*Asian Journal of Computer and Information Systems (ISSN: 2321 – 5658)*
*Volume 04– Issue 03, June 2016*

# The Prediction of Paediatric HIV/AIDS Patient Survival: A Data Mining Approach

Idowu Peter Adebayo[1*], Agbelusi Olutola[2], Aladekomo T. A.[3]

[1]Department Computer Science and Engineering
Obafemi Awolowo University, Ile-Ife Wesley Hospital
Osun State, Nigeria

[2]Department of Computer Science,
Rufus Giwa Polytechnic,
Owo, Nigeria

[3]Department of Paediatric and Child Health
Obafemi Awolowo University, Ile-Ife Wesley Hospital
Osun State, Nigeria

[*]*Corresponding author's email: paidowu1 [AT] yahoo.com*

**ABSTRACT---** ***This research requires the development of predictive model for determining the survival of Paediatric HIV/AIDS patients who are receiving antiretroviral drugs in the South-western Nigeria. The WEKA software was used in developing the predictive model using Naïve Bayes' Classifier. Naïve Bayes' Classifier was used to predict the length of survival of HIV/AIDS patients based on variables like CD4 count, viral load, opportunistic infection and nutritional status. The result shows that Naïve Bayes' Classification can predict the survival of paediatrics HIV/AIDS patient with an accuracy of 60% to 100% based on selected dependent variables.***

## 1. INTRODUCTION

Epidemic diseases have highly destructive effects around the world and these diseases have affected both developed and developing nations. Disease epidemics are common in developing nations especially in Sub Saharan Africa in which HIV/AIDS is the most serious of all (Idowu, 2009). The human immunodeficiency virus (HIV) is a type of virus called a retrovirus, which infects humans when it comes in contact with tissues such as those that line the vagina, anal area, mouth, eyes or through a break in the skin (Eric et al, 2012). While acquired immunodeficiency syndrome is a disease caused by HIV. It alters the immune system making people much more vulnerable to infections and diseases (Med News, 2012). As at present there is no cure for HIV but the disease is managed with antiretroviral therapy (ART) or with optimal combination of ART which is known as Highly Active Antiretroviral therapy (HAART). ART means treating retroviral infections like HIV with drugs. The drugs do not kill the virus, however they slow down the growth of the virus (Rosma et al, 2012). HAART refers to the use of combinations of various antiretroviral drugs with different mechanisms of action to treat HIV. The epidemic of HIV/AIDS affects two classes of people: the paediatrics and the Non-paediatrics individual. The non-paediatric patient are patient that are above 15 years of age while the Paediatric patients which form the main target of this research work are patients whose ages are less than 15 years or a length-based weight of 36 kg or less. Patients who are known to be less than 15 years of age but whose weight exceeds 36 kg may still be considered paediatric patients (San Mateo, 2012). If the virus is diagnosed, earlier use of medicines is recommended before the CD4 cell count drops to a dangerous level. The CD4 count measures the number of CD4 cells in a sample of your blood drawn by a needle from a vein in your arm. CD4 cells are responsible for signalling other immune system cells to fight an infection in the body. They are also the prime target of HIV, which can cause the number of these cells to decrease over time. CD4 cells means that the immune system will no longer function like it is supposed to. The "c" and the "d" in CD4 stand for "cluster of differentiation," and refer to the cluster of proteins that make up a cell surface receptor.

## 2. PROBLEM DEFINITION

Despite the fact that HIV/AIDS is one of the incurable diseases that is militating against a number of children most especially in low income country like Nigeria, there is no means of forecasting the survival of pediatrics HIV/AIDS patient in Nigeria and this has brought about the need to develop an effective and efficient model that can be used to predict survival, hence this research work.

The Specific Objectives of this paper is to identify survival variables for HIV/AIDS pediatric patients in the South Western Nigeria and formulate a predictive model using supervised learning technique (naïve bayes'classifier) data mining based on variables identified. Also the prediction performance of naïve bayes'classifier was also tested in the paper**.**

## 3. REVIEW OF RELATED LITERATURE

Many Researchers had worked on the prediction of HIV/AIDS survival using different types of variables like CD4 count, CD8 count and viral load but in this paper, opportunistic infection like tuberculosis, cancer and nutritional status of infected patient was considered. Some of the researcher and the result of their finding are:

Rosma (2012) measure the predictive ability of two models using Data mining predictive techniques (regression and fuzzy neural network techniques) to determine the length of survival of AIDS patient based on their CD4, CD8 and viral load counts, the researcher compared the performance of both techniques to predict the survival of HIV/AIDS and discovered that both FuReA and fuzzy neural network models were able to predict the survival of HIV/AIDS with an accuracy of 60% to 100% based on selected dependent variables.

Moore D. (2006) used CD4 count to determine the prognostic value of baseline in terms of patient survival. Kaplan-Meier methods and Cox proportional hazards regression were used to model the effect of baseline CD4 strata, CD4 percentage strata and other prognostic variables on survival. He discovered that only CD4 can be used to determine the survival of paediatric HIV/AIDS patient.

Matthias & Margaret (2010) used Survival models to construct a prognostic model based on five clinical predictors and discovered that CD4 cell count , low body weight and severe anaemia can be used as prognostic factors to predict survival of HIV-1 infection in paediatric HIV/AIDS patient.

Puthanakit T. (2012) also used CD4 counts and early initiation of ART to predict survival of children between aged 1 and 12 years with CD4 counts between 15% and 24% and the result of his prediction confirm that early initiation of ART before 1 year of age and CD4 count are factors that can significantly improve survival in pediatrics HIV/AIDS patients.

## 4. RESEARCH METHODOLOGY

Extensive study on related existing body of knowledge on HIV/AIDS, Prediction of Diseases, Data Mining and prediction of HIV/AIDS were carried out. The study adopted the descriptive and exploratory designs that allow the collection of data from a sample population. Personal interview (both structured and unstructured) was also used to collect paediatric HIV/AIDS data from two tertiary health institutions in South Western Nigeria. (Federal medical Centre in Owo, Ondo State and Obafemi Awolowo Teaching Hospital, Ile- Ife in Osun state) in order to identify survival variables for pediatrics HIV/AIDS patients with a view to formulate survival predictive model using data mining approach.

The predictive model were developed using supervised learning techniques (naive bayes classifier in WEKA's environment. The performance of this model was tested and we found out that it can predict the survival of paediatrics HIV/AIDS patient with an accuracy of 60% to 100% based on selected dependent variables.

### Data Mining Technique

Data mining is a process of selecting, exploring and modelling a set of data in order to discover unknown patterns or relationships which provides clear and useful results to the data analyst. The goal of predictive data mining is to derive models that can use patient specific information to predict the outcome of interest. Predictive data mining methods can be used for medical diagnosis, prognosis, treatment planning and also for general screening purposes. Criteria for good predictive data mining technique include: Good performance, transparency, ability to deal with missing data and noise (outliers), ability to work with small sample and ability to explain the decisions being made (Kononenko, 2001). In this research a predictive data mining technique based on naïve bayes classifier was used in the prediction of HIV/AIDS survival.

### Supervised Learning

Machine learning is generally classified as Supervised and Unsupervised learning. Supervised learning methods allow for the classification of data into a given/known output class. Thus, all supervised learning methods can be applied to classification/prediction and forecasting/regression problems. Thus, allowing new data to be classified based on the results of a training set. Unsupervised learning methods are basically used to understanding classes or clusters among data. The prediction model being developed in the research uses supervised learning technique in performing the required classification problem.

Every classification problem is usually a two-step process involving a model construction followed by the model usage; this can then be followed by an estimation of the model's accuracy.

i. **Model construction:** this involves the description of a set of pre-determined classes of data. Each sample/Tuple is assumed to belong to a predefined class, as determined by a class label attribute. The set of tuples used for the model construction is called the training set. This model is represented based on the classification algorithm/classifier used.

ii. **Model usage:** this involves the process of classifying future or unknown objects using the developed model. The known label of the test sample is compared with the classified result from the model in order to determine the accuracy of the model. The accuracy rate is the percentage of test set samples that are correctly classified by the model. The test set is independent of the training set so as to avoid over-fitting. If the accuracy is acceptable, then the model can be used to classify new data. But, if the test set is used to select models then it is called a validation (test) set.

**Data partitioning methods (training and testing data)**

Before the development of a prediction model; the dataset were collected and split into two parts: training and test (validation) data. The training data were used by the classifier to develop the prediction model while the test data were used to confirm the accuracy of the model developed by the classifier using certain measures of accuracy depending on the training method used. The type of training method used in splitting/partitioning data goes a long way in determining the accuracy of the model. For the purpose of this study, the k-fold partitioning method was used in partitioning the data that was used in developing the prediction model. Thus, the 10-fold cross validation techniques was used for the research in partitioning the datasets used in training and testing.

The 10-fold cross validation technique follows a detailed number of steps:

i. The whole dataset is first divided into 10 parts;
ii. 9 parts (90%) are used for training and 1 part (10%) for testing;
iii. The process is repeated 10 times by keeping 1 part for training from the first part to the last part.
iv. The predicted results of the 10 test parts are used to evaluate the performance of the prediction model developed for the study

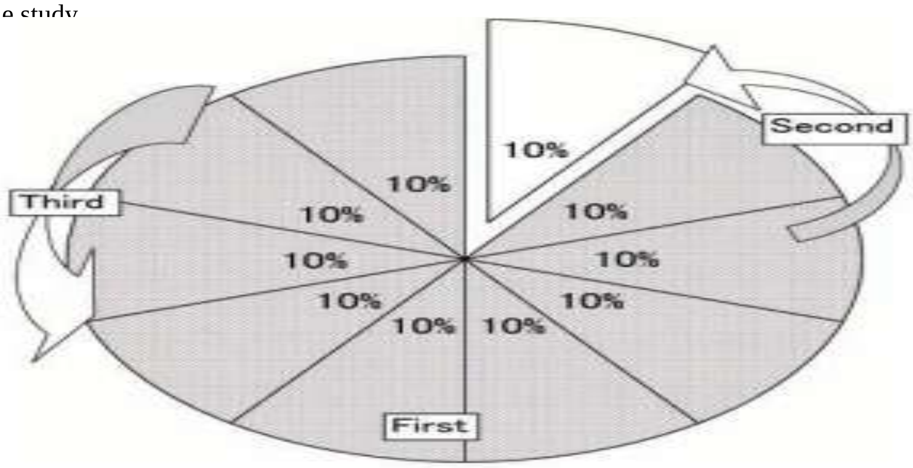


**Fig (1) 10 Folds Cross Validation Method**

## 5. PREDICTION MODEL

**Naïve Bayes' Classification model**

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This Classification is named after Thomas Bayes (1702-1761) who proposed the Bayes Theorem.

Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

The Bayes' Theorem:

$$P(Class|Data) = \frac{P(Data|Class) * P(Class)}{P(Data)}$$

Where

P(Class): Prior probability of class
P(Data): Prior probability of training data
P(Class|Data) : Probability of Class given the data
P(Data|Class) : Probability of Data given the class.

The naïve bayes' classification is based on the Bayesian theorem, it is particularly suited when the dimensionality of the inputs is high. Parameter estimation for naive Bayes models uses the method of maximum likelihood. In spite over-simplified assumptions, it often performs better in many complex real-world situations.
Mathematically, the Naïve Bayes' Classification is expressed as follows:

$$\mathbf{P(X|Ci) = \prod_{k=1}^{n} P(X_k|Ci)}$$

$$\mathbf{P(X|C_i) = P(x_1|Ci)* P(x_2|Ci)* P(x_3|Ci)*……….*P(x_k|Ci)}$$

Limitations of the naïve bayes' prediction model

i. Naïve bayes' classification efficiency is usually limited by large datasets; and
ii. Due to its simplistic nature, solving of more complex classification problems is not possible with naive Bayesian classifier.

## 6. RESULTS AND DISCUSSIONS

Simulation program was used in performing the assessment of the performance of the prediction model developed for determining the survival of paediatric HIV/AIDS patients. Prediction model were developed using the naïve bayes' and the performance of this model were tested using various classifier evaluation parameters. Parameters used in testing the prediction model is shown in fig (3) below

### The WEKA Simulation Environment

WEKA® is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from a code (as a jar file in the library of a Netbeans® java project. WEKA contains tools for data-processing, classification, regression, clustering, association rules and visualization.

Some of the main features of WEKA contain:

i. 49 data preprocessing tools;
ii. 76 classification/regression algorithms;
iii. 8 clustering algorithms;
iv. 15 attribute/subset evaluators and 10 search algorithms for feature selection;
v. 3 algorithms for finding association rules; and
vi. 3 graphical user interfaces.
    - The Explorer (exploratory data analysis - preprocessing, attribute selection, learning, visualization);
    - The Experimenter (experimental environment – testing and evaluating machine learning algorithms); and
    - The Knowledge Flow (new process model inspired interface – visual design of KDD process)

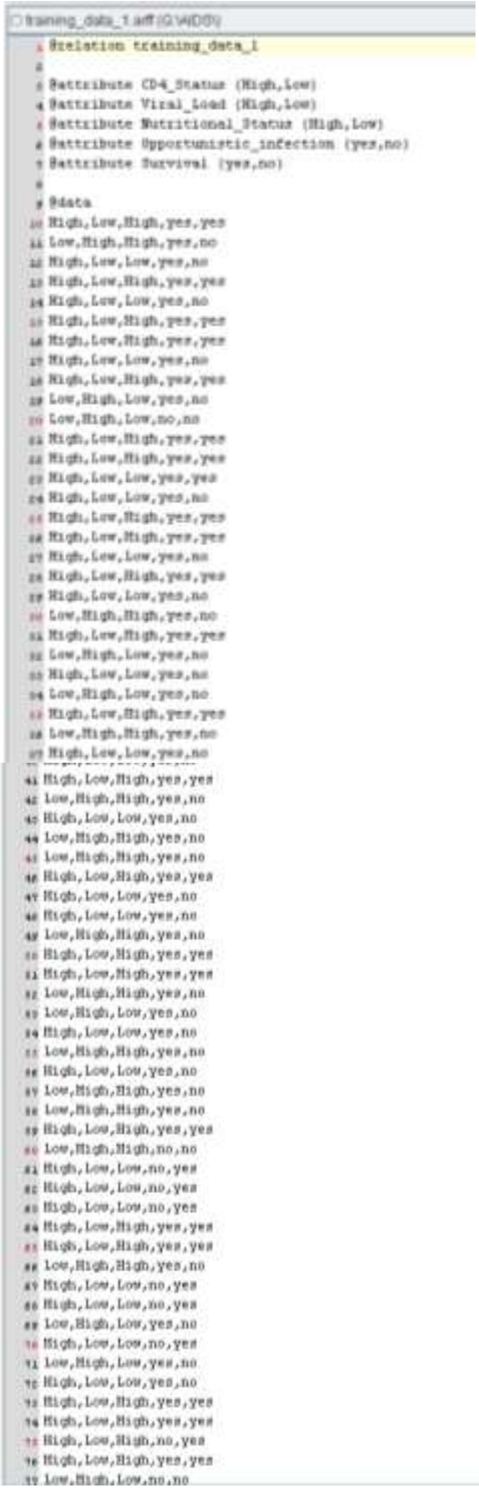
### Setting up the Simulation Environment

As stated earlier, the simulation environment of the prediction model uses Naïve Bayes' classifier is the WEKA 3.7 software environment. Before implementing the classifiers, the system was adjusted to perform a 10-fold cross validation training procedure on the 137 dataset provided by the Federal Medical Centre and OAUTHC (2). The default critical region for WEKA is 0.95 (5% level of significance).

The results of the prediction was then presented on a confusion matrix from which evaluation parameters like: recall, precision, accuracy, F-measure, TP and FP rates and area under ROC curves. WEKA also calculates the various errors experienced by the classifier used. After developing the prediction model; t-test is applied on the prediction model to determine the performance of the classifiers using a 10-fold cross validation training procedure on a 10 runs; this yielded 100 result set for each classifier (10 runs with 10-fold cross validation) alongside with the evaluation results.

### Prediction Modelling Simulation Results and Discussions

The prediction modelling results of the classifiers on the 137 dataset provided were presented and discussions provided. The classifier also was evaluated using the parameters stated in (fig3) to determine the performance of the classifier.



```
@relation training_data_1

@attribute CD4_Status (High,Low)
@attribute Viral_Load (High,Low)
@attribute Nutritional_Status (High,Low)
@attribute Opportunistic_infection (yes,no)
@attribute Survival (yes,no)

@data
High,Low,High,yes,yes
Low,High,High,yes,no
High,Low,Low,yes,no
High,Low,High,yes,yes
High,Low,Low,yes,no
High,Low,High,yes,yes
High,Low,High,yes,yes
High,Low,Low,yes,no
High,Low,High,yes,yes
Low,High,Low,yes,no
Low,High,Low,no,no
High,Low,High,yes,yes
High,Low,High,yes,yes
High,Low,Low,yes,yes
High,Low,Low,yes,no
High,Low,High,yes,yes
High,Low,High,yes,yes
High,Low,Low,yes,no
High,Low,High,yes,yes
High,Low,Low,yes,no
Low,High,High,yes,no
High,Low,High,yes,yes
Low,High,Low,yes,no
High,Low,Low,yes,no
Low,High,Low,yes,no
High,Low,High,yes,yes
Low,High,High,yes,no
High,Low,Low,yes,no
High,Low,High,yes,yes
Low,High,High,yes,no
High,Low,Low,yes,no
Low,High,High,yes,no
Low,High,High,yes,no
High,Low,High,yes,yes
High,Low,Low,yes,no
High,Low,Low,yes,no
Low,High,High,yes,no
High,Low,High,yes,yes
High,Low,High,yes,yes
Low,High,High,yes,no
Low,High,Low,yes,no
High,Low,Low,yes,no
Low,High,High,yes,no
High,Low,Low,yes,no
Low,High,High,yes,no
Low,High,High,yes,no
High,Low,High,yes,yes
Low,High,High,no,no
High,Low,Low,no,yes
High,Low,Low,no,yes
High,Low,Low,no,yes
High,Low,High,yes,yes
High,Low,High,yes,yes
Low,High,High,yes,no
High,Low,Low,no,yes
High,Low,Low,no,yes
Low,High,Low,yes,no
High,Low,Low,no,yes
Low,High,Low,yes,no
High,Low,Low,yes,no
High,Low,High,yes,yes
High,Low,High,yes,yes
High,Low,High,no,yes
High,Low,High,yes,yes
Low,High,Low,no,no
```

**Fig (2) Section of the pre-processed dataset used in training**

| Parameters | Description |
|---|---|
| Training Data | The dataset used in developing the prediction model by the classifier algorithm. |
| 10-fold cross validation | |
| Validation Data | The dataset used in validating the efficiency of the model |
| Testing Data | Future dataset used for testing the effectiveness of the model (it can also be a fraction of the training data). |
| True Positives/Negatives | Correctly classified instances (Yes for Yes or No for No) |
| False Positives/Negatives | Incorrectly Classified instances (Yes for No or No for Yes) |
| Recall/Sensitivity/TP rate | Portion of positives predicted that is positive. |
| FP rate | Portion of negatives predicted that is negative |
| Precision | Portion of the predicted positive that is positive |
| Accuracy | Portion of total prediction that is correct |
| F-measure/score | Precision $^*$ Recall |
| Area under ROC | A measure of the effectiveness of the classifier model chosen with an interval; 0<Area under ROC curve<1. Effective classifiers must have an area under ROC curve > 0.5. |

**Fig (3) Classifier Evaluation Parameters**

**Naïve Bayes' classifier results and discussions**

The classification of each training data was performed via the implementation of the naïve bayes' classification algorithm which calculates the probability and manipulates them into the necessary results. A typical demonstration of how this is achieved is shown as follows:

Consider the training data provided and we intend to classify the following data, X:

X = (CD4 count=High, Viral Load=Low, Nutritional Status=High, Opportunistic infection=yes)

Will any patient with the Tuple X survive AIDS?

Data X = $(x_1, x_2, x_3, x_4)$ = (CD4 count, viral load, nutritional status, opportunistic infection)

a. Determine the output classes
   Survival, C = (C1, C2) = (Yes, No)

b. Determine the probability of each class
   P(C1) = P(Survival = yes) = 59/137 = 0.43066

   P(C2) = P(Survival = no) = 78/137 = 0.56934

c. Determine the probability of each class given the label of the variable
   P(x1|C1) = P(CD4 count = High|Survival= yes) = 53/59 = 0.89831
   P(x1|C2) = P(CD4 count = High|Survival = no) = 41/78 = 0.52564

   P(x2|C1) = P(viral load = Low|Survival= yes) = 53/59 = 0.89831
   P(x2|C2) = P(viral load = Low|Survival = no) = 41/78 = 0.52564

   P(x3|C1) = P(nutritional status = High|Survival= yes) = 43/59 = 0.72881
   P(x3|C2) = P(nutritional status = High|Survival = no) = 19/78 = 0.24359

   P(x4|C1) = P(opportunistic infection = yes|Survival= yes) = 38/59 = 0.64407
   P(x4|C2) = P(opportunistic infection = yes|Survival = no) = 64/78 = 0.82051

d. Determine the probability of each class given a tuple, X

   P(X|Survival = yes) = P(CD4 count = High|Survival= yes)* P(viral load = Low|Survival= yes)*P(nutritional status = High|Survival= yes)*P(opportunistic infection = yes|Survival= yes) = 0.89831*0.89831*0.72881*0.64407 = 0.37879

Similarly,

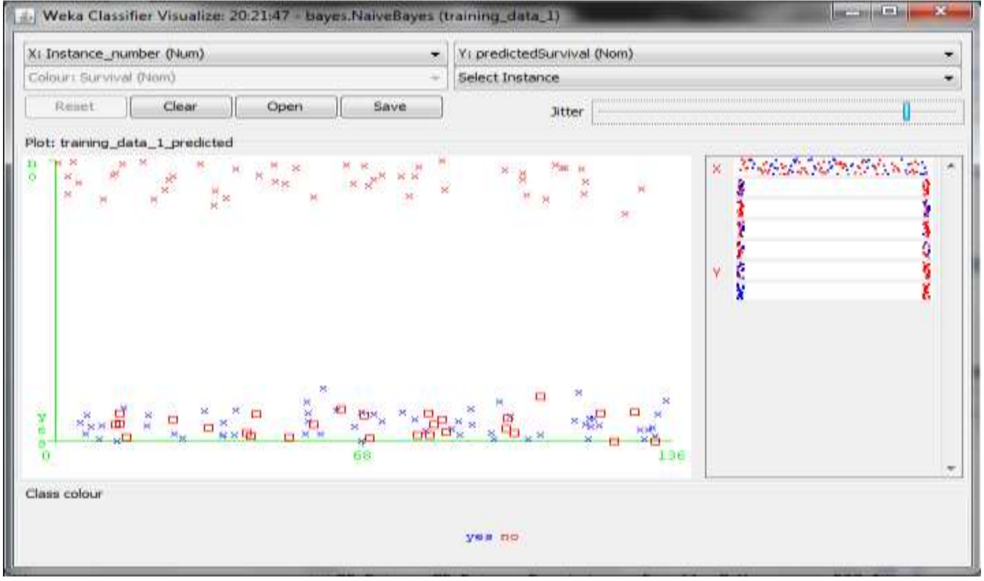
P(X|Survival = no) = 0.52564*0.52564*0.24359*0.82051 = 0.05522

e. Find the class Ci that maximizes P(X|Ci)*P(Ci)
P(X|Survival=yes)*P(Survival=yes) = 0.37879*0.43066 = 0.16313
P(X|Survival=no)*P(Survival=no) = 0.05522*0.56934 = 0.03144

Survival = max(P(yes), P(no)) = max(0.16313, 0.03144)
Therefore, the survival of HIV/AIDS for a paediatric patient with data X is YES.

After using the Naïve Bayes' Classifier to train the data and validate the model developed using 10-fold cross validation, it was discovered that the naïve bayes' prediction model made 109 correct and 28 incorrect classifications of the output of the survival. Figure (4) below shows the graph of the results of the classification made by the naive bayes' prediction model; blue crosses identify YES and red crosses identify NO while the boxes show misclassifications.

From the results of the analysis made on the dataset using naïve bayes' classification (fig 2) in developing the predictive model of AIDS survival for paediatric patients; the following were discovered: that out of 59 dataset which had survival of YES, all were classified as YES and of 78 dataset which had a survival of NO, 28 were classified as YES and so classified as NO (see Figure (5).



Weka Classifier Visualize: 20:21:47 - bayes.NaiveBayes (training_data_1)

X: Instance_number (Num)

Y: predictedSurvival (Nom)

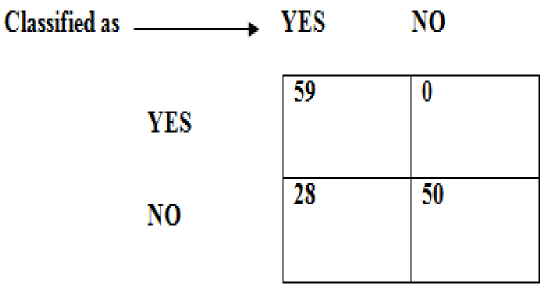
Colour: Survival (Nom)

Select Instance

Plot: training_data_1_predicted

Class colour

Fig (4) Graph showing the results of the naïve bayes' classification



Classified as ⟶ YES NO

| | YES | NO |
|---|---|---|
| YES | 59 | 0 |
| NO | 28 | 50 |

Fig 5 Confusion matrix of the results of Naïve Bayes' classification

These results were used to determine the precision and recall of the predictive model developed using naïve bayes' classifier for each output classes (YES and NO); the precision is a fraction of retrieved instances that are relevant while recall is the fraction of relevant instances that are retrieved. Therefore, the naïve bayes' classifier was able to predict the survival of YES more accurately than the survival of NO and about 68% of the survival of NO was correctly classified.

| Key Fold | Number of training instances | Number of testing instances | Number correct | Number incorrect | Number unclassified |
|---|---|---|---|---|---|
| 1 | 123 | 14 | 14 | 0 | 0 |
| 2 | 123 | 14 | 9 | 5 | 0 |
| 3 | 123 | 14 | 14 | 0 | 0 |
| 4 | 123 | 14 | 9 | 5 | 0 |
| 5 | 123 | 14 | 11 | 3 | 0 |
| 6 | 123 | 14 | 12 | 2 | 0 |
| 7 | 123 | 14 | 9 | 5 | 0 |
| 8 | 124 | 13 | 9 | 4 | 0 |
| 9 | 124 | 13 | 13 | 0 | 0 |
| 10 | 124 | 13 | 9 | 4 | 0 |

**Fig 6 Summary of the analysis of naïve bayes' model**

## 7. REFERENCES

1. Eric et al, 2012: Human Immunodeficiency Virus Management
2. P.A Idowu, N. Okonronkwo and E.R Adagunodo(2009): Spatial Predictive Models for Malaria in Nigeria, journal of Health Informatics in Developing Countries, 3(3): pp 9-17. New Zealand.
3. Kononenko (2001) Machine learning for medical diagnosis: History, state of the art and perspective, Invited paper, Artificial Intelligence in Medicine - ISSN 0933-3657, vol. 23, no. 1, pp. 89-109.
4. Matthias & Margaret (2010): New models predict short-term survival of HIV patients starting antiretroviral therapy in sub-Saharan Africa.
5. Med News 2012: The management of Human Immunodeficiency Virus
6. Moore et al (2006) "CD4 percentage is an independent predictor of survival in patients starting antiretroviral therapy with absolute CD4 cell counts between 200 and 350 cells/microL" in the Journal of HIV Medical Vol 7, pp 383-8.
7. Puthanakit et al (2012) "Early versus deferred antiretroviral therapy for children older than 1 year infected with HIV (PREDICT): a multicentre, randomised, open-label trial" in Lancet infectious Diseases Journal; 12(12):933-41. doi: 10.1016/S1473- 3099(12)70242-6.
8. Rosma et al (2012) "The prediction of AIDS survival: A Data Mining Approach" in proceeding of the WSEAS international Conference on Multivariate Analysis and its application in Science and Engineering.
9. Sam Mateo (2012):