International Journal of Immunology
2018;6(1): 5-16
http://www.sciencepublishinggroup.com/j/iji
doi: 10.11648/j.iji.20180601.12
ISSN: 2329-177X (Print); ISSN: 2329-1753 (Online)

A Predictive Model for the Risk of Mental Illness in Nigeria Using Data Mining

Mhambe Priscilla Dooshima[1], Egejuru Ngozi Chidozie[1], Balogun Jeremiah Ademola[1], Olusanya Olayinka Sekoni[2], Idowu Peter Adebayo[1]

[1]Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria

[2]Department of Computer Science, Tai Solarin University of Education, Ijebu Ode, Nigeria

Email address:

paidowu@oauife.edu.ng (I. P. Adebayo)

[*]Corresponding author

To cite this article:

Mhambe Priscilla Dooshima, Egejuru Ngozi Chidozie, Balogun Jeremiah Ademola, Olusanya Olayinka Sekoni, Idowu Peter Adebayo. A Predictive Model for the Risk of Mental Illness in Nigeria Using Data Mining. *International Journal of Immunology*. Vol. 6, No. 1, 2018, pp. 5-16. doi: 10.11648/j.iji.20180601.12

Received: October 24, 2017; **Accepted**: November 9, 2017; **Published**: January 23, 2018

Abstract: This study identified the risk factors for mental illness and formulated a predictive model based on the identified variables. The study simulated the formulated model and validated the model with a view to developing a model for predicting the risk of mental illness. Following the review of literature in order to understand the body of knowledge surrounding mental illness and their corresponding risk factors, interview with mental experts was conducted in order to validate the identified variables. Naïve Bayes' and the Decision Trees' Classifiers were used to formulate the predictive model for the risk of mental illness based on the identified and validated variables using the WEKA software. Data was collected from 30 patients with an almost equal distribution of no, low, moderate and high risk of mental illness cases. The results showed that there were three classes of risk factors associated with mental illness, namely: biological factors, psychological factors and environmental factors. The results further showed that the formulation with Decision Trees Classifiers revealed the most relevant variables for the risks of mental illness such as losing anyone close. C4.5 decision trees algorithm with an accuracy of 83.3% outperformed the Naïve Bayes' algorithm which had an accuracy of 76.7%. The study concluded that the variables identified by the C4.5 Decision Trees algorithm can assist mental health experts to apply the rules deduced by the algorithm for the early detection of mental illness.

Keywords: Mental Illness, Predictive Modeling, Machine Learning, Risk Classification

1. Introduction

The World Health Organization (WHO) defines mental health as 'a state of well-being in which the individual realizes his or her own abilities to cope with the normal stresses of life and work productively and fruitfully, and be able to make contribution to his or her community [1]. Mental illness refers to all of the diagnosable mental disorders which are characterized by abnormalities in thinking, feelings or behaviours. Mental illness is closely related to vulnerability, both in its causes and in its effects. Globally, 14% of the global burden of disease is attributed to mental illness – with 75% of those affected being found in low-income countries – which includes a broad spectrum of diagnoses, from common mental illnesses such as anxiety and substance abuse, to severe illnesses like psychosis.

A study conducted in Uganda revealed that the term *depression* is not culturally acceptable amongst the population while another study conducted in Nigeria found that people responded with fear, avoidance and anger to those who were observed to have a mental illness [2]. The stigma linked to mental illness can be attributed to lack of education, fear, religious reasoning and general prejudice [3, 4]. According to a study by the Grand Challenges in Global Mental Health Initiative, the biggest barrier to global mental health care is the lack of an evidence-based set of primary prevention intervention methods [5, 6]. This indicates that mental health is one of the most under-resourced areas of public health in Africa, even though mental health problems are on the rise.

Thus, in African countries this area of public health requires more attention than it is currently receiving. In most parts of Africa, the family remains an important resource for the support of patients with mental health problems [7]. Abuse of psychoactive substances is a mental health problem with strong social origins. In particular, the sources of problems due to the use of alcohol and the means of curtailing them are often found in the social fabric [8]. The financial and human resources in the African Region are insufficient to address adequately the burden of mental health disorders. Patients are admitted for short period and left in the care of the relatives [9]. Only 56% of African countries have community-based mental health facilities and only 37% of the countries have mental health programmes for children, while only 15% have programmes for the elderly. The rise in the number of individuals who present with mental health problems places an even greater burden on an already under-resourced health care service [10].

Data mining involves the identification of unseen patterns in information stored in database using machine learning algorithms. Data mining has a great potential to enable healthcare systems to use data more efficiently and effectively thereby reducing the likely costs associated with making decisions [11]. Data mining techniques are very useful in healthcare domain. They provide better medical services to the patients and helps to the healthcare organizations in various medical management decisions. Classification is one of the most popularly used methods of Data Mining in Healthcare sector. It divides data samples into target classes. The classification technique predicts the target class for each data points. With the help of classification approach a risk factor can be associated to patients by analyzing their patterns of diseases.

According to a study by the Grand Challenges in Global Mental Health Initiative, the biggest barrier to global mental health care is the lack of an evidence-based set of primary prevention intervention methods [5]. Machine learning algorithms provide means of obtaining objective unseen patterns from evidence-based information especially in the public health care sector. Therefore, there is a need for the development of a predictive model for the classification of the risks of mental illness based on information regarding the associated risk factors, hence this study.

2. Related Works

Deziel [12] analyzed the mental health of engineering students using classification and regression techniques. The data was collected by conducting an online survey and on campus survey during a seven-day period in the Fall semester, targeting undergraduate Engineering students from all programs. 312 responese were received, which corresponded to 5.6% of the Engineering students' population, and six were discarded due to missing responses and/or values out of bounds, with 70% of the respondents being male and 71% percent were single. The model was developed for five components of mental health for which the accuracies were 80% for ability to enjoy life; 84% for resilience; 80% for balance; 83% for self-actualization and 75% for flexibility. Overall, engineering students rated their ability to enjoy life more highly than other mental health aspects.

Kipli *et al.* [13] detected depression from structural MRI scans to diagnose the mental health of patients. They investigated performances of four Feature Selection algorithms such as One R, Support Vector Machine (SVM), Information Gain (IG) and Relief. Finally, they concluded that the Support Vector Machine (SVM) Evaluator in combination with Expectation Maximization (EM) classifier and the Information Gain (IG) Evaluator in combination with Random Tree Classifier achieve the highest accuracy. Seixas *et al.* [14] proposed a Bayesian Network (BN) Decision Model for diagnosis of dementia, Alzheimer's disease and Mild Cognitive Impairment. The BN model was considered as it is well suited for representing uncertainty and causality. Network parameters were estimated using a supervised learning algorithm from a dataset of real clinical cases. Model was evaluated using quantitative methods and Sensitivity Analysis, which showed better results when compared to most of the other well-known classifiers.

Agbelusi [15] performed a comparative analysis of three supervised machine learning algorithm to the prediction of the survival of pediatric HIV/AIDS patients. The machine learning algorithms used were naïve Bayes' Decision Trees and Multi-layer Perceptron without the application of feature selection algorithms to identify relevant features. Rather than base features used in the study to predict HIV/AIDS survival, a larger number of features monitored in HIV/AIDS patients could have been identified with feature selection methods used in identifying the relevant features for HIV/AIDs survival. Idowu [11] developed a predictive model for the survival of pediatric Sickle Cell Disease (SCD) using clinical variables. The predictive model was developed with fuzzy logic using three (3) clinical variables while the rules for the inference engine were elicited from expert pediatrician. The fuzzy logic-based model was not validated using live clinical datasets. Moreover, relevant variables for SCD survival could have been easily identified using feature selection methods from a larger collection of variables monitored for pediatric SCD survival.

Bhakta and Sau [16] worked on the development of a predictive model for the prediction of depression among senior citizens of India using machine learning classifiers. Data was collected from a slum at Bagbazar, Kolkata, which is the service area of Bagbazar Urban Health and Training Centre (UHTC) and is also the urban field practice area of Department of Community Medicine, R. G. Kar Medical College and Hospital, Kolkata, India. Data was collected from 1st April 2016 to 30th April 2016 from 60 senior citizens using Geriatric Depression Scale (GDS) to collect training data. Five Classifiers were compared with respect to four metrics – Accuracy, ROC area, Precision and Root Mean Square Error (RMSE). The machine learning algorithms used were naïve Bayes (NB), logistic regression

(LR), Multi-layer Perceptron (MLP), Support Vector Machines (SVM) and Decision Trees (DT). The results showed that SVM had highest accuracy and precision while the NB had the ROC and lower RMSE.

Sumathi and Poorna [17] developed a prediction model for the risk of mental health problems among children using machine learning techniques. The study identified eight machine learning techniques and compared their performances on different measures of accuracy in diagnosing five basic mental health problems. A data set consisting of 60 cases was collected for training and testing the performance of the techniques. 25 attributes were identified as being important for diagnosing the problem from the documents. The attributes were reduced by applying Feature Selection algorithms over the full attribute data set. The accuracy over the full attribute set and selected attribute set on various machine learning techniques were compared. It was evident from the results that the three classifiers, Multilayer Perceptron, Multiclass Classifier and LAD Tree, produced more accurate results, and there is only a slight difference between their performances over full attribute set and selected attribute set.

3. Methods

The methodological approach of this study composes of a number of methods which include the identification of the required variables for the risk of mental illness; the collection of historical datasets about mental illness risk cases about patients; formulation of the predictive models using the supervised machine learning algorithms proposed; the simulation of the predictive models using the WEKA simulation environment; and the performance evaluation metrics applied during model validation for the predictive models.

3.1. Data Collection

For the purpose of this study, data was collected from 30 patients located in the south-western part of Nigeria using structured questionnaires that consisted of four (4) main sections, namely: demographic information consisting of 12 questions, biological factors consisting of six (6) questions, psychological factors consisting of five (5) questions and environmental factors consisting of five (5) questions. The information collected from the respondents using the questionnaire was stored in a spreadsheet application – Microsoft Excel of the Microsoft Office 2013. The information collected consisted of the risk factors associated with the mental illness for each patient as proposed by the mental health expert. A description of the attributes contained in the dataset is presented in Table 1.

Table 1. *Identified Variables for the Risk of Mental Illness.*

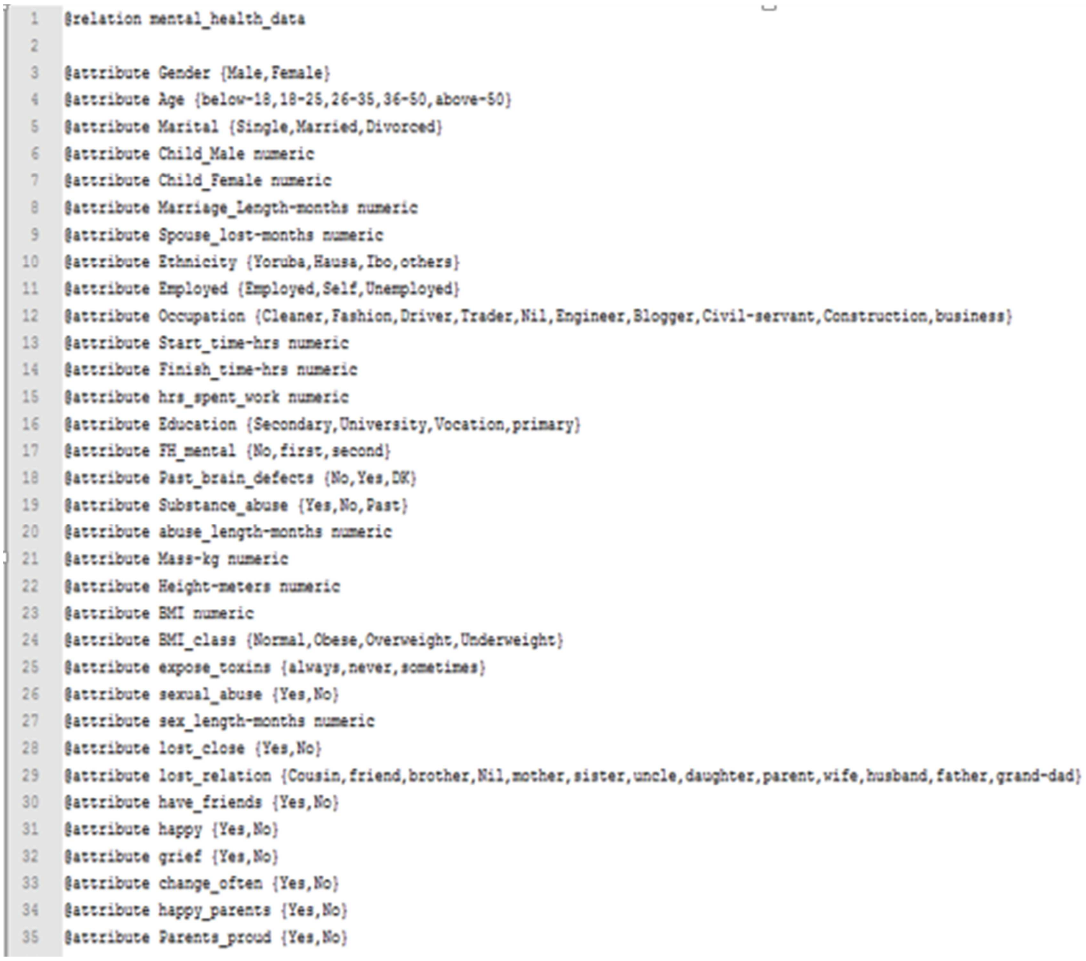
Categories	Risk Factors	Labels
Demographic	Gender	Male, Female
	Age (years)	< 18, 18 – 25, 26 – 35, 36 – 50, > 50
	Marital Status	Single, Married, Divorced, Separated
	Number of Children	Numeric
	Marriage Length (years)	Numeric
	Ethnicity	Yoruba, Hausa, Ibo
	Occupation	Unemployed, Self, Employed
	Education	Primary, Secondary, University, Vocation
Biological	Family History	First, Second, None
	Past Brain Defects	Yes, No, Don't know
	Substance Abuse	Yes, No, Past
	Length of Abuse (years)	Numeric
	Body Mass (Kg)	Numeric
	Height (meters)	Numeric
	Body Mass Index (Kg/m^2)	Numeric
	Exposure to toxins	Always, Sometimes, Never
Psychological	Sexually Abused	Yes, No
	Length of Abuse (years)	Numeric
	Lost anybody close	Yes, No
	Relationship with lost one	Nominal
	Do you have friends	Yes, No
Environmental	Happy with oneself	Yes, No
	Feelings of grief/depression	Yes, No
	Often Change jobs	Yes, No
	Happy with parents	Yes, No
	Parents proud of one	Yes, No
Target Class	Risk of Mental Illness	No, Low, Moderate, High

3.2. Data-Preprocessing

Following the collection of data from the 30 patients alongside the attributes (34 risk factors) alongside the risk of mental illness, the data collected was checked for the presence of error in data entry including misspellings and missing data. Following this process, there was no error in misspellings but there were missing data in the cells describing some records. The data was transformed into the attribute file format (.arff) for the purpose of the development of the predictive model for the risk of mental illness using the simulation environment. Figure 1 shows a screenshot of the format of the. arff used for model development in the Waikato Environment for Knowledge Analysis (WEKA) – a light-weight java application composed of a suite of supervised and unsupervised machine learning tools. The arff file is composed of three parts, namely:

a. The relation name which section contains the tag @relation *mental_health_data,* used to identify the name of the relation (or file), contains the data needed for simulation. This section is located at the first line of the file and the tag 'name' following @relation must always be the same as the file name else the file loader of the simulation environment will cease to open the file. This section is followed in the next line by the attribute names section;
b. The attribute names section which contains the tag @attribute *attribute_name label,* was used to identify the attributes that describe the dataset stored in the. arff file needed for simulation. Each attribute name alongside its labels is stated following the @relation tag on each line. The label can be a set of values inserted between brackets or a descriptor (e. g. date, numeric etc.). The last attribute is identified as the target class (risk of mental illness) while the previous attributes are the risk factors for the risk of infertility. This section is followed in the next line by the data section; and
c. The data section which contains the tag @data followed in the next line by the values of the attributes for each record of the risk of infertility is separated by a comma. Each value was listed on a row for each record in the same order as the attributes were listed in the attribute names section. The values inserted into each record must be the same values defined in each respective attribute; if there is an error in spelling or a label not defined is inserted then the file loader of the simulation environment will fail to load the file.

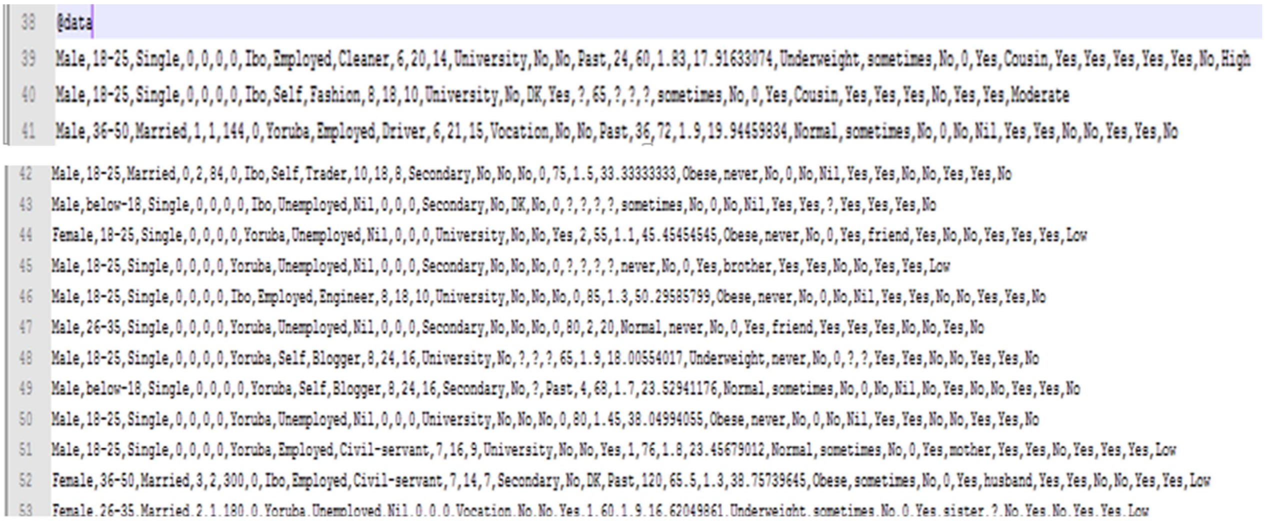
The dataset collected for the purpose of the development of the predictive model for the risk of mental illness was stored in. arff in the name *mental_health_data. arff* while the number of attributes listed in the attribute section were 34 including the target attribute. Following this, the values of the risk factors for the record of the 30 patients considered for this study was provided.



```
@relation mental_health_data

@attribute Gender {Male,Female}
@attribute Age {below-18,18-25,26-35,36-50,above-50}
@attribute Marital {Single,Married,Divorced}
@attribute Child_Male numeric
@attribute Child_Female numeric
@attribute Marriage_Length-months numeric
@attribute Spouse_lost-months numeric
@attribute Ethnicity {Yoruba,Hausa,Ibo,others}
@attribute Employed {Employed,Self,Unemployed}
@attribute Occupation {Cleaner,Fashion,Driver,Trader,Nil,Engineer,Blogger,Civil-servant,Construction,business}
@attribute Start_time-hrs numeric
@attribute Finish_time-hrs numeric
@attribute hrs_spent_work numeric
@attribute Education {Secondary,University,Vocation,primary}
@attribute FH_mental {No,first,second}
@attribute Past_brain_defects {No,Yes,DK}
@attribute Substance_abuse {Yes,No,Past}
@attribute abuse_length-months numeric
@attribute Mass-kg numeric
@attribute Height-meters numeric
@attribute BMI numeric
@attribute BMI_class {Normal,Obese,Overweight,Underweight}
@attribute expose_toxins {always,never,sometimes}
@attribute sexual_abuse {Yes,No}
@attribute sex_length-months numeric
@attribute lost_close {Yes,No}
@attribute lost_relation {Cousin,friend,brother,Nil,mother,sister,uncle,daughter,parent,wife,husband,father,grand-dad}
@attribute have_friends {Yes,No}
@attribute happy {Yes,No}
@attribute grief {Yes,No}
@attribute change_often {Yes,No}
@attribute happy_parents {Yes,No}
@attribute Parents_proud {Yes,No}
```

Figure 1a. *Arff file containing identified attributes.*



```
@data
Male,18-25,Single,0,0,0,0,Ibo,Employed,Cleaner,6,20,14,University,No,No,Past,24,60,1.83,17.91633074,Underweight,sometimes,No,0,Yes,Cousin,Yes,Yes,Yes,Yes,Yes,No,High
Male,18-25,Single,0,0,0,0,Ibo,Self,Fashion,8,18,10,University,No,DK,Yes,?,65,?,?,?,sometimes,No,0,Yes,Cousin,Yes,Yes,Yes,No,Yes,Yes,Moderate
Male,36-50,Married,1,1,144,0,Yoruba,Employed,Driver,6,21,15,Vocation,No,No,Past,36,72,1.9,19.94459834,Normal,sometimes,No,0,No,Nil,Yes,Yes,No,No,Yes,Yes,No
Male,18-25,Married,0,2,84,0,Ibo,Self,Trader,10,18,8,Secondary,No,No,No,0,75,1.5,33.33333333,Obese,never,No,0,No,Nil,Yes,Yes,No,No,Yes,Yes,No
Male,below-18,Single,0,0,0,0,Ibo,Unemployed,Nil,0,0,0,Secondary,No,DK,No,0,?,?,?,?,sometimes,No,0,No,Nil,Yes,Yes,?,Yes,Yes,Yes,No
Female,18-25,Single,0,0,0,0,Yoruba,Unemployed,Nil,0,0,0,University,No,No,Yes,2,55,1.1,45.45454545,Obese,never,No,0,Yes,friend,Yes,No,No,Yes,Yes,Yes,Low
Male,18-25,Single,0,0,0,0,Yoruba,Unemployed,Nil,0,0,0,Secondary,No,No,No,0,?,?,?,?,never,No,0,Yes,brother,Yes,Yes,No,No,Yes,Yes,Low
Male,18-25,Single,0,0,0,0,Ibo,Employed,Engineer,8,18,10,University,No,No,No,0,85,1.3,50.29585799,Obese,never,No,0,No,Nil,Yes,Yes,No,No,Yes,Yes,No
Male,26-35,Single,0,0,0,0,Yoruba,Unemployed,Nil,0,0,0,Secondary,No,No,No,0,80,2,20,Normal,never,No,0,Yes,friend,Yes,Yes,Yes,No,No,Yes,No
Male,18-25,Single,0,0,0,0,Yoruba,Self,Blogger,8,24,16,University,No,?,?,?,65,1.9,18.00554017,Underweight,never,No,0,?,?,Yes,Yes,No,No,Yes,Yes,No
Male,below-18,Single,0,0,0,0,Yoruba,Self,Blogger,8,24,16,Secondary,No,?,Past,4,68,1.7,23.52941176,Normal,sometimes,No,0,No,Nil,No,Yes,No,No,Yes,Yes,No
Male,18-25,Single,0,0,0,0,Yoruba,Unemployed,Nil,0,0,0,University,No,No,No,0,80,1.45,38.04994055,Obese,never,No,0,No,Nil,Yes,Yes,No,No,Yes,Yes,No
Male,18-25,Single,0,0,0,0,Yoruba,Employed,Civil-servant,7,16,9,University,No,No,Yes,1,76,1.8,23.45679012,Normal,sometimes,No,0,Yes,mother,Yes,Yes,No,Yes,Yes,Yes,Low
Female,36-50,Married,3,2,300,0,Ibo,Employed,Civil-servant,7,14,7,Secondary,No,DK,Past,120,65.5,1.3,38.75739645,Obese,sometimes,No,0,Yes,husband,Yes,Yes,No,No,Yes,Yes,Low
Female,26-35,Married,2,1,180,0,Yoruba,Unemployed,Nil,0,0,0,Vocation,No,No,Yes,1,60,1.9,16.62049861,Underweight,sometimes,No,0,Yes,sister,?,No,Yes,No,Yes,Yes,Low
```

Figure 1b. *Arff file containing identified attributes.*

3.3. Model Formulation

Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take input of a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. Supervised machine learning algorithms make it possible to assign a set of records (mental illness risk indicators) to a target classes – the risk of mental illness. Equation 1 shows the mapping function that describes the relationship between the risk factors and the target class – risk of mental illness.

$$\varphi: X \rightarrow Y \qquad (1)$$

$$defined\ as: \varphi(X) = Y$$

The equation shows the relationship between the set of risk factors represented by a vector, *X* consisting of the values of *i* risk factors and the label *Y* which defines the risk of mental illness – low, moderate and high risk of mental illness as expressed in equation 2. Assuming the values of the set of risk factors for an individual is represented as $X = \{X_1, X_2, X_3, \ldots\ldots, X_i\}$ where X_i is the value of each risk factor, i = 1 to i; then the mapping φ used to represent the predictive model for mental illness risk maps the risk factors of each individual to their respective risk of mental illness according to equation 2.

$$\varphi(X) = \begin{cases} No\ risk \\ Low\ risk \\ Moderate\ risk \\ High\ risk \end{cases} \qquad (2)$$

Supervised machine learning algorithms are Black-boxed models, thus it is not possible to give an exact description of the mathematical relationship existing among the independent variables (input variables) with respect to the target variable (output variable – risk of mental illness). Cost functions are used by supervised machine learning algorithms to estimate the error in prediction during the training of data for model development. Although, the decision trees algorithm is a white-boxed model owing to its ability of being interpreted as a tree-structure.

3.3.1. Naïve Bayes' Classifier

Naive Bayes' Classifier is a probabilistic model based on Bayes' theorem. It is defined as a statistical classifier. It is one of the frequently used methods for supervised learning. It provides an efficient way of handling any number of attributes or classes which are purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data. Let X_{ij} be a dataset sample containing records (or instances) of *i* number of risks factors (attributes/features) alongside their respective risk of mental illness, *C* (target class) collected for *j* number of records/patients and $H_k = \{H_1 = No, H_2 = Low, H_3 = Moderate, H_4 = High\}$ be a hypothesis that X_{ij} belongs to class C. For the classification of the risk of infertility given the values of the risk factor of the jth record, Naïve Bayes' classification required the determination of the following:

a. $P(H_k|X_{ij})$ – Posteriori probability: is the probability that the hypothesis, H_k holds given the observed data sample X_{ij} for $1 \leq k \leq 4$.
b. $P(H_k)$ - Prior probability: is the initial probability of the target class $1 \leq k \leq 4$;
c. $P(X_{ij})$ is the probability that the sample data is observed for each risk factor (or attribute), *i*; and
d. $P(|X_{ij}|H_k)$ is the probability of observing the sample's attribute, X_i given that the hypothesis holds in the training data X_{ij}.

Therefore, the *posteriori* probability of an hypothesis H_k is defined according to Bayes' theorem as follows:

$$Split(T) = -\sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \qquad (3)$$

T is the set of values for a given attribute X_i.

$$P(H_k|X_{ij}) = \frac{\prod_{i=1}^{n} P(X_{ij}|H_k)P(X_i)}{P(H_k)} \; for\ k = 1, 2, 3, 4 \qquad (4)$$

3.3.2. Decision Trees Algorithm

The theory of a decision tree has the following parts such as a root node which is the starting point of the tree and branches connect nodes showing the flow from question to answer. Nodes that have child nodes are called interior nodes. Leaf or terminal nodes are nodes that do not have child nodes and represent a possible value of target variable given the variables represented by the path from the root. The rules are inducted by definition from each respective node to branch to leaf**.** Given a set X_{ij} of j number of cases, the decision trees algorithm grows an initial tree using the divide-and-conquer algorithm as follows:

a. If all the cases in X_{ij} belong to the same class or X_{ij} is small, the tree is a leaf labeled with the most frequent class in X_{ij}.
b. Otherwise, choose a test based on a single attribute X_i with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition X_{ij} into corresponding subsets according to the outcome for each case, and apply the same procedure recursively to each subset.

ID3 (Iterative Dichotomiser 3) Decision Trees Algorithm is a classification tree used in the concept of information entropy, which provided a method for measuring the number of bits for each attribute and the attribute that yields the most Information Gain (IG) becomes the most important attribute and should thus go to the top of the tree. The C4.5 decision trees algorithm builds decision trees from a set of training dataset, X_{ij} the same way as ID3, using the information entropy. For this study, the C4.5 decision trees algorithm was used for the formulation of the predictive model for the risk of mental illness due to its advantages over the ID3 Decision Trees Algorithm and its ability to handle continuous and discrete attribute, missing values, attributes with differing costs and prune trees after creation.

The two criteria used by the C4.5 decision trees in developing its decision trees are presented in equations (4) and (5) defined as the Information Gain and the split criteria respectively. Equation (4) is used in determining which attribute is used to split the dataset at every iteration while equation (5) is used to determine which of the selected attribute split is most effective in splitting the dataset after attribute selection by equation (4).

$$IG(X_i) = H(X_i) - \sum_{t\epsilon T} \frac{|t|}{|X_{ij}|} \cdot H(X_i) \qquad (5)$$

Where:

3.4. Performance Evaluation

In order to evaluate the performance of the supervised machine learning algorithms used for the classification of the risk of mental illness, there was the need to plot the results of the classification on a confusion matrix (Figure 2). A confusion matrix is a square which shows the actual classification along the vertical and the predicted along the horizontal. Correct classifications were plotted along the diagonal from the north-west position for the low cases predicted as No (A), low (F), followed by moderate predicted as moderate (K) and high predicted as high (P) on the south-east corner (also called true positives and negatives). The incorrect classifications were plotted in the remaining cells of the confusion matrix (also called false positives). These results are presented on confusion matrix of 4 x 4 matrix table owing to the four (4) labels of the output class:

No	Low	Moderate	High	
A	B	C	D	No
E	F	G	H	Low
I	J	K	L	Moderate
M	N	O	P	High

Figure 2. *Diagram of a Confusion Matrix.*

Also, the actual no cases are A+B+C+D, actual low cases are E+F+G+H, actual moderate cases are I+K+K+L while actual high are M+N+O+P and the predicted no are A+E+I+M, low are B+F+J+N, predicted moderate are C+G+K+O and predicted high are D+H+L+P. The developed model was validated a number of performance metrics based on the values of $A - P$ in the confusion matrix for each predictive model. They are presented as follows.

a. Accuracy: the total number of correct classification

$$Accuracy = \frac{A+F+K+P}{total\_cases} \qquad (6)$$

b. True positive rate (recall/sensitivity): the proportion of actual cases correctly classified

$$TP_{no} = \frac{A}{A+B+C+D} \qquad (7)$$

$$TP_{low} = \frac{F}{E+F+G+H} \qquad (8)$$

$$TP_{moderate} = \frac{K}{I+J+K+L} \qquad (9)$$

$$TP_{high} = \frac{P}{M+N+O+P} \qquad (10)$$

c. False positive (false alarm/1-specificity): the proportion of negative cases incorrectly classified as positive

$$FP_{no} = \frac{E+I+M}{actual_{low}+actual_{moderate}+actual_{high}} \qquad (11)$$

$$FP_{low} = \frac{B+J+N}{actual_{no}+actual_{moderate}+actual_{high}} \qquad (12)$$

$$FP_{moderate} = \frac{C+G+}{actual_{no}+actual_{low}+actual_{high}} \quad (13)$$

$$FP_{high} = \frac{D+H+L}{actual_{no}+actual_{low}+actual_{moderate}} \quad (14)$$

d. Precision: the proportion of predictions that are correct

$$Precision_{no} = \frac{A}{A+E+I+M} \quad (15)$$

$$Precision_{low} = \frac{F}{B+F+J+N} \quad (16)$$

$$Precision_{moderate} = \frac{K}{C+G+K+O} \quad (17)$$

$$Precision_{high} = \frac{P}{D+H+L+P} \quad (18)$$

4. Results

This section presents the results of the methods that were applied for the development of the predictive model for the risk of mental illness. The results presented were that of the data collection, model formulation and simulation results using the WEKA software following the results of the model validation of the predictive model for mental illness.

4.1. Data Description

For this study, data was collected from 30 patients using the questionnaires constructed for this study among which; the risk of mental illness was identified. Table 2 gives a description of the number of patients with their respective risk of mental illness from the 30 patient records selected for model formulation and validation which were stored in the file mental health_data.arff. The table shows that out of the 30 patients considered 9 (30.00%) had low risk of mental illness; 11 (36.67%) had moderate risk of mental illness while 10 (33.33%) had high risk of mental illness. It was observed that the highest case presented was for respondents with moderate risk of mental illness while the least case was presented for respondents with low risk of mental illness.

***Table 2.** Distribution of mental illness risk among historical dataset.*

Hypertension risk	Frequency	Percentage (%)
No	8	26.67
Low	7	23.33
Mild	7	23.33
High	8	26.67
Total	30	100.00

Tables 3 and 4 give a description of the nominal and numeric data collected from all 30 respondents selected for the study; they both show the distribution of the values of each attributes defined for the dataset collected from the respondents. Based on the information presented in Table 3, the gender, age, marital status, ethnicity, employment, occupation and education were the nominal data collected using the questionnaire from the demographic data. Majority of the respondents were male with a proportion of 66.67% of respondents owing to a ratio of 2:1 for the male and female respondents selected for the study. Information about the ages of respondents also revealed that the majority of respondents belonged to the age group of 18 – 25 years (43.44%) followed by 26 – 35 years (26.67%) with no (0%) respondents falling within the age group above 50 years. 56.67% of the respondents were single with 50% of the respondents been Yoruba while 50% of respondents were unemployed and 23.33% employed respondents.

***Table 3.** Description of Historical Data for nominal data.*

Questionnaire Section	Variable Name	Labels	Frequency (%)
Demographic	Gender	Male	20 (66.67)
		Female	10 (33.33)
	Age	Below 18	5 (16.67)
		18 – 25	13 (43.44)
		26 – 35	8 (26.67)
		36 – 50	4 (13.33)
		Above 50 years	0 (0.00)
	Marital Status	Single	17 (56.67)
		Married	8 (26.67)
		Divorced	5 (16.67)
	Ethnicity	Yoruba	15 (50.00)
		Hausa	1 (3.33)
		Ibo	12 (40.00)
		Others	2 (6.67)
	Employment	Employed	7 (23.33)
		Self-Employed	8 (26.67)
		Unemployed	15 (50.00)
	Occupation	Cleaner	2 (6.67)
		Fashion	1 (3.33)
		Driver	1 (3.33)
		Trader	3 (10.00)
		Engineer	1 (3.33)
		Blogger	2 (6.67)
		Civil Servant	3 (10.00)
		Construction	1 (3.33)
		Business	1 (3.33)

Questionnaire Section	Variable Name	Labels	Frequency (%)
Biological Factors	Education	Nil	15 (50.00)
		Secondary	12 (40.00)
		University	12 (40.00)
		Vocation	5 (16.67)
		Primary	1 (3.33)
	Family History	No	27 (90.00)
		First Generation	2 (6.67)
		Second Generation	1 (3.33)
	Past brain defects	No	21 (70.00)
		Yes	3 (10.00)
		Don't Know	4 (13.33)
		Missing	2 (6.67)
	Substance Abuse	Yes	11 (36.67)
		No	12 (40.00)
		Past	6 (20.00)
		Missing	1 (3.33)
	BMI Class	Normal	9 (30.00)
		Underweight	3 (10.00)
		Overweight	3 (10.00)
		Obese	9 (30.00)
		Missing	6 (20.00)
	Exposure to toxins	Always	3 (10.00)
		Never	10 (33.33)
		Sometimes	16 (53.33)
		Missing	1 (3.33)
Psychological Factors	Sexual Abuse	Yes	3 (10.00)
		No	27 (90.00)
	Lost relation	Cousin	2 (6.67)
		Friend	3 (10.00)
		Brother	2 (6.67)
		Mother	1 (3.3)
		Sister	1 (3.3)
		Uncle	1 (3.3)
		Daughter	1 (3.3)
		Parents	2 (6.67)
		Wife	1 (3.3)
		Husband	2 (6.67)
		Father	2 (6.67)
		Grand-dad	1 (3.3)
		Nil	10 (33.33)
		Missing	1 (3.3)
	Lost anyone close	Yes	19 (63.33)
		No	10 (33.33)
		Missing	1 (3.33)
	Have friends	Yes	23 (76.67)
		No	6 (20.00)
		Missing	1 (3.33)
Environmental Factors	Happy with oneself	Yes	20 (66.67)
		No	10 (33.33)
	Feelings of grief or depression	Yes	16 (53.33)
		No	13 (43.33)
		Missing	1 (3.33)
	Change Jobs/Schools often	Yes	8 (26.67)
		No	22 (73.33)
	Parents happy with oneself	Yes	23 (76.67)
		No	7 (23.33)
	Parents proud with oneself	Yes	19 (63.33)
		No	11 (36.67)

The results of the biological factors reveal that majority of the patients had no family history of mental illness (90%), no past brain defects (70%), normal or Obese BMI class (30%), sometimes exposed to toxins (53.3%) and did not partake in substance abuse (90%). The results of the psychological factors revealed that majority of the patients had lost someone close (63.3%) consisting of grand-parents (33.3%) and have friends (76.7%). The results of the environmental factors revealed that out of the patients (66.7%) were happy with themselves; (53.3%) had feelings of grief or depression; (73.3%) did not change jobs/schools; (76.6%) often were happy with parents and (63.35%). parents were proud of them.

According to table 4, the distribution of the numeric information collected from the demographic data was presented using the maximum, minimum, average (mean) and standard deviation. Based on the information provided, it was observed that the maximum number of male and female children born by couples were 3 and 2 respectively with some respondents bearing no children.

Table 4. *Description of Historical Data for numeric data.*

Questionnaire Section	Variable	Minimum	Maximum	Mean	Standard Deviation
Demographic	Number of Male child	0.0	3.0	0.30	0.90
	Number of female child	0.0	2.0	0.40	0.68
	Time before separation (months)	3.0	375.0	48.90	92.88
	Spouse lost (months ago)	3.6	18.0	0.70	3.31
	Start work time (hours)	5.0	10.0	3.87	4.02
	Finish work time (hours)	12.0	24.0	9.50	9.88
	Work time spent (hours)	8.0	16.0	5.63	6.13
	Mass (Kg)	50.2	86.0	67.95	9.45
	Height (metres)	1.1	2.0	1.59	0.25
	BMI (kg/m^2)	16.6	50.3	29.20	10.10
Biological Factors	Substance abuse (months ago)	30.0	120.0	8.20	24.74
Psychological Factors	Sexual abuse length (months ago)	30.0	120.0	8.20	24.74

The result presented in table 4 showed that the mean time before separation by divorced couples was 48.9 months (about 4 years) with some spouse being lost by the widowed as recent as 3 months ago. There were respondents who start work as early as 5.00 am or latest by 10.00 am and those who close form work as early as 12 noon and as late as 12 midnight owing for minimum working hours of 8 and maximum working hours of 16. The average working hours of employed respondents is about 6 hours. Among respondents who partook in substance abuse in the past, the minimum time is 30 months ago and the maximum time is 120 months ago. The results also showed that 90% of the respondents have never been sexually abused, but out of those that have been sexually abused, the minimum time was 30 months while the maximum time was 120 months ago.

4.2. Simulation Results

Two different supervised machine learning algorithms were used to formulate the predictive model for the risk of mental illness, namely: Naïve Bayes' and Decision Trees Classifiers. They were used to train the development of the prediction model using the dataset containing 30 patients' risk factor records. The simulation of the prediction models was done using the Waikato Environment for Knowledge Analysis (WEKA). The C4.5 Decision Trees Algorithm was implemented using the J48 Decision Trees Algorithm available in the trees class and the Naïve Bayes' Algorithm was implemented using the Naïve Bayes' Classifier available in the Bayes class all available on the WEKA environment of classification tools. The models were trained using the 10-fold cross validation method which splits the dataset into 10 subsets of data – while 9 parts are used for training the remaining one is used for testing; this process is repeated until the remaining 9 parts take their turn for testing the model.

4.2.1. Results of the Naïve Bayes' Classifier

Using the Naïve Bayes' Classifier to train the predictive model developed using the training data via the 10-fold cross validation method. Figure 3 shows the graphical plot of the predictions made by the Naive Bayes' Classifier algorithm on the dataset, each class of mental illness is represented using a specific colour and each correct classification is represented with a star while each misclassification is represented as a square.

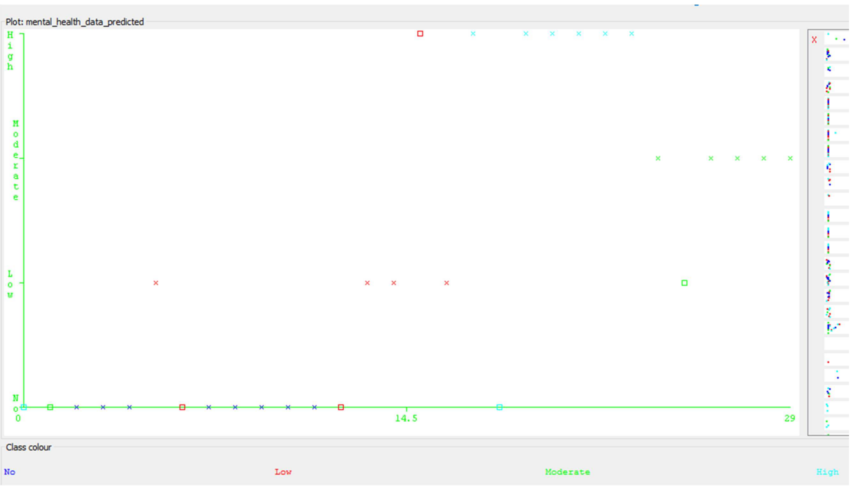
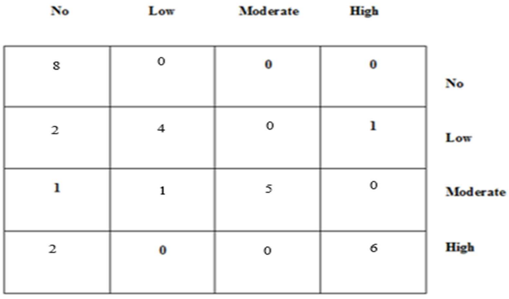


Figure 3. *Screenshot of Naïve Bayes' Classification Results.*

The results presented in figure 3 were used to evaluate the performance of the Naive Bayes Classifier algorithm and thus, the confusion matrix was determined as shown in figure 4. From the confusion matrix shown in figure 4, the following sections present the results of the model's performance. Out of the 8 actual no cases, all were correctly classified, out of the 7 actual low cases, there were 4 correct classifications with 2 misclassified as no risk and 1 misclassified as high risk; out of the 7 moderate risk cases, there were 5 correct classifications with 1 misclassified as no risk and 1 misclassified as low risk while out of the 8 high cases, there were 6 correct classifications with 2 misclassified as low risk. Therefore, there were 23 correct classifications out of the 30 records considered for the model development owing for an accuracy of 76.67%.



No	Low	Moderate	High	
8	0	0	0	No
2	4	0	1	Low
1	1	5	0	Moderate
2	0	0	6	High

Figure 4. *Confusion matrix of performance evaluation using naïve Bayes.*

4.2.2. Results of the C4.5 Decision Trees Classifier

Using the C4.5 Decision Trees Classifier to train the predictive model developed using data via the 10-fold cross validation method. Figure 5 shows the graphical plot of the predictions made by the C4.5 Decision Trees Classifier algorithm on the dataset, each class of mental illness is represented using a specific colour and each correct classification is represented with a star while each misclassification is represented as a square. The results presented in figure 5 were used to evaluate the performance of the C4.5 decision trees classifier algorithm and thus, the confusion matrix determined as shown in figure 6.

From the confusion matrix shown in figure 6, the following sections present the results of the model's performance. Out of the 8 actual no cases, 7 were correctly classified as no while 1 was misclassified as low risk, out of the 7 actual low cases, there were 5 correct classifications with 1 misclassified as moderate risk and 1 misclassified as high risk; out of the 7 moderate risk cases, there were 6 correct classifications with 1 misclassified as no risk while out of the 8 high cases, there were 7 correct classifications with 1 misclassified as moderate risk. Therefore, there were 25 correct classifications out of the 30 records considered for the model development owing for an accuracy of 83.33%.

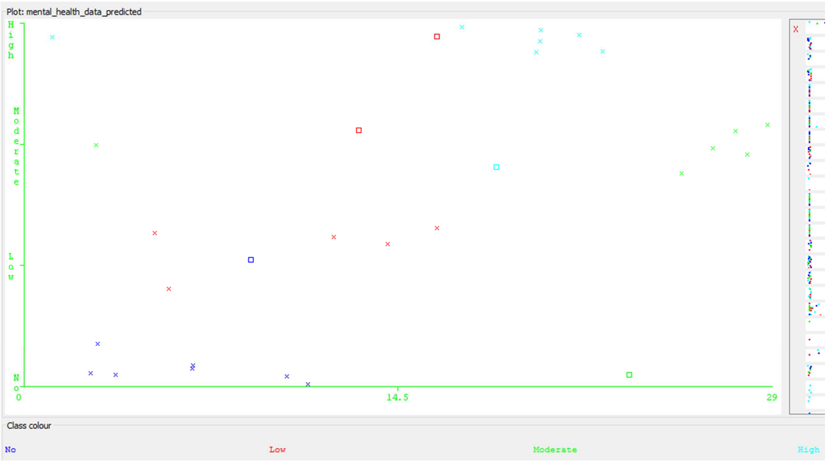
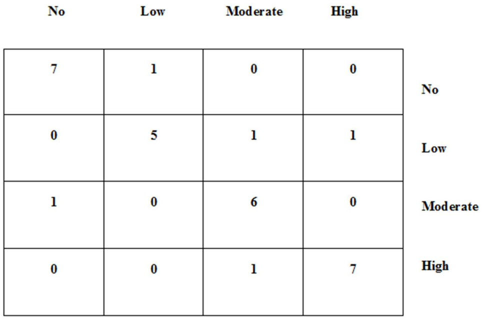


Figure 5. *Screenshot of C4.5 decision trees Classification Results.*



No	Low	Moderate	High	
7	1	0	0	No
0	5	1	1	Low
1	0	6	0	Moderate
0	0	1	7	High

Figure 6. *Confusion matrix of performance evaluation using C4.5 decision trees.*

The Decision Tree that was plotted from the simulation of the predictive model using the C4.5 decision trees is presented in figure 7. The main risk factors considered with the highest gain ratio by C4.5 were parents been proud of the person, losing anybody close and ethnicity. Using the decision tree in Figure 7, the following rules were deduced and can be used to predict the risk of mental illness based on the values of the three identified risk factors. There are 5 rules and are presented as follows:

a. IF (Are parents proud="No") AND (lose someone close="No") THEN (Mental illness Risk = Moderate);
b. IF (Are parents proud="No") AND (lose someone

close="Yes") THEN (Mental illness Risk = High);

c. IF (Are parents proud="Yes") AND (lose someone close="No") THEN (Mental illness Risk = No);

d. IF (Are parents proud="Yes") AND (lose someone close="Yes") AND (ethnicity="Yoruba/Others") THEN (Mental illness Risk = Low); and

e. IF (Are parents proud="Yes") AND (lose someone close="Yes") AND (ethnicity="Ibo/Hausa") THEN (Mental illness Risk = Moderate).

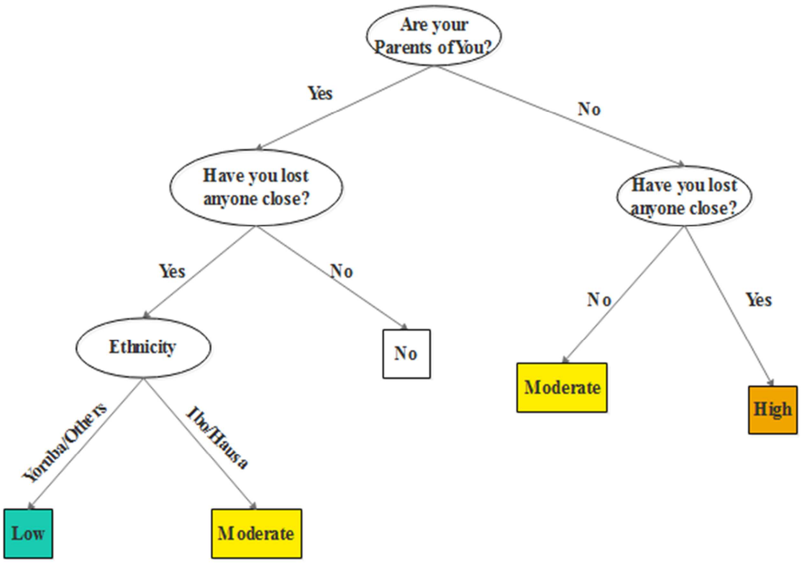


Figure 7. *Graphical plot of the C4.5 decision tree for mental illness risk.*

4.3. Discussions

The results of the performance evaluation of the C4.5 and Naïve Bayes' algorithms are presented in Table 5. The true positive rate which gave a description of the proportion of actual cases that were correctly predicted which showed values of 0.875, 0.714, 0.857 and 0.875 respectively for no, low, moderate and high risk cases by the C4.5 decision trees algorithm and 1.000, 0.571, 0.714 and 0.750 for the naïve Bayes classifier. Thus, the decision trees algorithm showed equal capacity to correctly classify the actual no high cases of mental illness better than the moderate and low risk cases while the Naïve Bayes' Classifier had the ability to correctly classify the no risk cases better than the C4.5 Decision Trees but not as good as C4.5 for the other risks of mental illness.

Table 5. *Summary of Validation Results for C4.5 and naïve Bayes' classifiers.*

Performance Metrics	Risk Labels	C4.5 DT	Naïve Bayes
True Positive (TP) rate/recall/sensitivity	No	0.875	1.000
	Low	0.714	0.571
	Moderate	0.857	0.714
	High	0.875	0.750
False Positive (FP) rate/false alarm	No	0.045	0.227
	Low	0.043	0.043
	Moderate	0.087	0.000
	High	0.045	0.045
Precision	No	0.875	0.615
	Low	0.833	0.800
	Moderate	0.750	1.000
	High	0.875	0.857

The false positive rate which gave a description of the proportion of predicted cases that was incorrectly classified showed values of 0.045, 0.043, 0.087 and 0.045 for the no, low, moderate and high risk cases respectively for the C4.5 Decision Trees Algorithm while the naïve Bayes' classifier had values of 0.227, 0.043, 0.000 and 0.045 respectively for no, low, moderate and high risk cases. The Naïve Bayes' Classifier was able to correctly distinguish moderate cases better than the C4.5 Decision Trees algorithm and showed equal capacity to distinguish no and high cases as equal as the C4.5 Decision Trees algorithm. The precision which gave a description of the proportion of the predicted cases that was correctly classified showed values of 0.875, 0.833, 0.750 and 0.875 for no, low, moderate and high cases respectively while the Naïve Bayes' classifier showed values of 0.615, 0.800, 1.000 and 0.857 respectively for no, low, moderate and high risk cases. The Naïve Bayes' classifier was able to provide better prediction results of the risk of mental illness compared to the C4.5 decision trees algorithm based on the precision.

In general, the C4.5 Decision Trees Algorithm was able to predict the risk of mental illness than the Naïve Bayes' Classifier in addition to the identification of the relevant variables that can be used for the early detection of mental illness. The number of variables identified by the C4.5 decision trees algorithm was three (3) which shows more capacity of determining the risk of mental illness.

5. Conclusions

This study focused on the development of a prediction model using identified risk factors in order to classify the risk of mental illness in selected respondents for this study. Historical dataset on the distribution of the risk of mental illness among 30 respondents was collected using questionnaires following the identification of associated risk factors of mental illness from expert medical practitioners. The dataset containing information about the risk factors identified and collected from the respondents was used to formulate predictive models for the risk of mental illness using C4.5 Decision Trees and Naïve Bayes' Classifier algorithm. The predictive model development using the algorithms were formulated and simulated using the WEKA software.

The results of the study revealed the variables that were identified by the Decision Trees algorithm as relevant for identifying the risk of mental illness in respondents. The Decision Trees algorithm was observed to show a better accuracy compared to that of the Naïve Bayes' classifier using the training dataset presented in the study. The C4.5 Decision Trees' model developed was capable of predicting the risk of mental illness with an accuracy of 83.3% compared to 76.7% of the Naïve Bayes' Classifier.

Following the development of the prediction model for mental illness risk classification, a better understanding of the relationship between the attributes relevant to mental illness risk was proposed. The model can also be integrated into existing Health Information System (HIS) which captures and manages clinical information that can be fed to the mental illness risk classification prediction model, hence improving the clinical decisions affecting mental illness risk and the real-time assessment of clinical information affecting mental illness risk from remote locations.

References

[1] World Health Organization (2011a). Political Declaration of the High-level Meeting of the General Assembly on the Prevention and Control of Non-Communicable Diseases. 66th Session of the Unites Nations General Assembly. New York: WHO.

[2] Gordon, A. (2013). Mental Health Remains an Invisible Problem in Africa. Think Africa Press. Retrieved from http://thinkafricapress.com on May 12, 2017.

[3] Arboleda-Florez, J. (2002). What Causes Stigma. World Psychiatry 1 (1): 25–26.

[4] Corrigan, P. W., Druss, B. G. and Perlick, D. A. 2014. The Impact of Mental Illness Stigma on Seeking and Participating in Mental Health Care. Psychological Science in the Public Interest, 15 (2) 37–70: sagepub.com/journalsPermissions.nav.

[5] Fournier, O. A. (2011). The Status of Mental Health Care in Ghana, West Africa and Signs and Progress in the Greater Accra Region. Berkeley Undergraduate Journal 24 (3): 1–6.

[6] Naifeh, J. A., Colpe, C. L. J., Aliaga, P. A., Sampson, N. A., Heeringa, S. G., Stein, M. B., Ursano, R. J., Fullerton, C. S., Nock, M. K., Schoenbaum, M. and Zaslavsky, A. M., 2016. Barriers to initiating and continuing mental health treatment among soldiers in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Military medicine, 181 (9), p.1021.

[7] Hanlon, C., Wondimagegn, D. and Alem, A. (2010). Lessons Learned in Developing Community Mental Health Care in Africa. World Psychiatry 9 (3): 185–189.

[8] World Health Organization (2011b). WHO African Regional Ministerial Consultation on Non-communicable Diseases. Brazzaville, Congo. Brazzaville: WHO Regional Office for Africa.

[9] World Health Organization (2015). Mental health atlas 2014. World Health Organization, Reterived from http://apps.who.int/iris/bitstream/10665/178879/1/9789241565011_eng.pdf. Accessed on March 20th, 2016.

[10] Njenga, F. (2002). 'Focus on Psychiatry in East Africa'. British Journal of Psychiatry (181): 354-59.

[11] Idowu, P. A., Aladekomo, T. A., Williams, K. O. and Balogun, J. A. (2015). Predictive Model for Likelihood of Survival of Sickle Cell Anemia (SCA) among Pediatric Patients using Fuzzy Logic. Transactions in Networks and Communications 31 (1): 31-44.

[12] Deziel, M., Olawo, D., Truchon, L. and Golab, L. (2013). Analyzing the Mental Health of Engineering Students Using Classification and Regression.

[13] Kipli, K., Kouzani, Z. and Hamid, I. R. (2013). Investing Machine Learning Techniques for Detection of Depression Using Structural MRI Volumetric Features. International Journal of Biosciences, Biochemistry and Bioinformatics 3 (5): 444–448.

[14] Seixas, F. L., Zadrozny, B., Laks, J., Conci, A., & Saade, D. C. M. (2014). A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer′ s disease and mild cognitive impairment. Computers in biology and medicine, 51, 140-158.

[15] Agbelusi, O. (2014). Development of a Predictive Model for Survival of HIV/AIDS Patients in South-Western Nigeria. Unpublished MPhil Thesis submitted to the Department of Computer Science and Engineering, Obafemi Awolowo University.

[16] Bhakta, I and Sau, A. (2016). Prediction of Depression among Senior Citizens using Machine Learning Classifiers. International Journal of Computer Applications 144 (7): 11–16.

[17] Sumanthi, M. R. and Pooma, B. (2016). Prediction of Mental Health Problems among Children Using Machine Learning Techniques. International Journal of Advanced Computer Science and Applications 7 (1): 552–557.