A Flood Hazard Prediction and Risk Assessment Model Based on Machine Learning Approach

Jiakang Zhang *

School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China, 400065

* Corresponding Author Email: zjkhandsome123@163.com

Abstract. With the acceleration of global climate change and the frequent intensification of human activities, the frequency and intensity of flood disasters are rising, and the global flood prevention situation is becoming more and more critical. Therefore, in-depth investigation of the causes of floods and their wide-ranging impacts has become an important issue that needs to be urgently addressed in the fields of environmental science and disaster management. Given the high cost of comprehensive data collection and processing of all potential flood prediction indicators, accurate selection of key indicators becomes the key to improve prediction efficiency. With the help of data analysis and machine learning techniques, this study aims to predict the flood risk level and the probability of flood occurrence through scientific methods, and to propose effective prevention and response strategies. In the selection of key indicators, this paper adopt a combination of strategies: on the one hand, the Spearman correlation coefficient and Cramér's V are used to screen out the indicators that are closely related to the occurrence of floods; on the other hand, the risk of flood events is subdivided into high, medium, and low levels through K-means clustering method, and important features are further identified with the help of logistic regression, random forest, to construct the model for predicting and assessing the risks of floods. The results show that the optimized feature selection method, especially the application of the combined risk classification and random forest algorithm, not only significantly improves the prediction accuracy and generalization ability of the model, but also speeds up the model operation and enhances the interpretability of the decision-making process. This innovative method provides a more efficient and reliable technical support for flood risk assessment, disaster prevention and mitigation, which helps to reduce the negative impacts of flood disasters and safeguard people's lives and properties.

Keywords: Machine Learning, K-means Clustering, Random Forest, Logistic Regression.

1. Introduction

Flooding is a natural disaster in which the volume of water in rivers and lakes increases rapidly or the water level rises rapidly as a result of natural factors such as heavy rainfall, snowmelt and storm surges. It not only poses serious damage to infrastructure, but also directly threatens the safety of human lives and the stability of the ecological environment. Therefore, in-depth study of the causes and impacts of floods has become an important problem to be solved in the field of environmental science and disaster management.

Traditional flood prediction methods mainly rely on hydrological and meteorological models. Examples include the Xin'an jiang Model, Muskingum, and so on. The Xin'an jiang Model is a rainfall runoff model researched and published by Prof [1]. Zhao ren jun's team in the 1960s and 1970s. There are many parameters in the model, which are mainly divided into four levels: evapotranspiration, stream production, diversion and confluence, and the specific values are closely related to the local meteorological climate and topography. Because of the many parameters, the important step in the design of Xin'an jiang Model is the tuning process of parameters. Zuo xiang's team improved the Xin'an jiang Model by optimizing the rate setting of the model parameters, which achieved better results, but still has the disadvantage of slower convergence [2]. The Muskingum model is an aggregate hydrological forecasting model proposed by McCarthy in 1939. Same as the Xin'an jiang Model, the problem of the Muskingum model is that the optimization process of many parameters is complicated, and it is difficult to find the global optimum. In this regard, based on the ripple

phenomenon in nature, Ting hui wang's team proposed a new heuristic intelligent optimization algorithm - ripple algorithm, which improves the solution speed and accuracy, but due to the huge number of parameters, its practicability is low [3]. In recent years, with the development of big data technology and machine learning field, data-driven flood prediction techniques have gradually emerged into the limelight. Such techniques skillfully integrate rich historical data resources with advanced algorithms, such as random forests, support vector machines and neural networks, which greatly enhance the accuracy and efficiency of flood prediction. Nevertheless, current research still faces challenges in the effective screening of key features and deep optimization of model performance, which to some extent restricts the promotion of these models in practical applications.

To summarize, in order to overcome these shortcomings, this paper proposes a high-precision prediction model of flood occurrence probability based on machine learning. First, a combination of Spearman correlation coefficient and Cramér's V is used to screen key indicators that are highly correlated with flood occurrence. Then, K-means clustering method is used to classify the risk level of flood events into high, medium, and low levels, and the most important features are selected by logistic regression model to construct a risk level early warning assessment model. Finally, based on the random forest machine learning model, the flood occurrence probability prediction model is constructed and the model is optimized by mean square error and coefficient of determination. The goal of this study is to improve the accuracy and efficiency of flood prediction by optimizing the feature selection method and model structure, and to provide valuable experience and reference for future flood prediction and prevention [4].

2. The basic fundamental of model

In this paper, logistic regression model is selected as the core and combined with K-means clustering algorithm to predict the flood risk level, and random forest model is also selected to predict the probability of flood occurrence and optimize the model prediction results.

2.1. Logistic Regression Model

Logistic regression is a supervised learning algorithm widely used in classification problems. It is derived from statistics and aims to model the relationship between a dependent variable and one or more independent variables. Unlike linear regression, logistic regression does not directly predict values, but rather estimates the probability that a sample belongs to a particular category [5]. This is usually accomplished through a Sigmoid function (or log odds function), which is capable of mapping any real number between 0 and 1. The formula is as follows:

$$g(y) = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\cdots\theta_n x_n)}} = \frac{1}{1+e^{-\theta^T x}} \quad (1)$$

2.2. Random Forest Model

Random forest is an integrated learning method to accomplish classification or regression tasks by constructing multiple independent decision trees, which has the advantage of reducing the risk of overfitting and improving the stability of the model [6]. Each tree is trained by randomly selecting subsamples from the original data (Bootstrap sampling) and randomly selecting features. The final prediction result is obtained by voting (classification task) or averaging (regression task) of all trees. The process of random forest splitter is shown in Figure 1:

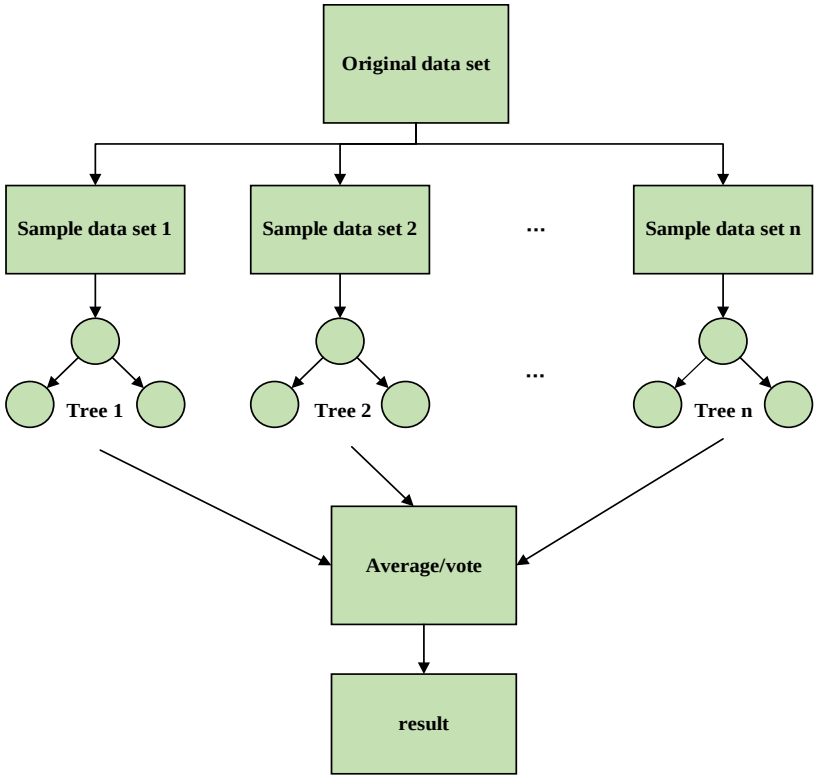


Figure 1. Random Forest Splitter Process

3. Results

3.1. Data description

The data in this paper is sourced from the Kaggle platform, containing over 1 million records related to flood events. This dataset serves as a robust resource for constructing a flood occurrence probability prediction model using data analysis and machine learning techniques, providing a scientific basis for flood risk assessment and preventive measures. The training set comprises 1,048,573 entries, while the test set includes 745,306 entries, maintaining a 7:3 ratio. Variable Y represents flood probability, and variable X includes 21 indicators, which are numbered 1 to 21 for conciseness, as shown in Table 1:

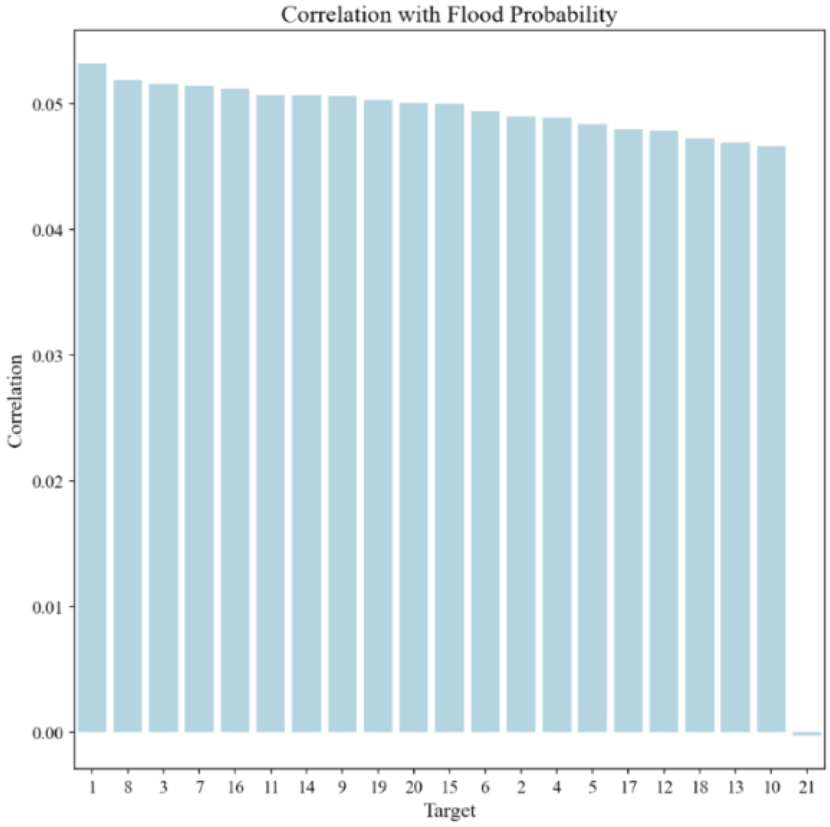
Table 1. Index of each index

Target	Monsoon intensity	Topographic drainage	Stream management	Deforestation	Urbanization
Number	1	2	3	4	8
Target	Climate change	Quality of dams	Siltation	Agricultural practice	Erode
Number	6	7	8	9	10
Target	Ineffective disaster prevention	Drainage system	Coastal vulnerability	Landslide	Watershed
Number	11	12	13	14	15
Target	Deterioration of infrastructure	Population score	Wetland loss	Insufficient planning	Policy factors
Number	16	17	18	19	20

3.2. Correlation analysis results

In order to filter out key indicators that influence the prediction of flood risk levels, this paper applied two separate correlation measures to each indicator in the training set: for categorical variables,

this paper used Cramer's V coefficients to capture the strength of the nonlinear correlation between them and the probability of flooding, whereas for variables that may be categorized in a continuous or ordinal manner, this paper used the Spearman's rank correlation coefficient [7]. Subsequently, based on these quantitative values, this paper ranked the correlations to clearly show the impact of each indicator on the probability of flood occurrence. This is shown in Figure 2:



Number	Target
1	Monsoon intensity
2	Topographic drainage
3	Stream management
4	Deforestation
5	Urbanization
6	Climate change
7	Quality of dams
8	Siltation
9	Agricultural practice
10	Erode
11	Ineffective disaster prevention
12	Drainage system
13	Coastal vulnerability
14	Landslide
15	Watershed
16	Deterioration of infrastructure
17	Population score
18	Wetland loss
19	Insufficient planning
20	Policy factors
21	Id

Figure 2. Correlation of indicators with probability of flooding

As shown in Figure 2, this paper can draw the following conclusions:

Firstly, the top five indicators positively correlated with flood occurrence probability are, in order: monsoon intensity, sedimentation, river management, dam quality, and infrastructure deterioration. These indicators show significant correlation and align with actual factors contributing to floods. Secondly, indicators such as erosion, coastal vulnerability, wetland loss, drainage system efficiency, and population score exhibit weaker correlations with flood probability. This does not diminish their importance; rather, it may reflect dataset limitations, the specificity of the analysis method, or interactions with more strongly correlated indicators. Lastly, regarding the negative correlation with ID numbers, it should be noted that these are unique identifiers assigned to flood events in this study and do not indicate any relationship with flood occurrence probability [8].

3.3. Results of flood risk level predictions

Through correlation analysis, this paper identified some key indicators that affect the prediction results, and in order to construct a flood risk level prediction model that is both accurate and efficient, this paper decided to use a logistic regression model as the core. However, before inputting the key metrics into the model for training, an integral step is to calculate the weights of these metrics in the model. In order to realize this part, this paper adopted the Principal Component Analysis (PCA) method to calculate the weights of each indicator [9]. As shown in Table 2:

Table 2. Weights of indicators under the logistic regression model

Target	Monsoon intensity	Topographic drainage	Stream management	Deforestation	Urbanization
Weight	0.051012	0.050882	0.050839	0.050795	0.050719
Target	Climate change	Quality of dams	Siltation	Agricultural practice	Erode
Weight	0.050675	0.050627	0.050607	0.050569	0.050542
Target	Ineffective disaster prevention	Drainage system	Coastal vulnerability	Landslide	Watershed
Weight	0.050438	0.050421	0.050396	0.050340	0.050317
Target	Deterioration of infrastructure	Population score	Wetland loss	Insufficient planning	Policy factors
Weight	0.050245	0.050240	0.050231	0.050101	0.040005

As shown in Table 2, the table shows in detail the weights assigned to each indicator. Among these indicators, the indicators with larger weights imply that they occupy a more important position in the training of the model, e.g., monsoon intensity, topographic drainage, river management, and deforestation. Therefore, these indicators with larger weights should be considered as important factors affecting the probability of flood occurrence when constructing the flood risk level prediction model.

Next, this paper used the K-means clustering algorithm to cluster the data in the training set [10]. With this algorithm, the data were scientifically classified into three different risk categories: high-risk, medium-risk, and low-risk, and in order to facilitate the differentiation, this paper used three colors as the identifiers of these three risk categories respectively. As shown in Figure 3:

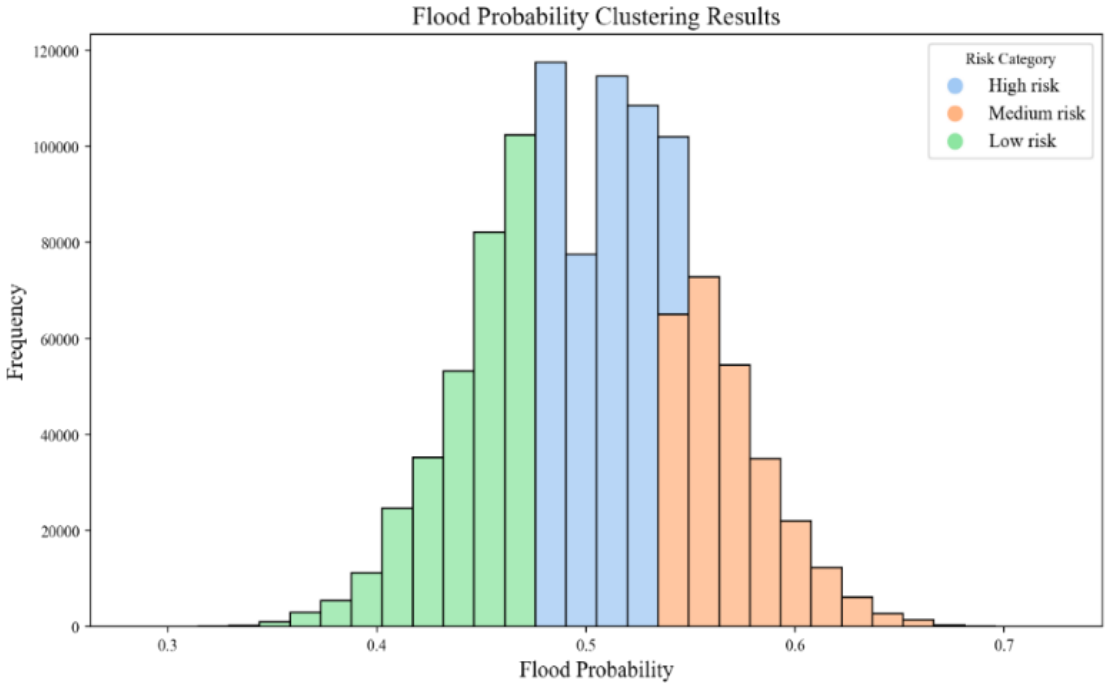


Figure 3. Classification of flood risk levels

Table 3. Risk level index number

Target	High risk	Medium risk	Low risk
Number	1	0	2

This figure visualizes the distribution of the flood risk classes in a pattern that closely fits the properties of a normal distribution, showing that the data fluctuates around a central value and that the number of data points decreases as it extends towards the ends.

Subsequently, this paper replaced the three risk categories using 0, 1, and 2 to facilitate data processing. After the model training is completed, the next step is to predict the risk class for the flood data in the test set, and some of the results are shown in Table 4:

Table 4. Partial predictions of flood risk classes

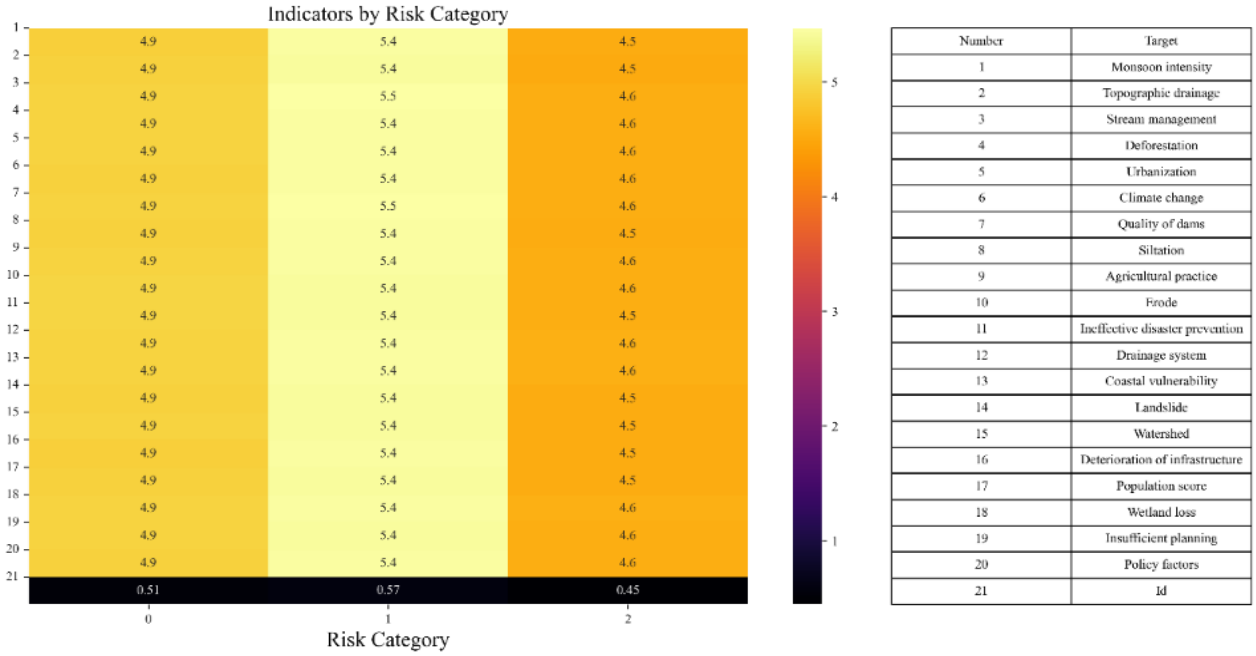
Predicted risk category	0	2	0	1	0	2	2	0	1	0
Actual risk category	0	2	0	1	1	2	2	0	1	0

Table 5. Results of logistic regression model assessment

	Category	Precision	Recall	F1-score
	0	0.69	0.71	0.70
	1	0.85	0.81	0.83
	2	0.71	0.70	0.71
Maro avg		0.75	0.74	0.75
Weight avg		0.74	0.73	0.74
Accuracy				0.73

As can be seen in Table 5, the overall precision of the flood risk class prediction model reaches 0.73. The prediction precision is 0.83 for the high-risk class, 0.71 for the medium risk class, and 0.70 for the low-risk class.

Next, this paper analyzed the mean characteristics of each indicator in each risk category to identify the significant characteristics of the high, medium, and low risk categories [11]. We calculated the mean value of each indicator in different risk categories and plotted heat maps to demonstrate these characteristics, as shown in Figure 4:



Number	Target
1	Monsoon intensity
2	Topographic drainage
3	Stream management
4	Deforestation
5	Urbanization
6	Climate change
7	Quality of dams
8	Siltation
9	Agricultural practice
10	Erode
11	Ineffective disaster prevention
12	Drainage system
13	Coastal vulnerability
14	Landslide
15	Watershed
16	Deterioration of infrastructure
17	Population score
18	Wetland loss
19	Insufficient planning
20	Policy factors
21	Id

Figure 4. Characteristics of indicators for different risk categories

As shown in Figure 4, the distribution of each key indicator under different risk categories is demonstrated. The depth of color in the figure reflects the value of the indicator, with darker colors indicating higher values. Through observation, this paper finds that there are significant differences in the indicators between different risk categories. Specifically, the indicators of Risk Category 1 are generally higher, showing its high-risk characteristics; while the average value of the indicators of Risk Category 2 is relatively lower, reflecting a more moderate degree of risk.

Taken together, this paper concludes that high-risk flood events are often characterized by ineffective disaster prevention, deteriorating infrastructure, deforestation, increased monsoon intensity,

and dam quality concerns. Medium-risk flood events have moderate performance on indicators such as erosion, landslide risk, policy implementation, urbanization, and wetland loss. Low-risk flood events are low on coastal vulnerability, topographic drainage capacity, inadequate planning, climate change impacts and deforestation.

3.4. Flood probability prediction results

In order to predict the probability of flood occurrence, this paper built a random forest model separately and then used the data in the training set for training and subsequently predicted the flood probability in the test set.

Before building the random forest model, this paper makes the following assumptions:

Assume that all decision trees are independent.

Assume that all samples are equally likely to be used by the decision trees.

Assume that each decision tree makes decisions with the features it chooses.

Assume that the overall result is generated in a voting manner based on the decision trees.

Combining the results of correlation analysis and the results of flood risk level prediction, this paper constructed a random forest model and predicted the flood probability in the test set after training [12].

Table 6. Random Forest Model Evaluation Results (Before Improvement)

	MSE	RMSE	MAE	MAPE	R^2
Train	0.001	0.026	0.021	4.198	0.741
Cross-validation	0.001	0.026	0.021	4.198	0.74
Test	0.001	0.026	0.021	4.204	0.742

As shown in Table 6: In order to evaluate the accuracy of the random forest model, this paper introduced a series of metrics including MSE (Mean Square Error), RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error).

The R^2 of the random forest model for the test set is 0.742, in view of this, this paper appropriately increased the proportion of the training set during the training process, but as the proportion of the training set is higher, the accuracy is lower, the reason could be that the training sample size is large, the data features are more, the network level is too little, and the features are not adequately trained leading to worse prediction results.

Thus, this paper introduced K-layer cross-validation, i.e., the original dataset is randomly divided into K subsets, one of which is used as the validation set, and the remaining K-1 subsets are used as the training set for K times of model training and validation. The overfitting phenomenon is reduced to some extent and the model performance is less sensitive to the division of data [13]. The improved prediction results are shown in Table 7:

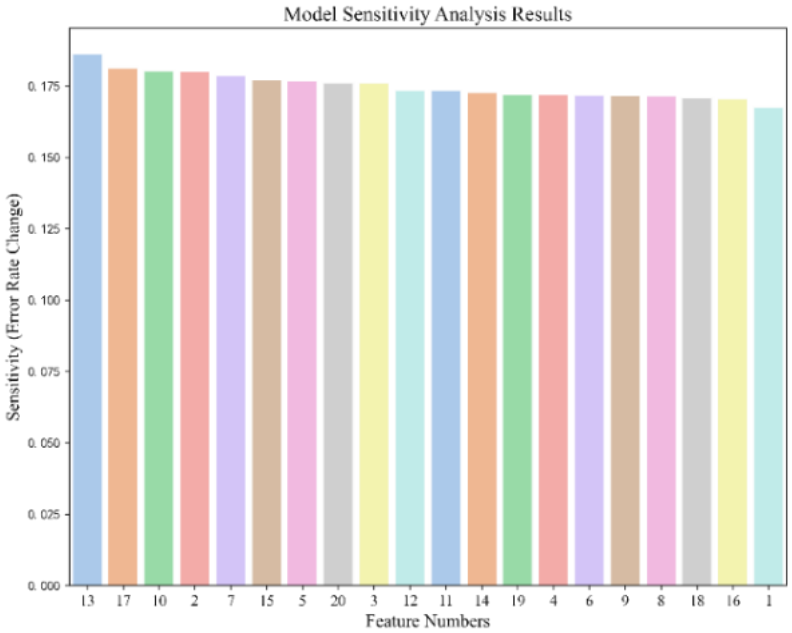
Table 7. Random Forest Model Evaluation Results (Improvement)

	MSE	RMSE	MAE	MAPE	R^2
Train	0	0.02	0.016	3.22	0.845
Test	0	0.02	0.016	3.214	0.846

Through continuous trials and improvements, there is a small increase in the accuracy of the test set, at which point the training ratio is: 0.8 and the cross-validation is: 50%.

3.5. Sensitivity analysis results

After completing the prediction of the flood risk class and the probability of flood occurrence, a sensitivity analysis is carried out next. Through this method, this paper is able to assess whether the prediction model is reliable and stable [14]. This paper selected 20 indicators for a systematic sensitivity analysis and quantitatively assessed the sensitivity of each indicator. As shown in Figure 5:



Number	Target
1	Monsoon intensity
2	Topographic drainage
3	Stream management
4	Deforestation
5	Urbanization
6	Climate change
7	Quality of dams
8	Siltation
9	Agricultural practice
10	Erode
11	Ineffective disaster prevention
12	Drainage system
13	Coastal vulnerability
14	Landslide
15	Watershed
16	Deterioration of infrastructure
17	Population score
18	Wetland loss
19	Insufficient planning
20	Policy factors

Figure 5. Visualization of indicator sensitivity

As shown in Figure 5, this paper draws the following conclusions:

Indicators such as coastal vulnerability, population score, erosion, topographic drainage, and dam quality have higher sensitivity and have a greater influence on the results while making model predictions. Indicators such as siltation, wetland loss, infrastructure deterioration, and monsoon intensity have lower sensitivity and have less impact on the prediction of model results. In summary, in the process of practical application, this paper needs to consider these indicators comprehensively and find a balance point, that is, the model has a certain sensitivity to the changes of these indicators, but can maintain a certain degree of stability and robustness [15].

4. Conclusion

The flood prediction method proposed in this paper improves the effectiveness of flood risk management in several aspects. First, using the K-means clustering method and logistic regression model, a flood risk level prediction model was established, and the overall prediction accuracy reached 73%. Secondly, this paper used a combination of Spearman's correlation coefficient and Cramér's V to select key indicators, and predicted the probability of flooding using the random forest model, which had a fitting degree of 74.2%. Finally, this paper introduces cross-validation to optimize the model, and the degree of fit is improved to 84.6%. The accuracy and robustness of the prediction of flood occurrence probability are significantly improved. In summary, the research results of this paper not only make significant progress in the accuracy and reliability of flood prediction, but also improve the efficiency and effectiveness of flood risk management in practical applications.

References

[1] Dong Y P, W H, et al. Xin'an River model and its application considering snowmelt and soil freezing and thawing--an example of runoff simulation in the upper Yalong River [J]. Progress in water sciences, 2024, 17: 1 - 13.

[2] Zuo X, Zhao X X, Ye R L, et al. Research on parameter rate determination of Xin'an River model based on improved adaptive genetic algorithm [J]. China Rural Water Conservancy and Hydropower, 2023, (11): 10 - 18+26.

[3] Wang T H, Gu Z H, Zhou T, et al. The ripple algorithm and its application to parameter optimization of the Muskingum model [J]. Journal of Water Resources and Water Transportation Engineering, 2024, (02): 81 - 90.

[4] Cai R, Xu W H. Socio-economic loss risks of future coastal flooding disasters in coastal cities of China [J]. China Population, Resources and Environment, 2022, 32 (08): 174 - 184.

[5] Shi Y L, Chun Y, et al. Study on the early warning model of lung cancer risk based on logistic regression of tongue features [J/OL]. China Journal of Traditional Chinese Medicine Information, 1 - 8 [2024 - 09 - 27].

[6] Wang Y B. Design of Early Warning Algorithm for Power Grid Protection under Strong Convective Weather Based on Improved Random Forest Algorithm [J]. Electronic Design Engineering, 2024, 32 (19): 182 - 186.

[7] Wang D L. Trend analysis of vegetation changes based on Spearman rank coefficients [J]. Journal of Applied Science, 2019, 37 (04): 519 - 528.

[8] Gao Y Q, Wang H, et al. Flood hazard risk assessment based on spatial information grids [J/OL]. Advances in water conservancy and hydropower science and technology.1 - 16 [2024 - 10 - 28].

[9] Wang B J, Hu A S, et al. Long-distance pipeline risk evaluation based on principal component analysis and random forests [J]. China Water Transportation (second half of the month), 2024, 24 (10): 145 - 147.

[10] Liu H D, et al. A study of a K-Means clustering algorithm based on improved differential evolution[J]. Modern Electronic Technology, 2024, 47 (18): 156 - 162.

[11] Guo S H. Impact analysis of watershed characteristics on the occurrence of different types of floods in dry and wet zones of China [J]. China Rural Water Conservancy and Hydropower, 2023, (12): 26 - 32.

[12] Zhang F, Zhang Y Y, Chen J X, et al. Evaluation of Multiple Machine Learning Models for Simulating Different Flood Type Characterization Indicators [J]. Progress in geosciences, 2022, 41 (07):1239 - 1250.

[13] Shao M L, Qi D Y. An Improved Random Forest Boost Multi-Label Text Categorization Algorithm [J]. Computer Applications and Software, 2022, 39 (11): 215 - 221+303.

[14] Zhang S. Flood hazard prediction and risk assessment based on data analytics and machine learning [J]. Information Technology and Informatization, 2024, 1 (08): 189 - 194.

[15] Li N, Wang Y, Wang J, et al. Flood risk management theory and technology[J]. China flood control and drought relief, 2022, 32 (01): 54 - 62.