Research on flood warning model based on dimensionality reduction and integrated neural network

Qianle Zhang, Kaihao Huang, Boyi Xiao *

South China University of Technology, Future Tech College, Guangzhou, China, 511442

* Corresponding Author Email: ft1328678827@mail.scut.edu.cn

Abstract. Flood disasters have long-term impacts on human societies and ecosystems, and population growth and urbanization lead to changes in surface conditions that increase the frequency and magnitude of floods. Therefore, understanding and predicting flood probability and constructing a flood warning mechanism are urgent problems. In this paper, a flood warning mechanism based on the probability of flood occurrence is proposed to construct a high-performance and highly interpretable model in a step-by-step manner using machine learning and deep learning methods. First, the correlation level is elucidated by the PCA dimensionality reduction method. Secondly, the K-means clustering method is used to replace the qcut binning method, and the classification model is built by combining the multidimensional data point distribution and classified by polynomial logistic regression. Finally, the integrated neural network model is constructed, the visual model performance is used for full feature optimization, and the performance of the classification model is optimized to achieve the flood prediction effect through the mean square error as the combination of the weights of the integrated model.

Keywords: Deep learning, Flood forecasts, Data downgrading.

1. Introduction

Flooding is a phenomenon of rapid rise in the water level of rivers, lakes and reservoirs caused by natural factors such as heavy rainfall, rapid melting of snow and ice, storm surges, etc., and is a type of natural disaster. It is directly linked to rapid population growth. With the population surge, human activities such as expansion of arable land, enclosure of lakes and deforestation continue to change the state of the land surface, and these changes not only affect the conditions of surface convergence but also exacerbate the disasters caused by floods.

The impact of floods on human societies in years of increased precipitation, and the severity of disasters, are likewise significantly influenced by human behavior. Historically, deforestation has played a significant role in exacerbating flood disasters. For example, soil erosion due to deforestation in the upper reaches of the Yangtze River is extremely serious, affecting an area of 350,000 square kilometers, with an average annual soil loss of up to 2.5 billion tons, and a large amount of this lost soil is silted up in rivers, lakes, and reservoirs, with annual siltation of about 2 billion tons.

Liu Mingjin, Zhang Zhiyong et al. used the linear regression model in sklearn library, and initially established a flood prediction model by analysing the real-time rainfall data and adding the parameters such as continuous rainfall, rain time and continuous rain days to optimise the model [1]. Lu Yanjie, Ni Qian et al. used a variety of highly interpretable Copula models to analyse the risk of flooding in a specific river basin [2]. Li Jiayu used a combination of hierarchical analysis and entropy weighting method subjectively and objectively to calculate the weights of the flood influencing factors so as to grade the target risk probability [3]. Seydi et al. (2023) compared the effectiveness of various machine learning algorithms such as Random Forests (RF), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) for flood sensitivity mapping [4].

However, these studies still face some limitations, such as the weak ability to deal with non-linear relationships, the scope of applicability is limited to specific regions, and multiple models are not jointly unified.

In this paper, SVG, XGBoost and other neural network methods are organically combined, and the introduction of flood hierarchical warning methods provides subdivided warning levels in the

early warning system, which effectively improves the accuracy and practicability of the prediction, and provides a new way of thinking for the establishment of a more accurate flood early warning research system in the future.

2. A deep learning-based approach for flood prediction model construction

2.1. Data preprocessing

The dataset used in this paper is from [Kaggle Dataset] (https: //www.kaggle.com/datasets/aritra100/flood-prediction-dataset/data), specifically "Flood Prediction Dataset ". This dataset includes scores for specific influences in different areas, as well as the probability of flooding in those areas.

Considering that the risk assessment and prediction model of flood disaster is built and evaluated on top of a data set with a large volume of data, the data set totals more than 1.7 million, to cope with the data quality and data missing problems that may arise under the condition of large volume data set, the experiment adopts the approach of data preprocessing to pre-check the data of database data to ensure the performance of the subsequent model and the accuracy of the algorithm.

First of all, the null value of the data set for the test, through the pandas comes with the test function of the original data screening, and secondly, the possible existence of anomalous data in the data set for the denoising process.

Observation of the original data set on the distribution of data on different indicators (shown in Figure 1), in line with the law of normal distribution, so the project uses the Z-core method to set up a threshold for the identification of outliers and denoising process[5], after processing the data statistical box plot shown in Figure 2.

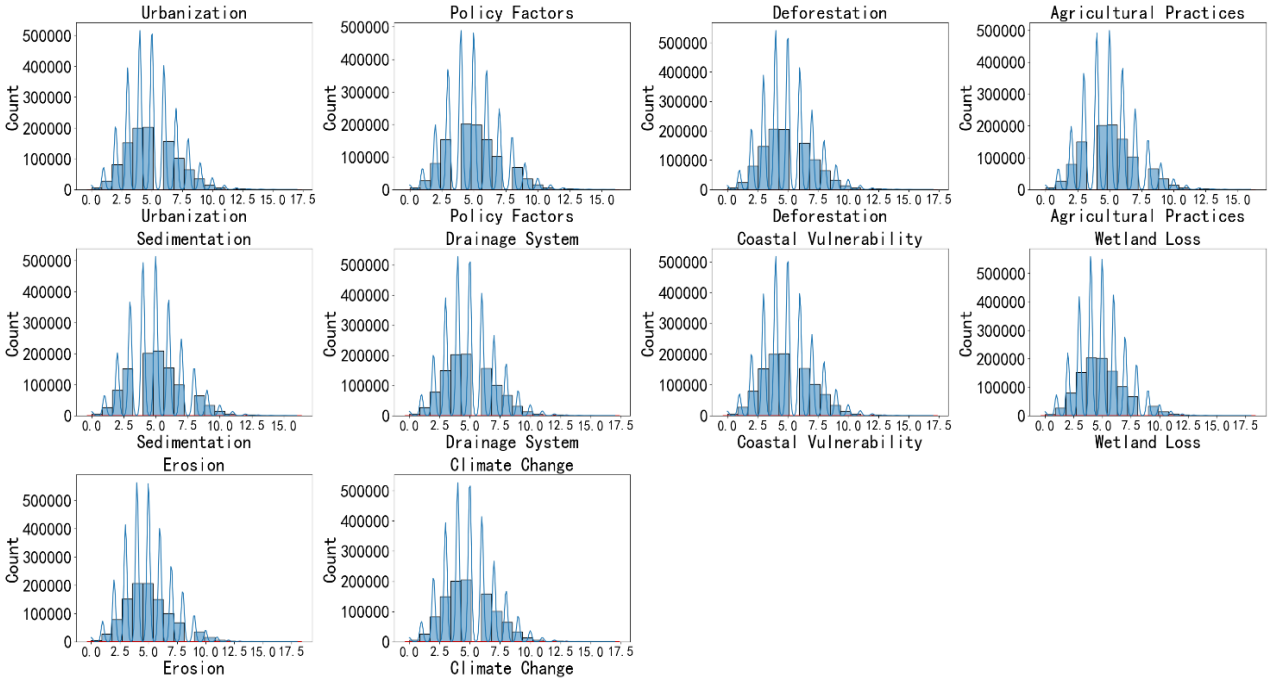


Figure 1. Bar chart of data distribution by indicator**.**

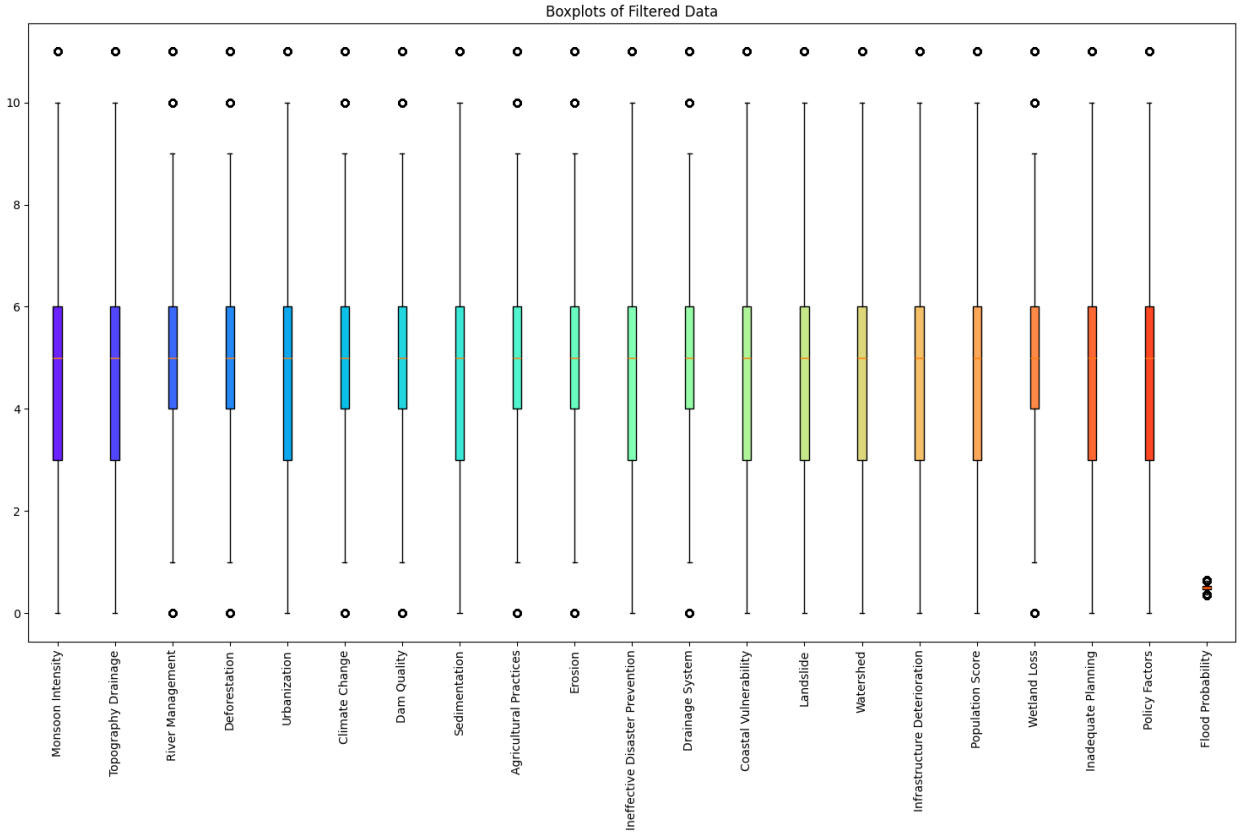


Figure 2. Box line plot of data distribution for each indicator after outlier treatment.

The project uses noise-processed data for model training and analysis to guarantee the accuracy and generalizability of the model. The training set data (train.csv) is divided before the experiment, which facilitates the direct implementation of the model performance test on the training set, to select and optimize the model.

2.2. Feature Selection and Flood Risk Early Warning Model Construction

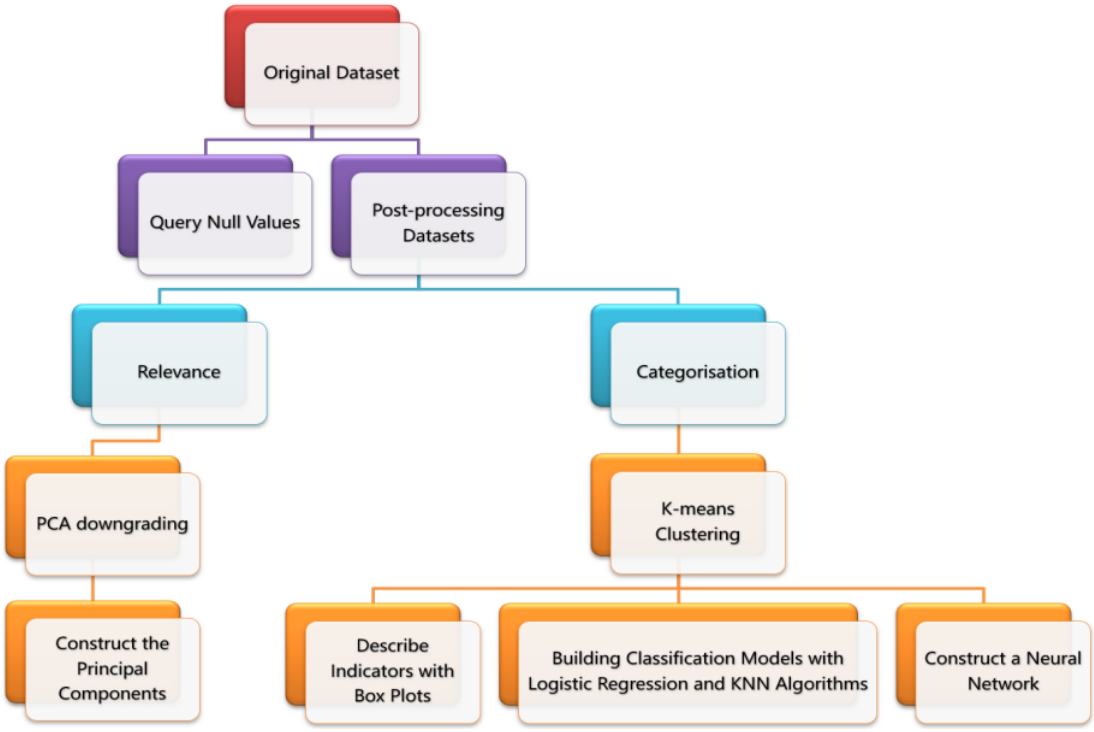


Figure 3. Data downscaling and flood warning model construction.

PCA (Principal Component Analysis) is a statistical method that transforms data to a new coordinate system through orthogonal transformation [6], and downscaling to simplify the data structure and reduce the computational cost of model training to achieve the purpose of feature selection. In the specific implementation process, PCA is affected by the scale of the data, you need to first standardize the data, so that each feature has a mean of 0 and a standard deviation of 1. On this basis, the original features are fitted by a specific weight to form a new principal component. After PCA processing, the original high-dimensional space of the data points is converted to a two-dimensional plane, which allows you to more intuitively observe the distribution of the data. Because the relationship with the features after the dimensionality reduction process is converted to a link with the fitted principal components, it is not possible to directly derive the link between the target variable and the original features from the data distribution graph, so in the process of solving the correlation, the experiment needs to analyze the link between the principal components and the original features, which can be derived by solving the weighting ratio.

To further carry out data feature selection to establish a reasonable flood prediction model, the probability of flood occurrence will be divided into three categories of high, medium and low-risk events, i.e., the categories of flood events are judged by the high and low probability of occurrence of flood events. However, considering the strict probabilistic statistical behavior of flood probability in the original data set. Or it as a probabilistic statistical behavior is greatly influenced by chance, for example, two data points with similar original probability, one of which has a flood event shortly after, leading to a significant increase in its probability. So the flood probability should not only be divided based on the probability prediction statistics, but also need to consider the situation where all the features of the data points are similar in terms of the probability of occurrence but have a large deviation, so the experiments here introduce the K-means clustering model to re-divide the dataset[7].

The K-means model divides the dataset into K clusters and calculates the Euclidean distance between each cluster to divide it so that the data points within each cluster are as similar as possible and the data points between different clusters are as different as possible. Figure 3 shows the schematic diagram of the whole data downscaling process.

Logistic regression is a statistical method widely used in classification problems, with high computational efficiency and realizability [8]. It is based on a linear model in the underlying principle, and at the same time, it has high sensitivity to the feature scale. The early warning and evaluation model of different risks of floods needs to predict the probability category of floods occurring in the belonging place in advance according to the attributes of the feature indicators of the specific location and predict the risk category belonging to it in advance, to play the role of early warning and evaluation. So the essence of the early warning evaluation model is a classification model built based on correlation indicators, and this experiment uses logistic regression for the construction of the classification model based on the two sets of eigenvalues in the first question.

2.3. Neural network flood prediction model construction

The prediction model of flood probability is constructed based on suitable indicators, based on the data feature indicators with high correlation with the occurrence of flood selected in section 2.2. In the experiment, firstly, the SHAP interpreter[9] and PCA processing method are adopted to calculate two groups of feature values for feature training, in the model selection experimental effect is better LightGBM, XGBoost, linear regression, and GBM model[4], through two groups of features four groups of machine learning machine for integrated training and prediction, due to the traditional linear regression algorithm on the complex nonlinear problem of the fit is weaker, so finally the integrated model is used to construct the flood prediction model[10]. Therefore, the integrated model is finally used to construct the flood prediction model.

On this basis, the existing models are integrated and processed to reach further optimized results. In the preparation stage of the integrated model, it is found that there exists a superior model LightGBM[11] for the model GBM, and at the same time, the model performance of GBM itself is

insufficient compared with the other three, so the model SVR[12] is used to replace the integration[13]. The integration method is shown in Figure 4.

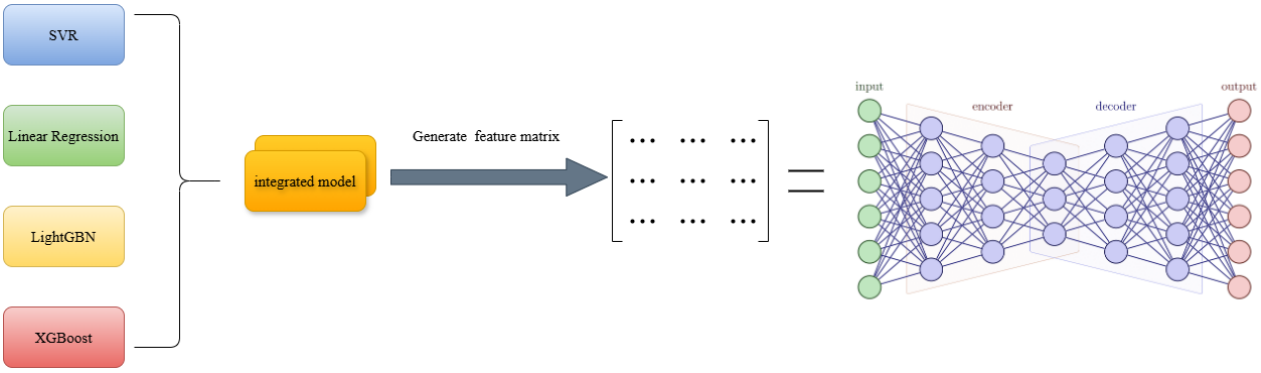


Figure 4. Integrated learning neural network construction.

3. Results

3.1. Feature Selection and Analysis of Flood Risk Early Warning Model Results

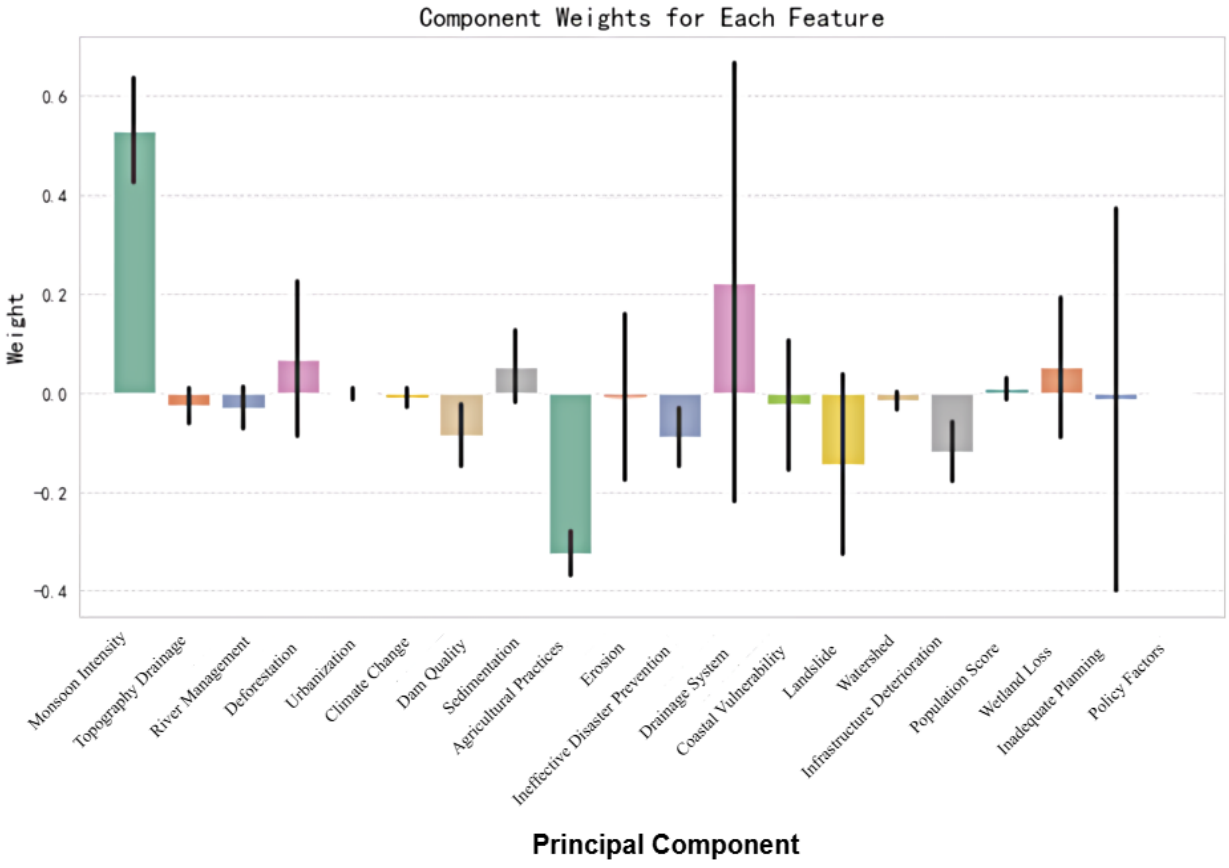


Figure 5. Distribution of PCA-treated compared to principal component data.

Figure 5 shows the features lie in the weight share of the principal components is a scalar with positive and negative values, and the negative correlation is also an important factor affecting the target variable, so the experiment uses a single feature relative to the absolute value of the share of the two principal components for the summation of the resulting combination of the summation of the value as a measure of the level of weights and correlation of the first ten features of the combination of the weight share of figure 6 shown below.

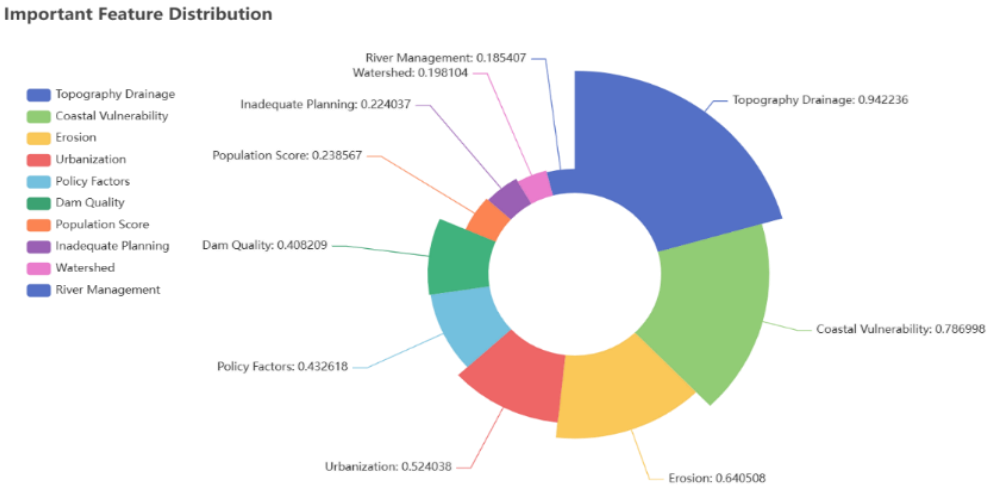


Figure 6. Percentage of portfolio weights for the top ten features.

The distribution of the data based on the same indicators for the three risk levels after the classification was completed showed a more significant level of difference, and the differences in the characteristics of the indicators in each clustering model were analyzed in the experiment by constructing box-and-line plots.

The ten features that yielded a strong correlation in the first question were selected in the training of the K-means model to prevent features with weak correlation from influencing the model, making the model present an overfitting state on a specific dataset and affecting the model generalizability.

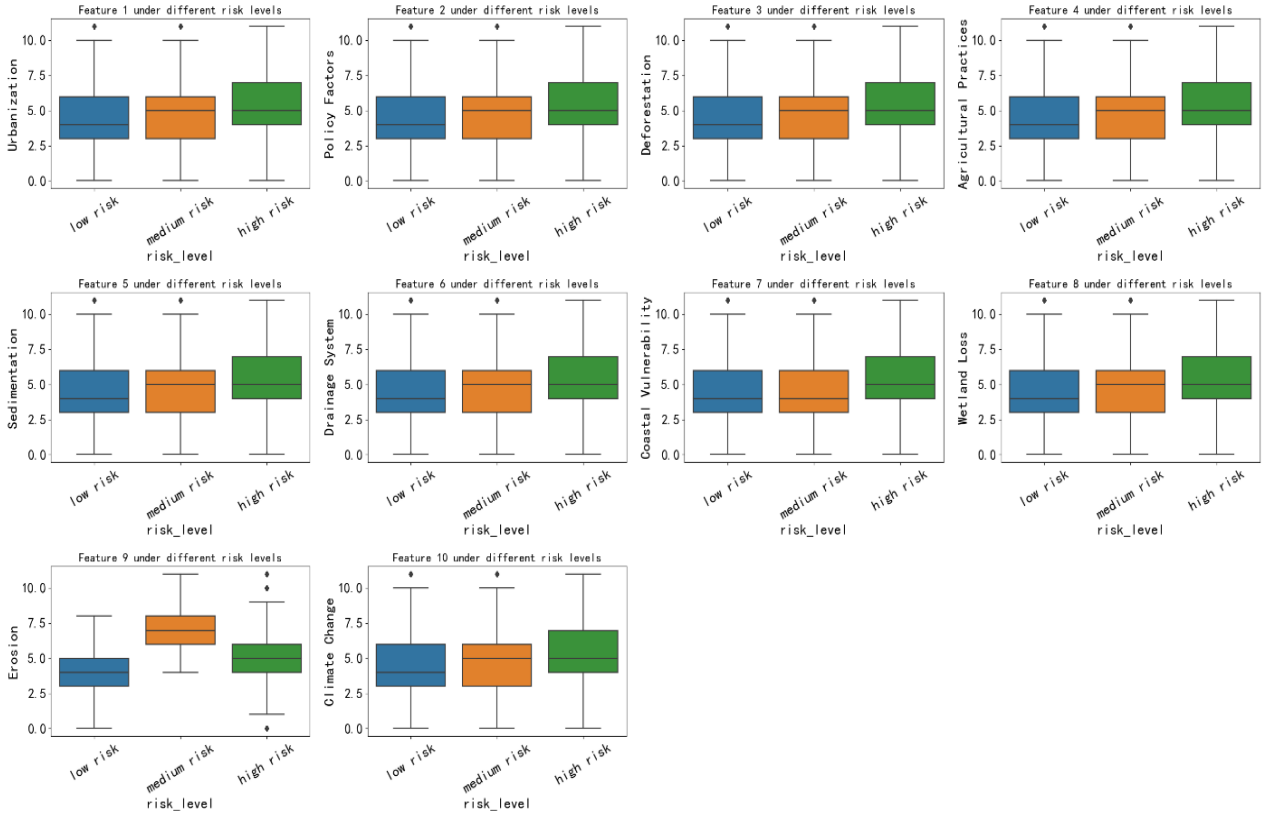


Figure 7. Characteristics of each important indicator in the case of K-means clustering.

Figure 7 shows the correlation between each feature and the overall, K-means clustering case of each important indicator of the overall features still shows a high level of positive correlation, only in a few features there is a difference.

The early warning and evaluation model of different risks of flooding needs to predict the probability category of flooding in the attributed place in advance according to the attributes of

specific location features and indicators and predict its risk category attribution in advance, to play the role of early warning and evaluation. Therefore, the essence of the early warning evaluation model is a classification model based on the correlation index, and this experiment uses logistic regression for the construction of the classification model based on two groups of characteristic values.

The following figure shows the display of the correlation coefficient of the classification model.

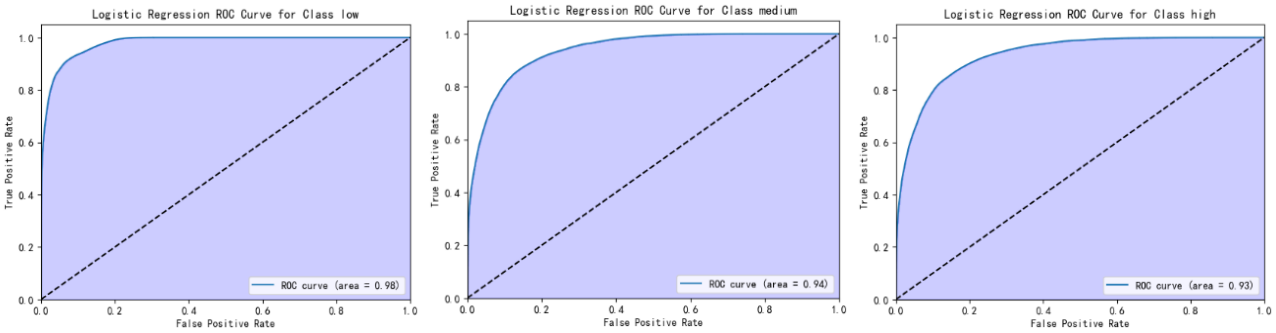


Figure 8. ROC curves for low, medium and high-risk logistic regression classification models under the SHAP interpreter.

Meanwhile, Figure 8 shows the model ROC curve correlation, the final value of the ROC curve is close to 1.0, and its curve enclosing area is also close to 1. The AUC value is high, and the model performance is good, which can achieve accurate classification. The model performance is excellent and the classification results are accurate.

3.2. Flood prediction model performance and analysis

By ensuring that the model is not only affected by a single variable and overfitting affects the model performance, the corresponding model performance results are shown in Figure 9:

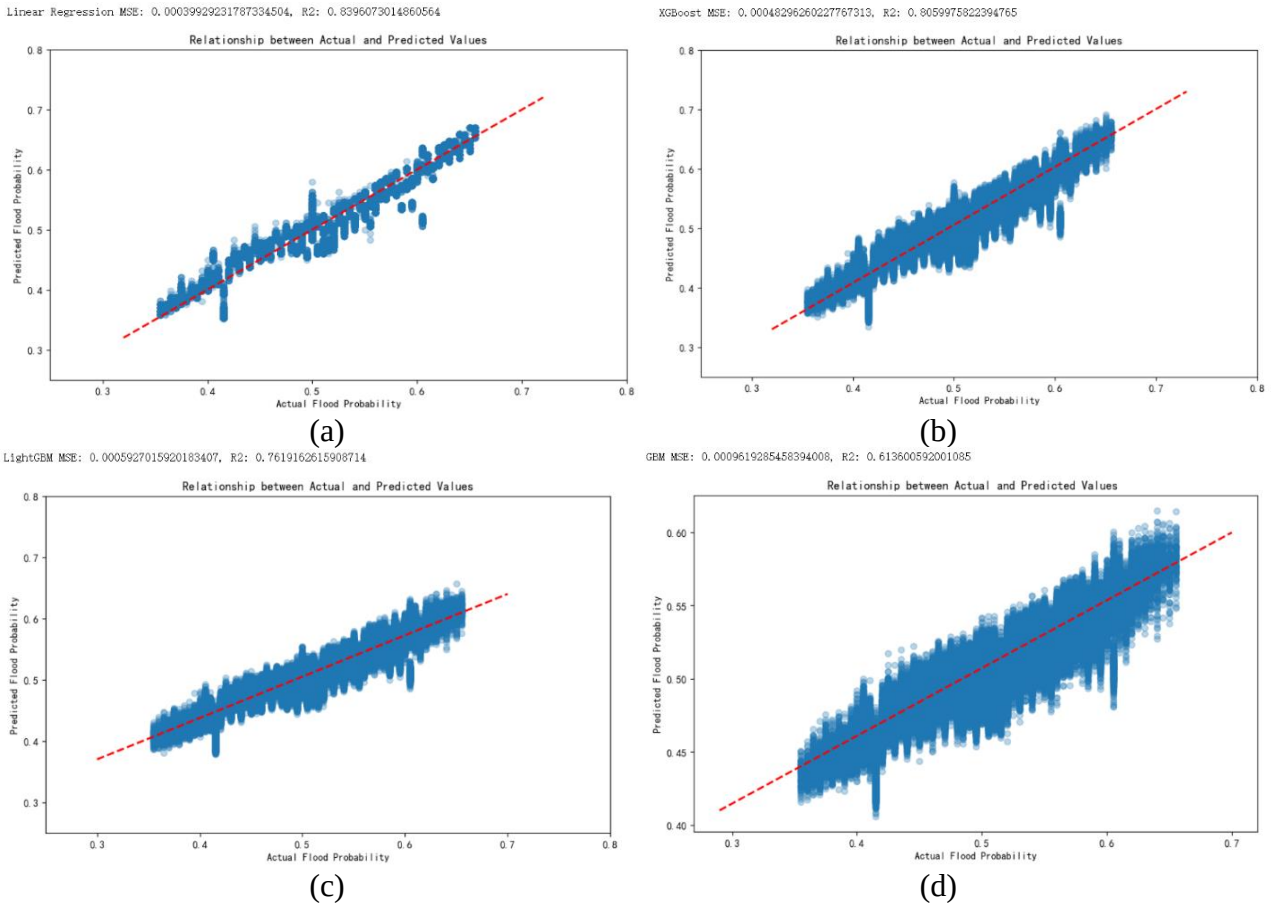


Figure 9. Full-featured model performance. (a)Linear (b) XGBoost (c)LightGBM (d)GBM).

As shown in Figure 9, the absolute coefficients of the models were significantly improved after optimization, and all four models were elevated to more than 0.6, while the highest absolute coefficient reached 0.83, which was a good fit. The absolute coefficient values calculated by the single model are first used as the weights for generating the parallel model, and the integrated model is trained to effectively synthesize the advantages of multiple base models. The feature matrix is created on the predicted values generated by the integrated model, which is then put into the neural network as a parameter for training to produce the final optimized model, and the results are shown in Figure 10:

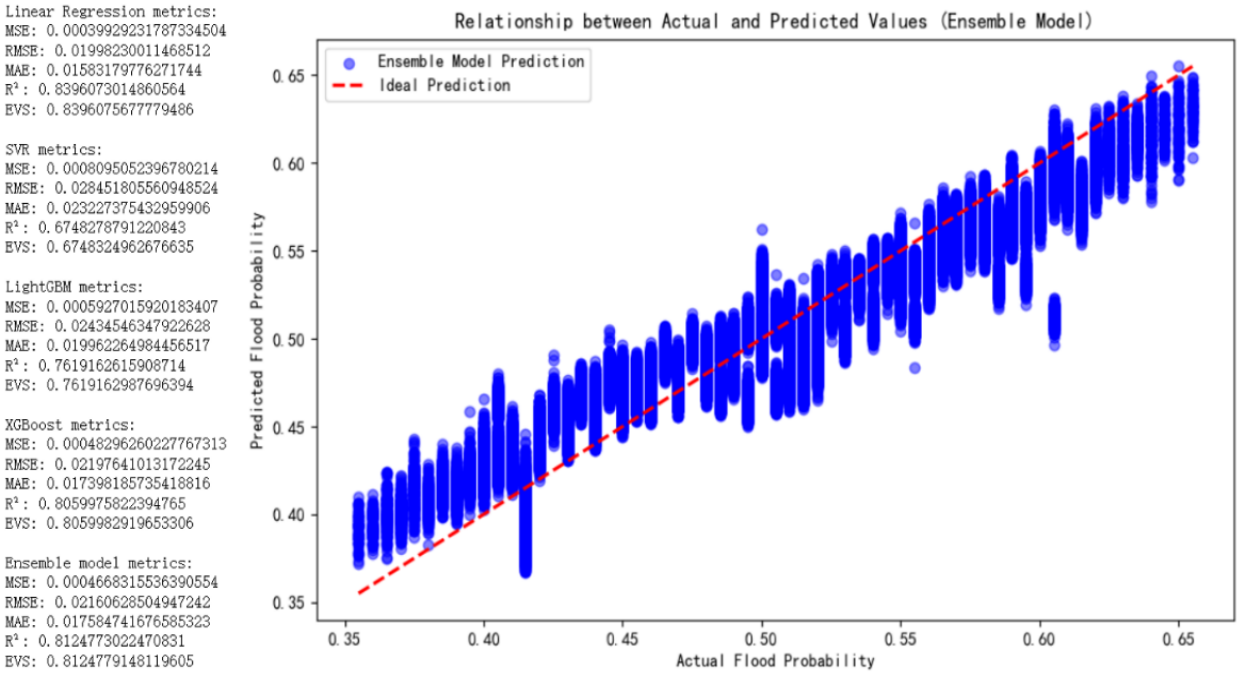


Figure 10. Final integrated model synthesized neural network performance plot.

As can be seen from the figure, the final integrated learning neural network possesses a better prediction ability with an absolute coefficient of 0.81, which possesses a better performance. However, it is worth noting that the prediction model trained using full features can easily achieve better training results on the current training set, but this effect may be due to model overfitting, i.e., the model is only highly adapted to the current dataset and does not present excellent results for other datasets. Therefore, after the experiment is finished training the above model it is necessary to observe the data distribution in the fourth question to avoid the abnormal situation of model overfitting.

If the condition restriction limits the number of feature values to five, the pre-processing needs to be modified. From the results of the first question, it can be concluded that the correlation level of each indicator is low compared to the target variable, so the indicators ranked at the back side of the importance actually have less impact on the model, and over-consideration of these indicators may even affect the overall model performance. The RTF selector is used in the experiment to gradually trim out the unimportant indicators and limit the final indicators to five for model training.

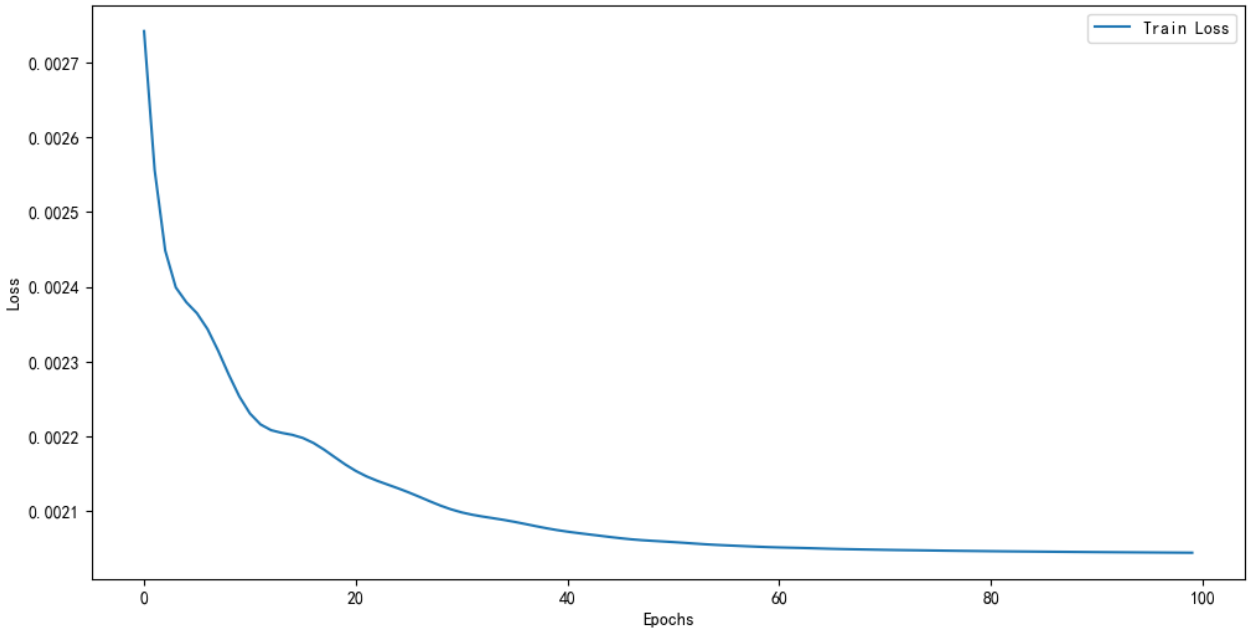


Figure 11. Five-feature neural network training process.

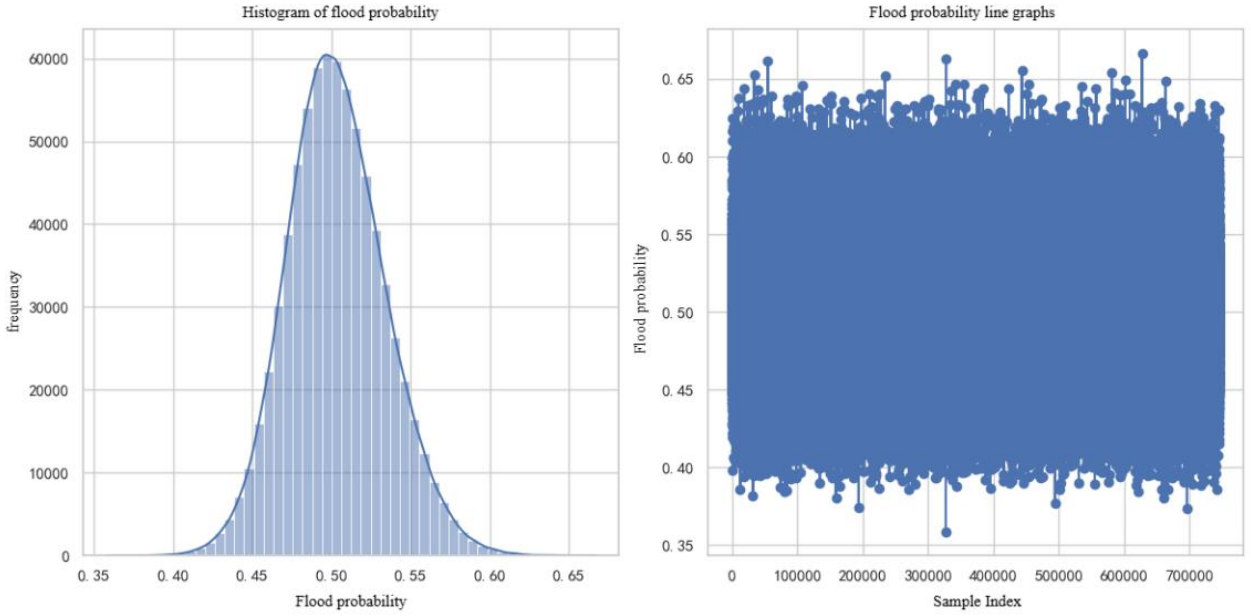


Figure 12. Distribution of projected data.

According to the analysis of the model results, Fig. 11 shows that the model mean square error is less than 0.002, and Fig. 12 exhibits that the data distribution of the prediction set demonstrates strong normality, and the K-S test statistic is less than 0.001, which is in line with the general law of data distribution of large datasets, and side-by-side reflecting the performance status of the model.

4. Conclusions

In this paper, a flood warning mechanism and a flood prediction model based on the probability of flood occurrence are proposed by combining machine learning and deep learning methods, aiming at gradually constructing a high-performance and highly interpretable model to cope with the prediction and warning of flood disasters. The PCA dimensionality reduction method is used to process the high-dimensional data and extract the effective features of the data. The risk warning model is analyzed and built by K-means clustering according to the sublinear type. Multiple models such as linear model, XGBoost, LightGBM, and GBM are integrated by multi-model integrated

learning method to improve the accuracy and reliability of prediction. However, due to the limited performance of the computing equipment, which results in the model not being able to fit completely, there is still room for performance improvement. In future research, extending the training time and optimizing the training process are expected to further improve the accuracy and stability of the model. In conclusion, this study has made significant progress in the flood warning mechanism and provided valuable methods and references for related fields.

References

[1] Liu Mingjin,Zhang Zhiyong,Wang Bin,et al. Research on flash flood prediction model based on machine learning and big data[J]. Sichuan Water Conservancy, 2022, 43(06):107-113.

[2] Lu Yanjie, Ni Qian, Gan Fuwan, et al. Comprehensive analysis of rain and flood encounters in the upper Yujiang section of Pinglu Canal based on multiple Copula functions[J/OL]. Hydrology, 2024:1-9.

[3] Li Jiayu. Probabilistic analysis model of flood overtopping risk in small basin water conservancy projects [J]. Journal of Wuhan University of Technology (Information and Management Engineering Edition), 2023, 45(04):527-531.

[4] Seydi ST, Kanani-Sadat Y, Hasanlou M, Sahraei R, Chanussot J, Amani M. Comparison of Machine Learning Algorithms for Flood Susceptibility Mapping[J]. Remote Sensing, 2023, 15(1):192.

[5] Hussain I M, Islam N S H. The Great Flood and its Aftermath in Kashmir Valley: Impact, Consequences and Vulnerability Assessment [J].Journal of the Geological Society of India, 2021, 97(6):661-669.

[6] Iulia A, Horațiu Ș, Cristian M, ET al.Mapping social vulnerability to floods. A comprehensive framework using a vulnerability index approach and PCA analysis [J]. Ecological Indicators, 2023,154.

[7] Chengshuai L, Caihong H, Chenchen Z, et al. Research on Urban Storm Flood Simulation by Coupling K-means Machine Learning Algorithm and GIS Spatial Analysis Technology into SWMM Model [J]. Water Resources Management, 2024, 38(6): 2059-2078.

[8] Wu Y, Zhang Z, Qi X, et al. Prediction of flood sensitivity based on Logistic Regression, eXtreme Gradient Boosting, and Random Forest modeling methods. [J]. Water science and technology: a journal of the International Association on Water Pollution Research, 2024, 89(10): 2605-2624.

[9] Wang M, Li Y, Yuan H, Zhou S, Wang Y, Adnan R, Ikram A, Li J. An XGBoost-SHAP approach to quantifying morphological impact on urban flooding susceptibility [J]. Ecological Indicators, 2023, 156: 111137.

[10] Chu W, Zhang C, and Li H, et al. SHAP-powered insights into spatiotemporal effects: Unlocking explainable Bayesian-neural-network urban flood forecasting [J]. International Journal of Applied Earth Observation and Geoinformation, 2024, 131: 103972-103972.

[11] Xu, K., Han, Z., Xu, H. et al. Rapid Prediction Model for Urban Floods Based on a Light Gradient Boosting Machine Approach and Hydrological–Hydraulic Model[J]. International Journal of Disaster Risk Science, 2023, 14: 79–97.

[12] Liong S Y, Chandrasekaran S. Flood Stage Forecasting with Support Vector Machines [J]. JAWRA Journal of the American Water Resources Association, 2007, 38: 173-186.

[13] Pradhan B, Lee S, Dikshit A, Kim H. Spatial flood susceptibility mapping using an explainable artificial intelligence (XAI) model [J]. Geoscience Frontiers, 2023, 14(6): 101625.