EJTAS ISSN 2786-7447
EUROPEAN JOURNAL OF THEORETICAL AND APPLIED SCIENCES
OPEN ACCESS
PEER REVIEWED

Flood Endangered Area Classification Using the K-Nearest Neighbour Algorithm

Oghenevovwero Zion Apene
Centre for Satellite Technology Development, National Space Research and Development Agency, Abuja, Nigeria

JohnPaul A.C. Hampo
Department of Computer Science, Federal University of Technology, Owerri, Imo State, Nigeria

Clement Omamode Ogeh
Department of Computer Science, Delta State University, Abraka, Delta State, Nigeria

Suleiman Usman Hussein
Centre for Satellite Technology Development, National Space Research and Development Agency, Abuja, Nigeria

Suggested Citation

Apene, O.Z., Hampo, J.A.C., Ogeh, C.O. & Hussein, S.U. (2023). Flood Endangered Area Classification Using the K-Nearest Neighbour Algorithm. *European Journal of Theoretical and Applied Sciences, 1*(5), 1051-1061.
DOI: 10.59324/ejtas.2023.1(5).92

Abstract:

Preparing for the uncertainty of life is one aspect of the human existence that cannot be over emphasized. With the growth of technology especially the sophisticated nature of data mining and machine learning algorithms, these uncertainties can be predicted, planned and prepared for using existing variables and computer methodologies. The achievements and accomplishments of big data analytics over the past decade in diverse areas called for its implementation in meteorological and space data. Notably, enhancement of the proper management of life's uncertainties when they eventually occur. This research work focuses on the classification of areas within the Nigerian Geographical territory that are prone to flood using the K-nearest neighbour Algorithm as a classifier. Data from Nigeria Meteorological Agency (NiMET) on seasonal rainfall prediction and temperature of different stations and cities for over three (3) years (2014-2017) was used as a dataset which was trained and classified with the k-Nearest Neighbour algorithm of machine learning. Results showed that some areas are prone to flood considering the historic data of both rainfall and temperature.

Keywords: *K-Nearest Neighbour, Meteorological data, Weather Analysis, Data mining, Machine learning.*

Introduction

One of the most devastating natural disasters which causes massive loss of human life, property, infrastructure and agriculture thereby affecting the socioeconomic activities of a particular geographical zone is flood (Mosavi, Ozturk, & Chau, 2018).

Government agencies are therefore in constant search for methods of identifying risk prone areas for the development of policies that will prevent flood, protect high risk areas and help prepare for the eventualities of a flood disaster.

With the tremendous advancement in Computer science and information technology, Machine learning and Artificial intelligence have been

applied in various aspect of human endeavours and disaster management is not an exception. Flood prediction models have been developed to predict hydrological events such as flood high risk areas using hydrological time series data like rainfall, temperature, storm, soil water level, humidity etc. for water resource management strategies, policies, suggestions and analysis (Danso-Amoako, et al., 2012).

Flood prediction can be divided into long and short term to enable long and short term decision making systems by policy makers. Flood prediction is however, a very dynamic activity due to unstable nature or climate and weather conditions. Therefore, modern day flood prediction models are mainly specific data with simplified assumptions that can be expressed mathematically.

Background

Human industrialisation, exploration and mining activities among others degrades the ozone layer and bedrock hence, the good weather conditions of yester years are fast becoming a story. Weather is simply the atmospheric state specifying the temperature (hotness or coldness), humidity (wet or dry), cloud (clear or cloudy) and wind (clam or stormy) and precipitation, over a short period of time. When large water bodies move out of their boundary or bank into another reason is termed flood. Flood is caused mostly by human activities either in destruction of the water boundary or in exploration and gas flaring. Meteorological data are facts about the atmosphere. These data are geographical, geophysical and geochemical in nature.

Machine Learning

In the design and creation of machine learning methods for flood prediction, historic time series data are usually used. These data are collected using a variety of tools by weather scientist. Some of these tools including rain gages and remote sensing tools used for real time data capturing like Satellites, Airborne laser and weather radar (Hwa, n.a.).

Flood prediction models built based on radar rainfall data have proven to have higher accuracy (Maddox, et al., 2002). The historic data can be in years, months, weeks and days divided into sets to construct and evaluate models. To achieve that, the individual set data undergo the training stage of the model, validation, verification and testing stage of the model. The major Machine Language algorithm that have been used for flood prediction are: Artificial Neural Network (ANN), Neuro-Fuzzy Logic, Adaptive Neuro-fuzzy Inference System (ANFIS), Support Vector Machine (SVM), Wavelet Neural Network (WNN) and Multilayered Perceptron (MLP). The chart below describes the flow of activities in creating a Machine Learning Model.

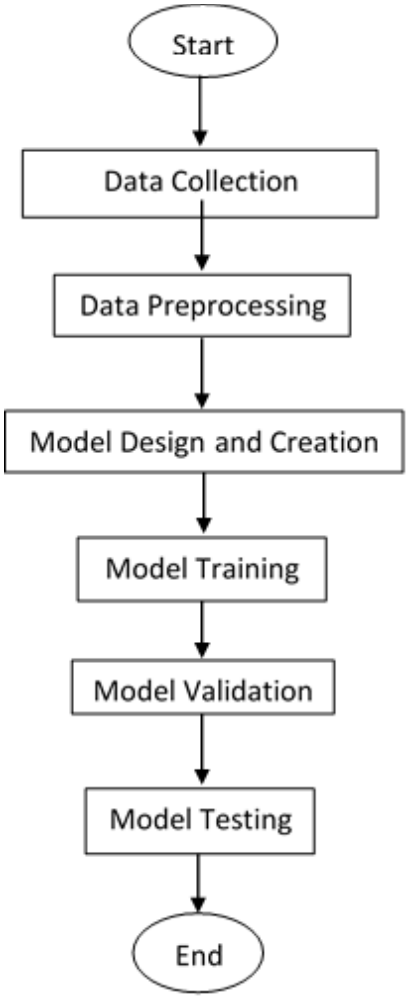

Figure 1. The Flow of Activities for Creating a Machine Learning (ML) Model
Source: Kumar, et al., 2019

In this study, the K- Nearest Neighbor (KNN) Algorithm was used to create the Machine Learning Classifier of flood prone areas in Nigeria using historic data of rainfall and temperature from NIMET.

Data Mining

The growth of information technology has resulted to vast amount of data generated and stored in large database and data warehouses. These data are useless, unless explored for productive decision making by management of various organization at different levels.

Data mining is a term that refers to the various processes involved in the extraction of relevant and objective information from a large amount of data via learning from experience and intuition. This is also referred to as knowledge discovery or knowledge mined from data (Kumar, et al., 2019; Point, 2015).

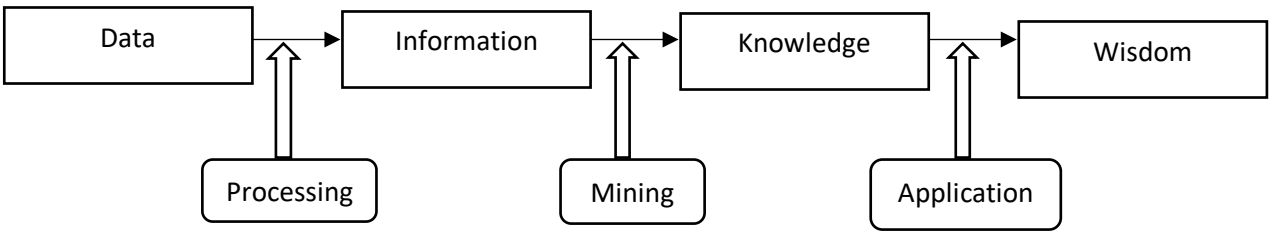

Figure 2. Relationship between Data, Information, Knowledge and Wisdom

Data mining involves processes that start with data as inputs that are processed into information. This is usually called the preprocessing stage that includes variable selections and data cleaning. The information derived is further explored for implicit patterns and rules that were previously unknown. This exploration is done in order to learn from experience the patterns existing in the data. This learning from experience is what is termed as Knowledge (Surbakti, 2015: Liew, 2013). The knowledge gathered are further applied in solving complex problems. This application of knowledge is termed wisdom. The knowledge derived from data mining can be applied in market analysis, prediction, fraud detection, time series mining, Optimizations etc.

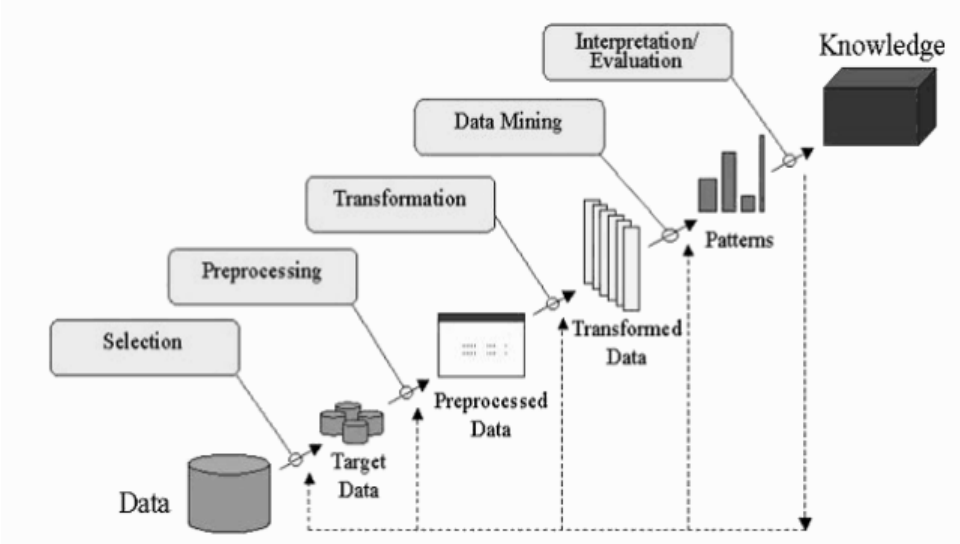

Figure 3. Knowledge Discovery process
Source: Nicholas, 2019

The major aim of data mining is to find patterns and rules that were previously unknown in order to make decisions that will help develop businesses through proper decision making. Data mining involves the following stages as opined by Ramageri, & Bharati (2010).

1. Data Exploration: This involves the refining and identification of data to be mined. This is known as the preprocessing stage.

2. Identification of Pattern: This involves the use of some techniques to extract patterns or rules from the preprocessed data. The patterns discovered is termed as knowledge.

3. Deployment of knowledge: This is the application of knowledge to give desired results.

Data Mining Techniques and Algorithm

Some of the techniques and algorithm used in in knowledge discovery are:

1. Classification: Classification is one of the most used data mining techniques. It involves the use of pre-classified samples to create a model that can classify the population of records at large bases on identified parameters. Fraud detection and credit-risk applications are typical examples or the classification problem. Algorithms are used to encode parameters into a model called a classifier (Ramageri, & Bharati, 2010). Example of these classification models are: Classification by decision tree induction, the K-Nearest Neighbour (K-NN) Classifier, Bayesian Classification Neural Networks, Support Vector Machines (SVM) and Classification Based on Associations.

2. Clustering: Clustering is the identification of objects of similar classes. The clustering techniques identifies dense and sparse regions in an object space and discovers the overall distribution pattern and correlations among data attributes. The grouping of customers based on purchasing patterns and the categories genes with similar functionality are examples of the clustering problem. Some clustering methods are: Partitioning Methods and Grid-based methods.

3. Prediction: Regression analysis can be used to model the relationship between one or more independent and dependent variables. In data mining independent variables are attributes already known and response to the variables are predicted. However most real-world problems are not easy to predict. Point (2015) for example, stock market prices and volumes of sale. Hence, complex techniques like the logistic regression, decision trees, or neural nets may be necessary to forecast future values. Examples of algorithms used for prediction are Neural Network (NN), Support Vector Machine (SVM) and Fuzzy logic. Hybrid algorithms can be used to improve the prediction capability of a model.

4. Association rule: Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis.

5. Neural Networks: Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques (Ramageri, & Bharati, 2010). These are well suited for continuous valued inputs and outputs. They are used in Recognition problem, prediction problems.

Related Papers

A study of UNISR, 2005 described flood as one-third of all the natural disasters in the world and along with its counteract, storm comprises 77% of economic loss caused by extreme weather events

In the study of (Elsafi, 2014), Artificial Neural Network (ANN) was used as a predicting tool for area where flood occurrence might likely

occur. Readings from different points along the Dongola Station, River Nile, Sudan was used between 1965 and 2003. The study used Artificial Neural Network (ANN) model to forecast flooding along the Nile River with data from the Sudan Ministry of Irrigation using daily readings from different stations. Root Mean Squared error was used to examine the performance of the model.

The study of Mosavi, Ozturk, & Chau (2018), empirically reviewed flood prediction using machine models. Findings from the study showed that the most commonly used machine learning method from 2008 to 2017 for flood prediction are Artificial Neural Network (ANN), Support vector machines (SVM), Multilayer perceptron (MLPs), Decision tree (DT), Adaptive euro-fuzzy inference system (ANFIS), Wavelet-based neural network WNN) and Ensemble prediction systems (EPS).

Michael & Patience, (2018) applied Artificial Neural Network model for flood prediction in Nigeria using Deep feed- forward neural network with two inputs, rainfall and temperature, back propagation algorithm was used to train the neural network. The study that concluded that flood prediction system is very important to help prepare and plan for flood likely events.

Similarly Kim, et al., (2016) presented a real-time forecasting model for after-runner storm surges along the Tottori Coast in Japan. They trained an artificial neural network (ANN) model using historical data of after-runner storm surges, sea surface temperature (SST), and wind direction and speed. The ANN model was found to be effective in real-time forecasting of after-runner storm surges and outperformed traditional statistical methods. The study's findings indicated that the ANN model can be useful in improving coastal disaster management and response by providing accurate predictions of after-runner storm surges with a 5-hour lead time.

Tehrany, Jones, & Shabani, (2019) compare the precision of two datasets for flood susceptibility analysis in Brisbane, Australia. The first dataset (DS1) is based on LiDAR-derived topographical and hydrological factors and the second (DS2) includes additional factors like geology, soil, LULC, and proximity to roads and rivers. DT and SVM machine learning models were used to analyze flood susceptibility and compares the outcomes. The study findings showed that adding extra conditioning factors beyond LiDAR-derived factors did not increase precision and DS1 alone was sufficient. The study also found that altitude is a major variable in flooding and the most accurate susceptibility map was produced by the DS1-DT model with a 91.22% success rate and 88.47% prediction rate.

Lai et al., (2016) propose a new approach to flood risk zoning using the ant colony algorithm based on rule mining (Ant-Miner) to map the regional flood risk in the Dongjiang River Basin in Southern China at a grid scale and improve accuracy and simplicity compared to traditional methods such as decision tree (DT) and random forest (RF) and fuzzy comprehensive evaluation (FCE). The results show that the Ant-Miner method has higher accuracy, generates simpler rules, and reduces implementation steps and computing time.The proposed Ant-Miner method according to the authors offers a novel approach for flood risk zoning, flood risk management, prevention, and reduction of natural disasters in the study area.

Xiong, (2019) presented a model for assessing and mapping flash flood vulnerability using Geographic Information System (GIS) and Support Vector Machine (SVM) techniques in their paper "A GIS-based support vector machine model for flash flood vulnerability assessment and mapping in China". Historical data on flash floods and environmental variables, such as topography and land use, were used to train the SVM model. Subsequently, the model was applied to several regions in China to assess flash flood vulnerability and produce maps of flood-prone areas. Findings from the study revealed that GIS-SVM model was effective in identifying flash flood-prone areas, and that topography and land use were important factors affecting flash flood vulnerability and concluded that GIS-SVM

model is a useful tool for flash flood risk assessment and management in China.

From the foregoing, based on the reviewed literature, the researchers applied ANN and SVM models mostly and a few other models for flood classification and prediction, and are not tailored to Nigeria environment, KNN also was not explored for flood classification. This study therefore focussed on flood endangered area classification with kNN algorithm using Nigeria meteorological data (rainfall and temperature) as significant input.

Methodology and Implementation

The k Nearest Neighbour (kNN) algorithm of machine learning is a supervised learning method which learns from labelled or known data. The counterpart of kNN is the kMeans algorithm but this is an unsupervised learning which learns from an unlabelled or unknown data. In classifying the flood endangered areas using kNN algorithm, we used data on temperature and rainfall from the Nigeria Meteorological Agency's website (NiMet, 2023).

Table 1. Supervised and Unsupervised Learning

S/N	Supervised	Unsupervised
1.	Decision Tree	k-Means
2.	k-Nearest Neighbour	Hidden Markov Model
3.	Naïve Bayes	Principal Model Analysis
4.	Artificial Neural Network	Gaussian Mixture Model
5.	Linear Discriminant Analysis	Hierarchical clustering
6.	Rule-Based Classifiers	Spectral clustering

The above table shows some examples of algorithms or methods used in supervised and unsupervised learning.

The data got was cleaned and transformed, producing different datasets for temperature and rainfall on Microsoft Excel CSV format which is comma delimited. A total of three (3) datasets was used and these are NiMET dataset on temperature for NiMET data collection stations from 2015 – 2017, NiMET dataset on Rainfall for NiMET data collection stations from 2014 – 2017 and NiMET dataset on Rainfall for States and Cities from 2014 – 2017. The datasets were exported to IBM SPSS Statistics version 22 and classified using kNN.

Due to lack of data, the temperature dataset consists of 137 samples; the rainfall dataset consists of 181 samples for data collection centres and 2166 samples for states/cities.

Expert Modeler and ARIMA (Autoregressive Integrated Moving Average) modelling was used also to forecast the temperature dataset. The Expert Modeler uses the best ARIMA model or Exponential Smoothing model on the dataset. The ARIMA (0,0,0), ARIMA (1,1,0) and ARIMA (1,0,1) models were used. The sequence plot for temperature was plotted. The results obtained are discussed below in the next subsection.

Result and Discussion

The datasets were experimented on using the aforementioned methodology and the following results were obtained.

Temperature

Figure 4 classified the target sites (data collection centre) into basically two clusters based on the values of the dataset as given by NiMET. The kNN model was trained with 70% of the dataset while 30% was used for the Holdout (testing). The number of k was taken as 4 and the selected predictors were 3. The model on the system we used took 30 microseconds to classify the dataset. Since, the values of the target have been grouped by the value range, the colour of the indicators in figure 4 shows classes. The two classes are representing high temperature and low temperature.

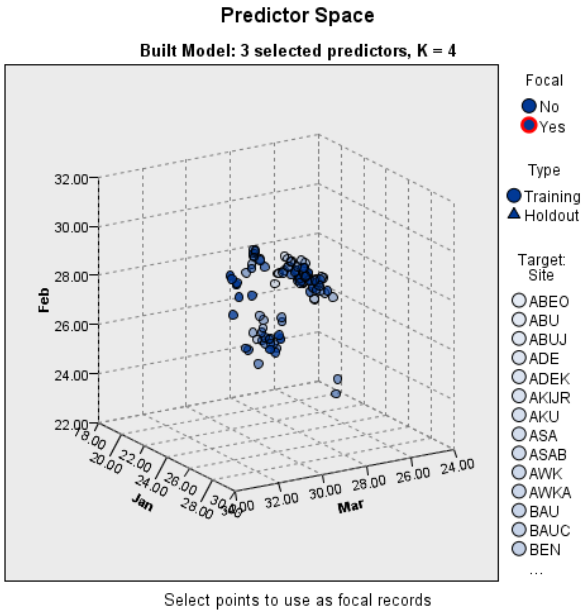

Figure 4. Classification of Temperature Dataset

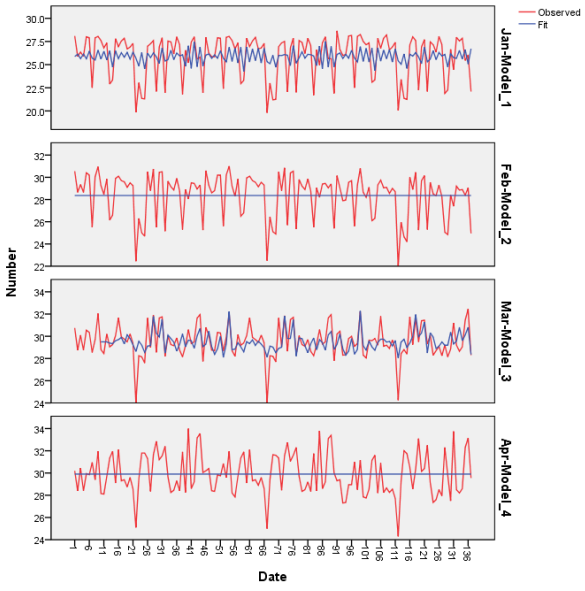

Figure 5. Expert Modeler on Temperature Dataset

The above figure shows the best fit temperature for the target sites used for the months under study. The best fit temperature is showed using the blue coloured lines. The Expert Modeler used ARIMA (1,0,1), ARIMA (0,0,0), ARIMA (0,0,9) and ARIMA (0,0,0) respectively for the months. The analysis was completed under 28 microseconds.

Using the ARIMA (1,1,0) model which ran under 28 microseconds, we have the figure below:

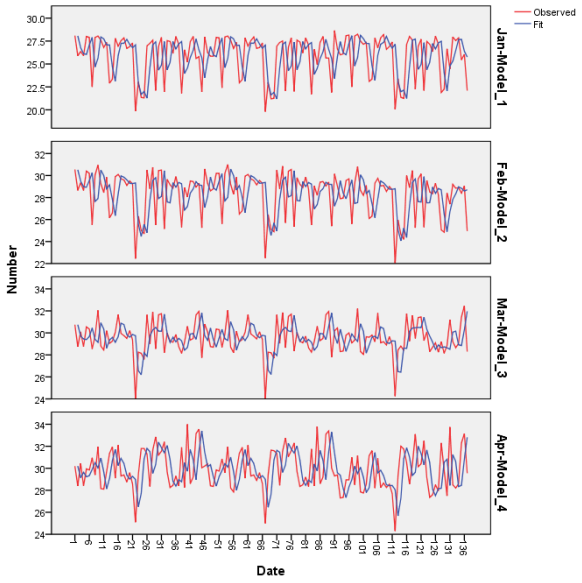

Figure 6. ARIMA (1,1,0) Model on Temperature Dataset

The temperature sequence diagram which completed in 6 seconds, 40 microseconds is thus:

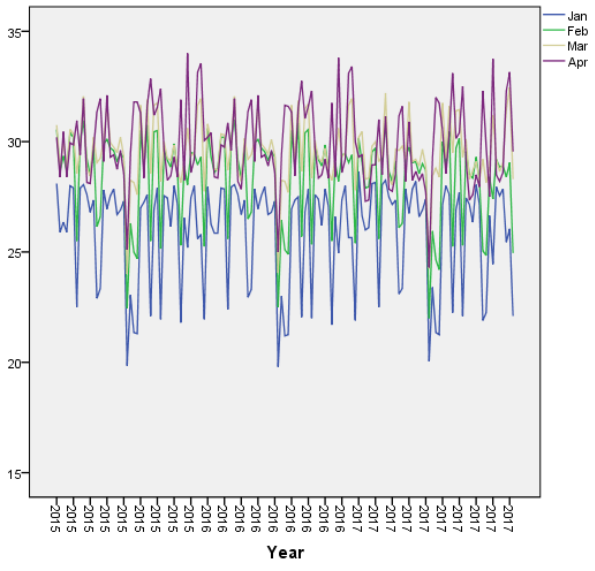

Figure 7. Sequence Chart for Temperature Dataset

Figure 7 depicts that the temperature of the Aba which is a data collection centre for the months of January to April, for three years (2015 to 2017). The temperature is higher in the month of April which means that in April there will be more rainfalls. Places with high temperature are flood endangered areas as there will be higher rainfall in those areas.

Rainfall

The classification for rainfall dataset for the data collection centre is given as:

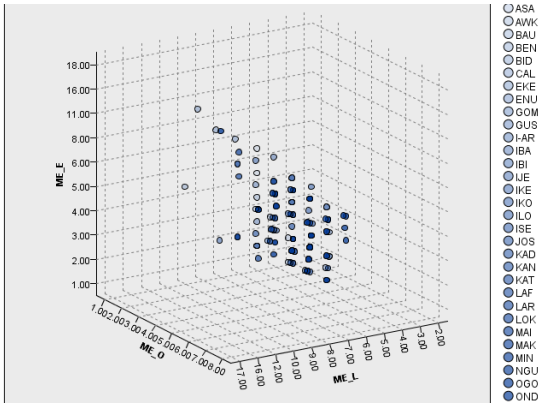

Figure 8. kNN on Rainfall Dataset (Data Collection Centres)

Figure 8 depict the classification of data collection centres as classified via the colour of the indications. The indicators with a darker colour are areas that are not flood endangered while those that have lighter colours are flood endangered areas. This is because such areas and centres have high rainfall amongst other contributors to flooding in a place.

The classification for rainfall dataset for the states is given as:

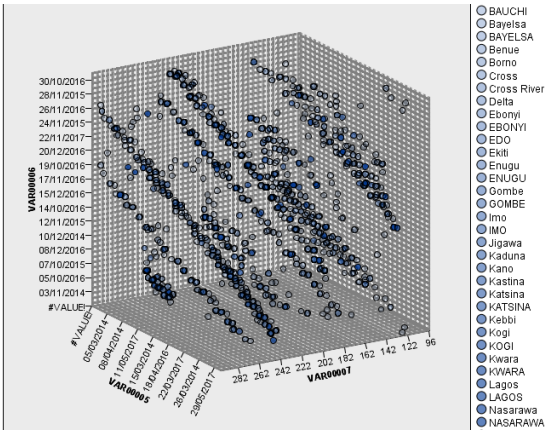

Figure 9. kNN on Rainfall Dataset (States)

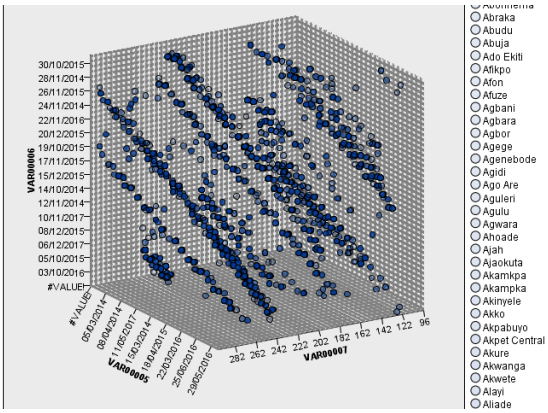

Figure 10. kNN on Rainfall Dataset (Cities)

Using Aba which is a city in Abia state, a forecast was made from the available rainfall dataset (2014-2017) till 2029 and the details are shown below.

Table 2. Aba Rainfall Forecast

Timeline	Values	Forecast	Lower Confidence Bound	Upper Confidence Bound
05/03/2014	2458			
16/03/2015	2269			
15/03/2016	2335			
08/03/2017	2505	2505	2505.00	2505.00
08/03/2018		2492.1058	2244.01	2740.20
08/03/2019		2518.3919	2262.60	2774.18
08/03/2020		2544.6779	2281.36	2808.00
08/03/2021		2570.964	2300.26	2841.66
08/03/2022		2597.25	2319.31	2875.19

08/03/2023		2623.5361	2338.49	2908.58
08/03/2024		2649.8222	2357.78	2941.86
08/03/2025		2676.1082	2377.19	2975.03
08/03/2026		2702.3943	2396.70	3008.09
08/03/2027		2728.6804	2416.31	3041.05
08/03/2028		2754.9664	2436.00	3073.93
08/03/2029		2781.2525	2455.78	3106.72
01/09/2029		2793.9994	2465.40	3122.60

Table 2 depict the forecast for Aba, showing the lower confidence bound and upper confidence bound. The forecast might be the exact as presented in forecast column or it might increase or decrease in the prediction but not exceeding the upper confidence bound or falling below the lower confidence bound. That is, the forecast must fall within the range of the lower confidence bound and the upper confidence bound.

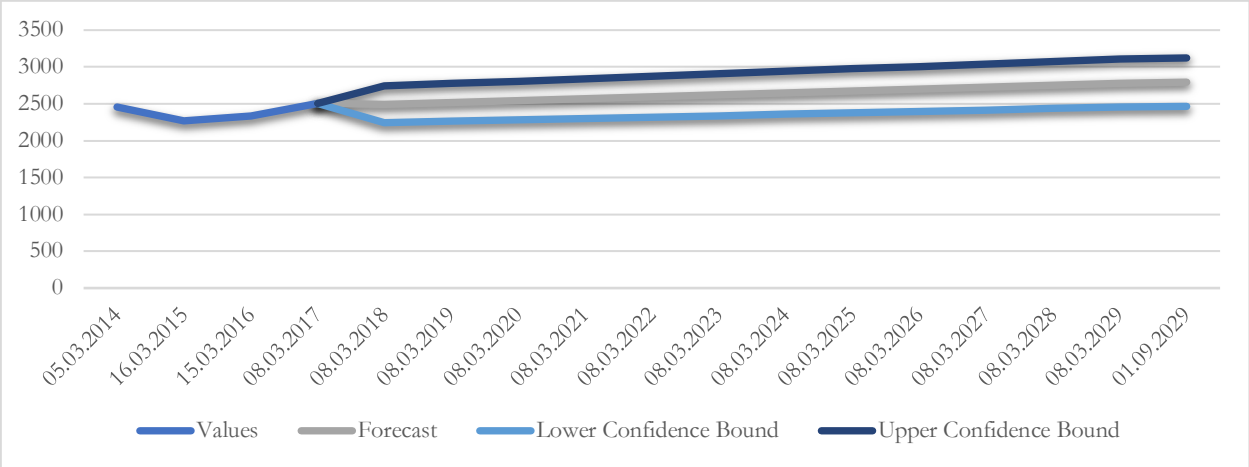

Figure 11. Aba Forecast Chart (2014 - 2029)

Figure 11 shows that Aba in Abia state will continue to experience high rainfall which is likely to result to flooding especially if the control measures are not put in place to curtail the increase in temperature and also a proactive measure in situations where rainfall increases undermining the high temperature control measures.

Conclusion and Further Studies

We have been able to analyse the data from the Nigeria Meteorological Agency. Though, the data covered only a few years, we applied kNN algorithm to classify the dataset of both temperature and rainfall to detect the flood endangered area.

The recommendation from this study is that government agencies, ministries, departments and Non-Governmental Organisations (NGOs) concerned should make readily available Geo-data to enable scientist develop Natural Disaster Warning models.

Furthermore, a forecast was made using SPSS and Excel which shows that there will be higher rainfall and higher temperature in the coming years hence, expect measures to check the occurrence of flood are to be made available by the government organisation concerned.

We therefore recommend the application of other learning algorithms to classify flood endangered area in the Nigerian meteorological space.

Constraints/Limitation

This research is limited to the data scrapped from the Nigeria Metrological Agency's periodic web publications and the use of SPSS and Excel for modelling and forecasting. The unavailability of data made the observations fewer and the span of data to be three years. The Nigeria weather (rainfall and temperature) dataset, as at the time of this research experiment, was not readily available in open data repository.

Acknowledgement

We wish to acknowledge the Nigeria Meteorological Agency and the developers of the software that was used, IBM and Microsoft.

Authors Statement

There is no conflict of interest between the authors on publishing this paper that was first presented in 2019 Centre for Satellite Technology Development (CSTD) Conference which was held in Obasanjo Space Centre, Abuja and organized by the Centre for Satellite Technology Development with the theme: "Exploration of Equatorial Orbit for Space Optimization".

References

Danso-Amoako, E., Scholz, M., Kalimeris, N., Yang, Q., & Shao, J. (2012). Predicting dam failure risk for sustainable flood retention basins: A generic case study for the wider Greater Manchester area. *Comput. Environ. Urban Syst., 36*, 423-433. https://doi.org/10.1016/j.compenvurbsys.2012.02.003

Elsafi, S.H. (20140. Artificial Neural Networks (ANNs) for flood forecasting at Dongola Station in the River Nile, Sudan. Alexandria Engineering Journal, 53(3), 655–662. https://doi.org/10.1016/j.aej.2014.06.010

Hwa, (n.a.). Destructive Floods: Tools meteorologist use to predict floods. Retrieved from https://extremeweatherwars.weebly.com/tools-meteorologists-use-to-predict-floods.html

Kim, S., Matsumi, Y., Pan, S., & Mase, H. (2016). A real-time forecast model using artificial neural network for after- runner storm surges on the Tottori coast, Japan. *Ocean Engineering, 122,* 44-53. https://doi.org/10.1016/j.oceaneng.2016.06.017

Kumar, V., Rajan, B., Venkatesan, R., & Lecinski, J. (2019). Understanding the Role of Artificial Intelligence in Personalized Engagement Marketing. *California Management Review, 61*(4), 135-155. https://doi.org/10.1177/0008125619859317

Lai, C., Shao, Q., Chen, X., Wang, Z., Zhou, X., Yang, B., & Zhang, L. (2016). Flood risk zoning using a rule mining based on ant colony algorithm. *Journal of Hydrology, 542*, 268-280. https://doi.org/10.1016/j.jhydrol.2016.09.003

Liew, A. (2013). DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and their Interrelationships. Business Management Dynamics, 2(10), 49–62.

Maddox, R. A., Zhang, J., Gourley, J. J., & Howard, K. W. (2002). Weather Radar Coverage over the Contiguous United States. *Weather and Forecasting*, *17*(4), 927-934. https://doi.org/10.1175/1520-0434(2002)017<0927:WRCOTC>2.0.CO;2

Madni, H.A., Anwar, Z. & Shah, M. (2017). *Data mining techniques and applications — A decade review.* Conference: 2017 23rd International Conference on Automation and Computing (ICAC), 1-7. https://doi.org/10.23919/IConAC.2017.8082090

Michael, E.B., & Patience, O. (2018). Flood Prediction In Nigeria Using Artificial Neural Network. *American Journal of Engineering Research, 7*(9), 15–21.

Mosavi, A., Ozturk, P., & Chau, K. (2018). Flood Prediction Using Machine Learning Models: Literature Review. *Water, 10*(11), 1536. https://doi.org/10.3390/w10111536

Nicholas, O. (2019). Data Mining Steps in KDD. Quantum Computing. Retrieved from

https://quantumcomputingtech.blogspot.com/2019/01/data-mining-steps-in-kdd.html

NiMet. (2023). Daily Weather for Nigeria. Retrieved from https://nimet.gov.ng/daily-weather

Point, T. (2015). About the Tutorial Copyright & Disclaimer.

Surbakti, H. (2015). Integrating Knowledge Management and Business Intelligence Processes for Empowering Government Business Organizations. *International Journal of Computer Applications, 114*, 36-43. https://doi.org/10.5120/19976-1874

Tehrany, M. S., Jones, S., & Shabani, F. (2019). Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. *Catena, 175*, 174-192. https://doi.org/10.1016/j.catena.2018.12.011

Xiong, J., Li, J., Cheng, W., Wang, N., & Guo, L. (2019). A GIS-Based Support Vector Machine Model for Flash Flood Vulnerability Assessment and Mapping in China. *ISPRS International Journal of Geo-Information, 8*(7), 297. https://doi.org/10.3390/ijgi8070297