Heard_FP_onNewspa
 2. Heard_FP_on_radio_lastfewmonths=No Heard_FP_onTV_lastfewmonths=No 15687 ==> Heard_FP
 3. Heard_FP_onTV_lastfewmonths=No Ever_been_tested_for_HIV=No 14615 ==> Heard_FP_onNew
 4. Place_Residence=Rural Heard_FP_onTV_lastfewmonths=No 18672 ==> Heard_FP_onNewspaper_
 5. Heard_FP_on_radio_lastfewmonths=No 17533 ==> Heard_FP_onNewspaper_lastfewmonths=No
 6. Heard_FP_onTV_lastfewmonths=No 21838 ==> Heard_FP_onNewspaper_lastfewmonths=No 20764
 7. Place_Residence=Rural 21080 ==> Heard_FP_onNewspaper_lastfewmonths=No 19642    <conf
 8. Ever_been_tested_for_HIV=No 17719 ==> Heard_FP_onNewspaper_lastfewmonths=No 16435
 9. Place_Residence=Rural Heard_FP_onNewspaper_lastfewmonths=No 19642 ==> Heard_FP_onTV_
10. Heard_FP_on_radio_lastfewmonths=No Heard_FP_onNewspaper_lastfewmonths=No 16837 ==> H
```

Table 6. Statistical association of knowledge, stigma, and risky sexual behaviours related attributes with ever being tested for HIV using chi- square, Ethiopia, 2013.

S. No	Attributes name	Category	Ever been tested for HIV		P-value
			Yes	No	
1	HIV related knowledge	Low	1,010	3,951	0.000
		High	3,611	6,320	
		Comprehensive	8,285	7,448	
2	HIV related stigma	No stigma	4,758	2,985	0.000
		Low	4,438	4,545	
		Moderate	3,543	9,428	
		High	167	761	
3	Risky sexual behaviour	No risk	5,273	6,513	0.000
		Some risk	6,835	9,291	
		High	798	1,915	
4	Heard family planning on mass media	Yes	8,591	6,686	0.000
		No	4,315	11,033	

N.B: 1. **HIV related knowledge**: was assessed using eight questions: (risk of getting HIV reduced by: Don't have sex at all, Having only one sex partner, use condom, healthy looking person can have HIV, can get HIV by sharing sharp materials, can get HIV by mosquito, can get HIV/AIDS by sharing food with AIDS person and can get by super natural). 2. **HIV related stigma index**: was assessed using five questions: (Would buy vegetables from vendor with AIDS person, Female teacher with HIV should continue to teach, Ever heard of AIDS, Willing to care for relatives with HIV and Would want HIV infection in family remain secret). 3. **Risky sexual behaviours:** was assessed using five questions: (Had any STI in last 12 months, had genital ulcer last 12 months, had genital discharge in last 12 months, wife justified asking husband to use condom if he has STI and Ever took alcohol drink).

of the analyzed system by discovering patterns and relationships in large datasets [37]. Predictive mining tasks, on the other hand, perform inference on the current data in order to make predictions [35]. Classification is the commonly used data mining technique for prediction, whereas association and clustering are considered as the descriptive data mining techniques [37].

3.8. Classification

It is learning function that maps (classifies) a data item into one of several predefined classes [19]. Basically classification involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories of classes. The term ‘mutually exhaustive and exclusive’ implies that each object must be assigned to precisely one class. For instance, a hospital system may want to classify people into those who are tested or not tested for HIV. Therefore, people should be classified as tested or not tested class [19].

3.9. Decision Trees

A decision tree is a tree-like structure, which starts from root attributes, and ends with leaf nodes. Generally a decision tree has several branches consisting of different attributes, the leaf node on each branch representing a class or a kind of class distribution. Decision trees are an approach of representing a sequence of rules that lead to a set or value. As a result, they are used for directed data mining, mainly classification. One of the main rewards of decision trees is that the model is quite reasonable since it takes the form of explicit rules [38]. The decision tree algorithm used in this research is J48 and random tree algorithms, which are an implementation of the C4.5 decision tree learner. This implementation produces decision tree models. Moreover, to extract rules from a decision tree, one rule is created for each path from the root to a leaf node. In order to make decision-tree models more readable, a path to each leaf can be transformed into an IF-THEN classification rules. The “IF” part of the classification rule consist of all tests on the path, where as the “THEN” part is the final classification. Therefore, this study used algorithms of decision trees to classify whether an individual was being tested for HIV or not. The tree is generated according to the information-gain measure, the procedures are briefly as:

a) Calculate:

$$I\left(s_1, s_2, \cdots, s_m\right) = -\sum_{i=1}^{m} p_i \log_2^{\left(p_i\right)}$$

$$p_i = \frac{s_i}{S}$$

where p_i is the probability that an arbitrary sample belongs to class c_i.

b) Calculate the entropy $E(A_i)$, which is the expected information based on the partitioning by attribute A_i:

$$E\left(A_i\right) = \sum_{j=1}^{q} \frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{S} I\left(s_{1j}, s_{2j}, \cdots, s_{mj}\right)$$

$$I\left(s_{1j}, s_{2j}, \cdots, s_{mj}\right) = -\sum_{i=1}^{m} p_{ij} \log_2^{\left(p_{ij}\right)}$$

where $p_{ij} = \frac{s_{ij}}{|S|}$, and $|S|$ is the number of samples in subset S_j.

Then the encoding information that would be gained by branching on A_i is:

$$Gain\left(A_i\right) = I\left(s_1, s_2, \cdots, s_m\right) - E\left(A_i\right)$$

The attribute A_i with the highest information gain is selected as the root node, the branches of the root node is formed according to different distinctive values of a_{ij}, $j = 1, \cdots, q$. The tree grows like this until if all the samples are all of the same class, and then the node becomes a leaf and is labeled with that class.

3.10. Naïve Bayes Classifiers

Naïve Bayes (NB) approach is a very popular classification method that does not use rules, a decision tree or any other explicit representation of the classifier. Rather, it uses the branch of Mathematics known as probability theory to find the most likely of the possible classifications [38]. Bayes theory is used to calculate a conditional probability $P(A|B,C)$ which the probability of *A* is given the probability of *B* and *C*. The conditional probability is also related to the joint probability of *A* and *B* is given by:

$$P(A|B,C)=\frac{P(A,B|C)}{P(B|C)}$$

For example, if *A* is the probability that a person tested for HIV, and *B* is the person's age, the $P(A|B,C)$ is a number between 0 and 1 describing the chance that the person being tested, taking into account their age, and the model for the test and risk factors (*C*).

3.11. Neural Network

A Neural network may be defined as "a model of reasoning based on the human brain" [39]. It is probably the most common data mining technique, since it is a simple model of neural interconnections in brains, adapted for use on digital computers. It learns from a training set, generalizing patterns inside it for classification and prediction. Neural networks can also be applied to undirected data mining and time-series prediction [40]. Neural networks are popularly applied in classification and prediction, as they have advantages such as high tolerance to noise, and the ability to classify unseen patterns. Neural networks are thus used as predictive data models. The basic model of a neuron is illustrated in **Figure 4** below.

The transfer function of the neuron is described by the relation:

$$y_i = F\left\{\sum_{i=1}^{n} w_i x_i\right\}$$

where $X_0 = 1$. The neuron's firing condition is: $\sum_{i=1}^{n} w_i x_i \geq \Theta_i$. Where the index i represent the index of neuron. The most popular nonlinear neurons are sigmoid, logsid and tansig functions. Furthermore, as the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. Hence, in this study, neural networks used 30,625 instances of the HIV testing dataset to train and test by adjusting the weights for each selected attributes that could affect for HIV testing and this enables to build the predictive model for HIV testing. X_i indicates inputs for the classifier, $W_i' s$ are weights assigned for each input and n is total number of inputs/attributes.

3.12. Logistic Regression

Logistic regression is an approach to prediction, like Ordinary Least Squares (OLS) regression. However, with logistic regression, the researcher is predicting a dichotomous outcome (being tested for HIV or not). Because the outcome variable is discrete, it cannot modeled directly by linear regression. Therefore, rather than predicting point estimates of the event itself, it builds the models to predict the odds of its occurrence. In two class

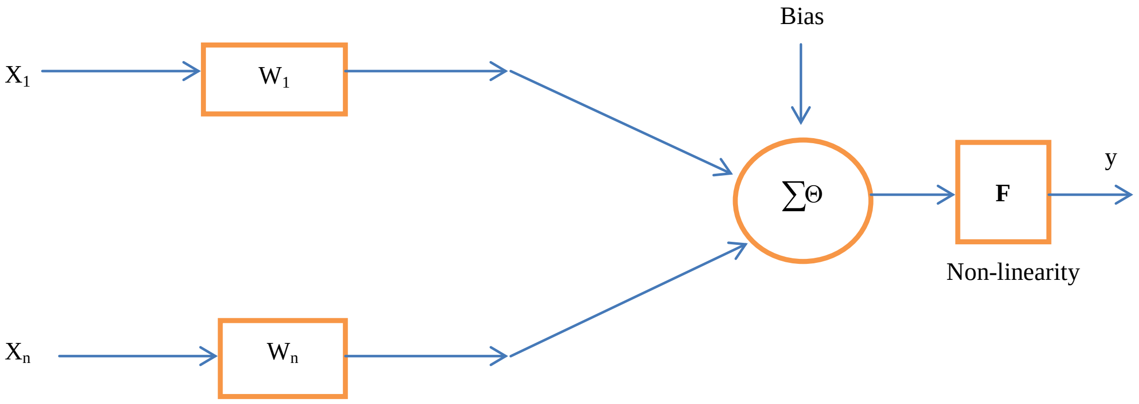


Figure 4. Basic neuron model.

problem, odds greater than 50% would mean that the case is assigned to the class designated as "1" and "0" otherwise. The binary outcome is the probability that being tested for HIV ($Y = 1$) and not tested ($Y = 0$).

Because the dependent variable is not a continuous one, the goal of logistic regression is a bit different, because we are predicting the likelihood that Y is equal to 1 (rather than 0) given certain values of X. That is, if X and Y have a positive linear relationship, the probability that a person will have a score of $Y = 1$ will increase as values of X increase. So, we are stuck with thinking about predicting probabilities rather than the scores of dependent variable.

The proposed logistic regression model is given by:

$$P\left(Y=1\right)=\frac{\mathrm{e}^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\varepsilon_i}}{1+\mathrm{e}^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\varepsilon_i}}$$

where: X's are the predictors such that: X_1 = age-group, X_2 = education, X_3 = residence, X_4 = region, X_5 = wealth index, X_6 = ethnicity, X_7 = HIV related stigma, X_8 = sex, X_9 = HIV related knowledge, X_{10} = Marital status and other predictors and the dependent variable is Y = being tested for HIV or not. Backward elimination procedure has been used for model development. Finally, 17 attributes were selected to fit the final model.

4. Association Rules

4.1. Apriori Approach

Association rules are one of the major data mining techniques. It is perhaps the most common form of local-pattern discovery in unsupervised learning systems [41]. Association rules are widely used in data mining to find patterns in data. Therefore, this study used association rules of data mining technique to examine which instances of HIV testing frequently occurred in a database and presents the patterns as rules among adults in Ethiopia.

4.2. Methods of Training and Testing

It has been stated that the classifiers rely on being trained before they can reliably be used on new data [42]. The more instances the classifier is exposed to during the training phase, the more reliable it will be as it has more experience. However, once trained, we would like to test the classifier too, so that we are confident that it works successfully. It has been also stated that, in order to predict the performance of a classifier on new data, we need to assess its error rate on an independent test set that played no part in the formation of the classifier [43]. The standard way of predicting the error rate of a learning technique is to use stratified 10-fold cross-validation. The data is divided randomly into 10 parts in which the class is represented in approximately the same proportions as in the full dataset. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. Thus the learning procedure is executed a total of 10 times on different training sets. Finally, the 10 error estimates are averaged to yield an overall error estimate [43]. The WEKA 3.7 tool provides a test options to test on the same set the classifier is trained on (use training set), to test on a user-specified test data (Supplied test set), test on k-fold cross validation, and to train on a percentage of the data and test on the remainder (percentage split). No matter which test option is used, the model that is output is always the one build from all the training data.

Therefore, in this study, 30,625 instances of the HIV testing dataset was used to train the classifiers using the training test option. Similarly, the instances of HIV testing were divided randomly into 10 parts using cross validation test parameter. In addition, the total instances of HIV testing were divided into 66% for training and 34% for testing using percentage split test parameter. Moreover, a user defined dataset were supplied to the supply test parameter in order to train the instances of HIV testing there by to predict whether an individual was being tested for HIV or not. Finally, training test parameter was selected among other testing parameters as it could achieve the best classification accuracy.

4.3. Methods of Analysis and Evaluation of the Model

The output of several experiments of classification models are analyzed and evaluated in terms of the details of the confusion matrix of the model. The complexity of the model in terms of the number of trees and leaves are also evaluated. The template confusion matrix of the model used to classify HIV testing is presented in **Table 7**.

Table 7. Summary of confusion matrix template.

		Predicted HIV test		Total
		Tested (T)	Not tested (NT)	
Actual HIV test	Tested (T)	True tested (TT)	False Not tested (FNT)	TT + FNT
	Not tested (NT)	False tested (FT)	True Not tested (TNT)	FT + TNT
Total		TT + FT	FNT + TNT	TT + FNT + FT + TNT

Notes: TT: The number of HIV tested clients that are classified as tested. FT: The number of HIV not tested clients that are classified as tested. FNT: The number of HIV tested clients that are classified as not tested. TNT: The number of HIV not tested clients that are classified as not tested. TT + FNT: The total number of actually HIV tested clients. FT + TNT: The total number of actually HIV not tested clients. TT + FT: The total number of predicted HIV tested clients. FNT + TNT: The total number of predicted HIV not tested clients. TT + FT + FNT + TNT: The total number of all clients.

The performance of the experiments is measured in terms of the details of the following measurements. These measurements include:

True tested Rate (TTR): The proportion of HIV tested clients that are correctly classified as tested.

$$\text{TTR} = \text{TT}/\text{TT} + \text{FT} \tag{1}$$

False Tested Rate (FTR): The proportion of HIV not tested clients that are erroneously classified as tested.

$$\text{FTR} = \text{FT}/\text{FT} + \text{TT} \tag{2}$$

False Not tested Rate (FNTR): The proportion of HIV tested clients that are erroneously classified as not tested.

$$\text{FNTR} = 1 - \text{TTR} \text{ or } \text{FNTR} = \text{FNT}/\text{TT} + \text{FNT} \tag{3}$$

True Not tested Rate (TNTR): The proportion of HIV not tested clients that are correctly classified as not tested.

$$\text{TNTR} = \text{TNT}/\text{TNT} + \text{TT} \tag{4}$$

Correctly Classified Instances (Accuracy): The proportion of clients that are correctly classified.

$$\text{Accuracy} = (\text{TT} + \text{TNT})/(\text{TT} + \text{FT} + \text{FNT} + \text{TNT}) \tag{5}$$

Incorrectly Classified Instances (Error Rate): The proportion of clients that are incorrectly classified.

$$\text{Error Rate} = (\text{FT} + \text{FNT})/(\text{TT} + \text{FT} + \text{FNT} + \text{TNT}) \qquad (6).$$

ROC (Receiver Operating Characteristic**):** ROC curves are a useful tool for comparing classification models [35]. The performance of the classifiers with different parameters is also compared by examining their ROC curve. They further said that ROC curve shows the trade-off between the TP rate (TT) and the FP rate (FT) for a given model. Moreover, models can be compared with respect to their speed, robustness, scalability, and interpretability which may have an influence on the model [36]. In summary, there are three measures for model performance evaluations: accuracy, sensitivity and specificity.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

where the variables: TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives respectively. In this study true positive (TP) refers to true test (TT), true negative (TN) refers to true not tested (TNT), false positive (FP) refers to false tested (FT) and false negative (FN) refers to false not tested (FNT) respectively.

4.4. Ethical Considerations

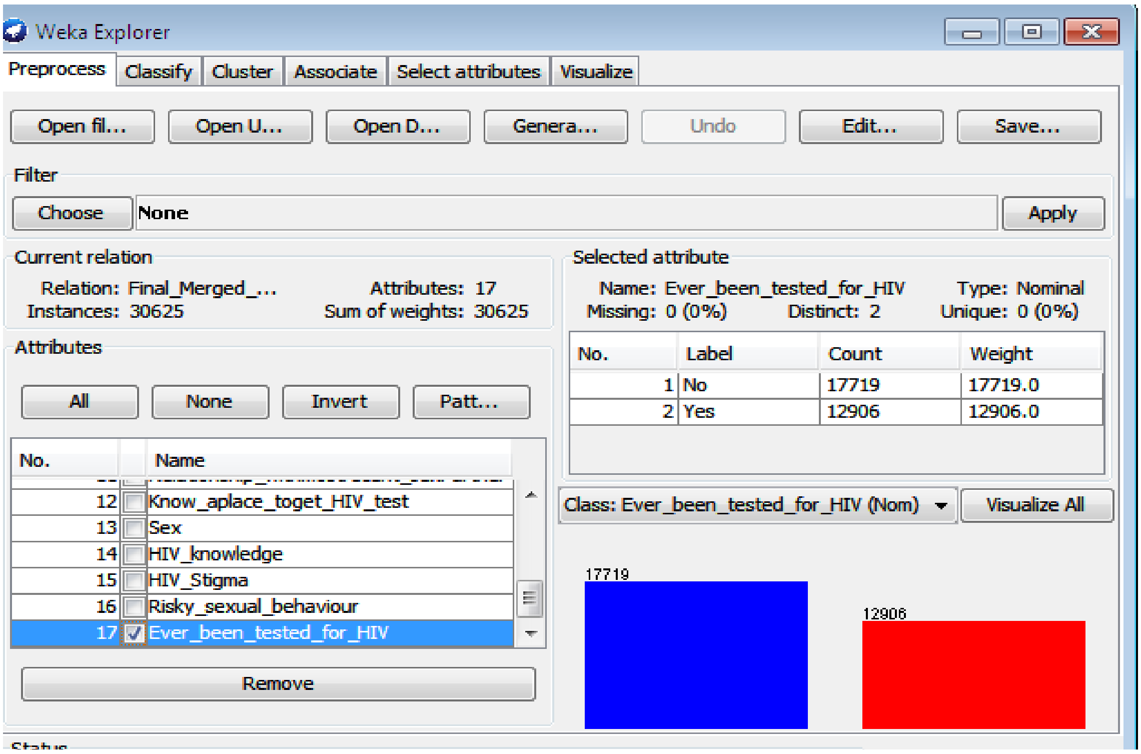
The study used Ethiopian Demographic and Health Survey 2011 dataset. The data were collected over 6-months (November 2010 - April 2011) by the Central Statistical Agency (CSA) under the auspices of the Ministry of Health. The Ethiopian Health and Nutrition Research Institute (EHNRI) was responsible for HIV testing from the dried blood samples (DBS). The researcher has been authorized by MEASURE DHS authority to work on EDHS 2011 of HIV testing dataset. Accordingly, the researcher has got clearance from the ethical clearance committee of the college of health sciences, department of Public Health in Mekelle University. I the investigator have treated the EDHS 2011 as confidential and no effort was made to identify any household or individual respondent interviewed in the survey.

In addition, the data set on HIV testing obtained from EDHS 2011 has not been passed on to other researchers without the written consent of DHS/EDHS. The principal investigator or user of the data is also intended to submit a copy of any reports/publications resulting from using the 2011 EDHS data sets. This report should also be sent to the attention of the EDHS data archive. Moreover, the research is meant for academic purpose and it was not attempted to harm anybody in any way.

5. Results

5.1. Experimentations

The classifiers used 30,625 instances for training. To build HIV test predictive model: decision tree (random tree and J48), Bayes (Naïve Bayes) and functions (logistic regression and artificial neural network) algorithms were used. Five of the experiments were done with varying testing parameters. The performances of the models were evaluated using training test option as it achieved the best accuracy particularly for this study. In all experiments, the 17 selected attributes (socio demographic factors, know place where to get test, HIV related stigma, HIV related knowledge, risky sexual behavior, frequency reading newspaper, heard family planning on mass media) were used. The outcome variable is a binary response which is ever been tested for HIV. List of the attributes used for this study are annexed (Table 2). This study used WEKA 3.7.7 data mining techniques and expertise to address the research problem (Figure 5).



Filter: None

Relation: Final_Merged_... Attributes: 17
Instances: 30625 Sum of weights: 30625

Name: Ever_been_tested_for_HIV Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	No	17719	17719.0
2	Yes	12906	12906.0

No.		Name
12	□	Know_aplace_toget_HIV_test
13	□	Sex
14	□	HIV_knowledge
15	□	HIV_Stigma
16	□	Risky_sexual_behaviour
17	☑	Ever_been_tested_for_HIV

Class: Ever_been_tested_for_HIV (Nom)

Figure 5. WEKA 3.7.7 explorer windows showing the number of attribute and instances.

5.2. Classification

In this study the predictive models were evaluated based on the three performance measures (accuracy, sensitivity and specificity). Using random tree algorithm of the decision tree the accuracy, sensitivity, and specificity were found to be 96%, 98% and 94.61% respectively. Whereas, using decision tree of J48 algorithm with training test option provided an accuracy of 79%, a sensitivity of 74% and specificity of 82% were demonstrated. Yet, using ANN model an accuracy of 77.54%, a sensitivity of 76.39% and specificity of 78.22% were identified. This shows that decision tree (random tree algorism) is the best model predictor of the other two (**Table 13**).

In addition, logistic regression has also been fitted in WEKA 3.7.7 and it achieved a classification accuracy of 73.97%, a sensitivity of 67.55%and a specificity of 79.44%. The results for each classifier (decision tree, Naïve Bayes, Artificial neural network and logistic regression) are presented in **Tables 8-13** below.

Table 8. Results from decision tree (J48) with different test parameters, Ethiopia, 2013.

Test parameters	TP	Precision	Recall	ROC	Class
Training	0.807	0.824	0.807	0.863	No
	0.763	0.742	0.807	0.863	Yes
Cross validation	0.757	0.781	0.757	0.799	No
	0.708	0.68	0.708	0.799	Yes
Percentile	0.752	0.778	0.752	0.798	No
	0.713	0.683	0.713	0.798	Yes
Supply	0.736	0.834	0.736	0.856	No
	0.796	0.685	0.796	0.857	Yes

Table 9. Results from decision tree (random tree) with different test parameters, Ethiopia, 2013.

Test parameters	TP	Precision	Recall	ROC	Class
Training	0.987	0.946	0.987	0.997	No
	0.923	0.982	0.923	0.997	Yes
Cross validation	0.758	0.719	0.758	0.704	No
	0.594	0.641	0.594	0.704	Yes
Percentile	0.747	0.71	0.747	0.694	No
	0.593	0.638	0.593	0.694	Yes
Supply	0.819	0.79	0.819	0.874	No
	0.698	0.736	0.698	0.874	Yes

Table 10. Results from Naïve Bayes with different test parameters, Ethiopia, 2013.

Test parameters	TP	Precision	Recall	ROC	Class
Training	0.758	0.755	0.758	0.804	No
	0.662	0.665	0.662	0.804	Yes
Cross validation	0.757	0.755	0.757	0.804	No
	0.662	0.665	0.662	0.804	Yes
Percentile	0.764	0.748	0.764	0.805	No
	0.656	0.676	0.656	0.805	Yes
Supply	0.843	0.741	0.843	0.84	No
	0.59	0.731	0.59	0.841	Yes

Table 11. Results from Neural Network (Multilayer Perception) with different test parameters, Ethiopia, 2013.

Test parameters	TP	Precision	Recall	ROC	Class
Training	0.848	0.782	0.848	0.871	No
	0.676	0.764	0.676	0.871	Yes
Cross validation	0.742	0.77	0.742	0.814	No
	0.696	0.662	0.696	0.814	Yes
Percentile	0.803	0.718	0.803	0.807	No
	0.579	0.688	0.579	0.807	Yes
Supply	0.753	0.801	0.753	0.827	No
	0.741	0.684	0.741	0.828	Yes

Table 12. Results from logistic regression with different test parameters, Ethiopia, 2013.

Test parameters	TP	Precision	Recall	ROC	Class
Training	0.743	0.794	0.743	0.832	No
	0.736	0.676	0.736	0.832	Yes
Cross validation	0.742	0.793	0.742	0.831	No
	0.734	0.675	0.734	0.831	Yes
Percentile	0.75	0.78	0.75	0.826	No
	0.719	0.683	0.719	0.826	Yes
Supply	0.749	0.815	0.749	0.846	No
	0.765	0.687	0.765	0.846	Yes

Table 13. Comparison of the different classifiers (training test), Ethiopia, 2013.

Classification Technique	Class: Ever been tested for HIV	Precision	Recall	ROC	Accuracy (%)
Decision tree (J48)	No	0.824	0.807	0.863	78.82%
	Yes	0.742	0.807	0.863	
Decision tree (Random tree)	No	0.946	0.987	0.997	96.03%
	Yes	0.982	0.923	0.997	
Naïve Bayes	No	0.755	0.758	0.804	71.74%
	Yes	0.665	0.662	0.804	
Logistic	No	0.794	0.743	0.832	73.97%
	Yes	0.676	0.736	0.832	
Neural network	No	0.782	0.848	0.871	77.54%
	Yes	0.764	0.676	0.871	

The complete set of results used for comparison of each model performance was prepared in a tabular format (**Table 14**). The detailed prediction results of the validation datasets for each predictive model using training test option are presented in a form of confusion matrixes. A confusion matrix is a matrix representation of the classification results. In a two-class prediction problem (such as ever been tested for HIV or not) the upper left cell denotes the number of samples classifies as true while they are truly (*i.e.*, true tested), and lower right cell denotes the number of samples classified as false while they were actually false (*i.e.*, true false or true not tested).

The other two cells (lower left cell and upper right cell) denote the number of samples misclassified. Particularly, the lower left cell denotes the number of samples classified as false while they were actually true (*i.e.*,

Table 14. Performance summary of different classifiers: using training test option, Ethiopia, 2013.

Testing criteria	Decision tree (J48)	Decision tree (random tree)	Naïve Bayes	Neural network	Logistic regression
Confusion matrix	14296 3423	17497 222	13423 4296	15024 2695	13157 4562
	3063 9843	995 11911	4359 8547	4183 8723	3409 9497
Accuracy (%)	78.82 %	96.03 %	71.74%	77.54%	73.97%
Sensitivity (%)	74.19%	98.17%	66.55%	76.39%	67.55%
Specificity (%)	82.35%	94.61%	75.48%	78.22%	79.44%
Area under the ROC (%)	86.3%	99.7 %	80.4%	87.1 %	83.2%
Computation time in seconds	3.94	0.6	0.12	**4194**	37.73

false not tested), and the upper right cell denotes the number of samples classified as true while they were actually false (*i.e.*, false tested). Once the confusion matrixes were constructed, the accuracy, sensitivity and specificity of each model were calculated using the respective formulas presented in the above.

5.3. The ROC Curve Analysis for the Classifiers

The area under ROC curve for HIV testing instances produced from the decision tree (J48 and random tree), functions (logistic regression and neural network) are annexed (**Figures 6-9**). The vertical axis (Y-axis) of ROC curve represents the true tested rate. The horizontal axis (X-axis) represents the false-tested rate. The HIV testing class value (Yes) gives the ROC accuracy of 86.3%, 99.7%, 83.2% and 87.1% respectively. This indicates given that the attributes as input, the classifiers are better than the random model to predict being an individual is tested or not because all the three classifiers (five algorithms) have a ROC curve values above 50%. The ROC curve analyses for all experiments displayed below showed that the curves moves sharply up from zero showing that there are more true tested than false tested rates. Then the curve starts to become more horizontal as it encounters less true tested and more false tested rates. The areas under the curve for the models are 86.3%, 99.7%, 83.2% and 87.1% which are closer to 1. Moreover, the decision trees and neural network classifiers look better because they achieved better accuracy than the other classifiers (logistic regression and naïve bayes).

The performances of each classifier to predict the model are presented below (**Figure 10**). The accuracy for the J48, random tree, Naïve Bayes, logistic and neutral network is 79%, 96%, 72%, 74% and 78% respectively. Moreover, the naïve bayes and logistic regression scored least ROC values than the other classifiers.

5.4. Logistic Regression

The final logistic regression predictive model of data mining technique used to predict whether an individual was being tested for HIV or not is given by here below:

$$P(y=0)=\Pi(x)=\frac{\mathrm{e}^{\beta_0+\beta_1*\text{place for HIV testing}_i+\beta_2*\text{WealthIndex}_i+\beta_3*\text{EdunLevel}_i+\cdots+\beta_{16}*\text{Sex}_i}}{1+\mathrm{e}^{\beta_0+\beta_1*\text{place for HIV testing}_i+\beta_2*\text{WealthIndex}_i+\beta_3*\text{EdunLevel}_i+\cdots+\beta_{16}*\text{Sex}_i}}$$

For sex, men = 1, women = 0;

$$OR_{\text{M/W}}=\frac{\Pi(x=1)}{1-\Pi(x=1)}\bigg/\frac{\Pi(x=0)}{1-\Pi(x=0)}=1.5447$$

Women were 36% more likely to have ever been tested for HIV than men while other predictors are holding constant.

For heard about family planning, yes = 1 and no = 0;

$$OR_{\text{no/yes}}=\frac{\Pi(x=1)}{1-\Pi(x=1)}\bigg/\frac{\Pi(x=0)}{1-\Pi(x=0)}=1.4174$$

Those participants who heard about family planning on mass media were 29.5% more likely to have ever

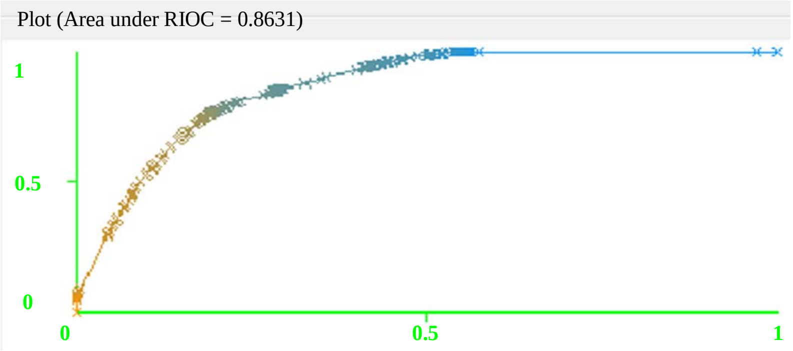


Figure 6. Decision tree (J48). The ROC curve analysis for the classifiers.

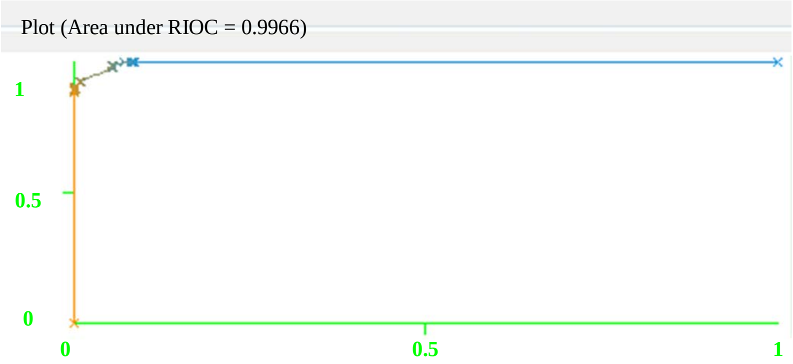


Figure 7. Decision tree (Randon tree). The ROC curve analysis for the classifiers.

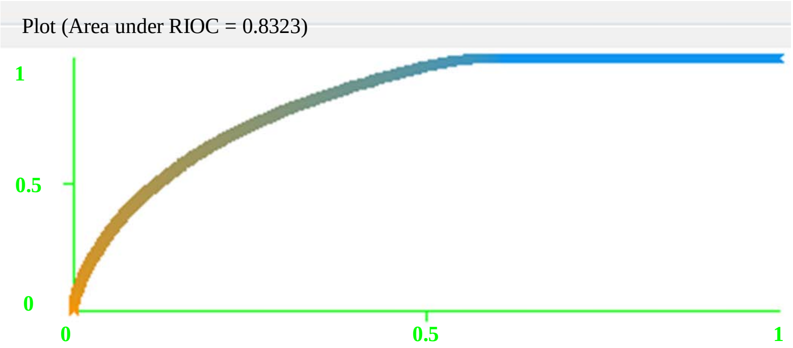


Figure 8. Logistic. The ROC curve analysis for the classifiers.

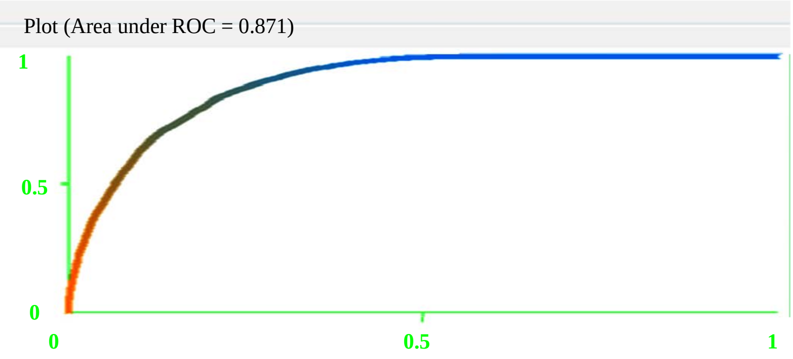


Figure 9. Neural nerwork (MLP). The ROC curve analysis for the classifiers.

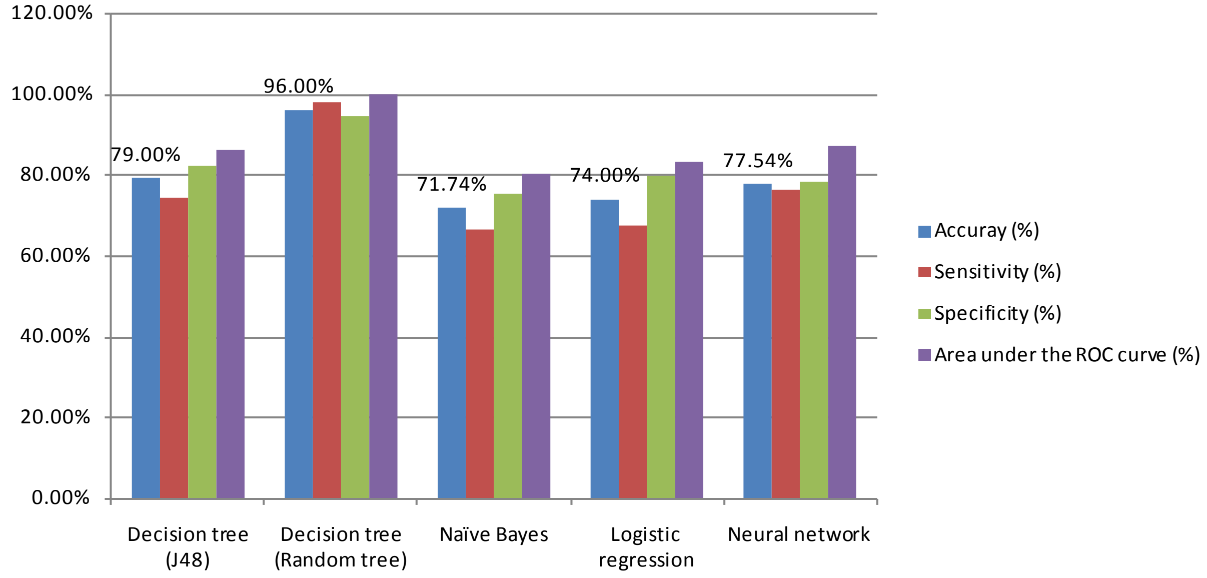


Figure 10. Measures of model performance evaluation, Ethiopia, 2013.

been tested for HIV compared to those who didn't hear about family planning on mass media while other predictors are holding constant. For educational level, Higher = 1 and take all other educational level as a reference category.

$$OR_{\text{Higher/Other}} = \frac{\Pi(x=1)}{1-\Pi(x=1)} \Big/ \frac{\Pi(x=0)}{1-\Pi(x=0)} = 0.654$$

Participants who were belonging to higher educational level were 35% more likely to have ever been tested for HIV compared to all other categories of education while other predictors are holding constant. For HIV related stigma, high = 1 and take all other stigma level (no stigma, low and middle) as a reference category.

$$OR_{\text{high/Other}} = \frac{\Pi(x=1)}{1-\Pi(x=1)} \Big/ \frac{\Pi(x=0)}{1-\Pi(x=0)} = 1.1578$$

Participants who have had high stigma towards to people with HIV/AIDS were 16% more likely to have never been tested for HIV compared to all other stigma while other predictors are holding constant. For knowing a place where to test for HIV: No = 0, Yes = 1;

$$OR_{\text{no/yes}} = \frac{\Pi(x=1)}{1-\Pi(x=1)} \Big/ \frac{\Pi(x=0)}{1-\Pi(x=0)} = 2458.9067$$

Those participants who didn't know the place where to get testing for HIV were 2459 times more likely not to have ever been tested for HIV than those who know the place where to test for HIV while other predictors are holding constant. For place of residence, urban = 1, rural = 0;

$$OR_{\text{urban/rural}} = \frac{\Pi(x=1)}{1-\Pi(x=1)} \Big/ \frac{\Pi(x=0)}{1-\Pi(x=0)} = 0.8887$$

The urban residents were 11% more likely to have ever been tested for HIV than rural while other predictors are holding constant. For region, Tigray = 1, all other regions = 0;

$$OR_{\text{Tigray/Others}} = \frac{\Pi(x=1)}{1-\Pi(x=1)} \Big/ \frac{\Pi(x=0)}{1-\Pi(x=0)} = 0.6383$$

Participants from Tigray were 36% more likely to have ever been tested for HIV than other regions in Ethiopia while other predictors are holding constant.

$$OR_{\text{Somali/Other}}=\frac{\Pi(x=1)}{1-\Pi(x=1)}\bigg/\frac{\Pi(x=0)}{1-\Pi(x=0)}=2.2961$$

Participants from other regions of Ethiopia were 56% more likely to have ever been tested for HIV than Somali region while other predictors are holding constant. For ethnicity, Nuwer = 1, all other ethnicity = 0;

$$OR_{\text{Nuwer/Others}}=\frac{\Pi(x=1)}{1-\Pi(x=1)}\bigg/\frac{\Pi(x=0)}{1-\Pi(x=0)}=1.7554$$

Participants from other ethnic group in Ethiopia (non-Nuwer) were 43% more likely to have ever been tested for HIV than Nuwer ethnic group of Gambella region while other predictors are holding constant.

In summary, the logistic regression predictive model of data mining technique showed that women were 36% more likely to have ever been tested for HIV than men. Participants from all regions in Ethiopia except for Diredawa were more likely not to have ever been tested for HIV compared to participants from Tigray. In addition, the urban residents were 12% more likely to have ever been tested for HIV than rural residents.

In this study, the Tigreans and Amara ethnic groups were less likely to have never been tested for HIV compared to all ethnic categories in Ethiopia (Afar, Nuwer, Oromo, Somalie, Sidama, Welaiyta, Guragie and others). Moreover, all educational categories were more likely to have ever been tested for HIV compared to the people who belong to no education category.

On the other hand, those who heard about family planning on mass media were more likely to get tested for HIV than those who didn't hear about family planning. Those participants who had read frequently newspaper were 17% less likely to have never been tested for HIV than those who didn't read newspaper. Furthermore, those who had HIV related stigma (moderate and high stigma) towards to a person who is infected with HIV were more likely to have never been tested for HIV than those who didn't have stigma towards to an infected individual.

5.5. Association Rules

Association rule mining was performed using apriori algorithm to discover the relationship of the selected attributes with being tested for HIV. Apriori in WEKA 3.7.7 starts with the upper bound support and incrementally decreases support by delta value. In most cases, it is sufficient to focus on a combination of support and confidence to quantitatively measure the quality of the rules. However, the real value of a rule, in terms of usefulness and action ability is subjective and depends heavily on the particular domain and business objectives. To conduct the association rules for this study, the machine used a minimum support of 95% and with 90% of confidence level based on the attributes it took as inputs. The 10 best rules are annexed (A).

The association rule extracted some interesting patterns and it demonstrated that those study participants who have not heard about family planning on mass media (either TV, radio or newspaper) were less likely to have ever been tested for HIV than those who heard about family planning (rules 3, 8). This association rule could demonstrate the direct relationship between family planning and HIV testing.

6. Discussion

This study examines four popular data mining algorithms (Decision tree, Naïve Bayes, Neural network, logistic regression) to build a model that predicts whether an individual being tested for HIV among adults in Ethiopia using EDHS 2011. The results of the experiment performance were evaluated based on their accuracy, sensitivity, specificity and area under the ROC curve. This study used these four different types of predictive models due to their popularity in the recently published literatures [15].

In this study, several experimentations with different testing options have been performed. The final experimentation results indicated that the decision tree (random tree algorithm) performed the best with accuracy of 96%, the decision tree induction method (J48) came out to be the second best with a classification accuracy of 79%, followed by neural network (78%). Logistic regression has also achieved the least classification accuracy of 74%. This implies that random tree, J48 algorithm, neural network and logistic regression predictive models were able to predict whether an individual was being tested or not for HIV given that wealth index, education level, ethnicity, residence, age group, region, knowing a place where to get testing for HIV, knowledge on family planning, knowledge related to HIV, HIV related stigma, risky sexual behaviors as an input with an accuracy of 96%, 79%, 78% and 74% respectively. List of the complete attributes used for prediction are annexed in (**Table 2**).

It has been stated that neural networks are known to be able to identify relationships even when some of the input data are very complex, ill-defined and ill structured [19]. In contrast to this, this study revealed that decision trees came out to be the best model predictor for HIV testing. This might be due to the fact that decision trees have especial attractive in a data mining environment for several reasons. First, all attributes used for predictive models of this study were categorized and this enables the resulting clasification model is easy to assimilate by humans [43] [44]. It has been also noted that the tree complexity has a crucial effect on its accuracy [43]. Third, decision trees can be constructed relatively fast and the accuracy of decisoin treesare comparable or supperior to other classification models [43] [44]. Therefore, the nature of the attributes used for this study might be convenient for decision trees to provide better predictive power. On the other hand, despite the numerous applications of neural network to classification in medicine, very little attention has been made to the HIV/AIDS prevention and planning [24]. Moreover, predictive models are basically evaluated based on three criteria: accuracy, specificity and sensitivity [15]. Hence, decision trees also achieved the higher values for each criterion for this study. This might be due to the nature of the attributes is convenient for decision trees and most importantly decision tree models are one of the predictive models that can predict the class value for a new case very quickly [45]. In addition, they are also very flexible, so that they can provide a powerful predictive tool [45].

Association rule mining was also performed using apriori algorithm to discover the relationship of selected predictors with HIV testing. Accordingly, this study has demonstrated that knowledge on family planning was positively associated with ever being tested for HIV. The association rule extracted from this study indicated that these people who have heard about family planning on mass media (radio, TV or newspaper) were more likely to have ever been tested for HIV. It is clear that both HIV testing and family planning services help clients avoid unwanted consequences of their sexual behavior which are HIV and unintended pregnancies respectively. In addition, this indicated that family planning and HIV testing services are interlinked and have many common objectives [46]. This is consistent with the findings of asystematic review which found that behavior that might lead to unintended pregnancies can also be a risk factor for HIV infection [46]. The authors further described family planning services are including contraceptive service provision, counseling, education and abortion [46]. Therefore, having knowledge on family planning may provide a wider opportunity to be tested for HIV.

This study has some limitations such as information about ever being tested for HIV was based on self-reported responses. Hence, the result of this study could be potentially affected by recall bias in case the test was offered a long time ago. In addition, the result could also be affected by the tendency of respondents preferring to give their desirable answers. Another limitation of this study is, the measurement used to calculate stigma index may not have been satisfactory due to limited questions to cover the various dimensions of HIV/AIDS-related stigma.

The other major limitation of this study is that its principal data source is a cross-sectional survey of the 2011 EDHS. Hence, study participants are not followed up; it might be difficult to know the changes on HIV testing practices of the study participants over time. Moreover, attributes were selected based on business process analysis and facts from literatures. It is also clear that data mining requires human intrusion to exploit the extracted knowledge. Hence, the patterns found in this study has to be evaluated by health experts (who have experience in the problem domain) to decide whether they are logical, actionable and novel to fuel new biological and clinical research directions.

7. Conclusions and Recommendations

7.1. Conclusions

This study showed that the predictive models: random tree, J48 algorithms, neural networks and logistic regression were able to predict whether an individual was being tested or not for HIV given that wealth index, education level, residence, HIV related stigma, knowledge on family planning, knowledge related to HIV, region, age group, and risky sexual behavior as inputs with an accuracy of 96%, 79%, 78% and 74% respectively. The least proportion of ever been tested for HIV was observed among study participants from Somali region, Ethiopia. On the other hand, those participants who were in the age group of 20 to 29 were more likely to have ever been tested for HIV compared to participants in other age groups.

In addition, the association rule extracted from this study indicated that those people who had heard about family planning on mass media were more likely to have ever been tested for HIV than those who hadn't heard about family planning on mass media (TV, radio or newspaper). Furthermore, the heterogeneity between clusters observed in this study could demonstrate the possible hypothesis development by characterizing the variation between subgroups concerning the HIV testing consequent (HIV status determination) among adults in Ethiopia. This could invite researchers to further study.

In conclusion, the results obtained from this research reveal that data mining is crucial in extracting relevant information for the effective utilization of HIV testing services which has clinical, community and public health importance at all levels. Furthermore, this study would also invite interested researchers to explore more on the application of data mining techniques in healthcare industry or else in related and similar settings for the future.

7.2. Recommendations

Based on the findings of this study, the following recommendations are forwarded.

- It is crucial to apply different data mining techniques for the same settings and compare the model performances based on accuracy, sensitivity, and specificity.
- Data mining is crucial in extracting relevant information for the effective utilization of HIV testing services which has clinical, community and public health importance at all levels.
- In order to encourage people to get testing for HIV, the HIV/AIDS prevention and control programs in Ethiopia should focus on reducing HIV related stigma, improving educational level and creating awareness of the society on HIV testing through mass media at large.
- Integrating family planning services with HIV testing could improve the proportion of the citizens that could be tested for HIV.
- Targeting on Somali region and Nuwer ethnic group (Gambella) while designing for HIV testing services would greatly reduce the risk of HIV/AIDS.
- Focusing on adults in the age group of 20 to 29 years old for HIV testing services also would greatly reduce the risk of HIV/AIDS, which is evidenced in this group.
- It is highly important that future ethnographic research should investigate the observation found on Nuwer ethnic group by comparing with other ethnic groups in Ethiopia.
- The strengthening of the health programs on advocating the benefits of HIV testing through mass media (TV, radio or newspaper) might be helpful to reduce fear of stigma and discrimination amongst adults.
- Efficient distribution of health care facilities offering HIV testing services among urban and rural areas are required
- This study would also invite interested researchers to explore more on the application of data mining techniques in healthcare industry or else in related and similar settings for the future.
- Attributes are selected based on business process analyses and facts from literatures. However, it can also be possible to utilize other feature selection algorithms for the purpose of comparison.

Acknowledgements

I am very grateful to MEASURE DHS authority for they have offered me all the necessary data and documents which I needed for this work. I would also like to extend my heartfelt gratitude to Dr. Girma Tadesse for his constructive comments, critical readings, encouragements and suggestions which are really important for the accomplishment of this work. Extraordinary thanks go for my almighty God for being with me and granting me grace, guidance and strength throughout my work.

References

[1] Ababa, A. (2006) AIDS in Ethiopia: 6th Report. Federal Ministry of Health National HIV/AIDS Prevention and Control Office.

[2] The Voluntary HIV-1 Counseling and Testing Study Group (2000) Efficacy of Voluntary HIV-1 Counseling and Testing among Individuals and Couples in Kenya, Tanzania, and Trinidad: A Randomized Trial. *The Lancet*, **356**, 103-112. http://dx.doi.org/10.1016/S0140-6736(00)02446-6

[3] UNAIDS (2002) HIV Voluntary Counseling and Testing: A Gateway to Prevention and Care. Five Case Studies Related to Mother-to-Child Transmission of HIV, Tuberculosis, Young People, and Reaching General Population Groups.

UNAIDS Case Study.
http://www.unaids.org/en/media/unaids/contentassets/dataimport/publications/irc-pub02/jc729-vct-gateway-cs_en.pdf

[4] WHO (2006) Towards Universal Access: Part II. A Report on "3 by 5" and Beyond.

[5] Lawn, S.D., Myer, L., Orrell, C., Bekker, L.G. and Wood, R. (2005) Early Mortality among Adults Accessing a Community-Based Antiretroviral Service in South Africa: Implications for Programme Design. *AIDS*, **19**, 2141-2148. http://dx.doi.org/10.1097/01.aids.0000194802.89540.e1

[6] Ita, M. (1998) Counseling in Reproductive Health among Young People in the Shitta Community in Lagos State. *Abstracts of the XIIth International AIDS Conference*, Geneva, 28 June-3 July 1998, Abstract 60857.

[7] Valdiserri, R.O., Holtgrave, D.R. and West, G.R. (1999) Promoting Early HIV Diagnosis and Entry into Care. *AIDS*, **13**, 2317-2330. http://dx.doi.org/10.1097/00002030-199912030-00003

[8] Carpenter, C.C., Fischl, M.A., Hammer, S.M., *et al*. (1998) Antiretroviral Therapy for HIV Infection: Updated Recommendations of the International AIDS Society-USA Panel. *The Journal of the American Medical Association*, **280**, 78-86. http://dx.doi.org/10.1001/jama.280.1.78

[9] Quinn, T.C., Wawer, M.J., Sewankambo, N., *et al*. (2000) Viral Load and Heterosexual Transmission of Human Immunodeficiency Virus. *The Journal of the American Medical Association*, **342**, 921-929.

[10] Alwano-Edyegu, M.G. and Marum, E. (1999) Knowledge Is Power: Voluntary HIV Counseling and Testing in Uganda. UNAIDS, Geneva.

[11] Denning, P., Nakashima, A., Wortley, C. and SHAS Project Group (1999) High Risk Sexual Behaviors among HIV-Infected Adolescents and Young Adults. *Abstracts of the* 6*th Conference on Retroviruses and Opportunistic Infections*, Chicago, 31 January-4 February 1999.

[12] Assefa, Y., Jerene, D., Lulseged, S., Ooms, G. and Van Damme, W. (2009) Rapid Scale-Up of Antiretroviral Treatment in Ethiopia: Successes and System-Wide Effects. *PLoS Medicine*, **6**, e1000056.
http://dx.doi.org/10.1371/journal.pmed.1000056

[13] Kononenko, I. (2001) Machine Learning for Medical Diagnosis: History, State of the Art and Perspective. *Artificial Intelligence in Medicine*, **23**, 89-109.

[14] Koh, H.C. and Tan, G.J. (2005) Data Mining Applications in Healthcare. *Journal of Healthcare Information Management*, **19**, 64-72. http://www.ncbi.nlm.nih.gov/pubmed/15869215

[15] Delen, D., Walker, G. and Kadam, A. (2004) Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods. *Artificial Intelligence in Medicine*, **34**, 113-127.

[16] Wang, J., Hu, X.H. and Zhu, D. (2008) Applications of Data Mining in the Healthcare Industry.
http://www.irma-international.org/viewtitle/12924/

[17] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996) From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, **17**, 36-51.

[18] Hand, D., Mannila, H. and Smyth, P. (2001) Principles of Data Mining. The MIT Press, Cambridge.

[19] Azevedo, A. and Santos, M.F. (2008) KDD, SEMMA and CRISP-DM: A Parallel Overview. *Proceedings of the IADIS European Conference Data Mining*, Amsterdam, 24-26 July 2008, 182-185.

[20] Bigus, J.P. (1996) Data Mining with Neural Networks. McGraw-Hill, New York.

[21] Kurgan, L.A. and Musilek, P. (2006) A Survey of Knowledge Discovery and Data Mining Process Models. *The Knowledge Engineering Review*, **21**, 1-24.

[22] Mining Techniques in Health Care (2011).

[23] Refaat, M. (2007) Data Preparation for Data Mining Using SAS. Morgan Kaufmann Publishers, San Francisco.

[24] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000) CRISP-DM 1.0: Step-by-Step Data Mining Guide. SPSS, Copenhagen.

[25] Shouman, M., Turner, T. and Stocker, R. (2012) Using Data Mining Techniques in Heartdisease Diagnosis and Treatment. http://www.academia.edu/1543820/

[26] Nabney, I. (2003) Netlab: Algorithms for Pattern Recognition. Springer Verlag, Berlin.

[27] Leke-Betechuoh, B., Marwala, T., Tim, T. and Lagazio, M. (2006) Prediction of HIV Status from Demographic Data Using NNs. *Proceedings of the* 2006 *IEEE International Conference on Systems*, *Man and Cybernetics*, Taipei, 8-11 October 2006, 2339-2344.

[28] Abraham, T. (2005) Application of Data Mining Technology to Identify Determinant Risk Factors of HIV Infection and to Find Their Association Rules: The Case of Center for Disease Control and Prevention (CDC). Master's Thesis, Addis Ababa University, Addis Ababa.

[29] Asmare, B. (2009) Application of Data Mining Technology to Support VCT for HIV: A Case of Center for Disease Control and Prevention. Master Thesis, School of Information Science, Addis Ababa University, Addis Ababa.

[30] Lemuye, E. (2011) HIV Status Predictive Modeling Using Data Mining Technology. Master Thesis, AAU School of Information Science and Public Health, Addis Ababa.

[31] Lee, C.W. and Park, J.-A. (2001) Assessment of HIV/AIDS-Related Health Performanceusing an Artificial Neural Network. *Information & Management*, **38**, 231-238. http://dx.doi.org/10.1016/S0378-7206(00)00068-9

[32] Kwak, N.K. and Lee, C. (1997) A Neural Network Application to Classification of Health Status of HIV/AIDS Patient. *Journal of Medical Systems*, **21**, 87-97. http://dx.doi.org/10.1023/A:1022890223449

[33] Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. 2nd Edition, Morgan Kaufmann Publishers, San Francisco.

[34] Liu, H. and Motoda, H. (1998) Feature Selection for Knowledge Discovery and Data Mining. Springer, Berlin.

[35] Famili, A. and Turney, P. (1997) Data Preprocessing and Intelligent Data Analysis. Institute of Information Technology, National Research Council Canada.

[36] Chakrabarti, S., Cox, E., Frank, E., Hartmut, G.R., Han, J., Jiang, X., Kamber, M. and Witten, I. (2009) Data Mining: Know It All. Morga Kaufmann Publishers, Burlington, San Francisco.

[37] Kantardzic, M. (2003) Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, Hoboken.

[38] Brachman, R. J. and Anand, T. (1996) The Process of Knowledge Discovery in Databases.

[39] Zurada, J.M. (1992) An Introduction To Artificial Neural Networks Systems. West Publishing, St. Paul.

[40] Lu, H., Setiono, R. and Liu, H. (1996) Effective Data Mining Using Neural Networks. *IEEE Transactions on Knowledge and Data Engineering*, **8**, 957-961.

[41] Roberts, A. (2005) AI32: Guide to Weka.

[42] Witten, I.H. and Frank, E. (2005) Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition, Morgan Kaufmann Publishers, San Francisco.

[43] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and Regresion Trees.

[44] Shafer, J., Agrawal, R. and Mehta, M. (1996) SPRINT: A Scalable Parallel Classifier for Data Mining. *Proceedings of the* 22*th International Conference on Very Large Data Bases*, Mumbai, 3-6 September 1996, 544-555.

[45] Predictive Modeling. http://searchdatamanagement.techtarget.com/definition/predictive-modeling

[46] Spaulding, A.B., Brickley, D.B., Kennedy, C., Almers, L., Packel, L., Mirjahangir, J., Kennedy, G., Collins, L., Osbornee, K. and Mbizvo, M. (2009) Linking Family Planning with HIV/AIDS Interventions: A Systematic Review of the Evidence. *AIDS*, **23**, S79-S88.