Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 7, July -2017

DATA MINING USING WEKA TOOL FOR INTERNALLY DISPLACED PERSONS DATASET IN NORTHEASTERN NIGERIA

LAWAL S.N, ANANDE T.J, GENGER T.K.

Department of Electrical and Electronics Engineering, University of Agriculture, Makurdi, Nigeria
Department of Electrical and Electronics Engineering, University of Agriculture, Makurdi, Nigeria
Department of Electrical and Electronics Engineering, University of Agriculture, Makurdi, Nigeria

Abstract - *Decision making process has greatly been influenced by the data mining methods applications. The utilization of data mining algorithms and methods to analyze various types of data has shown great advantages in different sectors of life. Some of the data require little or no preparations before being processed. Others meanwhile require large pre-processing because of the complex nature of those data sets. To ensure efficiency in the data mining process, the dimensionality of these data has to be reduced before mining. The WEKA data mining tool allows for the process of data pre-processing in terms of feature or attribute selection to enhance the accuracy and efficiency of data. The study applied the classification algorithms and feature selections on the IDP data sets, to determine the accuracy and efficiency between the different classifications and thus suggest how this can help manage data in the rising humanitarian crisis. The study concludes by acknowledging the need for a better data understanding and suggesting the appropriately the feature selection and classification algorithms that can be applied in Data mining in the humanitarian sector in Nigeria.*

Keywords- *Data Mining; WEKA; IDPs;*

I. INTRODUCTION

Chen et al (1996) described Data mining as knowledge discovery in databases. Meaning, it involves a process of nontrivial extraction of some previously known information from databases in bid of making it coherent.

Data mining is only just one of the levels involved in the Knowledge discovery processes (Fayyad et al 1996) and thus have been incorporated in various institutions and disciplines. In marketing for example, it can be seen to create tools such as the market basket analysis (Agrawal et al. 1996); in financial institutions, to look out for fraud and prices (Major and Riedinger 1992); also in medical research, various predictions have been ascertained in different diseases (Palaniappan and Awang 2008). Methods for Data mining can basically be divided into two major categories; the supervised learning method and the unsupervised learning method (Han 2001). The supervised learning involves the algorithm using a training data set to understand the model parameters. Unsupervised learning contrarily makes use of the data itself to generate the model. There exists another type of model; the semi-supervised model, which has been made known as a hybrid option. To successfully achieve a unique outcome, the data set being studied has to be appropriately matched with the data mining algorithm.

The Humanitarian Industry constitute of a huge set of data and Blum and Langley (1997) identified that in dealing with such high level datasets, pattern recognition and machine learning are key factors.

II. LITERATURE REVIEW

2.1 Data Mining for Disaster Management

Data mining tools and techniques has been implemented recently in various areas, ranging from management to business intelligence. Hristidis, Chen and Li et al. (2010) identifies a survey in the use of data mining in the area of disaster management. Management of every disaster as recognized by (Hristidis, Chen and Li et al. 2010) involves the following stages;

- Prevention
- Advance warning
- Early detection
- Analysis of the problem, and assessment of scope
- Notification of the public and appropriate authorities
- Mobilization of a response
- Containment of damage
- Relief and medical care for those affected

All the above mentioned stages can be further categorized into the four basic stages in humanitarian operations; preparedness, mitigation, response and recovery. It aimed at looking at where technology fits into these stages. Zheng, Shen and Tang et al. (2011) also identified the three basic types of data mostly available for this data mining at these stages, to include spatial data, temporal data and spatio-temporal data.

2.2 Preparedness

Peng, Zhang and Tang et al. (2011) identify areas where Data mining methods can be used to recognize useful patterns for pre -incident detection and provide differentiated services. The major activities involved in preparedness include organizing, planning, training, evaluation and invoking corrective action. A lot of data is being collected during the pre-event preparations, data types such as street maps, evacuation plans, number of households, building design records and others (Jain and McLean 2006). Differentiated services referred to the levels of prioritization for preparations of disaster. The appropriate data mining method is implemented based on priorities in the magnitude of the events.

2.3 Mitigation

The process of disaster mitigation management assumes one of the most challenging roles in the decision making process as it is accompanied by missing and uncertain information (Kandel, Tamir and Rishe 2014). One major way of improving disaster preparedness and mitigation capabilities is applying vibrant techniques for collection and analysis of data. Kandel, Tamir and Rishe (2014) also suggested the use of fuzzy logic based techniques for a promising result. Stating its advantage as being able to keep account of possibility of occurrence through the "truth- value" information.

However, as stated by Ganguly and Steinhaeuser (2008) new data techniques and tools now vastly available for the analysis of data. Data mining algorithm provides the various representations from non- linear regression and classification to decision trees and rules. There exist no standard establish standard for choosing the type of data algorithm to use, but totally dependent on the type and availability of data (Ganguly and Steinhaeuser 2008), these thus further explains that the data sets available determines the suitable selection of data algorithm to be implemented.

Other techniques can also be considered, such as the use of spatial data mining process geospatial data to obtain the prediction model (Li, L. et al. 2012). It adopted the Bayesian Network, which is a type of data mining technique to analysis of vulnerability. The use of the BN technique involved two main factors to consider; factors related to occurrence of natural disasters and factors related to environmental and system resistance. This technique having considered the uncertainty of the occurrences of natural disasters was effectively analyzed for vulnerability assessment of human beings in terms of catastrophic events.

2.4 Global Internally Displacement and Data Collection

Over the year 2015, the global report on displacement reports that there exist 27.8 million new displacements over 127 countries, this estimate equals the population of London, New York, Paris and Cairo all put together. This values are categorized into two; conflict/ violence and disaster. Conflict and violence occurs in 28 countries with 8.6 million displacements, while disaster in 113 countries with 19.2 million displacements (Anon. 2016).

In Nigeria, the continuous activities of the insurgent group Boko Haram in the past six years has caused the displacement of millions, thus giving rise to a humanitarian crisis in the north east region of the country. Trends in the displacement have identified people moving en masse to urban areas in search of humanitarian assistance (Anon. 2016).

The global report on internal displacement reports that Internal displacement is reoccurring in Nigeria, especially, over the last five years and this is caused majorly by general violence, natural disasters, human right violation and conflicts (Ladan and Tawfiq 2015). The global overview report on Internal displacement 2015 reports that Nigeria witnessed its highest rate of displacement in 2014 due to continuous brutal attacks from the Boko Haram terror group which rose to about a million people in 2015 (Council 2016).

2.5 Data Mining Intervention

The International Organization for Migration IOM in conjunction with the Nigerian National Emergency Management Authority set up a tracking matrix for the collection and dissemination of data (Anon. 2016). The establishment of this matrix gave an insight on the nature of the evolving situations, "*The lack of a holistic understanding of displacement dynamics in Nigeria resulted in a fragmented and inadequate humanitarian and development response to those in need*" (Anon. 2015). This unprecedented crisis situation with a huge chunk of collected unprocessed data is continuously in the need for a better data analysis mechanism; this is to aid pattern recognition and give more appropriate and accurate discoveries (Anon. 2016). As much as the collection of data remains a great challenge in crisis related areas, the available data collected need to be sufficiently and adequately analyzed to obtain the closest possible patterns (Anon. 2015). As seen from the applications of data mining in different areas as discussed earlier, the humanitarian crisis in Nigeria has had little research done in this area. This brings about the questions of data analysis or data mining methods used in disaster management, how these methods can be properly initiated and integrated to the large sets of collected data to easily and most efficiently identify patterns.

Research paper by T. Akomolafe and Olutayo (2013) explored the use of WEKA tool for the prediction of likely occurrences of road accident along a major express way in Nigeria. WEKA is a collection of machine learning algorithms and data processing tools. It contains various tools for data pre-processing, classification, regression, clustering, association rules and visualization. There are many learning algorithms implemented in WEKA including Bayesian classifier, Trees, Rules, Functions, Lazy classifiers and miscellaneous classifiers. The algorithms can be applied directly to a data set. The research found out that if proper data is available and properly analyzed using the WEKA tool,

accident occurrence predictions can be inferred. It particularly made used of the decision tree feature in the WEKA tool, this was due to the type of data available.

Another study by V.A and A.A (2014) also tried to compare the use decision trees technique in comparison to artificial neural networks with categorical and continuous data respectively. Though the neural network analysis gave a close enough gave a good enough results, the decision tree proved to be the most reliable and accurate technique for use in the case of the categorical data set. These two different techniques as considered by V.A and A.A (2014) shows degree of affirmation to the earlier research done by (T. Akomolafe and Olutayo 2013) as in both cases, similar data sets where involved, also proving the effectiveness of the decision tree model.

Research is considering the need and importance of data mining and analysis in respect to disaster management. It is focused on discovering the link of techniques being earlier adopted in related fields, how successful these techniques have thrived in relation to the proposed technique in the area of disaster management. Table 1 gives a summary of papers reviewed and areas of findings with respect to data mining.

III. METHODOLOGY

3.1 Research Techniques and Procedures

The research will aim at creating a balance from results questionnaires to compare and balance the test results of proven secondary data.

The techniques and procedures will involve data collection, data analysis and will also look at possible limitations that the research data collection process faced.

3.2 Data Collection

For the objective of any given or stated research to be achieved, proper identification of right means of data collection is mandatory (Sarantakos, 1994). Various methods are available for data collection base on the idea of the research. It is very important to note the type of primary or secondary data to be used. The research implements both the application of primary data and secondary data. Further discussions on the types and relevance of the data type selection given below.

3.2.1 Primary Data

The main aim of primary data is to produce precise answers towards meeting the required research objectives stated out. Saunders *et al* (2012) described the process as a first-hand information collection method. It can usually be gotten directly from respondents in various ways. It could be in form of structured or semi structured questionnaires, observations, interviews etc. The population of the research matters in the research credibility. Saunders *et al* (2012) identified sampling as the method of getting the right group of individuals needed to collect data from. The primary research has quite a small sample as its targets involved members of organizations specific to the need and use of data and data analysis. An online survey was identified at the suitable format to collect this data due to fact that the research is being carried in the UK while the sample of study is in Nigeria. This is a huge barrier and thus, the electronic survey is the best alternative. The Bristol Online survey will help for the collection of these data.

3.2.2 Primary Data Sources

Non-governmental Organizations in Nigeria are largely involved in the humanitarian activities across the country. The National Emergency Management Agency NEMA has the sole responsibility of information regarding the conditions and data records of the internally displaced camps in the country. Members of staff of various NGOs alongside other relevant organizations views will be collated from questionnaires. Even though Saunders et al (2012) pointed out the need for a high response rate, it is not automatically true that a low sample will give a bias results but its tendency is higher.

3.2.3 Sampling

The purposive sampling method will be used to develop the sample of the research. Saunders et al (2012) identified this method as a non-probability sampling technique, where the sample members are actually selected due to their expertise and full knowledge of the research subject. The sample population identified as members of the stated organizations will be collected through questionnaires as with the help of a senior member of the organization already identified to help in facilitating the collection process. The members selected have a special relationship with data collection and management in the organizations which is the phenomenon under investigation.

3.3 Secondary Data

Zikmund *et al.* (2009) described secondary research as the use of non – human sources and existing information from previous reports which has previously been gathered for a purpose other than the need of the researcher. The process of literature review has helped identified and streamlined focus on understanding research gaps and areas which need to be concentrated on. Secondary data helps understand specifically the nature of the problems in relation to the research questions before comparing with primary data if need be (Saunders *et al* 2012).

3.3.1 Secondary Data Sources

The research through the literature review done identified little previous academic research done in the related areas considering the location of research, despite the challenge, articles from various other perspectives and locations from different backgrounds helped shaped the identification of data sources. The main source of secondary data for the study is from database of responsible organizations identified. Credible websites provide an archive data relating to internal

displacement. This data is being updated regularly by the responsible organizations for research purpose use as this. Well organized Microsoft excel format of the needed data has already been identified ready to be analyzed.

3.4 Use of Software for Data Management and Analysis

One of the aims of the study is to analyze the use of the proposed WEKA software for data mining through feature selection and classification methods. The WEKA software will be used to analyze the secondary data collected. It will look to identify patterns, reoccurrence of events as regard to the sample data. Which will then be further analyzed and discussed. The primary data will be analyzed using the SPSS tool. This will be aimed at understanding views to support the results of the secondary analysis.

3.4.1 Data Mining Process

As earlier stated, classification algorithms and feature selection methods aimed to be tried using the WEKA tool with IDPs dataset collected. Fayyad et al (1996) described that for any data mining exercises, some basic steps need to be followed for the preparation of the data to be mined. This process of preparations is known as the pre-processing, which will access missing values, attributes normalization and minimize dimensionality of the data sets to be analyzed and thus selecting the feature selection method.

The study will implement the Cross – Industry Standard Process for Data Mining (CRISP-DM) model to define a framework for the data mining process which involves six different stages;

1. Business process understanding and determination of the main goals of the process
2. Key data sources identification and collection
3. Data preparation for mining
4. Model technique selection
5. Results evaluation and comparison with different models against initial goals
6. Model deployment.

The framework will be an iterative process to ensure efficiency in results.

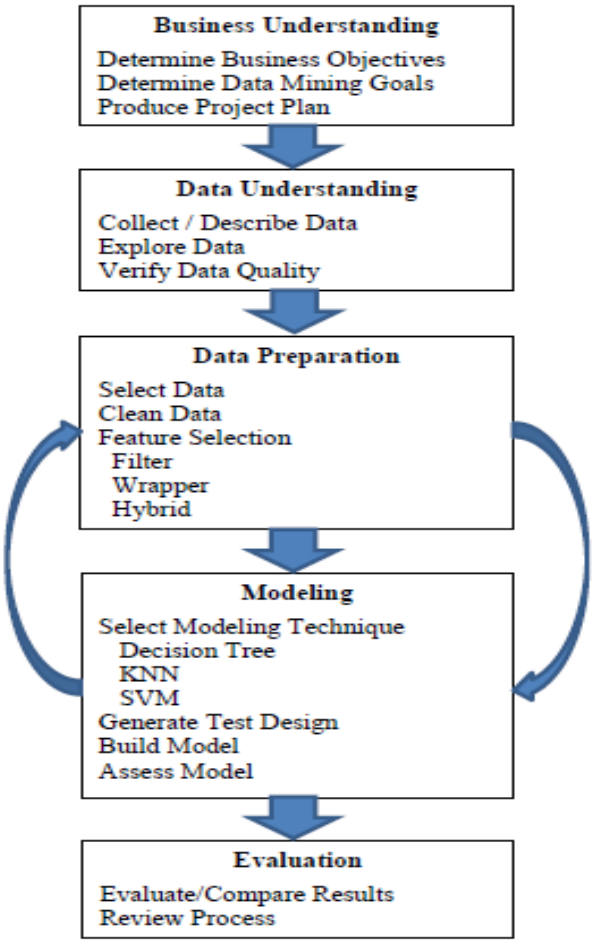

Figure 1 Research Framework

3.4.2 Data and Data Acquisition

The sets of data to be analyzed is the displacement tracking matrix data sets for internal displacement in North East Nigeria. This sets of data contain various attributes. The major objective is to apply feature selection method and also

classification algorithm to help further understand patterns in the data sets, to help humanitarian actors in planning rescue and aid plans.

3.4.3 Data Pre-processing

We briefly discus some of the performance measures as will be tested along the study.

Accuracy

This is a representation of the ratio of the correctly classified results. Accuracy can be calculated using the following equation

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

Similar to the accuracy, sensitivity and specificity can also be calculated using the following formula

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

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

F – Measures

The harmonic mean of recall and precision is known as the F – Measure. Recall refers to the number of positive classified examples over all the positive examples, while precision divides through by all the examples.

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F- Measure} = \frac{2\ X\ precision\ \ X\ recall}{precision\ +recall}$$

3.5 Validity and Reliability of data

The sampling population for this research was not randomly selected, but strictly based on full understanding of the subject matter and the activeness in the process of data analysis in the organization. This was ensured by the support of the management of the organizations as they carefully selected the responding parties. Due to this reason, a large number of sample respondents cannot be achieved. Ethical guidelines were carefully adhered to in the process the data collection and also, anonymity and confidentiality equally considered.

IV. RESULTS

The idea of this research study was to try and investigate the use of data mining in humanitarian industries in North East Nigeria, thereby applying the WEKA classification algorithm on relevant data sets. The research had the following research questions which informed then study:

I. What are the data mining techniques involved in the humanitarian relief process relating to internal displacement datasets?
II. Can the use of data mining tools efficiently facilitate humanitarian aid with internally displaced Persons dataset in North East Nigeria?

In a bid to understanding participants' perceptions and various experiences with different data mining tools with regards to data sets relating to Internally Displaced persons in Nigeria, a questionnaire was formulated and sent out to members of various organizations that are related to handling of such data sets.

The results findings in this chapter is strictly based on the analysis of the results of the survey questionnaire conducted. Further results of the use of Weka data tool will be discussed later.

The chapter will be segmented such that 4.1 discusses the nature of the various organizations involved and responsible for handling and analysis of datasets relating to Internal displacement in Nigeria. 4.2 shows the availability of data involved in data Mining as viewed by respondents. 4.3 discusses response regarding the process of data mining

4.1 Organizational Involvement

A total of twenty-six questionnaires were received from members of various organizations responsible for data analysis in the humanitarian sector relating to IDPs in Nigeria. Two of the questionnaires were not fully completed while the remaining twenty-four were fully completed. The questionnaire was not intended to all members of staff of this organizations, it was exclusively targeted at members of these organizations either working in their IT department or those specifically designated to handle the analysis of this data sets. Thus, the number of respondents was not expected to be very high, as most organizations have few people in this departments. However, a good number of responses was gained, Table 2 shows the participants and their relationship ratio to involvement in the handling of data.

Table 1 Organizational Participation

Is your organisation involved in the handling and analysis of IDPs data set in Nigeria

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	19	73.1	79.2	79.2
	No	5	19.2	20.8	100.0
	Total	24	92.3	100.0	
Missing	System	2	7.7		
Total		26	100.0		

In order to be certain about the responses and its credibility, respondents needed to identify whether or not they are involved in handling and analysis of IDPS dataset in Nigeria. Five out of those who completed the survey confirmed their non-involvement in handling and analysis of this data sets which makes up 19.2% of the total. Nineteen respondents making up 73.1% confirmed full association with the processes of analysis of IDPs data set and two respondents did not complete the questionnaire properly. The involvement of participants in this context means; they have directly been involved or are majorly responsible for the data mining processes in their various organizations. Due to the need for special skills required for these processes, it was predicted that the respondents number will not be very high, thus the reason for the low number of respondents

4.2 Availability of Data for Analysis

Table 2 Availability of data sets

IDPs Data set is readily available for analysis

		Responses		Percent of Cases
		N	Percent	
Analysis	Mostly Agree	4	17.4%	17.4%
	Agree	15	65.2%	65.2%
	Neither agree nor disagree	2	8.7%	8.7%
	Disagree	2	8.7%	8.7%
Total		23	100.0%	100.0%

To successively achieve the success of any data analysis, the availability of such data set has to be ascertained. The survey sought the opinions of the respondents on the views regarding the availability of access to the required datasets relating to IDPs data for adequate analysis. Over 80% of the respondents confirmed the readiness and availability of this sets of data for the required analysis. In a further bid to understand the relation between data analysis and the availability of data, the survey collected ideas on respondents' views of how the availability of data can affect the process of analysis.

Table 3 Availability of data for analysis

The availability of data affects the data analysis process

		Responses	
		N	Percent
response	Mostly Agree	6	26.1%
	Agree	10	43.5%
	Neither agree nor disagree	3	13.0%
	Disagree	3	13.0%
	Mostly Disagree	1	4.3%
Total		23	100.0%

Table 3 gives the response rate of participants, about 70% agree with the claim that the availability of data affects the process of the analysis, while less than 10% are either neutral in opinion or in disagreement.

These results give a reasonable understanding on the availability and usefulness of relative data sets to proposed analysis in IDPs data sets in Nigeria. The responses also acknowledge the reliability in the data collection process showing that most respondents agree to the fact that the process of data collection is reliable enough.

4.3 Data Mining Process

The survey looked to understand the present data mining tools employed in the various organizations involved in managing data related to IDPs in Nigeria. Question was asked to understand the present tools used in data mining and some of the tools identified are as followed; Microsoft Excel, SPSS, SQL Server, Adobe Illustrator, Tableau, SPATA, EPIDATA. These samples were particularly derived from the questionnaires collected to gain more insight on the present use of these tools. How effective the use of the present tools has been and also identify if need be for an alternative tool to be adopted in this sector.

The respondents were asked if they considered it necessary for the introduction of a new and simplified tool in the circumstance that the presently used proves inefficient. Over 80% of the respondents agreed to the claim that a more simplified tool was very much needed to aid the data mining process. The results also there is sometimes not adequate results gotten from the predicted values. As a lot of prior training is required for the use of these tools, respondents identified the desire for a less training involved tool which is also efficient enough.

Table 4 Need for Data mining technique

There is a need for a more simplified data mining technique

		Responses	
		N	Percent
Response	Mostly Agree	7	29.2%
	Agree	13	54.2%
	Neither agree nor disagree	2	8.3%
	Disagree	2	8.3%
Total		24	100.0%

One of the major findings identified is the time factor in analyzing data sets collected. Most of the respondents identified the use of their present software as time consuming and in a lot of ways not user friendly, thus the assumption for the high response rate on the need for a simplified process. For a more understanding on the needs of the IDPs, a more simplified and more efficient system need to be engage.

4.4 Technology in Humanitarian Response

Respondents fully identified the need for more involvement of technological factors in the area of humanitarian response. Survey went further to ascertain the fact that the introduction of more IT to humanitarian aid will help improve the situation. Furthermore, it was also identified that a new approach to the way data is being processed will help further understand more about the conditions of IDPs in Nigeria.

4.4.1 Results from Data Mining

This section will present the application of feature selection method and will further show the application of the classification algorithms on the data sets

4.4.2 IDPs Data Set

The IDP data sets used in the research was obtained from IOM DTM Nigeria repository (Anon. 2016). The sets contain records of crisis events in North east Nigeria. The data is available in the quarterly report presented by the International Organization for Migration. The data sets were segmented into various variables. This data sets where available in Microsoft xls format. The WEKA data mining tool only accepts the arff format. Different conversion processes were involved. Data sets had to be prepared for analysis.

4.4.3 Information Gain

The IDPs Data set was first and foremost tested with a feature selection known as the *Information gain*, this was to analyze the entropy of the features in each of the classes. This resulted in a list of features all ranked by their level of importance. The list as shown in the table below shows *information gain* scores as regard features in order of their importance.

Table 5 Information Gain on IDP data set

Rank	Information Gain	Attribute No	Description
1	0.258106	21	State_Name
2	0.075447	23	LGA_Name
3	0.073223	28	Est_no_of_households_displaced
4	0.053002	1	Est_no_of_individuals_displaced
5	0.045322	15	Reason_for_displacement-community clash
6	0.034355	24	Reason_for_displacement-Natural
7	0.023228	11	Other_reason_or_displacement
8	0.018985	29	latitude
9	0.014567	33	longitude

Osiris (2015) discussed that any Information gain value that has a value which is greater than zero implies some level of significance. In this study, only a total of nine features were ranked. Results shows that the attribute "state_name" has an Information gain of 0.258, which is about three times greater value than the next attribute, "LGA_name" and "Est_no_of_households _displaced".

4.4.4 Relief- F Feature Selection Method

The next feature selection method analyzed was the Relief-F feature selection method, the table below shows the result. The features were all ranked in descending order as regards to the used metric. It is observed from the result that attribute "Est_no_of_households" now takes the highest ranking with an approximate value of 0.045.

Table 6 Relief-F feature selection method

Rank	Information Gain	Attribute No	Description
1	0.044672416	28	Est_no_of_households_displaced
2	0.042455498	21	State_name
3	0.032983123	28	LGA_Name
4	0.030981916	1	Est_no_of_individuals_displaced
5	0.025322368	15	Reason_for_displacement-community clash
6	0.020355322	24	Reason_for_displacement-Natural
7	0.013228767	11	Other_reason_or_displacement
8	0.008985323	29	latitude
9	0.004567357	33	longitude

4.4.5 Correlation – Based Feature Selection

This is the final filter type used on the IDP data sets. This method analyzes all the features of the data set and gets a combination of features that can adequately analyze and give a good predictive result. This is done to help reduce redundancy between the features. In this test, the correlation based feature selection method reduced the number of attributes to five.

Table 7 Correlation Based Feature selection method

Attribute
State_name
Est_individuals_displaced
Reason_for_displacement
Latitude
Longitude

Looking closely at the table, it is noticed that the attributes "state_name" and "estimated_individuals_displaced" both had high rankings in the other feature selection methods.

Wrapper Selection method

This was the last feature selection method performed on the data set. During this test, feature reduction was applied using the classifier as part of the process. After the application of the wrapper selection method, Table 4.6 displays the results on each of the classification method selected.

Table 8 The Wrapper selection method

K - NN	Decision tree	SVM
Est_no_of_households_displaced	Est_no_of_households_displaced	Est_no_of_households_displaced
LGA_Name	LGA_Name	LGA_Name
State_name	State_name	State_name
Est_no_of_individuals_displaced	Other_reason_or_displacement	Est_no_of_individuals_displaced
Reason_for_displacement-community_clash	latitude	Reason_for_displacement-community_clash
Reason_for_displacement-Natural	longitude	Reason_for_displacement-Natural
Other_reason_or_displacement	Est_no_of_individuals_displaced	
latitude	Reason_for_displacement-community clash	
longitude	Reason_for_displacement-Natural	

It can be noticed that attributes like "state_name" and "est_no_of_households_displaced" maintain consistence in significance throughout the classification process. The next step is to run the classification algorithms on the data sets.

4.4.6 Decision Tree

As earlier proposed, the decision tree was tested using the J48 algorithm in WEKAwith the original data together also with the selections made by the feature selection methods; information gain, Relief-F and Correlation – Based selection methods. Different tests were taken, with the first using the default WEKA settings. A process which has not less than two instances per leaf with a 0.25 confidence factor. Results seen in the table below:

Table 9 Decision Tree classifier performance across IDP data sets

Data set	Accuracy	AUC	F- Measure	TP Rate	TN Rate
J48ALL	0.823	0.826	0.789	0.821	0.873
J48IG	0.821	0.827	0.827	0.801	0.831
J48RLF	0.858	**0.901**	0.858	0.834	0.882
J48CFS	0.821	0.882	0.881	0.802	0.840
J48WRP	0.862	0.886	0.814	0.839	0.886

J48ALL – Using all the Features
J48IG – Using the Features selected by Information Gain method
J48RLF – Using the Features selected by the Relief – F method
J48CFS – Using the Features selected by the Correlation – Based feature selection method
J48WRP – Using the Features of the Wrapper method

It was noticed that the reduction of feature set has improved the accuracy. But also, the AUC rates where all data sets were reduced, was seen to increase in all the cases. The wrapper method further increased the accuracy and also the F – Measure from 78.9.7% to 81.4%. it also increased the AUC from 82.6% to 88.6%. The wrapper method displayed the highest rate of accuracy, F- measure, and sensitivity/ specificity during the tests.

The test showed that the use of the wrapper feature selection method resulted to the confusion matrix for the J48 classification with the highest accuracy rate.

Table 10 Confusion Matrix for decision tree Algorithm using J48 wrapper data set

NO	YES	<--Classified as
437(TP)	82(FN)	No
65(FP)	484(TN)	Yes

The performance measure for the decision Tree algorithm is also shown below:

Table 11 Decision tree Algorithm performance measure using J48 Wrapper data set

Accuracy	0.842
Precision*	0.841
Recall*	0.842
F-Measure*	0.842
TP Rate	0.876
TN Rate	0.819

*Weighted average

4.4.7 K - Nearest Classifier (K-NN)

The K-NN algorithm test was performed by choosing *K* value of 1, 5 and 10, with parameter search with *k* values from 1 to 10 with an increment of 1. This was done in terms of Accuracy, AUC, F-Measure, TP Rate and TN rate evaluations.

Table 12 performance of K-NN classifier across the IDP data sets in terms of accuracy, AUC, F-Measure, TP Rate and TN Rate

Data Set	Accuracy	AUC	F-Measure	TP Rate	TN Rate
1NNALL	0.735	0.801	0.735	0.787	0.684
1NNIG	0.776	0.775	0.776	0.792	0.759
1NNRLF	0.807	0.807	0.807	0.810	0.804
1NNCFS	0.762	0.768	0.761	0.767	0.755
1NNWRP	0.796	0.801	0.796	0.806	0.787
5NNALL	0.765	0.87	0.779	0.787	0.684
5NNIG	0.735	0.898	0.765	0.835	0.782
5NNRLF	0.843	0.867	0.806	0.839	0.845
5NNCFS	8.809	0.865	0.843	0.052	0.810
5NNWRP	0.849	0.908	0.852	0.832	0.866

The test showed that wrapper feature selection method and a *k* value of 5 resulted with an accuracy and F- Measure of 85.2% with an AUC of 90.8%.

The confusion matrix for the K – Nearest Neighbor algorithm using $k = 5$ is shown in the table below.

Table 13 Confusion matrix for Nearest Neighbor using k = 5

No	**Yes**	**<- classified as**
399 (TP)	90 (FN)	No
71 (FP)	478 (TN)	yes

The table below shows the performance measures for the Nearest Neighbor after using the value k =5.

Table 14 performance measures for K-NN

Accuracy	0.824
Precision	0.824
Recall	0.824
F-Measure	0.824
TP Rate	0.843
TN Rate	0.832

Results showed the accuracy rate of 82.4% when using the parameter search with the *K* values between 1 to 10 and also AUC of 90.5% when using attributes selected by the Relief -F method of *k* = 5

4.4.8 Support Vector Machine

This was the last classification algorithm to be performed with the data sets. Just like the other classifications, it was done using the default WEKA parameters. Performance measures were tested through the data sets in terms of accuracy, AUC, F-Measure, TP rate, TN rate evaluation statistics. Table below show the result

Table 15 performance of the LibSVM classifier on IDP data sets across accuracy, AUC, TP rate, TN Rate, F-Measure evaluation

Data Set	**Accuracy**	**AUC**	**F-Measure**	**TP Rate**	**TN Rate**
LibSVMALL	0.785	0.785	0.784	0.860	0.71
LisSVMIG	0.814	0.814	0.813	0.872	0.756
LibSVMRLF	0.819	0.819	0.819	0.863	0.775
LibSVMCFS	0.785	0.785	0.783	**0.890**	0.680
LibSVMWRP	0.818	0.818	0.818	0.854	0.782

The best Accuracy rate for each classifier was determined after all the test was ran. Table 17 shows the J48 decision tree model with accuracy of 89.3%, then the Support Vector Machine with an accuracy of 83.8%, followed by the K- Nearest Neighbor with 84.9% accuracy all of them across a 10 – fold cross validation.

Table 16 Overall classification Accuracy on IDP data set based on individual runs

	K-NN	**Decision Tree**	**SVM**
IG	0.810	0.840	0.818
CFS	0.815	0.822	0.815
RLF	0.842	0.860	0.850
WRP	0.849	0.893	0.838

V. DISCUSSION

5.1 Discussion of Results

The study is guided by the following research questions, and it has been answered in the following way:

Research question: Can the proper adoption of data mining help humanitarian efforts?

The major organizations responsible for the analysis of humanitarian data sets with respect to Internal Displacement in Nigeria were involved in the study. It was ensued that from the responses collected, only personnel responsible for data analysis in the organizations responded to the questionnaire. The effective application of data mining is highly dependent on the availability of the data (Zheng, Shen and Tang et al. 2011). Results from analysis showed that the availability of data is a major determinant in the process of data mining for IDP data set. 80% of respondents in the research agreed to the claim that data availability helps faster analysis of data mining. They also agreed that there is adequate availability of data for analysis in the humanitarian sector in Nigeria. Lu, Bengtsson and Holme (2012) in a descriptive analysis of the 2010 earthquake in Haiti portrayed how the use of data mining for predictive analysis was used to help improved humanitarian efforts. Thus, it is seen that with the sufficient availability of data, and a proper implementation of data mining for pattern recognition, humanitarian efforts can be greatly improved, as there will be more understanding of events and occurring patterns. The feature selection and classification algorithms proved eminent from the experiments in reducing attributes of data sets thereby enhancing the accuracy of performance measures. This preprocess will greatly aid in the data mining of humanitarian data sets and thus aid humanitarian efforts.

Data Mining process understanding

The present tools used by humanitarian agencies for data mining was collected and recorded that due to complexity of use of some or most of the tools, a need for a more simplified approach was recognized. The research participants clearly identified the following reasons considered to be challenges for using the tools;

1. Complexity of use
2. Time consuming
3. Long training period

Over 80% of participants that recognized the need for a more simplified tool for analysis shared the opinion that a more user friendly and less time consuming tool will highly improve the timely response of the analysis.

T. Akomolafe and Olutayo (2013) discovered patterns of reoccurring accidents along some major roads in Nigeria using the WEKA data mining tool over a certain period of time. Data sets from IDP data was then tried using the following criteria and the results obtained discussed as follows;

Information Gain

The Information Gain was used to calculate the entropy of feature in each class, this was seen in rankings as from Table 6 for any information gain to show some level of significance, it has to show a value which is above zero, thus from the results, the attribute "state_name" had an information gain of 0.258, it had the highest value in the table this tells us how important this attribute is to the given set of data. Because the attribute possesses the highest value, it is derived to be the most significant. The next feature selection method applied was the Relief- F feature selection method and then followed by the wrapper selection method. Results were drawn and drafted in tables as seen from the results chapter.

After the selection methods were applied, the Decision tree classification algorithm was then further tried on the data which showed that by implementing the wrapper method, the accuracy and F-measure was further increased to 86.2% from initial 84.7%, the AUC also witnessed an increase from 86.2% to 89.9%. this accuracy level shows that the wrapper method has proved in terms of accuracy and sensitivity most reliable. (Figure 9) gives a visual representation of these accuracy graph. The experiment reveals that using the wrapper selection method, a J48 classification confusion matrix showed the highest accuracy rate (see Table 12)

Performance Evaluation

The analyses evaluate how the feature selection process affects the classification performance. Some of the performance measures used include Classifiers accuracy (ACC), Area under Curve (AUC), F- Measures, TP Rate and TN Rate. These performance evaluators were seen to be compared to each other at various times in the results of the classifications. Results at various algorithm classification displayed the performance measures thereby analyzing the percentage of each.

Chawla (2005) describes the confusion matrix as a table containing information regarding the real and the predicted classifications for any algorithm. Each of the evaluated gave a corresponding confusion matrix. This was to classify the actual value against the predicted value for all the classifications.

Accuracy

As stated earlier, the accuracy represents the correctly classified results in percentages. The higher the accuracy, the higher the performance of the classification model.

The True Positive Rate also known as sensitivity and the True Negative Rate (Specificity) of each of the classification algorithm was determine using the confusion matrix. This further strengthens the accuracy level.

It was observed from the results that there is a consistency in the selection of the significant attribute across the different selection methods.

VI. CONCLUSION

6.3 Conclusions

The study showed that feature selection method application used on the data sets displayed a mixed sets of results. The IDP data set as when applied the information gain, the Correlation Feature Selection and also the Relief – F methods was seen to have increased the accuracy of the results compared to the when using only the K – Nearest Neighbor classification algorithm. Similarly, the Correlation feature selection and the information gain methods displayed a bias to the positive class as indicated when the higher true positive rates where compared to the lower true positive rate values. This was as a result of the imbalance in the data set. The IDP data sets in Nigeria as identified continuously require better classification algorithm to determine more accurate situations and data mining results.

REFERENCES

[1] Aha, D. W. and Bankert, R. L. (1996) 'A comparative evaluation of Sequential feature selection Algorithms'. in *Lecture Notes in Statistics*. Springer Science + Business Media, 199–206
[2] Anon. (2015) *Global Ove view 2015no* [online] available from <http://www.internal-displacement.org/assets/library/Media/201505-Global-Overview-2015/20150506-global-overview-2015-en.pdf> [22 May 2016]
[3] Anon. (2016) *Publications* [online] available from <http://nigeria.iom.int/dtm> [13 August 2016]
[4] Anon. (2016) *DTM | International Organization for Migration* [online] available from <http://nigeria.iom.int/dtm> [11 August 2016]
[5] Anon. (2016) *Machine Learning Project at The University of Waikato in New Zealand* [online] available from <http://www.cs.waikato.ac.nz/ml/> [11 August 2016]
[6] Anon. (2016) *IDMC Grid 2016 - Global Report On Internal Displacement* [online] available from <http://www.internal-displacement.org/globalreport2016/#home> [11 August 2016
[7] Bryant, A. and Charmaz, K. C. (eds.) (2007) *The SAGE handbook of grounded theory*. Los Angeles: Sage Publications
[8] Burns, R. B. (2000) *Introduction to research methods*. 4th edn. Thousand Oaks, CA: SAGE Publications
[9] Council, I. (2016) *IDMC » Global Overview 2015: People Internally Displaced by Conflict and Violence*[online]availablefrom<http://www.internaldisplacement.org/publications/2015/global-overview-2015-people-internally-displaced-by-conflict-and-violence> [11 August 2016]
[10] Council, I. (2016) *IDMC » Nigeria IDP Figures Analysis* [online] available from <http://www.internal-displacement.org/sub-saharan-africa/nigeria/figures-analysis> [1 August 2016]
[11] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (n.d.) *Knowledge discovery and data mining: Towards a unifying frameworkno* [online] available from <http://www.aaai.org/Papers/KDD/1996/KDD96-014> [21 May 2016]
[12] Ganguly, A. R. and Steinhaeuser, K. (n.d.) *Data mining for climate change and impacts*. [online] 385–394. available from <http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=4733959> [21 May 2016]
[13] Gao, H., Barbier, G. and Goolsby, R. (2011) "Harnessing The Crowdsourcing Power of Social Media for Disaster Relief". *IEEE Intell. Syst.* 26 (3), 10-14
[14] Hristidis, V., Chen, S.-C., Li, T., Luis, S., and Deng, Y. (2010) 'Survey of data management and analysis in disaster situations'. *Journal of Systems and Software* 83 (10), 1701–1714
[15] Jain, A. and Zongker, D. (1997) 'Feature selection: Evaluation, application, and small sample performance'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (2), 153–158
[16] Jain, S. and McLean, C. (2006) "An Integrating Framework for Modeling and Simulation for Incident Management". *Journal of Homeland Security and Emergency Management* 3 (1)
[17] Kandel, A., Tamir, D. and Rishe, N. (2014) "Fuzzy Logic and Data Mining In Disaster Mitigation". *Improving Disaster Resilience and Mitigation - IT Means and Tools* 167-186
[18] Ladan and Tawfiq, M. (2015) *Strategies for adopting the national policy on IDPs and Domesticating in Nigeria the African union convention for the protection and assistance of IDPs in Africa by Muhammed Tawfiq Ladan: SSRN*. [online] available from <http://ssrn.com/abstract=2649377> [22 May 2016]
[19] Li, L., Wang, J., Leung, H., and Zhao, S. (2012) 'A Bayesian method to mine spatial data sets to evaluate the vulnerability of human beings to catastrophic risk'. *Risk Analysis* 32 (6), 1072–1092
[20] Lu, X., Bengtsson, L. and Holme, P. (2012) "Predictability of Population Displacement After The 2010 Haiti Earthquake". *Proceedings of the National Academy of Sciences* 109 (29), 11576-11581
[21] Peng, Y., Zhang, Y., Tang, Y. and Li, S. (2011) "An Incident Information Management Framework Based On Data Integration, Data Mining, And Multi-Criteria Decision Making". *Decision Support Systems* 51 (2), 316-327
[22] Saunders, M. N. K., Lewis, P., and Thornhill, A. (2012) *Research methods for business students (6th edition)*. 6th edn. Harlow, England: Financial Times Prentice Hall
[23] T. Akomolafe, D. and Olutayo, A. (2013) 'Using data mining technique to predict cause of accident and accident prone locations on highways'. *American Journal of Database Theory and Application* 1 (3), 26–38
[24] V.A, O. and A.A, E. (2014) 'Traffic accident analysis using decision trees and neural networks'. *International Journal of Information Technology and Computer Science* 6 (2), 22–28
[25] Zheng, L., Shen, C., Tang, L., Li, T., Luis, S. and Chen, S. (2011) "Applying Data Mining Techniques to Address Disaster Information Management Challenges On Mobile Devices". *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*
[26] Zikmund, W. G., Carr, J. C., Griffin, M., Babin, B. J., and Bab.., B. J. (2009) *Business research methods (with Qualtrics card)*. 8th edn. United States: South-Western Cengage Learning

