Journal of
Health & Medical Informatics

Mariam and Mariam, J Health Med Informat 2015, 6:6
DOI: 10.4172/2157-7420.1000208

Research Article Open Access

Application of Data Mining Techniques for Predicting CD4 Status of Patients on ART in Jimma and Bonga Hospitals, Ethiopia

Behailu Gebre Mariam[1] and Tesfahun Haile Mariam[2]*

[1]*Bonga College of Teachers Education, Bonga, P.O.B. 91, Bonga, Ethiopia*
[2]*Department of Health Informatics, Hawassa Health Science College, P.O.B 84 Hawassa, Ethiopia*

Abstract

Background: Many of the reports on HIV/AIDS shows that the number of ART registered patients are increasing from time to time. Despite those reports show increasing of patients' number, they did not try to make prediction of attributes based on the given attributes more than statistical explanation. This study concerned to use data mining techniques on ART data base. The main objective of the study is to apply data mining techniques for predicting CD4 status of patients on ART in Jimma and Bonga Hospitals.

Methodology: The study followed the CRISP-DM data mining methodology which has six phases: business understanding, data understanding, data preparation, model building, evaluation and deployment. For this study, data was taken from two hospitals of the south west of Ethiopia; Jimma and Bonga hospitals. Classification algorithm was used to predict CD4 status of the patients those who are following ART therapy. J48 is a technique used for building classification and PART is used to compare the result of J48 algorithm.

Results: The best performance achieved by J48 decision tree algorithm is a generalized decision tree pruning with reduced attributes. The model classifies instances correctly (88.79%) and incorrectly (11.21%). The weighted average precision of the model is 0.88 with recall of 0.89 and ROC area of 0.85. The model has 760 numbers of leaves and 916 tree size. The time taken to build the model is 0.05 seconds. The analysis of this model shows that the model is quit efficient to predict CD4 status of patients those who are following ART.

Conclusion: Classification done using J48 decision tree is the best model as compared to PART rule algorithm and that can be used for prediction. From the model built it is possible to conclude that attributes like: Eligible reason, ART status, ART start year, OA weight, OAWHO stage, Current regimen, Family planning, Functional status, Marital status, Past ARV are the most determining factors of CD4 status.

Keywords: Predictive modeling; ART data base; CD4 status; Data mining.

Introduction

Reports on HIV/AIDS shows that the number of ART registered patients are increasing from time to time [1]. Antiretroviral Therapy (ART) is treatment for AIDS that helps the body's immune system recover from the damage caused by infection with HIV. Although ART cannot cure AIDS, persons on ART will begin to feel better, eat more, and put on weight. Their bodies will recover the ability to fight infections. As persons on ART treatment become well, they can care for their children and return to household activities and productive life, which benefits the household and national economies [1].

CD4 cells are a type of lymphocyte and they are an important part of the immune system. They are sometimes called T-cells. HIV most often infects CD4 cells. The genetic code of the virus becomes part of the cells. When CD4 cells multiply to fight an infection, they make more copies of HIV. When someone is infected with HIV but has not yet started treatment, the number of CD4 cells goes down. This is a sign that the immune system is being weakened. The lower the CD4 cell count, the more likely the person will get sick. There are millions of different families of CD4 cells. Each family is designed to fight a specific type of germ. When HIV reduces the number of CD4 cells, some of these families can be wiped out. The person can lose the ability to fight off the particular germs that they were designed for. If this happens, the person might develop an opportunistic infection [2].

Although the epidemic is currently stable, HIV/AIDS remains a major development challenge for Ethiopia [3]. Poverty, food shortages, and other socio-economic factors amplify the impact of the epidemic. Many of the reports on HIV/AIDS shows that the number of ART registered patients are increasing from time to time. However those reports show that the increasing of patient's number, they did not try to make prediction of attributes based on the given attributes more than statistical explanation Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions [3].

The CD4 cell count remains the strongest predictor of HIV related complications, even after the initiation of therapy. The baseline pre-treatment value is informative: lower CD4 counts are associated with smaller and slower improvements in counts. However, precise thresholds that define treatment failure in patients starting at various CD4 levels are not yet established. As a general rule, new and progressive severe immune deficiencies are demonstrated by declining longitudinal CD4 cell counts that trigger patients to switch therapy [4].

A patient starting with low CD4 count may demonstrate slow recovery, but persistent levels below 100 cells/mm^3 represents significant risk for HIV disease progression. As a general principle,

***Corresponding author:** Tesfahun Haile Mariam, Lecturer, Department of Health Informatics, Hawassa Health Science College, P.O.B 84 Hawassa, Ethiopia, Tel: +251 462 205311; E-mail: tesfahunhailemariam@gmail.com

Received October 09, 2015; **Accepted** December 01, 2015; **Published** December 07, 2015

Citation: Mariam BG, Mariam TH (2015) Application of Data Mining Techniques for Predicting CD4 Status of Patients on ART in Jimma and Bonga Hospitals, Ethiopia. J Health Med Informat 6: 208. doi:10.4172/2157-7420.1000208

inter current infections should be managed; time should be allowed for recovery and the CD4 cell count should be measured before ART is switched. The CD4 cell count can also be used to determine when not to switch therapy, e.g. in a patient with a new clinical stage 3 even for whom switching is being considered or in a patient who is asymptomatic and under routine framework. In general, switching should not be recommended if the CD4 cell count is above 200 cells/mm^3 [5]. As we have seen in the above paragraphs CD4 cell count has many advantages on the patients ART follow up. Since it is difficult to measure the CD4 cell count every time, and will create unnecessary anxiety on patients; so that it is necessary to prepare a CD4 cell count status prediction model.

Data mining technology helps to find patterns in data that are valid, novel, useful, and understandable. The pattern which has discovered can help the health care service provider and decision makers in order to give knowledge full decisions. Predicting CD4 status from the ART dataset of patients who following ART is very important in order to reduce the complication of patients. This research is to find the pattern of attributes of the patients in order to build predictive model using data mining techniques. The model built helps to predict patient's CD4 status either will decrease or increase after a known period of time. Based on this predictive value, the patient and the concerned health service delivery body can control the situation before happening using different methods.

Thus, the goal of this research is to apply data mining techniques and predict CD4 status of patients who will develop low CD4 count status and high CD4 count status after three months using the data set of ART database. The research outcomes have a contribution to identify the important patterns of ART data set in order to make decision and intervention. In doing so, this research answers the following questions:

- What are the main attributes affecting CD4 status?
- Which data mining technique is suitable for predicting CD4 status of patients?

Methods/Data Mining Modeling

In this research CRIPS-DM (Cross Industry Standard Process for Data Mining) modeling was used as shown in Figure 1.

Data mining and knowledge discovery involves a complex process of identifying, understanding, and transforming data modeling and evaluation efforts. Naturally, any models that are useful would need to be deployed before making an impact in practice. CRISP-DM offers a general model for data/text mining projects that highlighting the key tasks involved. According to the CRISP-DM framework, the life cycle of a knowledge discovery process consists of six phases. CRISP-DM method is iterative in nature for the choice of subsequent phases often depends on the outcome of preceding phases [6].

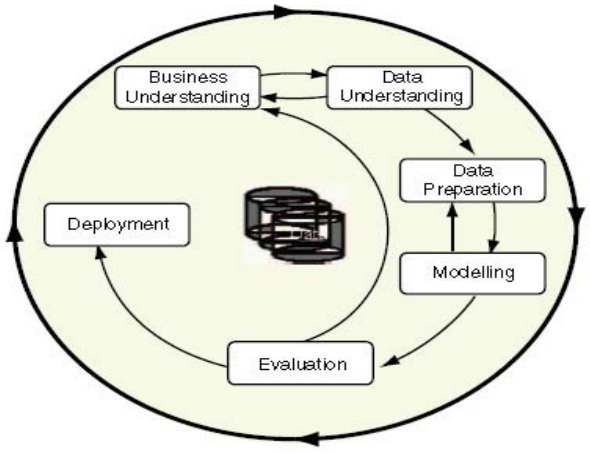

Figure 1: 1 Phases of the CRISP-DM modelSource [6].

The algorithm used by Weka is known as J48. It is a version of an earlier algorithm. Decision trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data. The researcher used J48 algorithm in this study, for it gives several options related to tree pruning. Pruning produces fewer, more easily interpreted results. More importantly, pruning can be used as a tool to correct for potential over-fitting. Thus, it classifies until each leaf is pure i.e. the data has been categorized as close to perfectly as possible so that the process ensures maximum accuracy on the training data.

Classification model decision tree, attribute ranking, confidence factor and incorrectly classified instances, measuring classifier accuracy for decision tree, binary decision tree model building, generalized decision tree model building, ROC Area curve analysis for different scenarios, rule generating using PART rule induction Algorithm were used in this study [7,8]. The Weka GUI provides a starting point for launching Weka's main GUI applications and supporting tools. It includes access to the four Weka's main applications: Explorer, Experimenter, Knowledge Flow and Simple CLI.

Understanding and pre-processing Jimma and Bonga hospitals dataset

Raw data description: The data sources for this research was Jimma, and Bonga hospitals ART data base. The total data of the two databases is 8438; obtained from 2003 to 2012. The numbers of cases are 6242 and 2197 in Jimma and Bonga hospitals respectively. The Patient and Register files in combination contain all the relevant attributes. Attributes from the files were:

Registration Date, Sex, Age, Religion ID, Marital Status ID, Educational Level ID, Occupation, ART Status, Functional Status, EL Date, Eligible Reason ID, ART Start Date, Family Planning, Pregnant YN, OA Weight, OAWHO stage, Current Regimen, Past ARV Treatment, and OACD4 as shown in Table1.

Eligibility reason deals about why the patient is eligible to the treatment of ART. This attribute has nominal value and they are: Clinical only, CD4, TLC, Transfer in (TI), Clinical and TLC, Clinical and CD4. The data originally entered as numeric values 0 to 6. Based on the Ministry of Health guide line the researcher recoded the numeric values 1 to Clinical only, 2 to TLC, 3 to CD4, 4 to Transfer in (TI), 5 to Clinical and TLC and 6 to Clinical and CD4.

Data preprocessing : The approach that gives an even simpler solution for elimination of missing values in this study is a formal and is often automatic replacement of missing values with some constant value. The first type of replacement is replacing a missing value with its mean for numeric data type (Table 2). The other replacement is done by replacing a missing value with mode for the nominal data type (Table 2).

Data transformation is converting data from different sources into common new format. Apply data reduction & data categorization/ binning to ease data mining. Discretization: Reduce data size by dividing the range of a continuous attribute into intervals. Interval labels can then be used to replace actual data values. The Discretization

No	Attributes	Meaning	Value	Data type
1	Sex	Sex of the patient	Female Male	Nominal
2	Age	Age of the patient	Numeric age values ranged as 0-14, 15-24, 25-49, 50-64	Numeric
3	Religion	The religion of the patient	1-Muslim 2-Orthodox 3-Protestant 4-Catholic 5-Other	Nominal
4	Marital Status ID	The marital status of the patient	1-Never married 2-Married 3-Separated 4-Divorsed 5-Widow/Widower	Nominal
5	Educational Level ID	Educational level of the patient	1-No education 2-Primary 3-Secondary 4-Teritiary	Nominal
6	ARTStatus	Status of ART care	OA-On taking ARV drug IN-In care for other disease EL-Eligible to be on ARV Drug ER-Eligible and ready to start ARV drug	Nominal
7	Functional Status	Functional level of the patient	W-working A-Ambulatory B-Bedridden	Nominal
8	Reason Eligible For ART	The reason for the patient eligible for ART	1-Clinical only 2-TLC 3-CD4 4- Transfer in(TI) 5-Clinical and TLC 6- Clinical and CD4	Nominal
9	ART Start Date	The date ART starts	From 2003-2012	Date
10	Family Planning	This attribute asks the usage of Family planning methods	Yes/No	Yes/No
11	Pregnant	Asks the patient pregnant or not on ART	Yes/No	Yes/No
12	OAWeight	The weight of the patient on ART	Numeric values ranged as:0-99,100-199, 200-349, 350-999,1000-2999	Numeric
13	OAWHO stage	WHO stages at which the patient is on ART	1-stage 1 2-sage 2 3-stage 3 4-stage 4	Nominal
14	Current Regimen	The current regimen the patient is taking on ART	1a(30)=d4T(30)-3TC-NVP 1a(40)=d4T(40)-3TC-NVP 1b(30)=d4T(30)-3TC-EFV 1b(40)=d4T(40)-3TC-EFV 1c=AZT-3TC-NVP 1d=AZT-3TC-EFV 2a, 2b, 2c, 2d. 4a, 4b, 4c and 4d. 5a, 5b, 5c, 5d Other	Nominal
15	Past ARV Treatment	Dose the patient took Past ARV Treatment	Yes for treatment No for not treatment	Yes/no
16	OACD4	The CD4 count of the patient currently on the ART	None zero positive number ranged as:0-49,50-99, 100-199,200-349,350-999	Numeric

Table 1: The Raw Data Attributes Description.

is performed using method of Equal-width (distance) partitioning. This method divides the range into N intervals of equal size. The method assign A and B as the lowest and highest values of the attribute. The width of intervals: W = (B –A)/N. Applying the formula the OA weight attribute discretized as shown in Table 3.

The other transformation done in this research is transforming the numeric value of OA CD4 count to nominal value. In order to transform the value, the researcher followed the baseline treatment guide line. Different scholars suggest that there should be a cut point for CD4 count to be low and normal. According to Mellors [5] classifies as: if CD4 count is below 200 cells/mm^3, it can be categorized to low CD4 status. Where as, if the CD4 count is above 200 cells/mm^3, it can be categorized to normal CD4 status. Table 4 below shows the transformed OA CD4 count.

Exploratory data analysis: The attribute's description, data type, use of frequency tables for the selected attributes was done on the dataset before experimental analysis using data mining. By using frequency tables, the exploratory data analysis was performed and the interpretation was done.

Sex: This attribute describes about the sex of the patient and it was described as Female and Male. Out of the total 8438 data, 5471 was females and 2536 was males. The missing value of the data was 431 and it holds 5.1% of the total data as shown in Table 5.

Age: The age attribute contains the age of the patients which starts from 0 to 64 and above 65 considering as a single category. This attribute has grouped into ranges of different values i.e. 0-14, 15-24, 25-49, 50-64, and above 65. Out of the total data, missing values was 0.38% (Table 6).

Religion: The religion attribute contains the religion of the patients

No	Attributes	Percentage of missing values	Replaced with	Data type	Remark
1	Sex	5.1	Female	Nominal	Mode
2	Age	0.38	28.8(25-49)	Numeric	Mean
3	Religion	6.33	Orthodox	Nominal	Mode
4	Marital status	5.84	Married	Nominal	Mode
5	Educational level	5.93	Primary	Nominal	Mode
6	Functional Status	19.48	W-working	Nominal	Mode
7	Reason Eligible For ART	1.25	Transfer in (TI)	Nominal	Mode
8	ART start Date	25.20	2006	Numeric	Mode
9	Family Planning	14.14	No	Yes/no	Mode
10	OA weight	30.11	48.5	Numeric	Mean
11	WHO stage	27.4	Stage3	Nominal	Mode
12	Current Regimen	34.83	1a(30)	Nominal	Mode
13	Past ARV treatment	3.68	No	Yes/no	Mode
14	OACD4	7.83	102	Numeric	Mean

Table 2: Summary of Missing Value Handled.

Label	Frequency	Percent
3-27	280	3.31
27.5-52	3415	40.47
52.5-77	2170	25.71
77.5-102	31	0.36
Missing	2541	30.11
Total	8438	100

Table 3: Discretized Result of Weight Attributes.

OA CD4 count: Nominal data type	
Value	**Frequency**
Low	6987
Normal	1451
Total	8438

Table 4: Transformed Result of OA CD4 Count.

is following. The data originally code as numeric data from 1 to 5. Based on the ministry of Health guide line the researcher recode the numeric codes to Muslim, Orthodox, Protestant, Catholic, and Other in 1, 2,3,4,5 respectively. The missing values of the data was 535 (6.33%) of the total data (Table 7).

Marital status: The marital status attribute shows the marriage status of the patients. The data originally coded as numeric data from number 1 to number 5. Based on the Ministry of Health guide line, the researcher recoded the numeric codes to nominal values of never married, Married, Separated, Divorced, Widow in 1, 2,3,4,5 respectively as shown in Table 8.

Educational level: The educational level attributes deals on the level of education of the patient. Originally the data coded as numeric data from 1 to 4. Based on the Ministry of Health guide line the researcher recoded the numeric code to nominal values as No education, Primary, Secondary, and Tertiary in 1,2,3,4 respectively. The data contains 501 missing values, which accounts 5.93% of the total data (Table 9).

Sex: Nominal data type		
Value	**Frequency**	**Percent**
Missing	431	5.1
Female	5471	64.8
Male	2536	30
Total	8438	100

Table 5: Statistical Summary of Sex Attribute.

Age: Numeric data type		
Value	**Frequency**	**Percent**
Missing	32	0.38
0-14	290	3.44
15-24	2191	26
25-49	5692	67.45
50-64	231	2.73
Above 65	2	0.02
Total	8438	100

Table 6: Statistical Summary of Age Attribute.

Religion: Nominal data type		
Value	**Frequency**	**Percent**
Missing	535	6.33
Muslim	2295	27.19
Orthodox	4489	53.19
Protestant	1012	11.99
Catholic	45	0.53
Other	59	0.69
Total	8438	100

Table 7: Statistical Summary of Religion Attribute.

Marital Status : Nominal data type		
Value	**Frequency**	**Percent**
Missing	493	5.84
Never married	1681	19.91
Married(inc.de facto)	3793	44.94
Separated	477	5.65
Divorced	1010	11.96
Widow	984	11.66
Total	8438	100

Table 8: Statistical Summary of Marital Status Attribute.

ATR status: The attribute ART status shows the status of the patient condition currently. This attribute has five nominal data values, those are OA-On taking ARV drug, IN-In care for other disease, EL-Eligible to be on ARV Drug, ER-Eligible and ready to start ARV drug. The attribute does not have missing values (Table 10).

Functional status: The functional status attribute shows the patients status of physical and mental wellbeing on the ART. This is important to know the patients status whether in working or bed ridden conditions. The attribute contains three nominal data values coded as W for working, A for ambulatory, and B for bedridden. The data contains 1644 missing values (19.48%) of the total data (Table 11).

Reason eligible for ART: Eligibility reason attribute deals why the patient is eligible to the treatment of ART. This attribute has nominal values like: Clinical only, CD4, TLC, Transfer in (TI), Clinical and TLC, Clinical and CD4. The data originally entered as numeric values 0 to 6. Based on the Ministry of Health guide line the researcher recoded the numeric values 1 to Clinical only, 2 to TLC, 3 to CD4, 4 to Transfer in (TI), 5 to Clinical and TLC and 6 to Clinical and CD4 (Table12).

ART start date (year): The start year derived from the ART start

EducationalLevel : Nominal data type		
Value	**Frequency**	**Percent**
Missing	501	5.93
No education	1538	18.22
Primary	2998	35.52
Secondary	2786	33.01
Tertiary	613	7.26
Total	8438	100

Table 9: Statistical Summary of Educational Level Attribute.

ART status : Nominal data type		
Value	**Frequency**	**Percent**
Missing	0	0
OA-On taking ARV drug	6832	80.95
IN-In care for other disease	1586	18.79
EL-Eligible to be on ARV Drug	20	0.23
Total	8438	100

Table 10: Statistical Summary of ART Status Attribute.

Functional status : Nominal data type		
Value	**Frequency**	**Percent**
Missing	1644	19.48
W-working	50900	60.31
A-Ambulatory	1357	16.08
B-Bedridden	347	4.11
Total	8438	100

Table 11: Statistical Summary of Functional Status Attribute.

Reasoneligible forART: Nominal data type		
Value	**Frequency**	**Percent**
Missing	2751	32.6
Clinical only	927	11.00
TLC	386	4.57
CD4	3683	43.6
Transfer in(TI)	136	1.6
Clinical and TLC	341	4.04
Clinical and CD4	214	2.53
Total	8438	100

Table 12: Statistical Summary of Reason Eligible for ART Attribute.

date of the patient. The year attribute describes the year the patient starts ART and it starts from the year 2003 to 2012. The data contains 2127 missing values (25.2%) of the total data as shown in Table 13.

Family planning: This attribute indicates that the usage of family planning method of the patients. Originally the data coded as 0 for not using and 1 for using methods of family planning. The researcher recoded the 0's and 1's code to No and Yes codes. The data contains 1196 missing values of the total data (14.17%) of the total data (Table 14).

Pregnant: The attribute pregnant indicates that whether the patient is pregnant or not. Originally the data coded as Yes and No for being pregnant and not pregnant respectively. The data do not have missing value. In this data there is no patient who is pregnant (Table 15)

OA weight: The OA weight attribute shows the weight of the patient on the ART service. The weight of the patient in the data starts from 3k.g to above 101k.g. This attribute has not grouped into ranges of different values originally in the database. In order to group the data in to different intervals, the researcher decides first to replace the missing values by mean weight and then group into different intervals by using discritization method. Table16 shows the statistically summary of the weight attribute.

WHO stage: The WHO stage attribute contains the stages of the patient in the treatment. The data originally coded as I, II, III, IV the researcher recode this attribute value in Stage 1, Stage 2, Stage 3 and Stage 4 respectively. The data contains 2313 missing values (27.4%) of the total data (Table 17).

Current regimen: Current regimen attribute describes the regimen type of the patient is taking. The regimen contains two line regimens firs line and second line. The lines also classifies as first line and second line adult and firs line and second line child. First line adult includes1a (30) or d4T (30)-3TC-NVP, 1a (40) or d4T (40)-3TC-NVP, 1b (30) or d4T (30)-3TC-EFV, 1b (40) or d4T (40)-3TC-EFV, 1c or AZT-3TC-NVP and 1d or AZT-3TC-EFV. Second line adult includes 2a , 2b, 2c, 2d. The first line child includes 4a, 4b, 4c and 4d. Second line child include 5a, 5b, 5c, 5d and other. The data contains 2147 missing values, which is 25.44% of the total data (Table 18).

Past ARV treatment: This attribute indicates that whether the patient had took or not any ARV treatment in the past. Originally the data coded as Yes and No for taken treatment and No for not taken treatment respectively. The data contains 311 missing values (3.68%) of the total data (Table 19).

OACD4count: OACD4 count describes the CD4 count values of the patient on ART. The CD4 count already grouped in the database as 0-49, 50-99,100-199,200-349,350-999 and 1000-2999. The data contains 661 missing (7.83%) of the total data. The numeric value of the CD4 count has transformed in to nominal values based on the base line guide of the treatment. Statistical Summary of OA CD4 Count attribute is shown in Table 20.

Experimentation, Analysis and Evaluation of Discovered Knowledge

Attribute ranking

By considering the importance of selecting the attribute, the researcher performed weka to select the best attribute. The implementation by Weka attribute ranking filter using information gain uses attribute Evaluator in supervised class of nominal values of 16 attributes. Attribute selection using entropy based information gain method of Weka, the top 10 determining attributes of the data set for predicting CD4 status of the patient on ATR care service are:Eligible Reasaon,ARTStatus,ARTStartyear,OAWeight,OAWHOstage,Current Regimn,FamiliyPlaning,FunctionalStatus,MaritalStatus,PastARVWith gain of 0.120,0.054,0.035,0.034,0.020, 0.010, 0.009, 0.005, 0.002, 0.002 in respectively. The digit of the gain is rounded off to the nearest number. The ranked attribute includes the attribute's indices number from top to least and the total attribute numbers was used in the comparison. The rank of the attribute shows the relevance of the attributes to the experimentation by excluding the least relevant attributes.

ART start Date (year): Numeric data type		
Value	**Frequency**	**Percent**
Missing	2127	25.20
2003	41	0.49
2004	158	1.83
2005	1091	12.92
2006	2678	31.73
2007	1344	15.92
2008	297	3.50
2009	341	4.02
2010	170	2.01
2011	190	2.25
2012	1	0.01
Total	8438	100

Table 13: Statistical Summary of ART Start Date (Year) Attribute.

Family planning: Nominal data type		
Value	**Frequency**	**Percent**
Missing	1196	14.17
Yes	2590	30.69
No	4652	55.12
Total	8439	100

Table 14: Statistical Summary of Family Planning Attribute.

Pregnant: Nominal data type		
Value	**Frequency**	**Percent**
Missing	0	0
Yes	0	0
No	8438	100
Total	8438	100

Table 15: Statistical Summary of Pregnant Attribute.

OA weight: Numeric data type		
Value	**Frequency**	**Percent**
Missing	2541	31.11
Mean	48.5	
Mode	50	
Minimum	3	
Maximum	101	

Table16: Statistical Summary of OA Weight Attribute.

WHO stage: Nominal data type		
Value	**Frequency**	**Percent**
Missing	2313	27.40
Stage 1	393	4.65
Stage 2	1442	17.08
Stage 3	2971	35.20
Stage 4	1319	15.63
Total	8438	100

Table 17: Statistical Summary of WHO Stage Attribute.

Classification model building

In classification model building the researcher intended to build four scenarios with all attributes like: Binary decision tree with pruning, Binary decision tree without pruning, generalized decision tree with pruning, generalized decision tree without pruning with all attributes for all scenarios. The researcher also intended to build a model based on the reduced (ten top attributes) [8]. Since algorithm selection is important to build the classification model, the researcher tried to implement decision tree model with pruning. The pruning step is achieved by splitting the data into sub-samples that are purer than the original. As discussed in [8], the ideal situation is each sub-sample consists of instances that have the same value for the class attribute i.e. completely pure nodes.

Confidence factor and incorrectly classified instances

Confidence factor has impact on the percentage of classifying instances in to correctly and incorrectly. The researcher conducted different experiments by changing the confidence factor values incrementally. As it can be seen from the experiment result table when the confidence factor increases, the incorrectly classified instance decreases. The researcher decided to take confidence factor 0.5 as a confidence factor for all the experiments; this is because confidence factor 0.5 has minimum incorrectly classified instances than the other experiments. The Table21 below shows that the pattern of confidence factors and incorrectly classified instances.

Current regimen: Nominal data type		
Value	**Frequency**	**Percent**
Missing	2147	25.44
1a(30)=d4T(30)-3TC-NVP	3469	41.11
1a(40)=d4T(40)-3TC-NVP	755	8.94
1b(30)=d4T(30)-3TC-EFV	406	4.81
1b(40)=d4T(40)-3TC-EFV	124	1.46
1c=AZT-3TC-NVP	781	9.25
1d=AZT-3TC-EFV	192	2.27
4a	37	0.43
4b	9	0.11
4c	172	2.03
4d	17	0.20
Other	329	3.89
Total	8438	100

Table 18: Statistical Summary of Current Regimen Attribute.

Past ARV treatment: Nominal data type		
Value	**Frequency**	**Percent**
Missing	311	3.68
Yes	780	9.23
No	7347	87.03
Total	8438	100

Table 19: Statistical Summary of Past ARV Treatment Attribute.

OACD4 count: Numeric data type		
Value	**Frequency**	**Percent**
Missing	661	7.83
0-49	3280	38.87
50-99	1060	12.56
100-199	1987	23.54
200-349	1286	15.24
350-999	151	1.78
1000-2999	14	0.16
Total	8438	100

Table 20: Statistical Summary of OA CD4 Count Attribute.

Confidence factor	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.4	0.45	0.5	0.55
Incorrectly classified instances	988	985	976	967	903	909	879	867	854	823	824	800

Table 21: Incorrectly Classified Instance for Different Confidence Factors.

Measuring classifier accuracy for decision tree

According to [9], it is important to check the appropriateness of the dataset for selecting certain validation method. In order to get the minimum percent to be used as learning purpose, the researcher has computed different experiments by varying the percent of k-fold cross validation as shown in Table 22. For the purpose of this research the experiment has started by taking samples from 10% to 50% as it has been shown on the Table 22. The precision of the experiments are nearly the same. Since there is no much difference on the precision of the experiments, the researcher decided to set the percentage of testing set into 50%.

Binary decision tree model building

The model built by Binary decision tree have four scenarios like: Binary decision tree with all attributes without pruning, Binary decision tree with all attributes with pruning, Binary decision tree with some selected attributes without pruning, Binary decision tree with some selected attributes with pruning. The Figure 2 shows the results using J48 when the binary classification is "True", this means the classifier is now using binary tree classification.

Generalized decision tree model building

Generalized decision tree model is one of the experiments selected by the researcher to be implemented. Generalized decision tree model is built by splitting into more than two sub trees. Even though it split more than two branches, they may be less important and may not have good knowledge. The model is built by setting the 'binary Split' to 'false'. The experiment built help for analyzing the comparison of the model with the other model (Binary). To build generalized decision tree model the setting of the Weka is changed to Binary split to "False" and the other setting of the Weka is remain the same to binary decision tree.

The model have four scenarios these are: Generalized decision tree with all attributes without prune, generalized decision tree with all attributes with prune, generalized decision tree with some selected attributes with prune, generalized decision tree with some selected attributes without prune. The prune and none prune done by setting the unpruned 'True' for nonprune one and unpruned to 'False' for prune one.

Experiment description

The experiment and analysis of classification models concerned on experimenting different scenarios of classification models using J48 decision tree algorithm. The scenarios were done by using the selected

Sample %	Accuracy
10	0.89
20	0.90
30	0.89
40	0.9
50	0.9

Table 22: Precisions of k- fold cross validation.

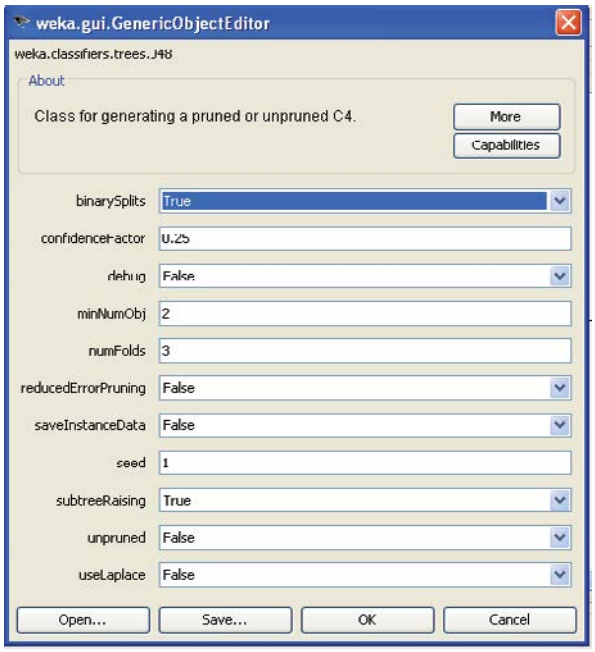

Figure 2: Binary decision tree classification.

confidence factor and setting the parameter into different values. The analysis of the model was concerned on selecting the best model from the experiments done. The selection of the model from the scenarios was based on comparing the result values of the analysis. The comparison parameters were number of leaves, tree size, time, correctly classified instances (CCI), incorrectly classified instances (ICI), True positive rate (TP), False Positive rate (FP), Precision, Recall and ROC Area. The comparison of the results ended with selecting the best model of the eight scenarios done. The next experiment is conducting PART rule generating algorithm with the same value of the previous model of J48 selected. The result values of the PART rule and the selected J48 model were compared to obtain the best rule of the experiment.

The experiment of the scenario includes: *Run information* which contains scheme, Relation, Instances, Attributes and Test mode. *Classifier model* contains number of leaves, size of the tree, time taken to build model. *Summary* with Correctly Classified Instances, Incorrectly Classified Instances and Total number of instances *Detailed Accuracy by Class includes* TP Rate, FP Rate, Precision, Recall, F-Measure ROC Area, Class and Confusion Matrix.

J48 Algorithms model building

The eight experiments for J48 decision tree scenarios are:

Scenario #1: Binary decision tree without pruning with all attributes

Scenario #2: General decision tree without pruning with all attributes

Scenario #3: Binary decision tree without pruning with reduced attributes

Scenario #4: General decision tree without pruning with reduced attributes

Scenario #5: Binary decision tree with pruning with all attributes

Scenario #6: General decision tree with pruning with all attributes

Scenario #7: Binary decision tree with pruning with reduced attributes

Scenario #8: General decision tree with pruning with reduced attributes

Figure 3 :In general the performance comparisons of 8 experiments were performed. As it has been described in summary of all experiments in Table 23, scenario 8, from all experiments became the best model as compared with all scenarios and the following rules were generated based on the outcome of the selected model. It was with confidence factor of 0.5 and a leaves of 760 with tree size of 916. The computation time for the model was 0.13. The CCI and ICI were 88.78% and 11.21% respectively. TP Rate, FP Rate, the precision, the recall, the ROC Area of the selected model were 0.89%, 0.45%, 0.88%, 0.89%, and 0.85% respectively. The researcher selected best rules that cover most of the data points in the study. After the rule extraction, the researcher turns back to domain experts to discuss up on the generated rules. Some of the rules generated by the model were presented in the following section:

Rules extracted from J48 decision tree

The decision tree of the model shows the rules of the model by traversing from the root node till the leaf. Below rules extracted from the tree and selected by the researcher as most important to predict CD4 status of the patient following ART are:

- When the patient is eligible to ART because of the CD4 count and if the patient started ART in the year 2003, the CD4 status of the patient for the next visit will be expected to Low.

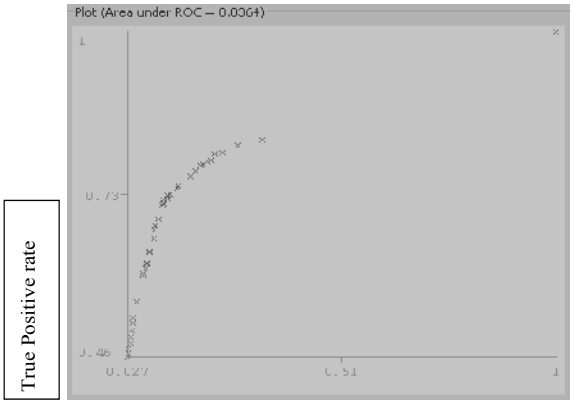

Figure 3

Experiment #	Type	Pruned	Leaves	Tree size	Time	CCI in %	ICI in %	TP Rate	FP Rate	Precision	Recall	ROC Area
1	Binary	No	649	1297	0.42	87.72	12.28	0.88	0.35	0.87	0.88	0.84
2	General	No	1797	2225	0.19	88.05	11.95	0.88	0.32	0.88	0.88	0.86
3	Binary	No	423	845	0.36	87.82	12.18	0.88	0.41	0.87	0.88	0.85
4	General	No	1264	1513	0.06	88.40	11.59	0.88	0.41	0.88	0.88	0.86
5	Binary	Yes	479	957	1.2	88.10	11.89	0.88	0.37	0.87	0.88	0.80
6	General	Yes	1093	1346	0.19	88.55	11.45	0.87	0.40	0.88	0.89	0.85
7	Binary	Yes	321	641	1	87.96	12.04	0.88	0.43	0.87	0.88	0.84
8	**General**	**Yes**	**760**	**916**	**0.13**	**88.78**	**11.21**	**0.89**	**0.45**	**0.88**	**0.89**	**0.85**

Table 23: General Summary of Experiments.

- When the patient is eligible to ART because of the CD4 count and if the patient started ART in the year 2004 that the marital status of the patient is one of Married, Divorced, Widow and Separate, the patient will develop a normal CD4 status.
- When the patient is eligible to ART because of the CD4 count and if the patient started ART in the year 2005, and the OA weight of the patient is between 3K.g to 52.5 K.g, the CD status of the patient for the next visit will be Low.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2005 and also the OA weight of the patient is between53K.g to 77 K.g and the OA WHO stage of the patient is stage 2 and if the patient is taking "1a40","1c" regimen then the patient will develop normal CD4 status.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2005 and the OAWHO stage of the patient is stage4 and the marital status is one of never married, Divorce, separate will probably develop low CD4 status. But if the patient is widow and current regimen the patient taking is "1C" will develop a normal CD4 status.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2005 and the OA weight of the patient is between 77.5K.g to 102K.g, then the patient will highly develop a low CD4 status.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2006 and the OA weight of the patient is 3K.g to 27.5K.g and the regimen taking is "1a30" the patient will develop low CD4 status for the next visit.
- If the eligible reason of the patient is CD4 and ART start year of the patient is 2006 and OA weight of the patient is between 28K.g to 52.5K.g, then the patient will probably develop low CD4 status for the next visit.
- When the eligible reason of the patient for the treatment is CD4 and if the patient started ART in the year 2007 and the OA weight is between 3K.g to 27.5K.g, then for almost all regimens the patient will develop low CD4 status.
- When the patient is eligible due to the reasons Clinical and TLC and current regimen that the patient is taking"1a30" and if the patient didn't take any past ARV treatment will develop a normal CD4 status for the next visit but for those patients who did take past ARV treatment will be expected low CD4 status for all ART start years.
- When the patient is eligible due to the reasons Clinical and TLC and the WHO stage of the patient is stage 1 and if the patient did not take past ARV treatment, then the patient will develop normal CD4 status but if the patient took treatment the CD4 status is expected to be Low.
- When the patient is eligible due to TI and the OA weight of the patient is in between 3K.g to 52.5K.g then the patient will develop a Low CD4 status for the next visit.
- When the patient is eligible because of clinical diagnosis only and if the patient is registered for ART in the year 2003 and 2004 and if the patient did not kook any family planning methods it is expected to develop a low CD4 status for the next visit.
- When the patient is eligible because of clinical diagnosis only and if the patient is registered for ART in the year 2006 and if the regimen the patient is taking is "1a30" and if the patient did not took past ARV treatment then a normal CD4 status will be expected.
- When the patient is eligible because of clinical diagnosis only and if the patient is registered for ART in the year 2006 and if the regimen the patient is taking is "1b30" and if the patient OA weight is in between 53K.g to 77K.g is expected to develop low CD4 status.
- When the patient is eligible because of clinical diagnosis only and the ART started year is 2008 and marital status is never married, the patient will develop a normal CD4 status for the next visit.
- When the patient is eligible to ART because of clinical diagnosis and the patient started ART in the year 2009 and the regimen that the patient is taking is "1a30" and "1c" then will develop a normal CD4 status and almost all the rest regimens develop low CD4 status.
- If the patient is eligible because of Clinical diagnosis and the ART start year is 2010, then the patient will develop a low CD4 status for the next visit.
- If the patient is eligible because of Clinical diagnosis and the ART start year is 2011 and the marital status of the patient is one of married, never married and widow ,then the patient will develop a normal CD4 status for the next visit.

Rule generating using PART rule induction algorithm

PART rule induction apply an iterative process of consisting first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set, to remain with no examples left to cover the process repeatedly iterate[7]. The PART rule induction algorithm experiment is done with reduced 10 attributes, and a confidence factor of 0.5. The result of the model of PART rule induction algorithm show that it has 259 rules generated, the time required for the computation is 0.44 second, CCI is 88.58% and ICI are 11.42%, TP Rate is 0.89% and FP Rate is 0.42%. The recall is 0.87% and ROC area is 0.89% it is enough to say the model is accurate. Some of the rules produced by PART rule induction algorithm and selected by the researcher are listed below:

- When the eligible reason of the patient is CD4 and the ART started year is 2006 and the OA Weight of the patient is between 28K.g to52.5K.g will be expected a low CD4 status.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2005 and the OA weight of the patient is between 77.5K.g to 102K.g, then the patient will highly develop a low CD4 status.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2006 and the OA weight of the patient is 3K.g to 27.5K.g and the regimen taking is "1a30" the patient will develop low CD4 status for the next visit.
- If the eligible reason of the patient is CD4 and ART start year of the patient is 2006 and OA weight of the patient is between 28K.g to 52.5K.g, then the patient will probably develop low CD4 status for the next visit.

- When the eligible reason of the patient for the treatment is CD4 and if the patient started ART in the year 2007 and the OA weight is between 3K.g to 27.5K.g, then for almost all regimens the patient will develop low CD4 status.
- When the patient is eligible due to the reasons Clinical and TLC and current regimen that the patient is taking "1a30" and if the patient didn't take any past ARV treatment will develop a normal CD4 status for the next visit but for those patients who did take past ARV treatment will be expected low CD4 status for all ART start years.
- When the patient is eligible due to the reasons Clinical and TLC and the WHO stage of the patient is stage 1 and if the patient did not take past ARV treatment, then the patient will develop normal CD4 status but if the patient took treatment the CD4 status is expected to be Low.
- When the patient is eligible due to TI and the OA weight of the patient is in between 3K.g to 52.5K.g then the patient will develop a Low CD4 status for the next visit.
- When the current regimen the patient is taking"1a30" and the ART start year of the patient is 2006 and if the patient did not took past ARV treatment, then low CD4 status is expected to develop.

To compare the best model from J48 decision tree algorithm and PART rule induction algorithm the following summary table displays the values of parameters used for comparison.

As it can be seen from the summary Table 24 of J48 and PART, precision of J48 is a little bit larger than PART. And also CCI of J48 is larger than PART's. Even though the precision and CCI of the J48 is larger to PART, the size of rules produced by PART is more manageable and readable than J48. Based on the comparison criteria J48 decision tree is better than part rule induction algorithm. The According to the result form the model created, Eligible Reason, ART start year, OA Weight, OA WHO stage, Marital Status, Current Regimen, Family Planning, Past ARV, Family Planning the most determining factors that can predict the CD4 status of patients those who are following ART therapy.

Conclusion

The ten top determining attributes are Eligible reason, ART status, ART start year, OA weight, OAWHO stage, Current regimen, Family planning, Functional status, Marital status and Past ARV consecutively. From the attributes ranked for predicting CD4 status of patients who are following ART, Eligible reason attribute is became the first determining attribute whereas educational level became the least determining attribute.

The best performance achieved by J48 decision tree algorithm is a generalized decision tree with pruning with reduced attributes. The model classifies instances correctly 88.79% and incorrectly classifies 11.21%. The weighted average precision of the model is 0.88 with recall of 0.89 and ROC area of 0.85. The model has 760 numbers of leaves and 916 size of tree. The time taken to build the model is 0.05 seconds. The analysis of this model shows that the model is quit efficient to predict CD4 status of patients following ART.

Performance measure	J48 decision tree	PART rule induction
Precision	0.88	0.86
CCI	88.78	87.45
Tree size/number of rules	916	267

Table 24: Summary of J48 and PART.

Classification done using J48 decision tree is the best model than PART rule induction algorithm. J48 algorithm is effective to predict the CD4 status of patients following ART. From the model built it is fund that attributes: Eligible reason, ART status, ART start year, OA weight, OAWHO stage, Current regimen, Family planning, Functional status, Marital status, Past ARV are the most determining factors for predicting CD4 status.

Recommendations

- Health facilities record keepers especially ART data clerks need to give attention in the recording of patient's information; this is because the records are important to predict the CD4 status of patients.
- Ministry of Health should work on a national standard data base format for recording ART data, because if the data bases are not in the same standard it is difficult for researchers to combine different health facilities record.
- Ministry of Health needs to give attention on the use of Data mining on the electronics Health records especially HIV/AIDS records.
- Since the patterns obtained are important to deploy and uses for decision support, the researcher leave open for the others investigators to work on the deployment of the model for Hospitals.
- Health facilities need to improve the accessibility of data for the researchers.

Acknowledgment

Our earnest gratitude goes Health and Medical sciences college, Addis Ababa University for proper review and approval of this paper. We would also like to extend our gratitude to data collectors for their patience to bring this meaningful information. Our special thanks also extended to Addis Ababa University and Bonga College of Teachers Education for financial support for this study.

References

1. Pathfinder International (2007) The Essentials of Antiretroviral Therapy for Health Care and Program Managers,Technical Guidance series number 5, USA.
2. ETC (AIDS Education and Training Center) (2011) aidsinfonet.Fact Sheets 124: 1.
3. Two Crows Corporation (2005) Introduction to Data mining and Knowledge Discovery. (3rd Edn) USA.
4. Gadelha A, Accacio N, Costa RL, Galhardo MC, Cotrim MR (2002) Morbidity and survival in advanced AIDS in RiodeJaneiro. Rev Inst Med Trop Sao Paulo 44: 179-186.
5. Mellors J, Munoz A, Giorgi JV, Margolick JB, Tassoni CJ (1997) Plasma viral load and CD4+lymphocytes As prognostic markers of HIV-1 infection. Ann Intern Med 126: 946-954.
6. Larose Daniel T (2005) Discovering knowledge in data: an introduction to data mining. John wiley and sons, USA.
7. Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques. (2nd Edn) Morgan Kaufmann, USA.
8. Han J, Kamber M (2006) Data Mining: Concepts and Techniques. (2nd Edn) Morgan Kaufmann, USA
9. Bramer M 2007 Principles of Data mining. Springer-verlag London limited, UK.