Risk Assessment and Classification Prediction for Water Treatment PPP Projects

Ruijia Yang (riga@hhu.edu.cn)
Hohai University Business School

Jingchun Feng
Hohai University Business School

Yong Sun
Guangzhou University

Research Article

Keywords: Water environment treatment, PPP projects, Government, Risk classification, Ensemble learning

Posted Date: September 1st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3199660/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Risk Assessment and Classification Prediction for Water Treatment PPP Projects

Ruijia Yang\textsuperscript{a}, Jingchun Feng\textsuperscript{a}, Yong Sun\textsuperscript{b,\,*}

\textsuperscript{a} Business School, Hohai University, Nanjing 211100, China
\textsuperscript{b} School of Public Administration, Guangzhou University, Guangzhou 510006, China

\* Corresponding author: School of Public Administration, Guangzhou University, Guangzhou, 510006, China. Institute of Rural Revitalization, Guangzhou University, Guangzhou, 510006, China. E-mail address: sunyong@gzhu.edu.cn

Abstract: Water pollution control is a crucial aspect of environmental safety and sustainable development. Public Private Partnerships (PPP) play a significant role in this control but are exposed to several risks. This study proposes a new risk classification prediction model for water treatment PPP projects to address these risks more effectively than traditional methods. The proposed model includes four key areas of risk: natural environment, ecological environment, socio-economic, and engineering entity. The study examines the correlation between these risk factors and project risk levels and develops an ensemble learning model based on Stacking for risk prediction. This model improves performance by using a weighted voting mechanism to adjust the importance of base learners. This model was tested using data from Phase I of the Jiujiang City water environment system project, demonstrating its effectiveness and accuracy. The proposed model outperforms other traditional machine learning models in terms of accuracy, macro-average precision, recall, and F1-score. Thus, it provides an effective method for risk classification prediction in water treatment PPP projects.

Keywords: Water environment treatment; PPP projects; Government; Risk classification; Ensemble learning
1 Introduction

The quality of any surface or groundwater body results from a combination of both natural and human processes. As population growth, industrial, and agricultural activities intensify, coupled with the threat of significant changes to the hydrological cycle due to climate change, the deterioration of water quality has escalated to a global concern (Li et al.2021; Su et al. 2022). Water environment treatment is essential for the protection of water resources, preservation of ecological equilibrium, and the promotion of sustainable economic growth. Water environment treatment projects, often implemented through Public-Private Partnerships (PPP), are gaining increasing attention due to their benefits such as relieving government fiscal pressures and effectively enhancing management efficiency.

In recent years, China has seen a consistent increase in investments in water environment treatment PPP projects. However, from a practical standpoint, many of these projects have failed to achieve their objectives due to inadequate risk management (Su et al.2022; Wang et al.2021). On the one hand, numerous risk factors are involved in the construction and maintenance of these projects, including local climate, hydrology, geology, flora and fauna, and the socioeconomic environment. These factors present high levels of uncertainty and danger, creating significant challenges for project risk management (Lu et al.2017). On the other hand, the primary motivation for social capital to participate in ecological protection and restoration is profit-driven. This is to say, with government policies' support and considering the market prospects and economic benefits of ecological restoration, investments are made to generate profits. However, compared to social capital, governments place greater emphasis on the subsequent environmental management effects of projects to maximize social and public benefits, while private investors tend to pursue short-term profit maximization without considering long-term operations (Tang et al.2021; Wang et al.2019). This conflict of interest often leads private investors to engage in opportunistic behaviors, such as providing low-quality products or services, resulting in inadequate treatment of polluted rivers. Such behaviors are more likely to occur when the government's project risk management is insufficient, posing potential threats to public safety and societal sustainability. These issues fundamentally stem from the public welfare and public attributes of PPP projects and water environment treatment. Decision-makers often overlook these attributes of water environment treatment projects during risk management, and the main creators and participants of these attributes are the government.

As PPP projects possess public welfare and public attributes, they cannot generate ideal returns. Social capital is therefore less willing to invest, and due to insufficient financial funds, the government has become one of the main bearers of risk.
in water environment treatment PPP projects. However, from the government's perspective, risk management must consider public risks more, which is also one of the core objectives of water environment treatment projects. Existing research rarely discusses the perspective of government management of water environment treatment PPP projects, and even less frequently analyzes the interrelationships between risk indicators (Xue et al. 2020; Liu et al. 2018). Therefore, there is an urgent need to construct risk assessment and prediction models for water environment treatment PPP projects that consider public welfare and public attributes, and analyze the risk factors of such projects from the government management perspective, thus providing more effective guidance for risk control.

Accurate risk forecasting is indeed a crucial factor in determining the success or failure of PPP project risk management. General methods for risk prediction in PPP projects can be categorized as qualitative, such as literature reviews, case studies, and questionnaires (Li et al. 2022); semi-quantitative, such as the analytical hierarchy process, game theory, and fuzzy comprehensive evaluation (Zhang et al. 2021); and quantitative, such as artificial neural networks and value for money analyses (Wang et al. 2021). Expert judgment data is one of the most frequently employed data sources in PPP risk management. However, the subjectivity and ambiguity inherent in this data source pose significant challenges to the accuracy of risk analysis and assessment, and there is no unified standard for judgment (Owolabi et al. 2020). The complexity and specificity of PPP projects for water environment treatment generate numerous and diverse data sources, requiring extensive monitoring, sampling, and analysis. This situation, characterized by multi-source and heterogeneous data, poses significant challenges to project risk management. Traditional PPP project risk assessment and prediction methods that rely on expert survey questionnaires are complex, lack credibility, and are not applicable to large-scale multi-source heterogeneous datasets. They are especially unsuitable for PPP projects involving extensive multi-source heterogeneous data for water environment treatment. Machine learning, a potential solution to these problems, has been widely applied in the PPP domain (Zheng et al. 2021). Owolabi et al. used machine learning models including regression trees, support vector machines, and deep neural networks to predict potential delays in PPP projects (Owolabi et al. 2020). While a small number of researchers have examined construction project risk classification using machine learning techniques, the majority have relied on single machine learning algorithms for risk classification, and the accuracy and generalizability of single algorithms on different datasets require improvement (Huang et al. 2022). Consequently, a number of researchers have used ensemble combination models to process multi-source heterogeneous data, yielding new insights (Chou et al. 2013).

This study aims to address the issue of government risk management in water environment treatment PPP projects, seeking
to improve the success rate of such projects. Based on the multi-source heterogeneous characteristics of risk data in these projects, this study designs a risk feature set from the government's perspective and discusses how specific indicators contribute to project risk. An empirical analysis is conducted using the Phase I comprehensive water environment treatment project in Jiujiang's central urban area. To predict risks in water environment treatment PPP projects, this research employs ensemble learning techniques that incorporate classifier combinations. The objective of this study is to expand the methodologies of engineering risk management, enhance the rationality, comprehensiveness, and accuracy of project risk prediction, and provide a robust guide for governments to manage risks in water environment treatment PPP projects. This research, therefore, holds significant implications for improving the quality of environmental public goods and services supply, as well as enhancing the capability and efficiency of water pollution prevention and control.

2 Methods

2.1 Ensemble Learning Method Based on Stacking

Ensemble learning is a technique for integrating multiple distinct algorithms or the same algorithm with varied parameters into a single model. Boosting, Bagging, and Stacking are methods of ensemble learning that are frequently employed (Wang et al. 2023). Stacking-based ensemble learning, which generates a new model by combining the prediction information from multiple models via self-organizing sampling or cross-validation, has superior performance compared to a single algorithm. Therefore, the Stacking method is used to build an ensemble learning model in this study (Chung et al. 2023).

Choosing precise and diverse machine learning classifiers is essential when constructing the base classifier. The diversity of the classifiers enhances the ensemble's ability to capture various aspects of the data, while the precision ensures reliable individual predictions. For this study, we selected diverse algorithms from six machine learning algorithms known for their excellent classification performance. These are the K-nearest neighbors (KNN), classification and regression tree (CART), linear discriminant analysis (LDA), Naive Bayes classifier (NB), Support Vector Machine (SVM), and Water Environment Treatment Project Risk Support Vector Machine (WETPR-SVM) classifier.

The KNN classifier is selected due to its ability to handle non-linear data, and it doesn't require any prior knowledge of the data distribution. CART was chosen for its interpretability and capability of handling both numerical and categorical data. The LDA was preferred for its ability to maximize the separability among known categories. The NB classifier is...
renowned for its simplicity and efficiency, especially when dealing with high-dimensional datasets. The SVM is well-regarded for its effectiveness in high dimensional spaces and its use of a subset of training points, making it memory efficient. The WETPR-SVM classifier was chosen for its tailored design to handle the specific task of risk classification in water environment treatment projects. The logistic regression algorithm was selected as the meta-classifier, a common choice in Stacking ensemble models, due to its strong interpretability and its ability to provide probabilities for outcomes, which can be beneficial in understanding the confidence level of the predictions.

As depicted in Figure 1, this study juxtaposes the combination model based on Stacking ensemble learning with traditional single models. Initially, we determine the risk feature set of water environment treatment PPP projects. Subsequently, we introduce risk data, label them, and preprocess them, dividing the dataset into training and test sets. Using a support vector machine classifier, classification is performed, and data fitting is accomplished with various kernel functions. Final prediction results are secured via an enhanced voting mechanism, and individual prediction models are fused via the ensemble learning mechanism to generate a comprehensive prediction model that serves as the ultimate risk grading prediction model for water environment treatment PPP projects. The model is then assessed using commonly used model evaluation algorithms in machine learning, continuously refining the performance of the training model through training effects.

Fig. 1 Research roadmap.
2.2 Model Interpretability Method

Due to the limited interpretability of machine learning in comparison to conventional generalized linear models, even though some models can measure feature importance, the specific effect scales of features are not intuitive. This issue is resolved by the emergence of the SHAP (Shapley Additive Explanations) model explanation. The SHAP explanation model interprets features by calculating each feature's contribution to the predicted value, and the values employed (SHAP values) can quantitatively characterize each feature's contribution (Nordin et al. 2023). The greater the SHAP value, the more the feature contributes to the predicted value (Li et al. 2022). This study calculates the contribution of multiple risk features in water environment treatment PPP projects using the SHAP explanation model, which has good computational performance and intuitive characteristics.

As shown in Equation (1), the SHAP value of a feature in the SHAP explanation model is the weighted sum of all possible feature value combinations.

\[
\phi_j(\text{val}) = \sum_{S \subseteq \{1, \ldots, p\}\backslash\{j\}} \frac{|S|!(p-|S|-1)!}{p!} (\text{val}(S \cup \{j\}) - \text{val}(S))
\]

Where: \(S\) is a subset of features used in the model, indicating that \(j\) is not included in the set \(S\); \(p\) is the number of features; \(\text{val}(S)\) is the prediction of feature values in set \(S\); \(\phi_j\) represents the contribution of the \(j\) feature to \(\text{val}\).

2.3 Improved Voting Mechanism

Ensemble learning, which combines multiple predictive models, often yields significantly improved accuracy and generalization performance compared to individual predictive models. This combination is typically achieved through voting mechanisms. However, risk prediction in water environment treatment PPP (Public-Private Partnership) projects is characterized by high complexity. Moreover, the presence of many missing and anomalous values in the data during certain periods may result in some base learners demonstrating higher accuracy or suitability for addressing specific issues. In such cases, traditional weighted voting strategies might not fully leverage the unique characteristics and variances of each base learner, potentially affecting overall prediction accuracy. Therefore, this paper proposes a weighted voting scheme designed to better accommodate the uncertainty and complexity in risk prediction for water environment treatment PPP projects, as illustrated in Figure 2. \(W_1 \sim W_n\) represent the predicted outcomes of each model, with values
of 1 (indicating stability) or -1 (indicating instability). \( W \) denotes the integrated result of the \( n \) predictive models, calculated as \( W = \sum_{i=1}^{n} \alpha_i \ast W_i \). When \( W > 0 \), the project is deemed stable; when \( W < 0 \), the project is considered unstable; and when \( W = 0 \), the project requires verification through alternative means or is conservatively categorized as unstable. The weighting factor \( \alpha_i \) represents the relative importance of each base learner during the voting process and can be learned through the training dataset.

A viable method is to employ cross-validation to assess the performance of different base learners and assign weights accordingly. More specifically, the accuracy scores of the base learners can be used as their weighting factors.

![Fig. 2 Schematic diagram of integrated learning voting mechanism.](image)

### 3 Experiments and Analysis

#### 3.1 Data Collection and Analysis

The majority of research on risks in public-private partnership (PPP) projects focuses on risk identification and classification, risk analysis and evaluation, and risk allocation and management strategies (Wang et al.2018). Extensive research has been conducted over the past decade to investigate risk management issues in PPP projects, identifying various types of risks, such as financial, operational, political, and environmental risks (Xu et al.2010). Water environment systems are dynamic, complex, open systems with temporal, spatial, and volumetric variations. This complexity results in distinct techno-economic characteristics of water environment treatment PPP projects compared to purely commercial PPP projects, including strong quasi-public interest, high difficulty in integrating governance technologies, complex assessment of governance effects, and difficult project coordination and collaboration (An et al.2018). Current research difficulty is identifying risk factors in water environment treatment PPP projects.
Using keywords or subject terms in both Chinese and English, such as "PPP", "risk", and "water environment treatment", a combined search was conducted in databases such as CNKI, ISI Web of Science, and ScienceDirect to find relevant literature for risk factor analysis. This resulted in a preliminary list of risk factors for water environment treatment PPP projects, as detailed in Table 1 below. The existing literature mainly discusses risks caused by the government, risks caused by social capital, and risks generated by the external environment. Government-caused risks primarily stem from government involvement in project management, including tax adjustments (Li et al. 2022; Liu et al. 2018), government intervention and credit issues (Wang et al. 2021; Cu et al. 2019), and inadequacies in existing laws, regulations, and regulatory systems (Su et al. 2022; Feng et al. 2022). Social capital-caused risks mainly arise from actual project construction and operation, such as completion risks (Su et al. 2022; Feng et al. 2022), construction technology risks (Zhang et al. 2021; Zhang et al. 2021), contract change risks (Su et al. 2022; Feng et al. 2022; Li et al. 2019), delay risks (Su et al. 2022; Li et al. 2019; Li et al. 2019), cost overrun risks (Su et al. 2022; El-Kholy et al. 2021), insufficient project revenue risks (Su et al. 2022; El-Kholy et al. 2021), dispute and infringement risks (Wang et al. 2019; Fu et al. 2023; Chou et al. 2013), and social capital change risks (El-Kholy et al. 2021; Wang et al. 2019). External environment risks refer to risks directly or indirectly caused by the external environment, including environmental damage risks (Su et al. 2022; An et al. 2018; Owolabi et al. 2020), geological condition risks (Feng et al. 2022; Cui et al. 2019), social stability risks (Li et al. 2021; Wang et al. 2019), public satisfaction (Li et al. 2020; Fu et al. 2023), inflation risks (Zhang et al. 2021; Wang et al. 2018), and force majeure (Li et al. 2018; Wang et al. 2018).

Although these summarized risk indicators can provide some reference for this study, they mostly analyze the project itself and do not consider the different categories of risks that different project participants should bear. Moreover, many of them involve qualitative data, which is often difficult to obtain comprehensively in practice.

<table>
<thead>
<tr>
<th>Risk Type</th>
<th>Specific Risk Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government-induced</td>
<td>Tax adjustment risk</td>
</tr>
<tr>
<td></td>
<td>Government intervention and credit issues</td>
</tr>
<tr>
<td></td>
<td>Inadequate legal and regulatory frameworks</td>
</tr>
<tr>
<td>Social capital-induced</td>
<td>Quality completion risk</td>
</tr>
<tr>
<td></td>
<td>Construction technology risk</td>
</tr>
<tr>
<td></td>
<td>Contract change risk</td>
</tr>
</tbody>
</table>

Table 1. Preliminary List of Risk Factors for Water environment treatment PPP projects.
<table>
<thead>
<tr>
<th>Risk Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schedule delay risk</td>
<td></td>
</tr>
<tr>
<td>Operational cost overrun risk</td>
<td></td>
</tr>
<tr>
<td>Project revenue shortfall risk</td>
<td></td>
</tr>
<tr>
<td>Dispute and infringement risk</td>
<td></td>
</tr>
<tr>
<td>Social capital change risk</td>
<td></td>
</tr>
<tr>
<td>Environmental damage risk</td>
<td></td>
</tr>
<tr>
<td>Geological condition risk</td>
<td></td>
</tr>
<tr>
<td>Social stability risk</td>
<td></td>
</tr>
<tr>
<td>External environment-induced</td>
<td></td>
</tr>
<tr>
<td>Public opinion risk (public satisfaction)</td>
<td></td>
</tr>
<tr>
<td>Inflation risk</td>
<td></td>
</tr>
<tr>
<td>Force majeure risk (political, natural conditions)</td>
<td></td>
</tr>
</tbody>
</table>

Although these aggregated risk indicators can provide some reference for this study, they are largely derived from an analysis of the projects themselves and do not consider the differentiation in risk categories borne by different project participants. Moreover, they involve many qualitative data that are often difficult to comprehensively obtain in reality. Additionally, due to the scarcity of research on risks in water environment governance projects, the preliminary list compiled leans more towards risks in general PPP construction projects. Literature analysis fails to capture the unique risk characteristics of water environment governance PPP projects. Therefore, the preliminary list is not entirely applicable to the context of this study.

The primary objectives of water environment treatment PPP projects, from the government's perspective, are to restore aquatic ecosystems, address water pollution issues, and maximize environmental, social, and economic benefits (Li et al.2022). From the perspective of government regulation, the public interest and public attributes of water environment treatment projects must be considered., Government departments bear the supervisory responsibility for project operation and ecological restoration, which is often overlooked or less emphasized by other stakeholders. Consequently, many risk incidents arise from this aspect (Li et al.2020). In light of these realities, this study includes continuous operation risks and natural ecological environment risks among the categories of risks that the government must prioritize in public-private partnership (PPP) projects for water environment treatment. In addition, discussing specific water environment treatment projects allows for the collection of as much subjective risk data as possible and the application of empirical analysis to validate the model's viability.
As a result, Jiujiang City, one of the first batch of demonstration cities for green development in China's Yangtze River Economic Belt, was chosen as the area of study. Jiujiang, the only city in Jiangxi Province located along the Yangtze River, has 152 kilometers of Yangtze River shoreline and two-thirds of the water surface and shoreline of Poyang Lake, China's largest freshwater lake. High-quality water resources are Jiujiang's most important ecological assets for high-quality development. In promoting the "Yangtze River Protection", Jiujiang is therefore saddled with significant responsibilities, heavy burdens, and numerous obstacles. This PPP-modeled project has a total investment of 76.99 billion yuan, prioritizes ecological and green development, and focuses on the management of the "Two Rivers" basin system. It is modeled after the Phase I project of the comprehensive water environment treatment in the central urban area of Jiujiang.

Project risk factor data from the past five years was gathered through various channels, such as the Chinese Government's public data platform (https://www.mee.gov.cn/), various monitoring stations in Jiujiang City, and interviews with the joint winning units (http://www.ctg.com.cn/sxjt/index2/index_html). Utilizing publicly available data along with enterprise data supports the scientific nature of this study in terms of data accessibility and authenticity. Expert interviews were also conducted based on the preliminary list of project risk factors (refer to Table 2 for expert information), with the aim of compensating for the scarcity of research on risks in water environment governance projects.

<table>
<thead>
<tr>
<th>Basic Information</th>
<th>Category</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Affiliated Unit</td>
<td>Government agencies</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Institutions of higher learning</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Water environment management enterprises</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>General PPP project enterprises</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Within 1 year</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1-3 years</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>3-5 years</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Over 5 years</td>
<td>4</td>
</tr>
<tr>
<td>Related Project Work or Research Experience</td>
<td>Very well understanding</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Better understanding</td>
<td>7</td>
</tr>
<tr>
<td>Degree of Understanding of Related Projects</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Adjustments were made to the preliminary list of risk factors for water environment treatment PPP projects using resources such as the Chinese government's public data platform, various monitoring stations in Jiujiang City, and joint bidding units. Five years' worth of project risk factor information was compiled, and the specific contents of various risk categories were categorized. This resulted in a set of twelve risk data features for water environment treatment PPP projects, covering natural environment, ecological environment, socio-economic, and project entity subsystems.

Indicators of evaluation are detailed in Table 3 below. The natural environment and ecological environment subsystems primarily reflect the government's natural ecological environment regulatory risks, whereas the socio-economic subsystem takes into account the significant public impact of water environment treatment PPP projects. The project entity subsystem, on the other hand, takes into account the government's primary role in the supervision of the sustainable operation of PPP projects, excluding engineering risks that the government does not share, such as cost overrun risks and construction technology risks.

Table 3. Risk Data Feature Table for Water environment treatment PPP projects.

<table>
<thead>
<tr>
<th>System Name</th>
<th>Risk Feature Name</th>
<th>Risk Feature-Related Evaluation Indicator Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Environment</td>
<td>Water environment</td>
<td>Hydro-sediment, Water quality, Water temperature, Water level, sediment</td>
</tr>
<tr>
<td></td>
<td>Acoustic environment</td>
<td>Noise</td>
</tr>
<tr>
<td></td>
<td>Atmospheric environment</td>
<td>Dust, Exhaust emissions, Local climate</td>
</tr>
<tr>
<td></td>
<td>Surface environment</td>
<td>Solid waste, Soil nutrients, Geology, Soil erosion, Soil salinization, Soil marshification, Landslides</td>
</tr>
<tr>
<td>Ecological Environment</td>
<td>Terrestrial organisms</td>
<td>Terrestrial animal and plant growth risks</td>
</tr>
<tr>
<td></td>
<td>Aquatic organisms</td>
<td>Safety risks of aquatic animals, Aquatic plants, Aquatic microorganisms</td>
</tr>
<tr>
<td></td>
<td>Livelihood security</td>
<td>Public satisfaction, Employment opportunities</td>
</tr>
<tr>
<td>Socio-economicSubsystem</td>
<td>Local economic development</td>
<td>Regional industry, Regional agriculture, Urban planning, Surrounding landscape, Regional economic risk</td>
</tr>
</tbody>
</table>
Using the equal interval method, the collected data were labeled with risk levels, including low, medium, higher, and high labels. There were collected a total of 927 risk data, including 12 risk features and 37 risk feature evaluation indicators. Among them, there were 789 low-level risk data, 86 medium-level risk data, 37 higher-level risk data, and 15 high-level risk data. The vast majority of the data had a low level of risk, indicating that the collected data belonged to a small sample of unbalanced data. Deep learning strategies exemplified by multilayer neural networks are unsuitable for this scenario (Tsai et al. 2023; Abdoli et al. 2023). Consequently, a conventional machine learning strategy was chosen for this research.

In order to analyze the relationships between various indicators and levels, the collected data risk was presented in the

Fig. 3 Relationship between Existing System and Government Perspective System.
form of variable scatter matrix plots. In this study, a total of 29 indicator data were collected. Here, only four indicators have been selected: water quality, local industrial economy, local climate, and soil erosion. These indicators were represented by pairwise coordinates to illustrate the connection between the various risk levels and project risk characteristics. Figure 4 demonstrates the outcomes. Under the four risk levels, the distribution of the four indicators, water quality, local industrial economy, local climate, and soil erosion, is relatively concentrated. In other words, there is a correlation between various risk indicators, as well as a correlation between the indicators and the risk classification of water environment treatment PPP projects.

![Fig. 4 Partial Feature Risk Level Correlation Analysis.](image)

### 3.2 Risk Feature Contribution Analysis

By analyzing the contributions of different feature indicators, we can determine the level of influence of different features on the model, allowing us to better explain and adjust the model during its construction. Figure 5 illustrates the distribution of SHAP values for risk indicator feature values. It is evident that water environment risk, operation and maintenance management risk, and local economic development have a greater impact on the prediction of risk levels for water
environment treatment PPP projects. In other words, the higher these three indicators are, the greater the probability that the project risk level will be high, which is consistent with the real-world scenario (Li et al. 2022). From the government's perspective, the primary objective of water environment treatment PPP projects is to strengthen water environment governance to achieve ecological restoration. Clearly, water environment risk is the most influential indicator of project risk, and achieving this objective is contingent upon the project's ability to be effectively operated and maintained. Therefore, operation and maintenance management risk is a prerequisite for the project's sustainable implementation. Moreover, the ultimate objective of the government's implementation of water environment treatment PPP projects is to improve or promote local economic development, which is directly related to the well-being of the entire society. Given the ascending order of objectives, it is reasonable that the contribution value of local economic development risk is ranked last.

Fig. 5 SHAP Values of Risk Features.
3.3 Dataset construction

Based on the determined risk feature set for water environment treatment PPP projects, we filled in the missing data values. The specific method begins with setting a threshold to determine whether a feature is missing or not. If the percentage of missing values for a feature exceeds this threshold, the feature is removed. In this study, the threshold for missing feature deletion is set at 80%. If the threshold is not exceeded, the KNN algorithm is used to locate the k nearest samples to the sample with the missing value, and the average value of their corresponding features.

To ensure model accuracy and eliminate the influence of dimensions, we standardized the original risk indicator data using a standardization algorithm (Wang et al.2019). Equation (2) represents the formula, where $x_i$ represents the ith evaluation indicator of the nth risk feature and $x_{std}$ represents the data for the standardized risk evaluation indicator.

$$x_{std} = \frac{x_i - \frac{1}{n} \sum_{i=1}^{n} x_i}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^{n} x_i^2\right)}}$$  

After processing missing features and dimensionless treatment, the distribution of feature values is between 0 and 1. The dataset is divided into training and testing sets in a 7:3 ratio, and a support vector machine classifier is used to classify the data.

3.4 Model Construction and Training

The final performance of the Stacking ensemble learning model is largely determined by the accuracy and similarity of the base classifiers. A superior classifier based on ensemble learning should adhere to the "good but different" principle (Chung et al.2023). Initially, we conducted experiments with the scikit-learn machine learning library for Python on the Jupyter Notebook platform. Six classification models were independently developed using machine learning: KNN classifier, CART classifier, linear LDA classifier, NB classifier, SVM classifier, and WETPR-SVM classifier. Individually, we trained them on the training set using cross-validation, random search, and learning curves to determine the optimal hyperparameter combination.

The Support Vector Machine (SVM) classifier was used to perform hyperparameter tuning (Chou et al.2013). By adjusting multiple SVM parameters and comparing the use of linear kernel function (LinearSVM), Gaussian kernel function
(RBFSVM), and polynomial kernel function (Sigmoid), the classification accuracy was continuously enhanced. The results are depicted in Figure 6, and the accuracy of the test set is shown in Table 3.

Figure 6 and Table 4 reveal that the Gaussian kernel function (RBFSVM) achieved the highest accuracy on the test set in this experiment, with a value of 0.9043. Comparatively, the test set precisions of the linear kernel function (LinearSVM) and the polynomial kernel function (Sigmoid) were 0.8191 and 0.8297, respectively. These results suggest that the Gaussian kernel function (RBFSVM) provides superior classification performance for this problem. This may be due to the fact that the Gaussian kernel function can map the data to a higher-dimensional space, rendering the data linearly separable in the higher-dimensional space. Given that water environment governance PPP project risk classification issues may involve complex nonlinear relationships, the Gaussian kernel function is more suited to addressing such issues.

Table 4. Test set accuracy of kernel functions.

<table>
<thead>
<tr>
<th>Kernel function</th>
<th>LinearSVM</th>
<th>RBFSVM</th>
<th>Sigmoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test set accuracy</td>
<td>0.8191</td>
<td>0.9043</td>
<td>0.8297</td>
</tr>
</tbody>
</table>

Fig. 6 Experimental results of kernel functions.
3.5 Model Comparison and Evaluation

This study's ensemble learning model ultimately completes the four-classification task for water environment governance PPP project risk. Therefore, we use accuracy (Accuracy), macro-average precision (Macro_P), macro-average recall (Macro_R), and macro-average F1 score (Macro_F1) as four indicators to evaluate the performance of the model.

Accuracy is the ratio of the number of project risk samples correctly classified by the model to the total number of project risk samples, which reflects the overall classification accuracy of the model (Choubin et al. 2023). The formula for calculation is depicted in Equation (3):

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + FN} \tag{3}
\]

The average precision across all classes is represented by the macro-average precision. Precision is the ratio of the number of risks correctly classified into a particular category to the number of risks classified into that category. Equation (4) demonstrates its formula for calculating:

\[
\text{Macro}_P = \frac{1}{n} \sum_{i=1}^{n} P_i \tag{4}
\]

The macro-average recall is the mean of all recall values across all classes. Recall is the proportion of risk samples correctly classified by the model for a particular project risk category relative to the total number of risk samples in that category. The formula for its calculation is shown in Equation (5):

\[
\text{Macro}_R = \frac{1}{n} \sum_{i=1}^{n} R_i \tag{5}
\]

To evaluate the performance of a classification model in practical applications, it is frequently necessary to consider the model's precision and recall in depth. As a result, the F1 score, which is the weighted harmonic average of the two, is
used as an evaluation metric. The formula for calculating the macro F1 score, which represents the mean of the F1 scores for all classes, is shown in Equation (6):

\[
\text{Macro}_n F1 = \frac{1}{n} \sum_{i=1}^{n} F_i
\]  

(6)

In Equations (3) - (6), \(TP\) represents the number of positive samples predicted as positive by the model; \(FP\) represents the number of negative samples predicted as positive by the model; \(FN\) represents the number of positive samples predicted as negative by the model; \(TN\) represents the number of negative samples predicted as negative by the model; \(n\) represents the number of risk feature categories, and \(P_i, R_i\) and \(F_i\) represent the precision, recall, and F1 scores of the model for different categories, respectively.

To gauge the performance of various classification algorithms, we utilized multiple methodologies to train the dataset and assess their effectiveness. LDA is the solitary linear algorithm in this mix; the rest are nonlinear. The relevant steps involved are as follows: (1) partition the training set; (2) appraise the algorithm models using 10-fold cross-validation; (3) generate six unique models for predicting new data; and (4) compare their classification accuracy. As depicted in Table 5, the WETPR-SVM model yields the highest Accuracy, Macro_P, Macro_R, and Macro_F1 scores, which are 0.9025, 0.9055, 0.9026, and 0.9021, respectively. These results indicate that the WETPR-SVM model developed in this study surpasses traditional singular machine learning classification models in terms of overall performance. It is proficient at solving the classification problem of water environment governance PPP project risk with enhanced accuracy and generalizability, and it possesses superior classification capacity. Thus, WETPR-SVM is chosen as the optimal model for predicting the risk classification of water environment governance public-private partnership projects.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Macro_P</th>
<th>Macro_R</th>
<th>Macro_F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>0.8532</td>
<td>0.8561</td>
<td>0.8537</td>
<td>0.853</td>
</tr>
<tr>
<td>CART</td>
<td>0.8251</td>
<td>0.8252</td>
<td>0.8249</td>
<td>0.8253</td>
</tr>
<tr>
<td>LDA</td>
<td>0.8469</td>
<td>0.8471</td>
<td>0.847</td>
<td>0.8469</td>
</tr>
<tr>
<td>NB</td>
<td>0.8778</td>
<td>0.8781</td>
<td>0.8779</td>
<td>0.8775</td>
</tr>
<tr>
<td>SVM</td>
<td>0.8467</td>
<td>0.8465</td>
<td>0.8469</td>
<td>0.8462</td>
</tr>
<tr>
<td>WETPR-SVM</td>
<td>0.9025</td>
<td>0.9055</td>
<td>0.9026</td>
<td>0.9021</td>
</tr>
</tbody>
</table>
Comparing box plot 7 reveals that the range of prediction accuracy for the WETPR-SVM model is the smallest at 0.07, indicating relatively high stability and consistent performance across various datasets. In contrast, the difference between the upper and lower quartiles of prediction accuracy for the NB model is 0.2636, representing the greatest variation in prediction accuracy, which may be attributable to its inability to effectively manage multi-source heterogeneous data, missing values, and class imbalance issues. The lowest lower bound of the median prediction accuracy for the CART model is 0.8251, which may be because it is a classification algorithm based on a decision tree that is sensitive to noisy data and overfitting issues. These results indicate that the WETPR-SVM model is superior for predicting the risk classification of water environment treatment PPP projects.

Fig. 7 Box plot of accuracy results for each classifier.

In conclusion, the ensemble learning-based approach proposed in this study can better utilize the benefits of various machine learning algorithms, overcome the limitations of single algorithms in dealing with multi-source heterogeneous data, missing values, and class imbalance issues, and improve prediction accuracy and generalization ability.

4 Conclusion

In this study, we proposed an effective method for risk assessment and prediction of water environment treatment PPP projects based on ensemble learning. Initially, we introduced a risk data feature set for water environment PPP projects, comprising four subsystems: the natural environment, ecological environment, socio-economic environment, and engineering entity. Secondly, we applied statistical analysis to evaluate the relationship between various feature indicators
and project risk levels, and to ascertain the contribution values of risk features. Subsequently, utilizing the concept of Stacking, we devised a two-layer ensemble learning model for classifying and predicting risks in water environment treatment PPP projects. Finally, we conducted an empirical analysis on the first phase of the comprehensive water environment system governance project in the central urban area of Jiujiang City, demonstrating the high effectiveness and accuracy of the risk assessment system and evaluation model established in this study.

The empirical investigation leads to the following conclusions:

(1) By assessing the risks of water environment treatment PPP projects and synthesizing previous research, we compiled a preliminary list of risk factors for these projects, which includes risks imposed by the government, social capital, and the external environment. From the viewpoint of government oversight, we determined the final set of risk features for water environment treatment PPP projects. The natural and ecological environment subsystems embody the government's regulatory role in natural ecological environments, the socio-economic subsystem indicates the substantial impact of water environment treatment PPP projects on public life and work, while the engineering entity subsystem reflects the government's regulatory role in the sustainable operation of water environment governance projects.

(2) By scrutinizing the correlation between risk feature indicators and risk levels, and by calculating the SHAP values of risk features, it becomes apparent that risk feature indicators correlate at different risk levels, and their contribution values to risk prediction are not uniform. Water environment risk, operational risk, and local economic development risk have higher contribution values. In recent years, water pollution in Jiujiang City has significantly affected urban development, and enhancing the water environment is the primary goal of governance projects. Water environment risk is the most direct risk faced by the project, operational risk pertains to the project's sustainability, and local economic development risk represents one of the crucial factors determining project success. These elements will be the focal points for future risk prevention and response strategies for water environment treatment PPP projects in Jiujiang City, China. Relevant government departments can utilize this as a foundation for augmenting the legal and policy system construction of water environment treatment PPP projects.

(3) To address the issue of missing and abnormal values in risk data for water environment treatment PPP projects during certain periods, we innovatively designed a weighted voting mechanism. By introducing weight factors to adjust the relative importance of base learners in the voting process, the model can better harness the differences between base learners, thereby enhancing the risk prediction accuracy for water environment treatment PPP projects.
(4) Regarding risk classification prediction models for water environment treatment PPP projects, the Accuracy, Macro_P, Macro_R, and Macro_F1 of the WETPR-SVM ensemble learning model developed in this research are 0.9025, 0.9055, 0.9026, and 0.9021, respectively. This model circumvents the subjectivity and ambiguity of traditional expert-based scoring methods and delivers higher prediction accuracy compared to conventional machine learning models.

There are still many areas for improvement in this study, and we suggest the following for future work. On the one hand, this study analyzes only one water environment governance project. In the next stage, we aim to collect more actual cases to enhance the risk feature set and further validate and optimize the prediction model. On the other hand, the ensemble learning model designed in this study does not provide sufficient diversity in combination types. Moving forward, we plan to explore the application of other advanced machine learning and artificial intelligence technologies in risk assessment and prediction for water environment treatment PPP projects to boost prediction accuracy.

**Ethical approval:** This study did not involve any experiments involving humans or animals, so ethical approval is not required. All research activities are conducted in accordance with the appropriate industry guidelines and regulations.

**Consent to participate:** All participants voluntarily agreed to participate in the study with a clear understanding of its purpose, procedures, and possible risks and benefits.

**Consent to publish:** All participants have explicitly given their consent to allow us to use and publish their data and/or images in the study.

**Author Contributions:** Conceptualization, Ruijia.Yang and Yong.Sun; methodology, Ruijia.Yang; software, Ruijia.Yang; validation, Ruijia.Yang and Yong.Sun; formal analysis, Ruijia.Yang and Jingchun.Feng; investigation, Ruijia.Yang and Yong.Sun; resources, Ruijia.Yang; data curation, Ruijia.Yang and Jingchun.Feng; writing—original draft preparation, Ruijia.Yang; writing—review and editing, Ruijia.Yang and Yong.Sun; visualization, Ruijia.Yang; supervision, Jingchun.Feng and Yong.Sun; project administration, Jingchun.Feng; funding acquisition, Jingchun.Feng. All authors have read and agreed to the published version of the manuscript.

**Funding:** Work was supported by the National Social Science Funds of China, grant number 17BGL156.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Availability of Data and materials:** The raw data cannot be made publicly available due to personal privacy concerns.

All the detailed steps of data processing and analysis are described in the article, and detailed information can be requested from the author.
References


