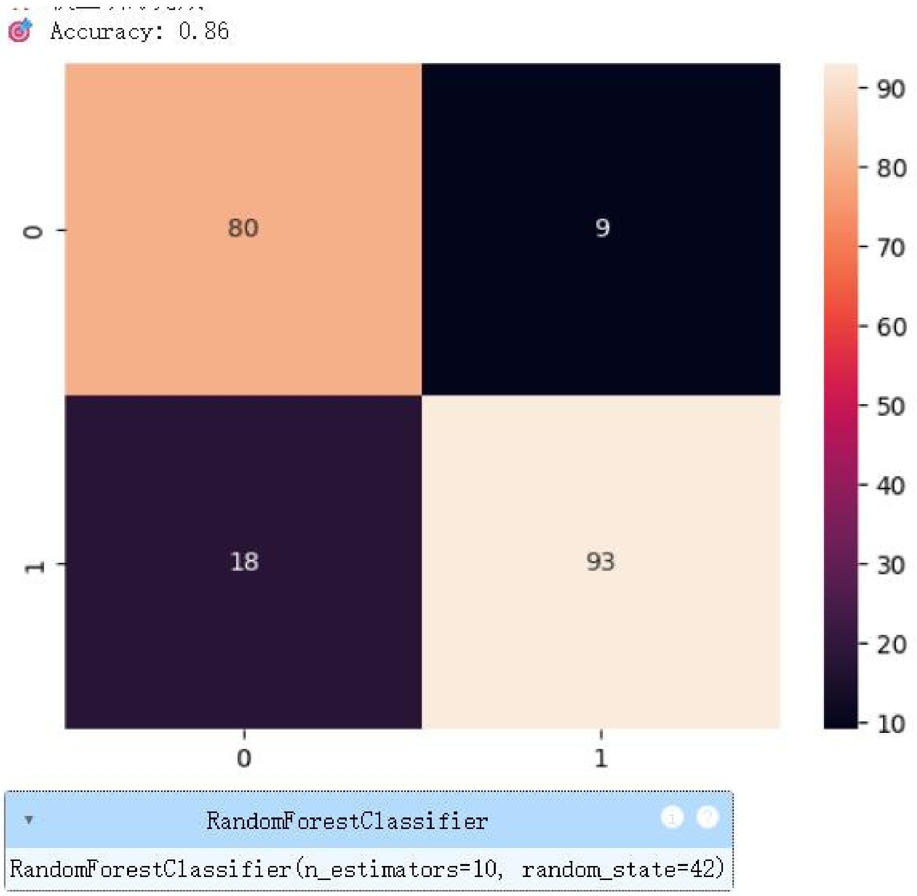
RESEARCH ON SOIL ENVIRONMENTAL AND HEALTH RISK ANALYSIS BASED ON MACHINE LEARNING.

ZHAO ZHIHAN

UNIVERSITI TEKNOLOGI MALAYSIA

4.1 Random Forest



```
Accuracy: 0.86
```

```
RandomForestClassifier(n_estimators=10, random_state=42)
```

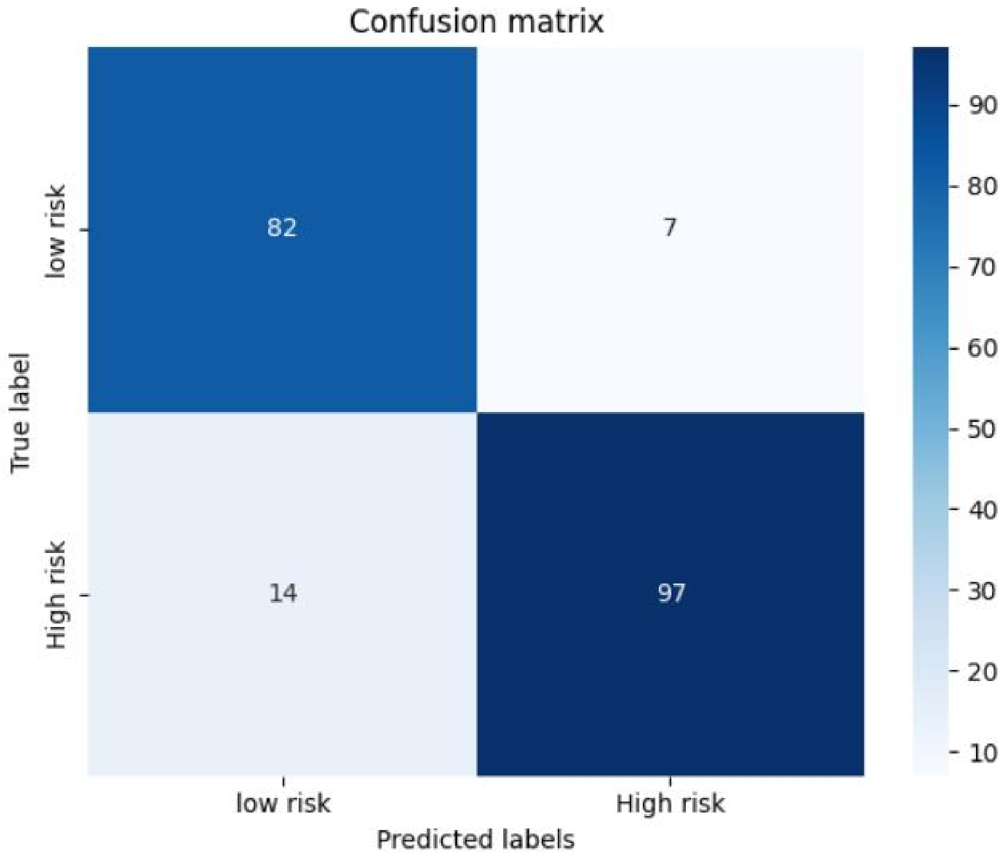
The confusion matrix and training results of the Random Forest classifier (Random Forest Classifier, with parameters n estimators=10 and random state=42) in the scenario of soil environment and health risk analysis. The overall accuracy rate of this model reaches 0.86.

After optimizing the parameters of the Random Forest model, through Grid SearchCV (grid search + 5 - fold cross - validation), within the preset ranges of number of decision trees, tree depth, and minimum number of samples for node splitting, the parameter combination that optimizes the accuracy is found. The optimal parameters are max depth = 10, min samples split = 5, and n estimators = 200. The final accuracy rate reached 0.91.

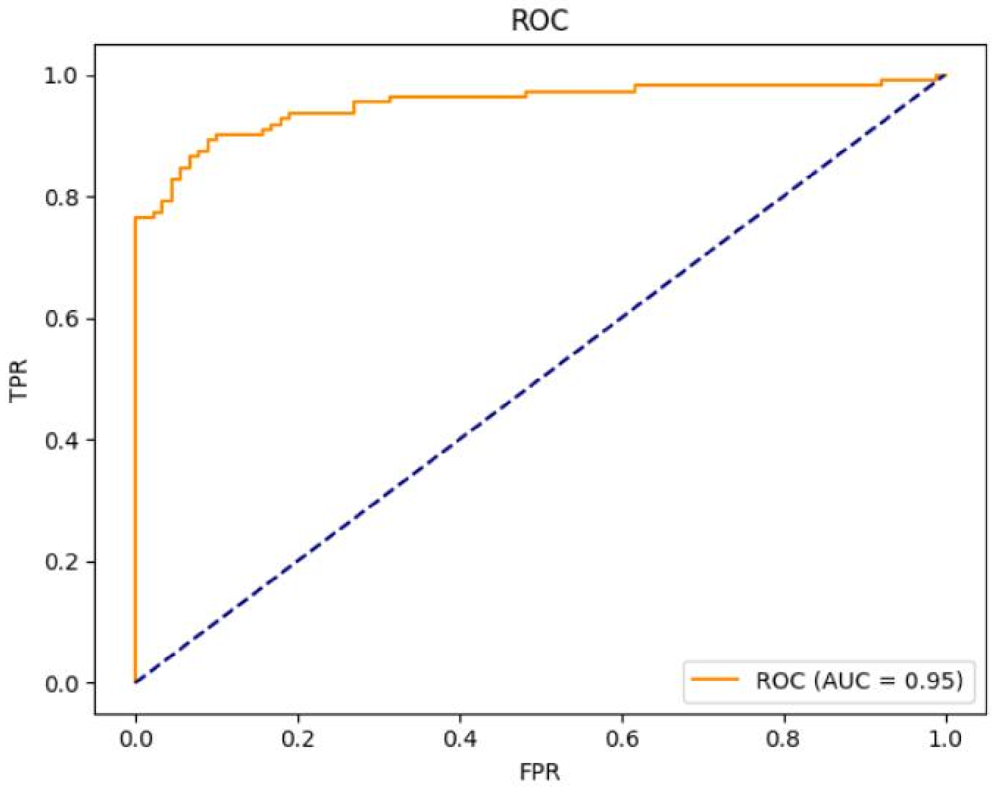
4.2 XGBoost

The performance of the XGBoost model was visually evaluated through the plot_model_metrics function, and the confusion matrix and ROC curve were output.

The rows represent the true risk labels , and the columns represent the model - predicted labels. The data shows that among the true low - risk samples, 82 cases were correctly predicted, and 7 cases were false positives . Among the true high - risk samples, 97 cases were accurately identified, and 14 cases were false negatives . This matrix intuitively presents the XGBoost model's ability to identify different risk categories. The accuracy of low - risk identification (precision = 82/(82 + 7) ≈ 0.921) and the accuracy of high - risk identification (recall = 97/(97 + 14) ≈ 0.874) indicate that the model has certain reliability in the soil health risk classification task, but there are still cases of misjudging high - risk samples, which need to be optimized in subsequent research.

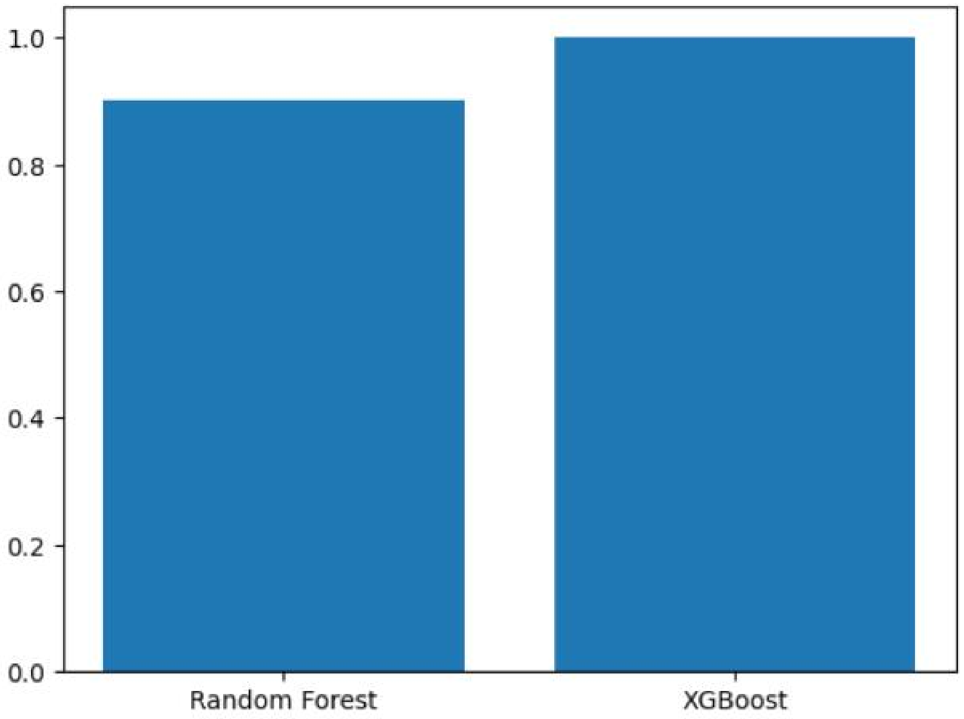


The horizontal axis represents the False Positive Rate (FPR), and the vertical axis represents the True Positive Rate (TPR). The Area Under the Curve (AUC) reaches 0.95. An AUC value close to 1 indicates that the XGBoost model has a strong ability to distinguish between high - and low - soil - health - risk samples and can effectively identify risk patterns. The trend of the ROC curve also reflects the model's performance in balancing the sensitivity and specificity of risk identification under different threshold settings, providing a reference for reasonably setting the risk determination threshold in practical applications.



4.3 Comparison between Random Forest and XGBoost models

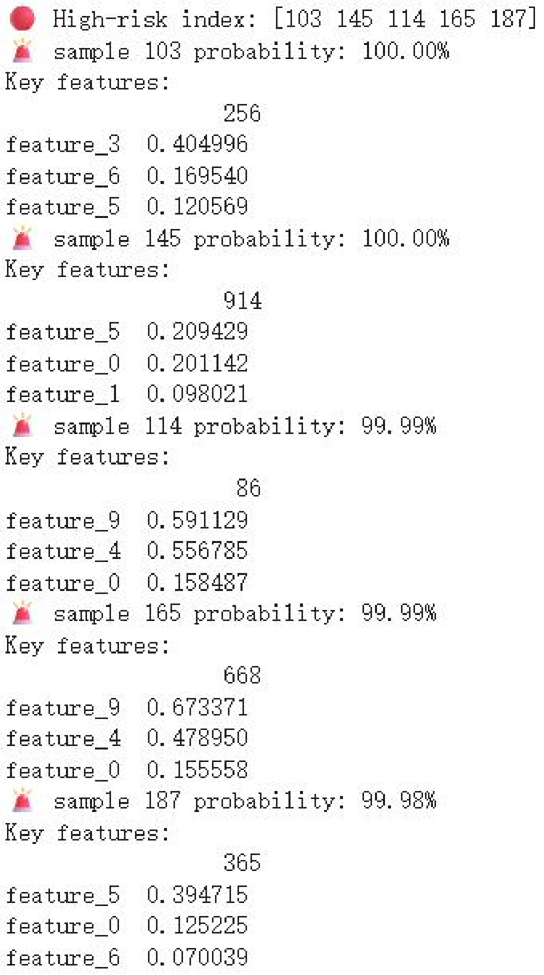
The XGBoost model was trained and tested, and then compared with the Random Forest model.



The final comparison results are as follows: Both models take very little time, and the Random Forest is slightly faster than XGBoost. In terms of accuracy, the Random Forest reaches 0.90, while XGBoost reaches 1.00 (the reason for the 100% accuracy may be that the dataset is too small and the training task is too simple). In the analysis of soil environment and health risks, the XGBoost model has a higher accuracy and performs better, indicating that it fits the complex relationships of soil characteristics better.

4.4 determine that the risk is high

The samples are divided into two parts: an 80% training set and a 20% test set, and the XGBoost is used to train the model. Then, the indexes of the 5 samples with the highest risks are selected. Combining with the feature importance of the model, the key influencing features and their corresponding contribution values of each high - risk sample are calculated and output. The results show that the risk probabilities of the 5 high - risk samples (indexes 103, 145, etc.) exceed 0.99.



```
🔴 High-risk index: [103 145 114 165 187]
🚨 sample 103 probability: 100.00%
Key features:
                 256
feature_3  0.404996
feature_6  0.169540
feature_5  0.120569
🚨 sample 145 probability: 100.00%
Key features:
                 914
feature_5  0.209429
feature_0  0.201142
feature_1  0.098021
🚨 sample 114 probability: 99.99%
Key features:
                  86
feature_9  0.591129
feature_4  0.556785
feature_0  0.158487
🚨 sample 165 probability: 99.99%
Key features:
                 668
feature_9  0.673371
feature_4  0.478950
feature_0  0.155558
🚨 sample 187 probability: 99.98%
Key features:
                 365
feature_5  0.394715
feature_0  0.125225
feature_6  0.070039
```

For sample 103, the features with high contribution correspond to "lead concentration" and "soil pH value", indicating that the soil environment of this sample facilitates the migration of heavy metals to organisms, potentially triggering health risks. If the key features of sample 114 involve "soil organic matter content" and "cadmium concentration", the combination of low organic matter and high cadmium concentration may lead to a significant increase in the bioavailability of cadmium, as the lack of organic matter fails to effectively adsorb heavy metals.