ISSN 2007-9737

Flood Prediction with Optimized Attributes and Clustering

Sarmistha Nanda, Chhabi Rani Panigrahi, Bibudhendu Pati

Rama Devi Women's University, Bhubaneswar,
India

{sarmisthananda, panigrahichhabi, patibibudhendu}@gmail.com

Abstract: An emergency is a situation that poses an immediate risk to health, life, property, or environment. Most emergencies require urgent intervention to prevent a worsening of the situation. So, it is always better to predict the emergency before its happening and to take action for optimizing the loss. In this work, we tried to predict the flood by analysing the month-wise rainfall index of a particular area. First, we tried to find the months which have more contributions towards predicting the flood. For this, we used Particle Swarm Optimization (PSO) as feature selection technique and then applied classification algorithms such as J48 and Random Forest (RF). The experimentation was done for both without and with feature selection on the considered dataset. The results obtained without feature selection indicate that 70.34% and 78.81% of data are correctly classified and with feature selection 66.10% and 76.27% respectively with J48 and RF classifiers. Then we removed the class attribute from the dataset to see the effect of results when the class is not available and we applied K-means and Density Based clustering techniques on the same dataset. It was observed from the results that K-means with manhattan distance approach and Density Based clustering without feature selection classifies accurately 72.03% and 72.88% of data respectively. Similarly, when K-means and Density Based clustering were used with feature selection, it was found that K-means and Distanced Based clustering result in correct classification of 70.03% and 68.64% of data. We had also compared the model building time for both classification and clustering techniques using without and with feature selection. It was noticed that although the accuracy percentage was decreased with feature selection in both the cases. However, the model building time was reduced by 29%, 50%, 78%, and 60% in case of j48, RF, K-means, and Density Based techniques respectively.

Keywords: Feature selection, PSO, clustering techniques, classification, Manhattan distance, emergency, prediction.

1 Introduction

Flood is one of the most frequent type of natural disaster and occur when an overflow of water submerges land that is usually dry. Floods are often caused by heavy rainfall, rapid snowmelt or a storm surge from a tropical cyclone or tsunami in coastal areas [16].

Floods can cause widespread devastation, resulting in loss of life and damages to personal property and critical public health infrastructure. Between 1998-2017, floods affected more than 2 billion people worldwide [1].

People who live in floodplains or non-resistant buildings, or lack warning systems and awareness of flooding hazard [15] are most vulnerable to floods. Floods are also increasing in frequency and intensity as the climate change is happening day by day. Flood prediction using Machine Learning (ML) algorithms is effective due to its ability to utilize data from various sources and classify and regress it into flood and non-flood classes [14].

ML methods have the potential to improve accuracy as well as reduce calculating time and model development cost [2]. Cloud-based application for natural disaster prediction and management that comprised natural or manmade catastrophes such as earthquakes, cyclones, and floods are also employed random forest regression to provide improved accuracy based on rain fall, temperature, cloud wind speed, and pressure, among other factors [28, 29].

In this work, both classification and clustering algorithms are used to predict the flood. The most contributing features are also selected using the feature selection algorithm. The results obtained with and without feature selection were compared and the changes were observed in terms of time

taken to build the model and the prediction accuracy.

1.1 Motivation

Previous studies suggest that personal flood experience is a major motivator for mitigation behaviour [9]. People who had not been harmed greatly underestimated the detrimental effects of a flood [12]. Based on the findings, it is possible to conclude that risk communication should not focus exclusively on technical factors.

According to recent research, social trust in people who manage a hazard is highly related to judgements about the hazard's risk and benefits [10]. When a person lacks knowledge about a hazard, social trust in the authorities in charge of handling the hazard determines perceived risks and benefits.

It is always preferable to anticipate an emergency situation and take steps to minimise loss. In this study, we attempted to anticipate floods by studying the month-by-month rainfall index of a specific area over a period of 118 years [13].

1.2 Contributions

The contributions towards the work are listed below.

- To predict the flood from the monthly rain fall index using supervised ML technique without and with optimized attributes.
- To predict the flood from the monthly rain fall index using unsupervised ML technique without and with optimized attributes.
- To compare the classes to cluster evaluation for identifying the best technique by evaluating the performance in terms of classification accuracy, and time consumption to build the model.

In this work, the related work is described in Section 2, methodology that is adopted is presented in Section 3, the dataset considered for the experiment and the environmental setup is described in Section 4, and result analysis is done in Section 5. Finally, Section 6 concludes the paper with certain future scopes.

1.3 Related Work

Chen *et al.* [30] used ML models such as Gradient Boosting Decision Trees (GBDT), eXtreme Gradient Boosting (XGBoost), and Convolutional Neural Network (CNN) for flood risk assessment, selecting twelve indices and using 2000 sample points for model training and testing before optimising the models using Hyperparameter.

The GBDT model has the maximum accuracy of 96.83%, although only 12 indices are insufficient for flood risk assessment. Jenifer *et al.* used Sentinel-1 SAR imagery to develop Otsu's thresholding technique in the Alappuzha region [31]. The raw SAR images was preprocessed using the SNAP software's Sentinel1 toolset.

The Otsu thresholding approach was then used to compute the threshold value in order to demarcate the water pixels in the SAR pictures in order to estimate the flooding in the region. The Area Under Curve (AUC) obtained by the authors was found to be 0.83, indicating that the classifier is excellent.

Ravansalar *et al.* [32] suggested a hybrid Wavelet Linear Genetic Programming (WLGP) model to estimate monthly streamflow in two gauging stations, which incorporates a discrete wavelet transform (DWT) and a Linear Genetic Programming (LGP).

The authors divided the original time series flow into sub-time series based on wavelet co-efficient. Sub-series were then applied with the LGP to anticipate streamflow one month in advance.

The authors utilised the Nash Coefficient to calculate efficiency, which was 0.877 and 0.817 for the Pataveh and Shahmokhtar stations, respectively. Jigaw *et al.* [33] combined the statistical method Regional Flood Frequency Analysis (RFFA) with Support Vector Regression (SVR).

Hydrometric data from Environment Canada's Hydrometric Database (HYDAT) were predicted using RFFA-SVR with a combination of different kernel functions (Linear, Polynomial, and Multilayer Perception kernels), but the radial basis kernel function outperformed all the kernel

functions with Nash Sutcliffe coefficient, with a coefficient of determination of about 0.7.

Lohani *et al.* [34] proposed a threshold subtractive clustering based Takagi Sugeno (TSC-T–S) fuzzy inference system which computes two cluster centers based on the hydrologica situation, i.e. one is frequent events and another is rare events.

A new evaluation model Peak Percent Threshold Statistics (PPTS) had also been proposed by the authors to evaluate the ability of forecasting model. The TSC-T–S has been compared with Self Organizing Map (SOM) and subtractive clustering based Takagi Sugeno fuzzy model (SC-T–S fuzzy model) and gave accurate forecast.

Damle *et al.* [35] presented Time Series Data Mining (TSDM), a method for characterising and predicting occurrences in complicated, nonperiodic, and chaotic time series that blends chaos theory and data mining.

Earthquakes, floods, and rainfall are examples of chaotic nonlinear systems, in which the interactions between variables in a system are dynamic and disproportionate, yet totally deterministic. Mosavi *et al.* [36] presented the most promising long-term and short-term flood prediction approaches.

The important developments in enhancing the quality of flood prediction models were also addressed. The most effective tactics for improving ML algorithms were identified by the authors and include hybridization, data decomposition, algorithm ensemble, and model optimisation.

As per thw authors, this survey can be used as a guideline for hydrologists and climate scientists in selecting the best ML method for the prediction task.

2 Methodology

In this work, we considered a month-wise rainfall dataset of 118 years of a particular region and then PSO search is applied for attribute selection. Then with the selected attribute different classification and clustering algorithms are applied and the results are compared.

The steps involved in the proposed methodology is described as follows.

a) Pre-Processing

It is one of the most important phases during the building of the ML model. Before passing the data to a model, it needs to be processed so that the performance can be enhanced [17].

This can be in terms of accuracy, processing time, or any other parameter. In this work, the redundant, irrelevant, and minimally contributing data were removed from the dataset to reduce the model building time can be reduced [18].

b) Feature Selection

It is a process by which the approximate to zero contributing features are eliminated from the dataset [6]. This helps to reduce the time consumption in building the model. Here the attribute selection was done with a subset having evaluation parameter with pull size one and the number of thread one.

In this work, PSO Search [5] was used for feature selection. It is an optimization technique based on population and can be implemented in many research areas. Kennedy and Eberhart [8] proposed this technique by getting the inspiration from fish schooling and flocking behaviour of birds.

A bird in the search space called particle is the solution of this problem. A group of particle called swarm tries to find its optimal position by moving in the search space. Each particle x_i has a velocity v_i and is represented as in Eqn. 1:

$$\begin{aligned} x_i &= (x_{i1}, x_{i2}, \ldots, x_{in}), \\ v_i &= (v_{i1}, v_{i2}, \ldots, v_{in}), \end{aligned} \tag{1}$$

where i is the particle number and n is the problem dimension or the number of unknown variables present in the problem.

A group of random particles are present in the PSO problem [11]. In each iteration of the problem solving, two best values are identified for each particle, i.e. pbest () and gbest().

The position and velocity of each particle can be updated by using the Eqn. 2:

$$\begin{aligned} v_i^k &= w v_i^k + c_1 r_1 \left(pbest_i^k - x_i^k\right) \\ &\quad + c_2 r_2 \left(gbest^k - x_i^k\right), \\ x_i^{k+1} &= x_i^k + v_i^{k+1}, \end{aligned} \tag{2}$$

where, v_i^k is the velocity of the ith particle at

the kth iteration, and x_i^k is the current solution (or position) of the ith particle at the kth iteration. c_1, c_2 are positive constants, and r_1, r_2 are two random variables with uniform distribution between 0 and 1.

In Eqn. 2, w is the inertia weight which shows the effect of the previous velocity vector on the new vector. An upper bound is placed on the velocity in all dimensions v_{max}. This limitation prevents the particle from moving too rapidly from one region in the search space to another.

This value is usually initialized as a function of the range of the problem. PSO can be used as a feature selection algorithm.

K-means: Unlike supervised learning, K-means clustering does not need labelled data for clustering [7,19]. K-means divides things into clusters that share commonalities and are distinct to the objects in another cluster[20]. The term 'K' refers to the number of clusters that will be produced. There is a method for determining the best or optimum value of K for a given set of data [21].

Manhattan distance approach: The Manhattan distance is the simple sum of the horizontal and vertical components, or the distance between two sites measured at right angles to each other [22]. The distance is calculated using the equation as presented in Eqn. 3:

$$d = \sum_{i=1}^{n} |q_x - p_x| + |q_x - p_y|. \quad (3)$$

Density Based Clustering: A cluster is a set of data objects scattered in the data space throughout a contiguous region with high density of objects in density-based clustering [24, 25].

Clusters based on density are separated from one another by uninterrupted regions of low object density [26]. Data items in low-density areas are often regarded as noise or outliers.

3 Dataset Details and Environmental Setup

The dataset considered in this experiment was downloaded from an open source data repository named Kaggle [3]. Here 118 year's of rainfall index data are available. These data are collected month-wise from the year 1900 to 2018.

By observing the pattern of rainfall index we tried to predict the flood. It is a labelled dataset that consists of two labels, i.e., "yes" for occurrence of flood and "no" for non-occurrence. The dataset details as well as the details of environment setup are presented in Table 1.

4 Result and Analysis

In this experiment, the classification techniques such as j48 and RF was applied to the considered dataset for both without and with feature selection algorithms. The performance was evaluated in terms of accuracy and model building time.

The clustering such as K-means and density based clustering was also implemented with the same set of data after removing the labelled field. Then the same experiment was also conducted for both with and without feature selection algorithms.

a) Classification without Feature Selection algorithm

The classification algorithms were implemented with 10-fold cross-validation. In this work, we applied j48 and RF classification algorithms on the considered dataset.

When J48 algorithm was applied, it was observed that while building the pruned tree the number of leaves generated was 11 and the size of the tree was 21.

The total time taken to build the model was recorded to be 0.07 second. Here, it was found that 70.34% of data were correctly classified. The confusion matrix of J48 algorithm without feature selection algorithm is given in the Table 2.

During the classification, all the evaluation parameters were observed and are listed in Table 3.

After the classification it was found that the kappa statistic value was 0.4058, mean absolute error value was 0.3104 and root mean squared error value was 0.5312. The confusion matrix for the same is given in Table 3.

Table 1. Dataset and environmental setup

Attribute	Values
Dataset Considered	Monthly Rainfall Index and Flood Probability [3] (data of 118 years)
Source	Kaggle open source data repository
Dataset last accessed	8th April 2023
Experimentation environment	WEKA version 3.9 [4]
Attribute Selection Algorithm	PSO Search
Clustering Algorithm	K-means clustering (manhattan distance approach), Density Based Clustering

Table 2. Confusion matrix for J48 algorithm without feature selection

	Flood	No Flood
Flood	45	15
No Flood	20	38

Table 3. Classification report for J48 algorithm without feature selection

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.750	0.345	0.692	0.750	0.720	0.407	0.702	0.659	Flood
	0.655	0.250	0.717	0.655	0.685	0.407	0.702	0.657	No Flood
Weighted Avg.	0.703	0.298	0.704	0.703	0.703	0.407	0.702	0.658	

Table 4. Confusion matrix for RF algorithm without feature selection

	Flood	No Flood
Flood	48	12
No Flood	13	45

RF classifier algorithm [27] was implemented on the same dataset with 100 numbers iteration and found that 78.81% of data were correctly classified.

The confusion matrix for RF algorithm without feature selection algorithm is given in Table 4.

Total time taken to build the model was 0.32 seconds and the accuracy of classification is given in Table 5.

In this classification, the kappa statistic value, mean absolute error, and root mean squared error was found as 0.576, 0.3375, and 0.3782 respectively.

b) Classification with Feature Selection Algorithm

In this case, first the PSO feature selection algorithm was applied to the dataset toselect the most contributory features. From the feature selection, it was found that JUN, JULY, SEPTEMBER are the months whose rainfall index is more responsible to predict the flood.

Table 5. Classification report for RF algorithm without feature selection

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.800	0.224	0.787	0.800	0.793	0.576	0.902	0.904	Flood
	0.776	0.200	0.789	0.776	0.783	0.576	0.902	0.901	No Flood
Weighted Avg.	0.788	0.212	0.788	0.788	0.788	0.576	0.902	0.902	

Table 6: Confusion matrix for J48 algorithm with feature selection

	YES	NO
YES	41	19
NO	21	37

Table 7. Classification Report for J48 algorithm with feature selection

TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0. 683	0. 362	0. 661	0. 683	0. 672	0. 322	0. 726	0. 663	YES
0. 638	0. 317	0. 661	0. 638	0. 649	0. 322	0. 726	0. 725	NO

Table 8. Confusion matrix for RF algorithm with feature selection

	YES	NO
YES	46	14
NO	14	44

Table 9. Classification report for RF algorithm with feature selection

TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0. 767	0. 241	0. 767	0. 767	0. 767	0. 525	0. 816	0. 820	YES
0. 759	0. 233	0. 759	0. 759	0. 759	0. 525	0. 816	0. 814	NO

After the feature selection, J48 classifier was implemented on the selected features. The total time taken to build the model was found to be 0.07 seconds.

The size of the tree created in J48 model creation is 11 where as the number of leaves is 11. Out of total 118 number of instances, 78 instances are correctly classified which is 66.10% of accuracy.

It was observed that, The values of Kappa statistic, Mean absolute error, Root mean squared error, and Relative absolute error obtained were 0.32, 0.35, 0.49, and 70.20 respectively.

The confusion matrix obtained after the implementation of J48 algorithm is given in Table 6. The values of TP rate, FP rate, Precision, Recall, F-Measure, MCC, ROC Area, and the class are listed in Table 7. We had also implemented the RF

classifier with 10-fold cross validation. The total time taken to build the model was found to be 0.16 seconds. 76.27% of data were classified properly where as 23.73% of data were incorrectly classified.

Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error were calculated and the value obtained were 0.53, 0.30, 0.43, and 63.82 respectively. The confusion matrix given in Table 8 and the values of other evaluating parameters are presented in Table 9.

c) Clustering without Feature Selection Algorithm

K-means clustering algorithm was applied on the considered dataset without feature selection. Two clusters were made, i.e. 1 for yes and 0 for no. Missing values were globally replaced with mean/mode. ManhattanDistance was taken into consideration during the cluster creation.

The total time taken to build the model was found to be 0.09 seconds and it has gone through 11 number of iteration. Out of 118 instances, 73 are clustered as no flood and 45 were clustered for predicting flood.

A total of 27.97% of the data are clustered incorrectly and rest are clustered correctly.The cluster instances are presented in Table 10 and the classes to cluster matrix is given in Table 11.

The Density Based clustering was also applied on the same dataset without any feature selection. It was observed that for cluster 0, the prior probability was 0.6 where as for cluster 1, the prior probability was 0.4.

The total time taken to build the model was 0.5 seconds. In this model, 27.11% of instances are incorrectly classified. The cluster instances are given in Table 12 and the classes to cluster matrix value is presented in Table 13.

d) Clustering with Feature Selection Algorithm

For better performance we applied feature selection using PSO where out of 12 attributes, 4 were selected. During the model creation, it has gone through iterations. Out of 118 instances, 43 instances were clustered as cluster 0 and 75 instances were as cluster 1.

The total time taken to build the model obtained was 0.02 seconds and it was observed that 29.66% of instances were clustered incorrectly.

Table 10. Clustered instances for K-means clustering without feature selection

No Flood	72 (61%)
Flood	46(39%)

Table 11. Classes to cluster matrix for K-means clustering without feature selection

	YES	**NO**
YES	36	24
NO	9	49

Table 12. Clustered Instances for K-means clustering without feature selection

No Flood	72 (61%)
Flood	46(39%)

Table 13. Classes to cluster matrix for K-means clustering without feature selection

	YES	**NO**
YES	37	23
NO	9	49

Table 14. Clustered instances of K-means with feature selection

No Flood	72 (61%)
No. Flood	46(39%)

Table 15. Classes to clusters matrix of K-means with feature selection

	YES	**NO**
0(YES)	34	26
1(NO)	9	49

Table 16. Clustered instances of density based clustering with feature selection

0	41 (35%)
1	77 (65%)

Table 17. Classes to clusters of density based clustering with feature selection

	YES	**NO**
YES	32	28
NO	9	49

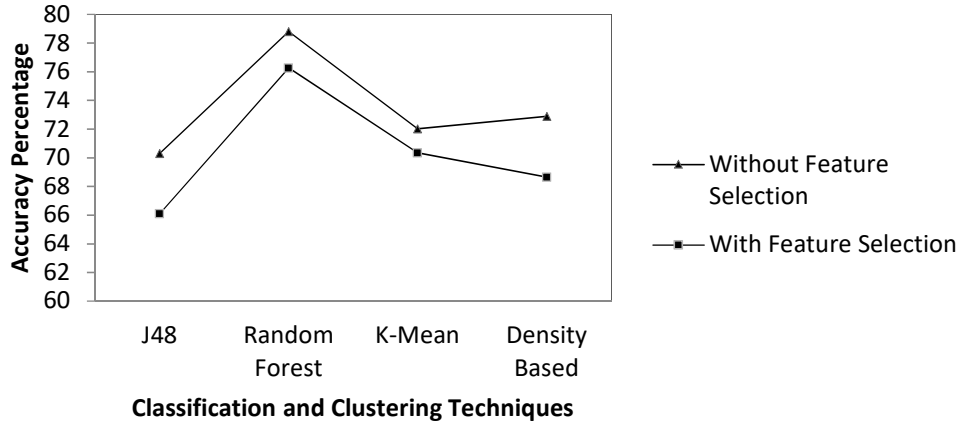

Fig. 2. Accuracy percentage with and without feature selection algorithm

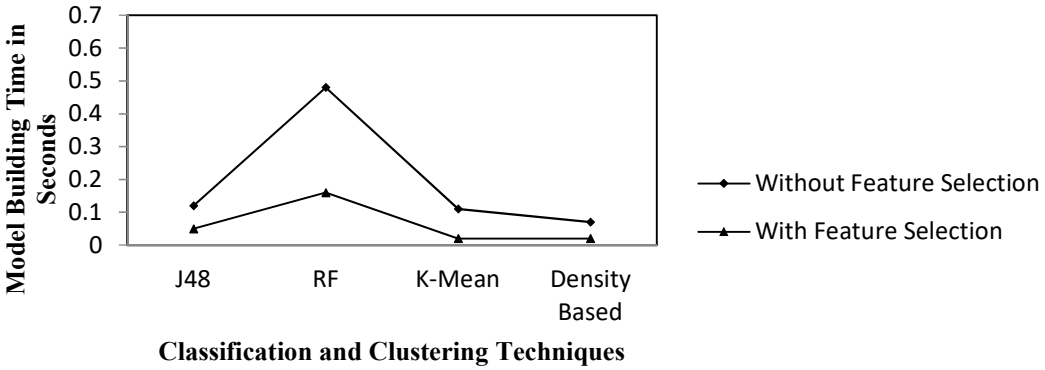

Fig. 3. Model building time comparision with and without feature selection algorithm

The cluster instances are presented in Table 14 and the classes to cluster matrix is given in Table 15.

Density based clustering: PSO algorithm was applied for feature selection and then the Density based clustering was implemented on the considered dataset.

In this model, out of 12 attributes 4 were selected as the most contributing attributes. It had gone through 5 iterations while building the model. Missing values were globally replaced with mean/mode.

According to cluster centroid from 118 instances 43 belongs to cluster 0 and 75 belongs to cluster1. Cluster 0 has the prior probability 0.4 cluster: 1 has prior probability: 0.6. The total time taken to build model obtained was 0.02 seconds and incorrectly clustered instances were 31.36 %.

The clustered instances of density based clustering with feature selection was given in Table 16 and the classes to clusters of density based clustering with feature selection was presented in Table 17.

Classes to cluster evaluation: The clustering accuracy performance was evaluated against the real classification. The accuracy percentage in both with and without feature selection is shown in Figure 2.

*Model building time comparision: T*he model building time comparision of both classification and clustering techniques with and without feature selection algorithm is shown in Figure 3.

From the graph, it is observed that the model building time reduces significantly with feature selection as compared to without feature selection.

5 Conclusion and Future Directions

In this work, we considered a dataset containing the rain fall details of a particular region. It contains the data of 118 years month wise. The classification techniques j48 and RF and the clustering techniques such as K-means and density based are impleted on the considered dataset. After this, the months minimally contributing for flood prediction were discarded using PSO feature selection method.

Then the same classification and clustering techniques were again applied and the results in both the cases were compared. It was observed that the accuracy percentage obtained was found to be higher in both clustering and classification when no features are discarded. But the model building time was reduced when the classifications and clustering techniques were applied on the selected features.

In future, if this type of model can be built and embedded in the cloud as a service which can be called as per requirement then alert can be made and proper management of the situation can be done.

References

1. **United Nations (2023).** Data application of the month: Machine Learning for Flood Detection. https://un-spider.org/links-and-resources/data - sources/daotm-floods-ml.
2. **Mukul (2023).** Flood prediction model. Kaggle, https://www.kaggle.com/mukulthakur177/kerela-flood/version/2.
3. **Holmes, G., Donkin, A., Witten, I. H. (1994).** Weka: A machine learning workbench. Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference, IEEE, pp. 357–361. DOI: 10.1109/ANZIIS.1994.396988.
4. **Fong, S., Biuk-Aghai, R. P., Millham, R. C. (2018).** Swarm search methods in weka for data mining. Proceedings of the 2018 10th International Conference on Machine Learning and Computing, pp. 122–127. DOI: 10.1145/3195106.3195167.
5. **Cai, J., Luo, J., Wang, S., Yang, S. (2018).** Feature selection in machine learning: A new perspective. Neurocomputing, Vol. 300, pp. 70–79. DOI: 10.1016/j.neucom.2017.11.077.
6. **Yadav, J., Sharma, M. (2013).** A review of K-mean algorithm. International Journal of Engineering Trends and Technology, Vol. 4, No. 7, pp. 2972–2976.
7. **Kennedy, J., Eberhart, R. (1995).** Particle swarm optimization. Proceedings of ICNN'95-international conference on neural networks, IEEE, Vol. 4, pp. 1942–1948.DOI: 10.1109/ICNN.1995.488968.
8. **Siegrist, M., Gutscher, H. (2008).** Natural hazards and motivation for mitigation behavior: People cannot predict the affect evoked by a severe flood. Risk Analysis: An International Journal, Vol. 28, No. 3, pp. 771–778. DOI: 10.1111/j.1539-6924.2008.01049.x.
9. **Siegrist, M., Cvetkovich, G. T., Gutscher, H. (2001).** Shared values, social trust, and the perception of geographic cancer clusters. Risk Analysis, Vol. 21, No. 6, pp.1047–1054. DOI: 10.1111/0272-4332.216173.
10. **Jaberipour, M., Khorram, E., Karimi, B. (2011).** Particle swarm algorithm for solving systems of nonlinear equations. Computers & Mathematics with Applications, Vol. 62, No. 2, pp. 566–576. DOI: 10.1016/j.camwa.2011.05.031.
11. **Lawal, Z. K., Yassin, H., Zakari, R. Y. (2021).** Flood prediction using machine learning models: a case study of Kebbi state Nigeria. IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1–6. IEEE. DOI: 10.1109/CSDE53843.2021.9718497.
12. **Syeed, M. M. A., Farzana, M., Namir, I., Ishrar, I., Nushra, M. H., Rahman, T. (2022).** Flood prediction using machine learning models. International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1–6. IEEE. DOI: 10.1109/HORA55278.2022.9800023.

13. **Zehra, N. (2020).** Prediction analysis of floods using machine learning algorithms (NARX & SVM). International Journal of Sciences: Basic and Applied Research (IJSBAR), Vol. 49, No. 2.
14. **Kanwar, B. P. S. (2022).** Development of flood prediction models using machine learning techniques. Missouri University of Science and Technology.
15. **Maspo, N. A., Harun, A. N. B., Goto, M., Cheros, F., Haron, N. A., Nawi, M. N. M. (2020).** Evaluation of machine learning approach in flood prediction scenarios and its input parameters: A systematic review. IOP Conference Series: Earth and Environmental Science, IOP Publishing, Vol. 479, No. 1, pp. 012038. DOI: 10.1088/1755-1315/479/1/012038.
16. **Kotsiantis, S. B., Kanellopoulos, D., Pintelas, P. E. (2006).** Data preprocessing for supervised leaning. International journal of computer science, Vol. 1, No. 2, pp. 111–117.
17. **Alshdaifat, E. A., Alshdaifat, D. A., Alsarhan, A., Hussein, F., El-Salhi, S. M. D. F. S. (2021).** The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. Data, Vol. 6, No. 2, pp. 11. DOI: 10.3390/data6020011.
18. **Na, S., Xumin, L., Yong, G. (2010).** Research on k-means clustering algorithm: An improved k-means clustering algorithm. Third International Symposium on intelligent information technology and security informatics. IEEE, pp. 63–67. DOI: 10.1109/IITSI.2010.74.
19. **Ahmed, M., Seraj, R., Islam, S. M. S. (2020).** The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, Vol. 9, No. 8, pp. 295. DOI: 10.3390/electronics9081295.
20. **Oyelade, O. J., Oladipupo, O. O., Obagbuwa, I. C. (2010).** Application of k-means clustering algorithm for prediction of students academic performance. arXiv preprint arXiv:1002.2425. DOI: 10.48550/arXiv.1002.2425.
21. **Szabo, F. E. (2015).** Manhattan distance. ScienceDirect. https://www.sciencedirect.com/topics/mathematics/manhattan-distance.
22. **Suwanda, R., Syahputra, Z., Zamzami, E. M. (2020).** Analysis of euclidean distance and manhattan distance in the K-means algorithm for variations number of centroid K. Journal of Physics: Conference Series IOP Publishing, Vol. 1566, No. 1. DOI 10.1088/1742-6596/1566/1/012058.
23. **Patel, K. A. (2016).** An efficient and scalable density-based clustering algorithm for normalize data. Procedia Computer Science, Vol. 92, pp. 136–141. DOI: 10.1016/j.procs.2016.07.336.
24. **Bhattacharjee, P., Mitra, P. (2021).** A survey of density based clustering algorithms. Frontiers of Computer Science, Vol. 15, pp. 1– 27.
25. **ArcGIS Pro (2023).** How density-based clusterind works. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm.
26. **Kulkarni, V. Y., Sinha, P. K. (2013).** Random forest classifiers: a survey and future research directions. International Journal of Advanced Computing, Vol. 36, No. 1, pp. 1144–1153.
27. **Alazawi, Z., Abdljabar, M. B., Altowaijri, S., Vegni, A. M., Mehmood, R. (2012).** ICDMS: an intelligent cloud based disaster management system for vehicular networks. Communication Technologies for Vehicles: 4th International Workshop, Nets4Cars/Nets4Trains, Vol 7266. DOI: 10.1007/978-3-642-29667-3_4.
28. **Nanda, S., Panigrahi, C. R., Pati, B. (2022).** An architectural framework to manage heterogeneous emergencies. Intelligent Systems: Proceedings of ICMIB 2021 Singapore: Springer Nature Singapore, Vol. 431, pp. 169–177. DOI: 10.1007/978-981-19-0901-6_16.
29. **Chen, J., Huang, G., Chen, W. (2021).** Towards better flood risk management: Assessing flood risk and investigating the potential mechanism based on machine learning models. Journal of environmental

management, Vol. 293. DOI: 10.1016/j.jenvman.2021.112810.

30. **Jacinth, J. J., Saravanan, S., Abijith, D. (2022).** Integration of SAR and multi-spectral imagery in flood inundation mapping–a case study on Kerala floods 2018. ISH Journal of Hydraulic Engineering, Vol. 28. pp. 480–490. DOI: 10.1080/09715010.2020.1791265.
31. **Ravansalar, M., Rajaee, T., Kisi, O. (2017).** Wavelet-linear genetic programming: a new approach for modeling monthly streamflow. Journal of Hydrology, Vol. 549, pp. 461–475. DOI: 10.1016/j.jhydrol.2017.04.018.
32. **Gizaw, M. S., Gan, T. Y. (2016).** Regional flood frequency analysis using support vector regression under historical and future climate. Journal of Hydrology, Vol. 538, pp. 387–398. DOI: 10.1016/j.jhydrol.2016.04.041.
33. **Lohani, A. K., Goel, N. K., Bhatia, K. K. S. (2014).** Improving real time flood forecasting using fuzzy inference system. Journal of hydrology, Vol. 509, pp. 25–41. DOI: 10.1016/j.jhydrol.2013.11.021.
34. **Damle, C., Yalcin, A. (2007).** Flood prediction using time series data mining. Journal of Hydrology, Vol. 333, No. 2-4, pp. 305–316. DOI: 10.1016/j.jhydrol.2006.09.001.
35. **Mosavi, A., Ozturk, P., Chau, K. W. (2018).** Flood prediction using machine learning models: Literature review. Water, Vol. 10, No. 11, pp. 1536. DOI: 10.3390/w10111536.

Article received on 11/06/2023; accepted on 25/10/2023.
Corresponding author is Chhabi Rani Panigrahi.