International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 5– No.7,May 2013 – www.ijais.org

Data Mining Techniques for Predicting Immunize-able Diseases: Nigeria as a Case Study

[1]Adebayo Peter Idowu
Dept. of Comp. Sc. & Engr.
Obafemi Awolowo University, Nigeria

[2]Bernard Ijesunor Akhigbe
Dept. of Comp. Sc. & Engr.
Obafemi Awolowo University, Nigeria

[3]Olajide Olusegun Adeosun
Comp. Sc. & Engr. Dept.,
Ladoke Akintola University of Technology, Ogbomoso, Nigeria

[4]Aderonke Anthonia Kayode
Dept. of Comp. Sc. & Engr.
Obafemi Awolowo University, Nigeria

[5]Adekemi Faidat Osungbade
Dept. of Comp. Sc. & Engr.
Obafemi Awolowo University, Nigeria

ABSTRACT
Disease rates vary between different locations particularly in the rural areas. While a database of diseases occurrence could be easily found, studies have been limited to descriptive statistical analysis, and are mostly restricted to diseases affecting adults. This paper therefore presents a Mathematical Model (MM) for predicting immunize-able diseases that affect children between ages 0 - 5 years. The model was adapted and deployed for use in six (6) selected localized areas within Osun State in Nigeria. Using the MATLAB's ANN toolbox, the Statistics toolbox for classification and regression, and the Naïve Bayesian classifier the MM was developed. The MM is robust in that it takes advantage of three (3) data mining techniques: ANN, Decision Tree Algorithm and Naïve Bayes Classifier. These data mining techniques provided the means by which hidden information were discovered for detecting trends within databases, and thus facilitate the prediction of future disease occurrence in the tested locations. Results obtained showed that diseases have peak periods depending on their epidemicity, hence the need to adequately administer immunization to the right places at the right time. Therefore, this paper argues that using this model would enhance the effectiveness of routine immunization in Nigeria.

General Terms
Data Mining, Public and Environmental Health, Information System

Keywords
Data mining techniques**,** Immunize-able diseases, MATLAB, Databases, Decision tree algorithm and Predictive model.

1. INTRODUCTION
Diseases are the abnormal medical conditions of organisms that impair bodily functions, and are associated with recognizable symptoms and signs. The causes of diseases may be related to external factors such as infectious disease or autoimmune disease in the case of internal dysfunctions [1]. Other related definitions of diseases, their mode of transmission and the type of diseases as classified by Global Health Council as examples of Immunize-able Diseases (IDs) are found in [2], [3] and [4]. Furthermore, IDs are diseases that can be prevented by inoculation or vaccination at the early age of an individual's life, and are also referred to as vaccine-preventable diseases [5]. When immunization is given between ages 0-5 years, a lasting immunity to diseases is thus provided [1]. In Nigeria, the predominant diseases which kill children that are considered in this study include: Yellow fever, Measles, Poliomyelitis, and Hepatitis B which are viral diseases; and Tuberculosis and Whooping cough which are contagious bacteria diseases.

In 2002 alone, over 25.5 million children worldwide died (before their fifth birthday), of vaccine-preventable diseases. In 2003 also, over 17 million children in Africa died of Immunize-able diseases. According to [6], in Nigeria alone (the country of context for this research), from 1998 to 2007 immunize-able diseases claimed over 22 million children. One wonders what the mortality rate is now, almost six (6) years after. According to the same report, the effect of high infant mortality rate caused by Immunize-able diseases is that it decreases life expectancy. Going by the figures quoted above, 35% of children born in Nigeria die before age five. In the same report, ten (10) in every 200 polio infections leads to irreversible paralysis (usually in the legs); thus leading to a lifelong deformation. For yellow fever; it is known to cause significant organ damage in two out of 20 children in the toxic phase [6].

However, existing vaccines against all immunize-able diseases have the potential to avert an estimated 65 million child deaths annually. But this will only happen if children receive all doses early enough in life to avoid exposure to these diseases. Between 1999 and 2002, for example, deaths due to measles infection were reduced by 30% globally through improved routine immunization and supplemental measles campaigns. In Nigeria, over six million deaths are prevented in children each year as a result of immunizations alone. But, there is so much more work to do before immunization would realize its full potentials for protecting children against diseases for which effective vaccines now exist [7]. Routine immunization continues to remain a serious concern for the Government of any country including Nigeria and development partners like WHO.

Although, most IDs are preventable with immunization, they are not curable. It is therefore worthwhile to develop a system that could be used to examine the trends of diseases and thus predict if they are immunize-able or not. The availability of a dataset makes the use of Data Mining (DM) and its techniques appropriate in this research and invaluable in determining relevant and interpretable patterns of disease trends. DM has been defined by several researchers e.g. [8]; [9]; and [10]. But, simply, DM is "the process of finding hidden information in a database" [10]. Furthermore, scores of literature exists where detailed discourse of DM and its techniques could be found, but some examples are e.g. [11]; [12]; [13]; [14]; [9]; and [10]; [15]. In this paper, three classification methods of

the DM based techniques are applied to propose a predictive model for immunize-able diseases. The methods are: The statistical-based Naive Bayes technique (NBT); the Decision Tree (DT) based algorithm; and Artificial Neural Network (ANN) based algorithm. Thus, the purpose of this paper is to present a triangulation of these techniques. As a result, this paper reports a Mathematical Model with a savvy of these techniques for IDs prediction using Nigeria as a case study. The paper advances thus: Section 2.0 contains a brief literature review; section 3.0 an overview of related Data mining techniques; section 4.0 a discussion of the paper's methodology; and section 5.0 contains the description of the components of the proposed model. In addition, in section 6.0 the result is presented, and in section 7.0 the study's conclusion is given.

2. BRIEF LITERATURE REVIEW

Data mining methods have been used severally to predict important clinical outcomes [16], [17]. Data mining (DM) techniques has been largely applied in clinical medicine. This has also been reviewed e.g. [18], [19]; [20]; and [21]. In literature the process of DM in the field of medicine and other domains, such as business, marketing and the economy has also been differentiated and elucidated e.g. [21] and [22]. The use of more than one DM techniques for prediction especially in the health related domain abounds, a review of few of these are as follows. [23] used a combination of Temporal Abstraction with (DT and Mining support a priori algorithm) data mining techniques. This approach was meant to analyze dialysis patients' biochemical data in order to develop a decision support system. Results obtained were helpful in predicting hemodialysis patients and suggesting immediate treatments to avoid hospitalization. In [24], the use of Naïve Bayes and WAC (weighted associative classifier) was reported. A prototype using these data mining techniques was obtained. The likelihood of patients getting a heart disease was consequently easily predicted using patients' medical profile. In both cases a triangulation of method was applied but restricted to two, but not for predicting immunize-able disease occurrence.

In health management, DM technique has also been applied. For example, [25] reported the use of DM methods to infer health care indices of individuals. As a result, they formulated the widely used Charlson Index using data mining techniques. [26] proposed a predictive cause of death model. The utility of the model helped in the generation of information that could accelerate progress towards MDG. The prediction of the survivability of various illnesses has been investigated. For example, [27], [28], [29] investigated survivability for breast cancer; [30] did the same for cervical cancer; and [31] for Crohn's Disease to name a few. In a related study [32], the potential use of classification based data mining techniques to mine massive volume of Immunization data was examined. However, the model was for predictive measles in children, using two classification DM techniques. Similarly, in [33], both the DT and NBC were used to design a predictive model to help predict the occurrence of measles outbreaks. The main difference between [32], and [34] research is the interchange of DM classification technique. Moreover, to the best of our knowledge, no work has been done in applying data mining techniques to develop predictive models for immunize-able diseases that affect children of 0-5 years old. Based on these literature reviews, this study integrates three DM classification techniques as earlier stated. As a result, in this paper were describe a predictive model that is triangularly put together using DM techniques such as NBT, DT, and ANN for predicting immunize-able diseases in infants.

3. MATERIALS AND METHOD

In this section we describe a triangulation of DM (see Section 3.1) and the methodology applied to achieve the goal of building the predictive model that is proposed in this paper (see Section 3.2).

3.1 Triangulation of Data Mining Techniques

A major role of the Expanded Program on Immunization (EPI) is to significantly contribute to the Millennium Development Goal (MDG). The goal is to half child mortality by 2015. Consequently, infant mortality would be reduced, and invariably life expectancy will increase [4]. Therefore, this paper argues that the art of predictiveness will have significant influence if introduced, to fostering the programme of immunization. In order to demonstrate this, we harness the potentials of ANN, NBT, and proposed a model that is predictive. The DT algorithm was formulated from decision trees. Its adoption and use in this paper is premised on the fact that DTs are reliable tools for predictive modelling. Its rule-based nature made exploiting its explicit set of "IF-THEN" inference rules (rather than abstract mathematical equations) possible. This advantage was an asset in interpreting relevant results.

The second technique in this triangularly approach is the Naive Bayes' Classifier (NBC). Its strong assumption as expressed by its theorem, which is the basis why its model is often described as an independent feature (since the NBC assumes mutually independent events) was the reason for its adaption and use in this study. Thus, the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [15]. So, from underlying evidence (as provided by the dataset used in this research) useful predictions are easily made. The Naive Bayes' theorem presents a probability model for prediction, and details of this theorem are evident in literature e.g. [15], [34]. The equation in (i) is a summary of the Naïve Bayes' algorithm, which is less complex and faster in modelling domains that are characterized by inherent uncertainty [35], hence its adoption in this research. The Naïve Bayes' Theorem states that:

$$P(A \mid B) = [P(B \mid A) * P(A)] / P(B) \quad \text{(i)}$$

Similarly, in practical terms the ANN offers a non-linear statistical data modelling tool that is usable for modelling complex relationships. These relationships are needed for finding patterns in data and between inputs and outputs. The back propagation network of the ANN is one of the most frequently used techniques in neural network learning algorithm. This algorithm is a supervised learning method where the training data contains examples of inputs together with the corresponding outputs. What the neural network does is to learn to infer the relationship between inputs and outputs. As recommended, the training data was taken from historical records [12]. Moreover, a proper training of the network, allows it to easily model unknown functions that relates the input variables to the output variables. This is subsequently used to make predictions where the output is not known [35]. Detailed description of ANN is not sparse in literature e.g. [35], [36], [8], [37].

These techniques (ANN, DT, and NBC) were adapted and used since they combine and work well with each other.

Despite the sophistication of these schemes in modelling extremely complex functions, they still have some limitations. Thus, caution is needed so as not to restrict DT's interpretation to a particular rule representation, else the predictive power of the model will be significantly restrict. For ANN, the cost of time and money is a major challenge, since a large diversity of training for real-world operation is required. In using the Naive Bayes' Classifier (NBC), there is need to painstakingly check the constraints on which the theorem's assumptions thrives, else false postulations would be upheld instead [36], [37].

3.2 Methodology

To satisfy the purpose of this paper, the methodology used to build the predictive model is the CRISP-DM. It is interpreted as the cross industry standard process for data mining methodology. Its choice for this paper is premised on first, its tools independent nature and secondly on its alloy of aspects of the academic and industrial model. It is one of the most widely used data mining methodology for knowledge discovery, and it consists of six phases: The business understanding, data understanding, data preparation, modeling, evaluation, and deployment phases. A detail of this methodology is available in e.g. [24]; [34]. Based on this methodology, the required tasks and methods needed for study were identified in order to predict the occurrence of six (6) immunize-able diseases, specifically in a randomly selected part of Nigeria. As a result, this study collected and analyzed data from six (6) selected immunize-able diseases, namely: Measles, Tuberculosis, Polio, Yellow Fever, Pertussis, and Hepatitis B. A Mathematical Model (MM) was consequently developed using three (3) data mining techniques (ANN, Decision trees, and Bayesian classifier**).** The focus was to compare the output of the models in order to suggest the best model for immunize-able disease outbreak prediction. The data used in this study was generated from quarterly records of cases of Immunize-able diseases that were reported in some selected local government health centres in Osun State of Nigeria. These centers are: Ife Central, Ife East, Ife North, Ilesha East, Ilesha West and Irewole.

In developing the predictive models, the three (3) data mining techniques provide a means of discovering hidden information. It also sufficed for detecting trends within databases in order to facilitate the prediction of future disease occurrence in such location. In this research, the MATLAB statistics toolbox allowed us to build the Classification and Regression Trees (CRT) from available data. The means by which we mapped observations about data items (previous cases of immunize-able diseases), and make useful conclusions about target values (future outbreaks of immunable diseases) were made possible using ANN and NBC using MATLAB toolbox.

4. THE PREDICTIVE MODEL

In the Predictive Model (PM), M was used to represent each of the month in a year. Hence, $M_1, M_2, M_3, ..., M_{12} \equiv$ 12 months of the year; then $L(i) =$ the actual outcome for each of the 12 months of the year, means that $i = 1 - 12$. Three consecutive years – 2007 to 2009 were considered for the experiment. Each of the 12 months was considered. In order to obtain useful monthly patterns, normalization was done to bring the original values of the parameters to values between 0 and 1. This was carried out using the Normalization formula (N_m) (see (ii)).

$$N_m = \frac{L_n(i) = L(i) - L_v}{L_p - L_v} \qquad (ii)$$

Where;

$L_n(i) = normalized\ \ outcome\ \ for\ \ month\,(i)$

$L(i) = outcome\ \ for\ \ month\,(i)$

$L_v = \min imum\, outcome\ \ for\ the\ years\ (2007 - 2008)$

$L_P = \max imum\, outcome\ \ for\ the\ years\ (2007 - 2008)$

After the monthly normalization;

$L(i) = 2007M_1, 2007M_2, ..., 2007M_{12}$ fell within the range from zero (0) to one (1) respectively. As a result, it was possible to estimate quarterly outcome. This was computed using the equation $L(i)$ (see *(iii)*).

$$L(i) = L_v + (L_p - L_v)L_n(i) \qquad (iii)$$

These models (*(ii)* and *(iii)*) were strengthened using the data mining techniques that are discussed further in sections 4.1, 4.2 and 4.3.

4.1 The Back propagation Network (BPN)

The Back Propagation Network (BPN) that is sometimes referred to as Multi-Layered Perceptions (MLPs) of the ANN was experimented with. This is because it learns by iteratively processing a set of training samples. These samples are compared to the network's prediction for each sample with the actual target value. For each training sample, their weights were modified to minimize the Mean Squared Error between the network's prediction and the actual target value. These modifications were made in the "backward" direction. That is from the output layer, through each hidden layer down to the first hidden layer like the one in Figure 1.

Since the BPN of the ANN is a feed-forward neural network structure; inputs were taken to the network, and were multiplied by the weights on the connections between neurons or nodes. As a result, the products were summed before they were passed through a threshold function to produce an output. In addition, the error between the output and the target (actual) were minimized by propagating the error back into the network using the BPN algorithm [38].

4.2 The Decision Tree

In DT learning technique ID3, ASSISTANT and C4.5 are the three common widely used algorithms [41]. In this research ID3 - a simple decision tree learning algorithm developed by [42] was used. The DT presented was therefore inducted using the ID3 (greedy) algorithm (For a detailed explanation of the greedy algorithm concept readers could consult [39], [40], [41]; and other relevant texts). The greedy algorithmic strategy involved representing the training samples as a single node to construct relevant decision trees. These trees started as a single node using the greedy algorithm in a top-down recursive divide-and-conquer manner.

As a result, an appropriate decision tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision is formed. For samples of the same class, the node then became a leaf and labeled with that class. These samples were partitioned accordingly, since the branch was created for each known value of the test attribute. This algorithm used the same process recursively to form a decision tree for the samples at each partition. Once an attribute occurred at a node, it was no longer considered in any of the node's descendents. The recursive partitioning stopped when any one of the following conditions was true: (i) All samples for a given node belong to the same class, (ii) there are no remaining attributes for further partitioning, and (iii) there are no samples left. A sample DT and the resultant DT are as presented in Figures 2 and 7 respectively.

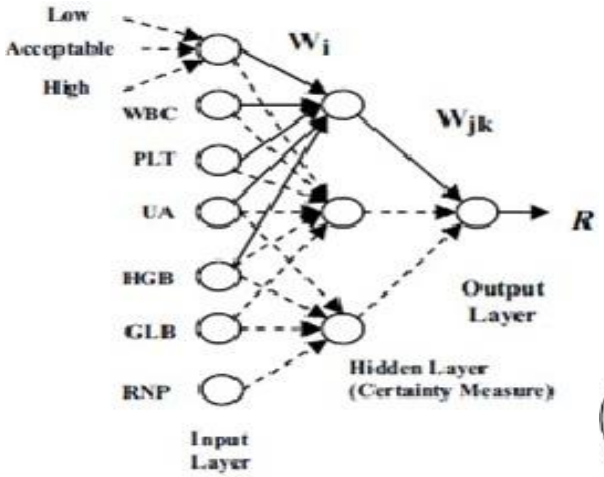

Figure 1: A Simple BPN of the ANN (Hemalatha and Megala, 2011)

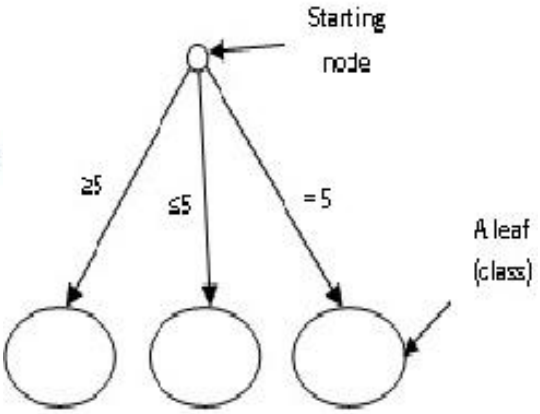

Figure 2: A Sample of the DT used in the study

4.3 The Naïve Bayesian Classifier

The naive Bayesian classifier was put to work thus: Each data sample was represented by an $n-d$imensional feature vector (see (*iv*)) below;

$$X = (x_1, x_2, ..., x_n) \qquad (iv)$$

In *(iv),* X depicts n measurements that is made on the sample dataset from n attributes - $A_1, A_2, \ldots, A_n$. To proceed, we assumed that there are m classes, say $C_1, C_2, ..., C_m$. As a result, given an unknown data sample - set X (that has no class label), the classifier will predict that X belongs to the class having the highest posterior probability conditioned on X. That is, the NBC assigned an unknown sample X to the class $C_i \Leftrightarrow$ *(v)* is valid (see (*v)).*

$$P(C_1 \mid X) \triangleright P(C_j \mid X) \qquad (v)$$

for $1 \leq j \leq m,\ j \neq i$; thus we maximized $P(C_1 \mid X)$.

The class C_i for which $P(C_1 \mid X)$ is maximized is called the maximum posteriori hypothesis. Therefore, by Bayes theorem the model of Naïve Bayes Classifier is described using the probability model in *(vi).*

$$P(H \mid X) = \frac{P(H \mid X)P(H)}{P(X)} \qquad (vi)$$

Hence,

$$P(x_k \mid C_i) = \frac{S_{ik}}{S_i} \qquad (vii)$$

As $P(X)$ is made constant for all classes, only $P(X \mid C_i)P(C_i)$ was maximized. Usually, if the class prior (preceding) probabilities are not known, it is commonly assumed that the classes are equally likely; that is $P(C_1) = P(C_2) = ... = P(C_m)$; therefore we maximized $P(X \mid C_i)$. Otherwise, we maximized $P(X \mid C_i)P(C_i)$. Note that the class prior probabilities may be estimated by

$$P(C_i) = \frac{S_i}{S} \qquad (viii)$$

Where,

S_i = is the number of training samples of class C_i, and S = is the total number of training samples.

Given a datasets with many attributes (like the ones employed in this research); to reduce the computations involved in evaluating $P(X \mid C_i)$ the naive assumption of the conditional class independence was made. With this we presumed that the values of the attributes are conditionally independent of one another, given the class label of the sample. This implies that there are no dependence relationships among he attributes. Thus,

$$P(X \mid C_i) = \prod P(x_k \mid C_i); \quad for \ (k = 1, ..., n) \qquad (ix)$$

As a result of *(ix)*, it was possible to estimate the required probabilities $P(x_1 \mid C_i); P(x_2 \mid C_i); ...; P(x_n \mid C_i)$ for each class from the training samples; Where if A_k is categorical;

Then,

$$P(x_k \mid C_i) = \frac{S_{ik}}{S_i} \qquad (x)$$

Where

S_{ik} = is the number of training samples of class C_i having the value x_k for A_k, and S_i = is the number of training samples belonging to C_i.

These models were simulated with respect to determining their predictive capabilities using the Matrix laboratory of MATLAB, since it offered a useful visualization capability that allowed the simulation process and results to be monitored. This approach is consistent with that of [44]. Some of the precautions taken were to first analyze the input data (disease data records) to remove redundant data. As a result, noise was eliminated from the input data, and thus increases the capabilities of the models to be better trained. Secondly, the time series nature of input data was preserved so as to capture the underlying regularity in the dataset for better predictability. With MATLAB it was easy to rescale values through the use of the pre-processing and post-processing functions. Ife Central, East, North; Ilesha West, east and Irewole are the sample locations considered for this work. The data obtained from these locations are from the National Health Management Information System (NHMIS) database. A summary of the dataset is shown in Table 1.

It is important to state that this is not the only values used. However, the values – historical record data in Table 1 is a snapshot of the entire time series data employed for the study. The others are on other five tables that are not included in this paper because of space.

Table 1. A Sample Historical Record for Ife Central used to develop models. Form disease database (January 2007 – December 2009)

Month	Measles	Polio	HB	Ptusis	TB	YF
Jan. 2007	12	4	6	2	5	11
Feb. 2007	3	1	3	1	11	0
Mar. 2007	2	2	11	0	7	1
Apr. 2007	3	1	9	3	1	6
May 2007	6	3	7	1	4	0
June 2007	1	1	2	7	19	1
July 2007	0	0	1	1	10	1
Aug. 2007	6	1	3	0	1	4
⋮	⋮		⋮		⋮	
Sept 2009	5	1	6	1	1	6
Oct. 2009	2	1	2	0	6	12
Nov. 2009	1	0	1	0	4	6
Dec. 2009	7	0	17	1	1	8

HB (Hepatitis B); Ptusis (Pertussis); TB (tuberculosis); YF (Yellow Fever); (Source: *NHMIS[1] unit, Osogbo, Osun State*)

5.0 RESULTS AND DISCUSSION

Based on literature review presented and the problem statement as analyzed in section 1, the need to triangularly adopt and combine the three DM techniques - ANN, DT and NBC in this paper were motivated as follows: (i) The enhancement of confidence in ensuing findings; (ii) the need to deepen and widen the understanding of immunization agencies in Nigeria, so that they can be more proactive in their life saving responsibilities; and (iii) the advantage of using diverse viewpoints – triangulation approach to cast light on a focus. This were believe would give rise to enlighten decision making regarding planning as suggested in [11]; and consistent with the recommendations of [43]. Moreover, the graphs presented buttresses this by revealing the trends in disease variation as predicted over the period of January to December (2007-2009).

From the equations of the three models that represent the best linear fit of data points of the training output, the following variables were defined and used. While *R* = Correlation Co-efficient (the strength of the relation between the input data and training data; the higher the strength was, the higher the value of R); *A* = Actual data; and *T* = Training data.

[1] *NHMIS: National Health Management Information System (NMHIS) Local Government Form prepared by the Department of health planning, research & Statistics Unit, Osogbo, Osun State, Nigeria.*

Furthermore, the equation of the best linear fit is A = (0.633) T + (2.31), R = 0.184. This is shown in Figures 3 and 4 (see sppendix). This model demonstrated that the predicted month with the lowest Disease Occurrence (DO) was October; while May and August are characterized with high DOs; and January is revealed as being the peak of DOs. In Figures 5 and 6 (see appendix), the model (equation) of the best linear fit was A = (0.562) T + (0.231), R = 0.563. The implication of this is that based on the probability of DO; the months of June, November, and December were predicted to be at the highest, while the lowest DOs were observed in October. On the ensuing DT model the codes - M, TB, Po, Yf, Pt, and Hb were used to represent the nodes based on the class of disease (measles, Tuberculosis, Polio, Yellow Fever, Pertussis, and Hepatitis B).

This is evident as shown in Figures 7 and 8 (see appendix), in the range classified by the tree using the rate of DO for the six (6) classes of diseases analyzed in this paper. As a way of comparison, from the results discussed so far and as observed from the graphs and the corresponding MM obtained, the error in the predictive ability of the models were minimized using the correlation coefficient (R). The R was used as a measure of how well the variation in output is explained by the targets values. As a result, a comparison of the R values for the various MM developed suggests that the Back Propagation Network (BPN) models were better tools for predicting the occurrence of immunize-able diseases in the selected locations. This is not to say that the other two models – the NBC and the DT are not good enough for IDs prediction. They are; we were only with respect to making available relevant information based on research findings interested in providing information on the best of the three models. Thus, triangularly the three (3) models were able to provide an alloy of results as reported in this paper for informed decision making in the promotion of EPI in Nigeria.

6. CONCLUSION

As earlier reported, the DM techniques provide an efficient medium of predicting disease trend over a period of time. As a result, an alloy of ANN, NBC Algorithm and the DT Algorithm were experimented with in the development of an MM model for IDs prediction. This study contribute to knowledge therefore by establishing that immunize-able diseases occurrence can be predicted using the efficiency of data mining tools in revealing hidden details in a database. We therefore strongly believe that information concerning IDs prediction will be useful in the efficient and effective administration of immunization. This we expect would reduce the rate of infant mortality; and easily provide predictive information for planned health actions particularly in the smooth promotion of a country's Expanded Program on Immunization (EPI). In this paper, three (3) DM techniques were therefore put together in order to investigate the occurrence of viral diseases such as: Yellow fever, Measles, Poliomyelitis, and Hepatitis B; and contagious bacteria diseases such as: Tuberculosis and Whooping cough.

As a result, with the DM techniques we have described a blend of mathematical and computational techniques that is able to aid the description, categorization and generalization of a given set of data as a means of developing a predictive model. So, according to [11], we assert that a focus on events and relationships between such events modelling is possible. This paper also argues that if a predictive model is introduced, the programme of immunization will be strengthened especially with the background information on the challenges analyzed in section 1. This is because; it is needful to give stakeholders useful information on future disease epidemic, so as to enable necessary and informed preparation to be made. In future, we hope to adapt the models presented into a working decision support system. We also intend to research more into how to mitigate the issue of privacy invasion in the use of data mining techniques, which is a key limitation in this work.

7. ACKNOWLEDGMENTS

Our sincere thanks go to all who contributed to this research effort one way or the other. And to the experts whose papers were reviewed to provide the premise for this paper we say a big thank you; and keep the flag flying.

8. REFERENCES

[1] Kumar, P., and Clark, M. L. (2007). Kumar and Clark's clinical medicine, 7th Edition, Saunders Ltd, England

[2] McWhinney, I. R. (1987). Health and disease: problems of definition. *CMAJ: Canadian Medical Association Journ -al, 136*(8), 815.

[3] Stedmans Medical Dictionary, (2000) 28th Ed,Wolters Kluwer Health Company.

[4] Laxminarayan R., Mills A.J., and Breman J.G., (2006). Advancement of global health: key messages from the Disease Control Priorities Project. Lancet, 367:1193-208. 2006.

[5] World Health Organization (2004a), "The global burden of disease": Public health programme, American journal of public Health 90,707-710.

[6] World Health Organization (2007), "Annual report and statistics on immunization and diseases" (http://www.who.int/ countries/nga/immunization/en/). Accessed May 13, 2010.

[7] World Health Organization (2004b), "Deaths by cause, sex and mortality stratum in WHO regions, estimates for 2002". http://www.who.int/research/en/).Accessed April20,2010.

[8] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). Advances in Knowledge Discovery and Data Mining, 4th Ed. Cambridge, MA: MIT Press.

[9] Thuraisingham, B. (2003). Web Data Mining Technologies and Their Applications in Business Intelligence and Counter-terrorism. CRC Press.

[10] Dunham, M. H. (2003). Classification. Data mining Introductory and Advanced Topics. Upper Saddle River, New Jersey, Prentice Hall: 93.

[11] Breiman, L., Friedman, J., Olsen, R., & Stone, C., (1984). Classification and Regression trees, Pacific Grove: Wadsworth.

[12] Smith, M. (1993). "Neural Networks for Statistical Modeling". Van Nostrand Reinhold Press.

[13] CRISP-DM. (2000). "CRoss Inductry Standard Process for Data Mining." Retrieved 1 Mar, 2009, from http://www.crisp-dm.org/Process/index.html

[14] Giudici, P. (2003) Applied Data Mining: Statistical Methods for Business and Industry, 2nd Ed. New York: John Wiley.

[15] Kotsiantis, S., and Pintelas, P., (2005). Logitboost of Simple Bayesian Classifier, Computational Intelligence in Data mining". Special Issue of the Informatica Journal, Vol. 29 (1), 2005: 53–59.

[16] Verduijn, M., Sacchi, L., Peek, N., Bellazzi, R., de Jonge, E., & de Mol, B. A. (2007). Temporal abstraction for feature extraction: A comparative case study in prediction from intensive care monitoring data. *Artificial Intelligence in Medicine*, *41*(1), 1-12.

[17] Silva, Á., Cortez, P. et al. (2008). "Rating organ failure via adverse events using data mining in the intensive care unit." Artificial Intelligence in Medicine 43(3): 179-193.

[18] Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R. (2005). Temporal data mining for the quality assessment of hemodialysis services, Artificial Intelligence in Medicine 34, pgs 25–39.

[19] Adlassnig, K.P., Combi, C., Das, A.K., Keravnou, E.T., Pozzi, G. (2006). Temporal representation and reasoning in medicine: research directions and challenges, Artificial Intelligence in Medicine, 38, pgs 101–113.

[20] Stacey, M. , and McGregor, C. (2007). Temporal abstraction in intelligent clinical data analysis: A survey. Artificial Intelligence in Medicine, 39 1–24.

[21] Bellazzi, R. and B. Zupan (2008). "Predictive data mining in clinical medicine: Current issues and guidelines." International Journal of Medical Informatics 77(2): 81-97.

[22] Pitt, E. (2009). Application of Data Mining Techniques in the Prediction of Coronary Artery Disease: Use of Anaesthesia Time-series and Patient Risk Factor Data. Unpublished Master's Thesis, Submitted to the School of Information Technology, School of Information Systems, Faculty of Science and Technology, Queensland University of Technology.

[23] Yeh, J.-Y., Wu, T.-H., and Tsao, C.-W. (2011). Using data mining techniques to predict hospitalization of hemodialysis patients. Decision Support Systems 50, 439–448 (Elsevier)

[24] Sundar, N.A., Latha, P., and Chandra, M.R. (2012). Performance Analysis of Classification Data Mining Techniques Over Heart Disease Data base. International journal of Engineering Science & Advanced Technology, Vol. 2, No. 3, pgs 470 – 478.

[25] Rajkumar, R., Shim, K. J., & Srivastava, J. (2010). Data Mining Based Predictive Models for Overall Health Indices. Technical Report submitted to the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, USA.

[26] Rao, C., Adair, T., and Kinfu, Y. (2011). Using Historical Vital Statistics to Predict the Distribution of Under-Five Mortality by Cause Clinical Medicine & Research, Vol. 9, No. 2, pgs 66-74

[27] Delen, D., Walker, G., and Kadam, A. (2005). Predict breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine, Vol. 34 (2), pages 113-127.

[28] Cruz, J.A., and Wishart, D.S. (2006). Applications of machine learning in cancer prediction and prognosis, Cancer Informatics, 2006(2):59–78

[29] Burke, H., Rosen, D., and Goodman, P. (1995), Comparing the Prediction Accuracy of Artificial Neural Networks and Other Statistical Models for Breast Cancer Survival, Advances in Neural Information Processing Systems, Vol. 7, pp. 1063--1067.

[30] Romeo, M., F. Burden, M. Quinn, B. Wood and D. McNaughton.(1998),"Infrared Microspectroscopy And Artificial Neural Networks In The Diagnosis Of Cervical Cancer.".U.S. National Library of Medicine National Institutes of Health ,Vol.44(1),pp179-87.

[31] Ifeachor, E., Sperduti, A.,and Starita, A., "Making the Distinction between Crohn's Disease and Ulcerative Colitis by Histopathological Examination: A Comparison of Human Performance, Logistic Regression and Adaptive Resonance Theory Mapping Neural Networks (ARTMAP)", In 3rd International Conference on Neural Networks and Expert Systems in Medicine and Healthcare, pp. 116--124.

[32] Hemalatha, M., and Megala, S. (2011). Mining Techniques in Health Care: A Survey of Immunization. Journal of Theoretical and Applied Information Technology, Vol. 25, No2, pgs 63-70.

[33] Assamnew, S. (2011). Predicting the Outbreak of Measles Occurrence in Ethiopia using Data Mining Technology. Unpublished M.Sc Thesis submitted to School of Public Health and School of Information Science, School of Graduate Studies, Addis Ababa University, Ethiopia.

[34] Viaene, S., Derrig, R. & Dedene, G. (2004). A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis. IEEE Transactions on Knowledge and Data Engineering, 16(5):12-620.

[35] Haykin, S. (1994). Neural Networks: A Comprehensive Foundation, 2nd Ed. New York: Macmillan Publishing.

[36] Bishop, C.M. (1995). Neural Networks for Pattern Recognition, Oxford: Oxford University Press.

[37] Han, Jiawei and Micheline Kamber (2007). Data Mining: Concepts and Techniques, 5th Ed. San Francisco, CA: Morgan Kaufmann publishers.

[38] Clarence, N.W.T (2002). 'An Artificial Neural networks primer with Financial applications. Examples in financial Distress Predictions and Foreign Exchange Hybrid Trading System', Bond University, Gold coast.

[39] Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. Wiley-IEEE Press.

[40] Witten, I. H., Frank, E., and Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

[41] Baradwaj, B.K., and Pal, S. (2011). Mining Educational Data to Analyze Students" Performance. International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, pgs 63-69.

[42] Quinlan, J. R. (1986). "Introduction of decision tree", Machine learn, 1: pp. 86-106.

[43] Olsen, W. (2004). Triangulation in Social Research: Qualitative and Quantitative Methods Can Really Be Mixed. Developments in Sociology, ed. M. Holborn, Ormskirk: Causeway Press.

[44] Sikander, M., and Mirza (2003) Introduction to Matlab: Resource book for students,1st Ed. , Pieas Publisher.

APPENDIX

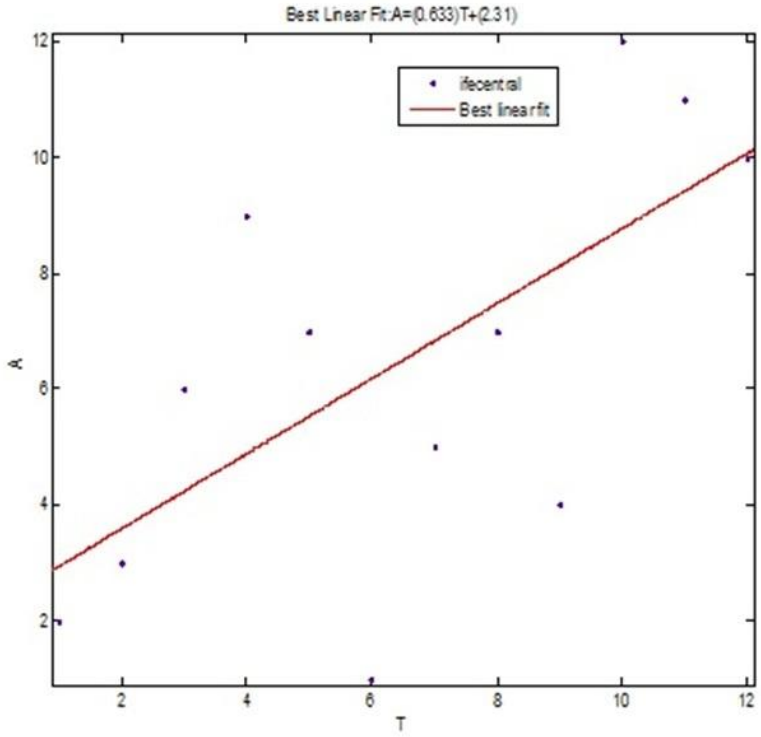

Figure 3: Best linear Fit of ANN model for Ife Central

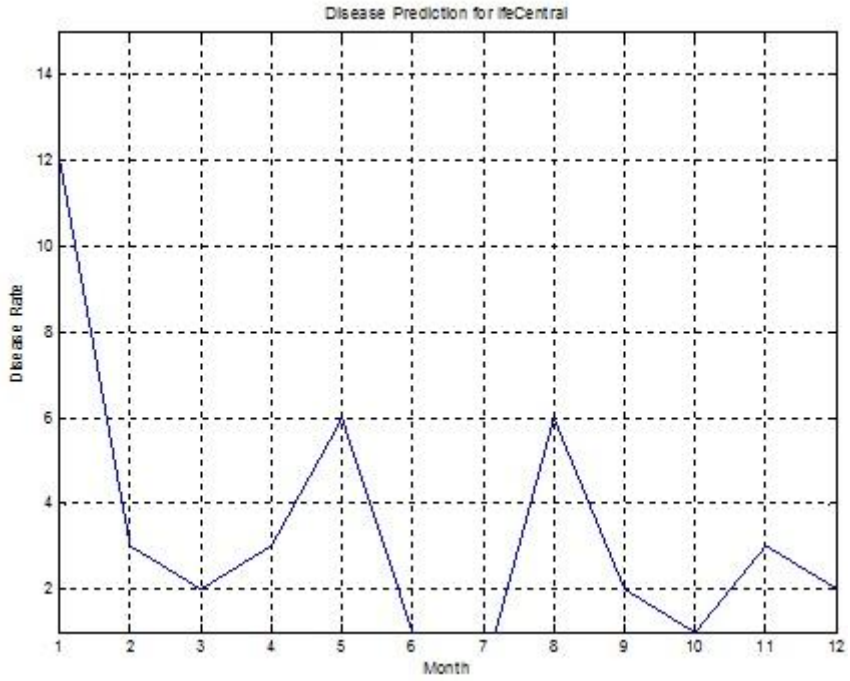

Figure 4: The Corresponding Graph the ANN model for Ife central

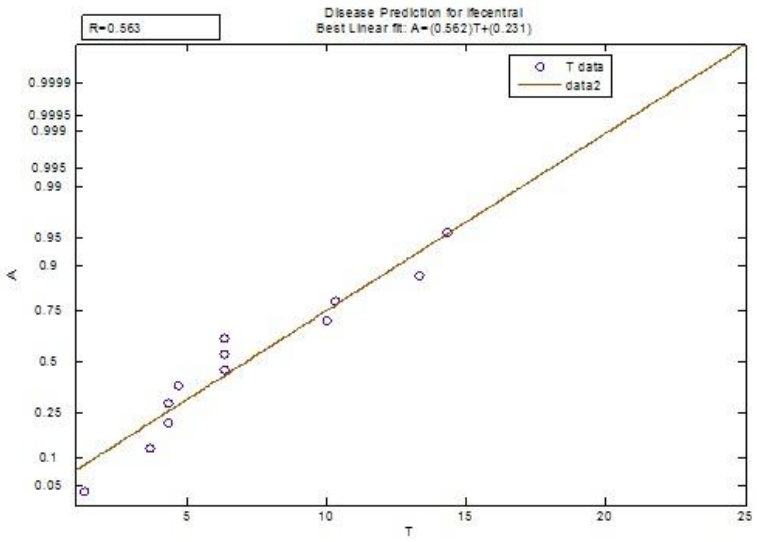

Figure 5: A Graph of the NBC model for Ife Central

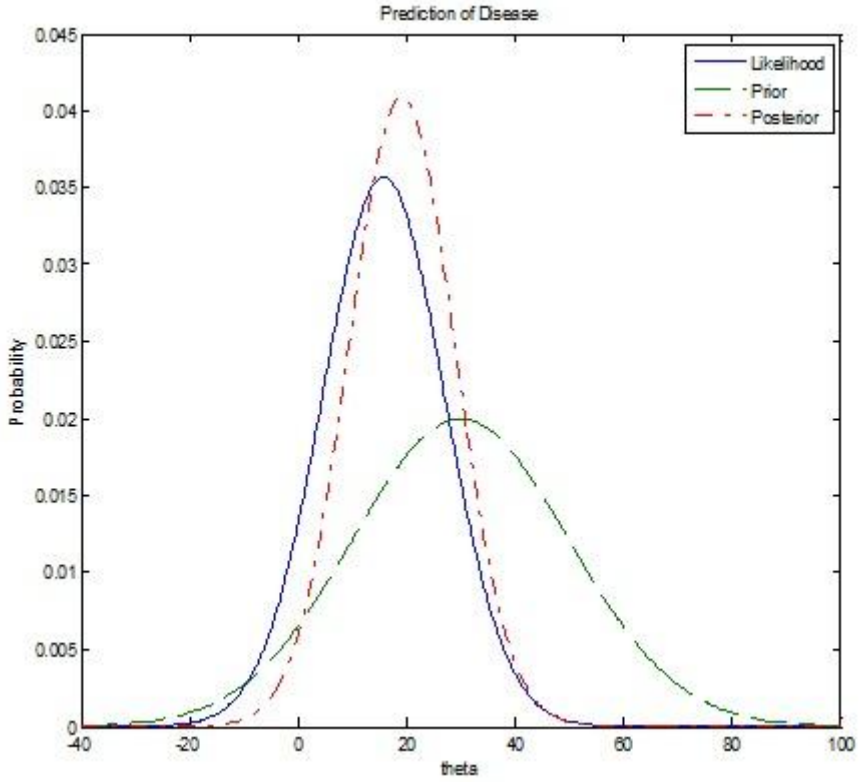

Figure 6: Graph of Disease Prediction/probability for NBC

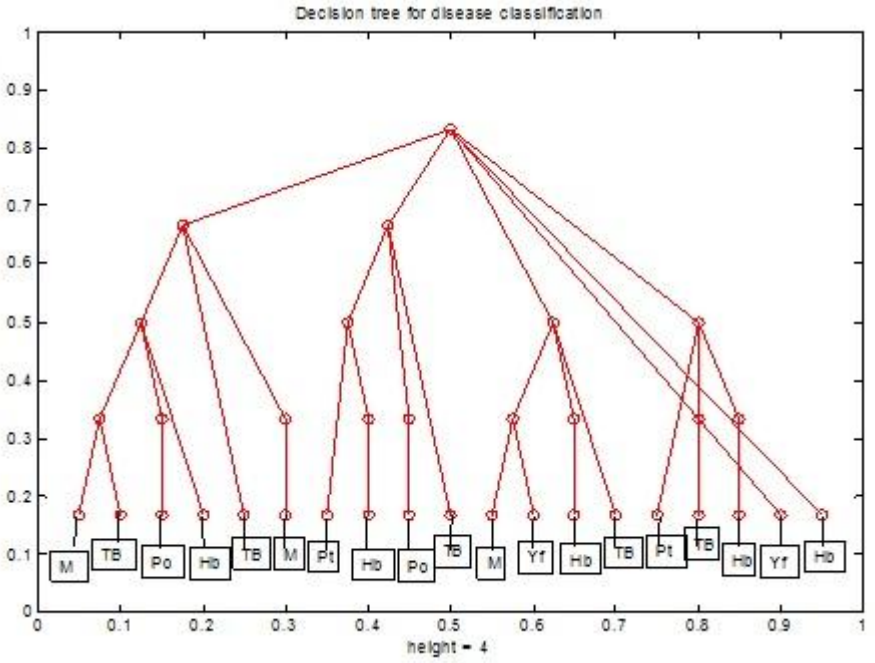

Figure 7: Decision tree Model

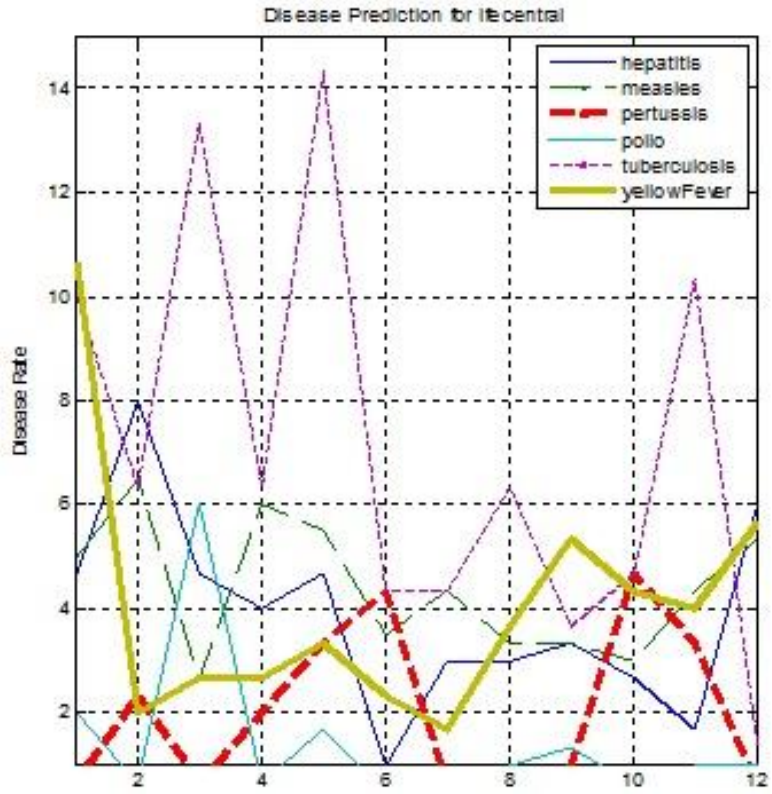

Figure 8: The Corresponding Graph of the DT model for Ife central