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

ZHAO ZHIHAN

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF THESIS

Author's full name	:Zhao Zhihan		
Student's Matric No.	:MCS241041	Academic Session	
Date of Birth	:19.07.2001	UTM Email	zhaozhihan@graduate.utm.my
Thesis Title	:	RESEARCH ON SOIL ENVIRONMENTAL AND HEALTH RISK ANALYSIS BASED ON MACHINE LEARNING	

I declare that this thesis is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the thesis belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this thesis for academic exchange.

Signature of Student:

Signature :
ZHAOZHIHAN

Date :
28.06.2025

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
NOOR HAZARINA HASHIM	MOHD ZULI JAAFAR
Date :	Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

Note: This page is only applicable for theses classified as Restricted/Confidential. Please delete this page from the template if your thesis is not classified as Restricted/Confidential. This letter should be written by a supervisor and addressed to Perpustakaan UTM. A copy of this letter should be attached to the thesis.

Date:

Librarian

Jabatan Perpustakaan UTM,

Universiti Teknologi Malaysia,

Johor Bahru, Johor

Sir,

CLASSIFICATION OF THESIS AS RESTRICTED/CONFIDENTIAL

TITLE: Click or tap here to enter text.

AUTHOR'S FULL NAME: Click or tap here to enter text.

Please be informed that the above-mentioned thesis titled _____ should be classified as RESTRICTED/CONFIDENTIAL for a period of three (3) years from the date of this letter. The reasons for this classification are

(i)

(ii)

(iii)

Thank you.

Yours sincerely,

SIGNATURE:

NAME: ZHAO ZHIHAN

ADDRESS OF SUPERVISOR:

"I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in term of scope and quality for the award of the degree of Master in (data science)"

Signature : _____

Name of Supervisor I :

Date :

Signature : _____

Name of Supervisor II :

Date :

Signature : _____

Name of Supervisor III :

Date :

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration Click or tap here to enter text. and Click or tap here to enter text.

Certified by:

Signature :
Name :
Position :
Official Stamp
Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

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

ZHAO ZHIHAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master in (data science)

School of Education
Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2025

DECLARATION

I declare that this thesis entitled *"Analysis of California's Decadal Wildfires: The Construction of Risk Management Models"* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	.....
Name	:	ZHAO ZHIHAN
Date	:	28 JUNE 2025

ACKNOWLEDGEMENT

We sincerely thank Dr Shahizan for assigning this research task to us. It has provided us with a valuable opportunity to apply academic writing skills to practical research. We are especially grateful to the teacher for their patient guidance in the design of the research framework and the expression of academic language. Your insights have greatly enhanced the logical structure and coherence of this article.

At the same time, I would like to express my gratitude to all the students for the constructive suggestions you provided during the class discussions. Your feedback on aspects such as data interpretation and result presentation has played a crucial role in improving the research content. Especially, I would like to thank you all for creating a collaborative learning environment during the research process, which has enabled us to make progress together through critical thinking.

The completion of this assignment would not have been possible without the supportive academic environment created by the writing course. I hope this paper can demonstrate the knowledge and skills acquired through the course guidance.

ABSTRACT

Global industrial and agricultural activities have exacerbated soil pollution, with pollutants entering the food chain through crops, posing a significant threat to public health. Traditional assessment methods lack the ability to integrate multi-source data (such as soil pollutants, meteorological factors, and health records), which limits the accuracy of spatial risk mode. This study proposes a machine learning framework that integrates 3,000 samples from the Kaggle "Soil Pollution and Health Impact Case" dataset (2023-2024). Through data preprocessing (KNN interpolation, outlier handling), feature engineering (interaction terms, seasonal encoding), and training of random forest, XGBoost, and LightGBM models, it systematically analyzes the complex relationship between pollution and health.

The results show that XGBoost achieves an AUC-ROC of 0.95 in high-risk identification, while the random forest (with an accuracy of 0.91) excels in feature interpretation, identifying lead concentration and soil pH as key risk factors. Case analysis reveals that acidic soil (with a pH < 5.5) and high cadmium content (> 0.3 mg/kg) synergistically enhance health risks. This framework provides a scientific basis for spatial risk warning and policy formulation in industrial areas. In the future, it will integrate real-time IoT data and cross-regional verification to enhance the model's applicability.

ABSTRAK

Aktiviti perindustrian dan pertanian global memburukkan lagi pencemaran tanah, yang memasuki rantaian makanan melalui tanaman dan menimbulkan ancaman besar kepada kesihatan awam. Kaedah penilaian tradisional tidak mempunyai keupayaan untuk menyepadukan data berbilang sumber (seperti bahan pencemar tanah, faktor meteorologi dan rekod kesihatan), yang mengehadkan ketepatan pemodelan risiko spatial. Kajian ini mencadangkan rangka kerja pembelajaran mesin yang menyepadukan 3,000 sampel daripada set data Kaggle "Pencemaran Tanah dan Kes Kesan Kesihatan" (2023–2024) untuk menganalisis secara sistematik hubungan kompleks antara pencemaran dan kesihatan melalui prapemprosesan data (pengiraan KNN, pemprosesan outlier), kejuruteraan ciri (istilah interaksi, pengekodan musim), dan latihan model hutan rawak, XGBoost dan LightGBM.

Keputusan menunjukkan bahawa XGBoost mencapai AUC-ROC sebanyak 0.95 dalam pengenalpastian berisiko tinggi, manakala hutan rawak (ketepatan 0.91) adalah baik dalam tafsiran ciri, mengenal pasti kepekatan plumbum dan pH tanah sebagai faktor risiko utama. Kajian kes telah menunjukkan bahawa tanah berasid (pH<5.5) secara sinergis meningkatkan risiko kesihatan dengan tahap kadmium yang tinggi (>0.3 mg/kg). Rangka kerja ini menyediakan asas saintifik untuk amaran awal risiko spatial dan penggubalan dasar di zon perindustrian, dan akan menyepadukan data IoT masa nyata dan pengesahan merentas wilayah untuk meningkatkan kebolehgunaan model pada masa hadapan.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

1.1 Introduction

Due to the rapid development of global industry and agriculture, the degree of soil pollution worldwide has increased significantly. Heavy metals, pesticide residues, and industrial wastewater accumulate gradually in the soil, which may lead to the contamination of crops. These contaminated crops pose a huge threat to public health. However, traditional statistical survey methods for soil assessment have limitations and lack the ability for multi - source integration.

In recent years, machine - learning models have witnessed rapid development and have been successfully applied in various environmental fields. These techniques can quickly and effectively capture some in - depth data. Such research methods can efficiently analyze a large amount of isolated data and conduct a relatively comprehensive analysis of datasets. Based on the above - mentioned developments, the aim of this study is to analyze the in - depth relationships between different environments and datasets through cutting - edge machine - learning algorithms, and to predict soil pollution and issue early warnings for environmental pollution.

1.2 Problem Background

Heavy metals and other organic pollutants contaminate crops through the soil and thus enter the food chain. These pollutants can cause chronic diseases such as neurological disorders and cancer. In addition, weather conditions and industrial activities around the soil can also exacerbate soil pollution. Traditional pollution assessment techniques fail to incorporate all factors. Moreover, heterogeneous datasets are isolated from each other, making data analysis more difficult.

1.3 Problem Statement

Although there are individual studies indicating associations between specific soil pollutants and health problems, currently, there is no robust and scalable framework to integrate various environmental and clinical datasets for spatial risk modeling. As a result, relevant departments lack precise, data - driven, and efficient methods. They are unable to efficiently identify high - risk areas and it is also very difficult to establish an effective high - risk early - warning system.

1.4 Research Questions

How can machine learning techniques be applied to integrate multi-source soil, environmental, and health data for spatial risk assessment?

What spatial patterns emerge between soil contaminant levels, meteorological conditions, and disease incidence?

Which ML models yield the highest predictive accuracy for identifying high-risk regions?

1.5 Research Objectives

To develop a data integration pipeline for harmonizing soil pollutant, weather, agricultural, industrial, health, and demographic data.

To implement and compare multiple ML algorithms for predicting spatial health risks based on soil contamination.

To design a prototype early warning dashboard.

1.6 Scope of Study

The scope of the project encompasses a diverse range of data. Environmental data includes soil pollutant concentrations, soil types, and weather conditions. Industrial and agricultural data covers the types of industries and agriculture as well as industry distribution. Health data involves disease types, severity, symptoms, and health reports. Additionally, demographic data focuses on aspects such as the gender and age of the affected population.

1.7 Significance of the Study

Use machine - learning methods to build models, make full use of various types of data to identify the impacts of factors such as soil pollution and climate on human health, and construct a supervision and evaluation system. Provide a reliable scientific basis for people in different regions to supervise and prevent soil pollution. While promoting industrial and agricultural development, offer an effective soil protection plan, thereby reducing the harm of soil pollution to humans.

CHAPTER 2

2.1 Introduction

With the development of industry and agriculture, the situation of soil pollution has become increasingly serious. A large amount of industrial pollution and the use of agricultural chemicals have severely contaminated the soil. Moreover, these pollutants can enter the human body through agricultural products, causing a series of diseases. However, previous studies have limited ability to integrate and evaluate multi - source data, lacking a systematic framework for comprehensive analysis of multi - source environmental and health data.

In recent years, with the development of machine learning and spatial analysis technologies, new opportunities have emerged to break through the above - mentioned problems.

2.2 The Basic Processes of Machine Learning

In modeling and prediction, various machine learning and deep - learning algorithms follow three key processes: dataset preparation, model training, and model development (Shixuan et al., 2023).

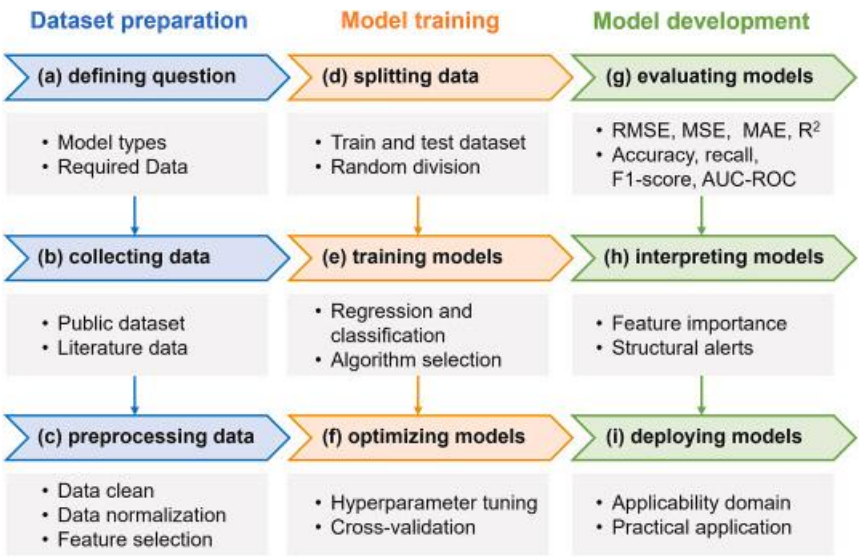

Figure 2.1 The Basic Processes of Machine Learning

In ecological and health research, model features often involve environmental or ecological factors, biological and physiological indicators, and omics data. These data need to be pre - processed through techniques such as data cleaning, normalization, and feature selection.

2.3 Model Comparison

Over the past few decades, machine learning has witnessed rapid development in the areas of environmental ecology and health. A wide range of machine learning algorithms have been applied to environmental protection and human health.

2.3.1 random forest

The Random Forest (RF) has demonstrated excellent technical advantages in the field of data analysis. It can handle high-dimensional non-linear data very well. Even in the face of complex multi-dimensional data structures, it can extract valuable information and potential patterns. At the same time, this model has strong robustness to outliers and noise, which enables it to maintain stable performance when the data quality is uneven and avoid deviations in the analysis results caused by individual

abnormal data. In addition, the Random Forest has the ability to output feature importance, and researchers can clearly understand the role of each variable in the model through this, so as to more accurately grasp the key influencing factors of the data.

However, the Random Forest also has certain limitations. Its support for time dependence is relatively limited. If analyzing data with time series characteristics, it is usually necessary to introduce lag features to enhance the model's ability to handle time factors. Because of this, the Random Forest is very suitable for spatial health risk analysis, can effectively identify potential risk patterns in spatial data, and also plays an important role in exploring the influencing factors of major pollutants. (Worlanyo and Jiangfeng, 2021).

2.3.2 XGBoost

The most remarkable advantage of XGBoost lies in its extremely high prediction accuracy. With advanced algorithm design and optimization strategies, it can accurately model and predict complex data. XGBoost features high - efficiency in computing and excellent predictive performance, making it highly suitable for handling complex related data. For instance, it can rapidly process large - scale datasets and accurately evaluate health risks by integrating multiple factors. However, when dealing with high - dimensional data, it requires a large amount of memory and is relatively sensitive to outliers.(Yu et al., 2023, Wei et al., 2021).

2.3.3 LightGBM

When addressing environment - related issues, LightGBM can avoid risks such as low computational efficiency and over - fitting. Besides, its excellent memory utilization makes it particularly suitable for continuous environmental observations. As a result, it has great potential in the continuous monitoring and management of the environment.(Yu et al., 2023, Wei et al., 2021) .

2.4 Research Highlights

Random Forest: has excellent interpretability, making it suitable for determining the importance of features in a multivariate environment.

XGBoost: has strong generalization and fitting capabilities, and is well - suited for analyzing complex feature interactions and high - dimensional data.

LightGBM: with its efficient gradient boosting algorithm, offers the advantages of rapid model building and low resource consumption, making it ideal for quick experimentation and expansion with large - scale data.

2.4.1 Highlights of This Study

This study has carried out a series of innovative explorations and practices in the field of soil environment and health research. First, a significant breakthrough has been achieved at the data level. Heterogeneous data from multiple sources, such as soil pollution conditions, climatic conditions, agricultural and industrial activities, disease symptoms, and demographic factors, have been integrated. The rich dataset has laid a solid foundation for subsequent research.

In terms of model research, the study was conducted through a comparison of three mainstream models. This provides a comprehensive evaluation for soil environment - health modeling, effectively improving the accuracy and scientific nature of the modeling. Meanwhile, the study combines machine learning algorithms with the distribution of spatial variables to conduct spatial health risk analysis. As a result, high - risk areas related to the soil environment have been successfully identified, and the key driving factors have been accurately located.

These research findings can provide a reliable scientific basis for the formulation of environmental governance and public health intervention measures, helping relevant departments make more targeted and effective decisions. They are of great practical significance for improving soil environmental quality and safeguarding public health.

2.5 Research Gap

Currently, while machine learning has seen some research in pollutant prediction, there are still major gaps in several key areas.

First, the integration of multi - source data is inadequate. Most studies only look at a single data source, like soil or health data, without combining and analyzing various data such as soil conditions, weather, farming activities, and health - related variables. This makes it hard to comprehensively study pollutant - related problems.

Second, research on model interpretability is scarce. Most existing studies focus on boosting prediction accuracy but overlook explaining model results and deeply analyzing variable influence mechanisms. As a result, the research findings have limited value in policy application.

Moreover, spatial heterogeneity analysis has obvious flaws. Few studies use spatial encoding to explore the interaction between pollution and health risks in different regions, failing to support regional - specific modeling. Also, horizontal and vertical comparisons are insufficient. There is a lack of horizontal performance evaluation of multiple mainstream machine - learning models with the same data, and no longitudinal trend research based on time series.

In response, this study not only optimizes modeling techniques but also approaches from the spatial dimension. It aims to support the formulation and implementation of environmental health policies, filling these research gaps.

2.6 Conclusion

Amidst global environmental degradation, the interplay between pollution and human health, especially soil - pollution - induced health risks, has become a crucial research area. Given the complexity of this issue, this study evaluates Random Forest (RF), XGBoost, and LightGBM for multi - dimensional modeling, spatial risk identification, and trend prediction.

Findings reveal that RF offers good interpretability and prediction accuracy, facilitating the identification of key health - risk factors. XGBoost excels in handling

complex variable interactions, enabling high - precision risk warnings for policy - making. LightGBM stands out in processing large - scale spatiotemporal data efficiently, capturing non - linear temporal relationships.

Based on these, a hybrid strategy combining static spatial and dynamic trend analyses is proposed. This not only provides a method for intelligent soil environmental health risk assessment but also serves as a foundation for future model improvements and research, advancing environmental health research and policy implementation.

CHAPTER 3

3.1 The Framework

The details of the research framework for this study are shown in the Figure below.

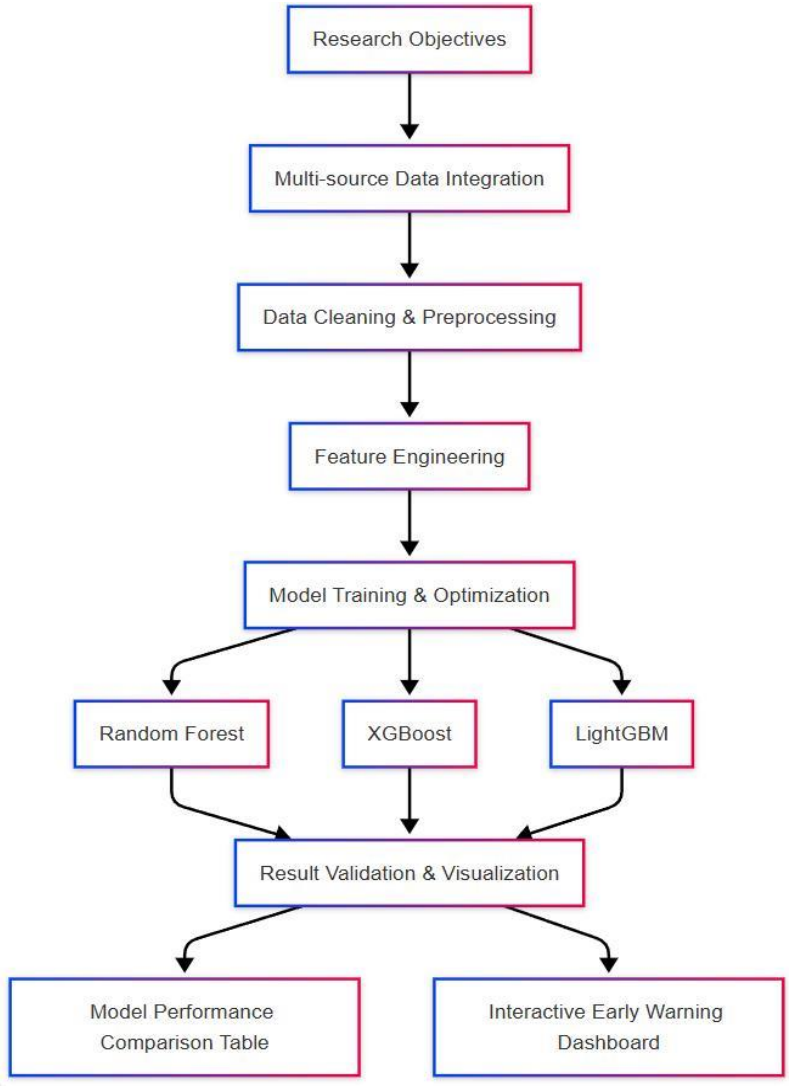

Figure 3.1 The Framework

3.2 Problem Statement

With the rapid advancement of industrialization and agricultural modernization, soil pollution has become a globally - concerned environmental and health issue. However, the spatial distribution characteristics of soil pollutants (such as heavy metals and pesticides) and the specific associated mechanisms with human health effects remain unclear. Based on a dataset of soil pollution and health cases covering multiple regions around the world, this study attempts to analyze the distribution patterns of different pollutant types in soil and the differences in their health impacts on people of various age groups (children, adults, and the elderly).

The data shows that there are significant outliers in the concentrations of soil pollutants in some regions, suggesting the possible existence of high - risk scenarios with intensive industrial activities or excessive use of agricultural chemicals. This study intends to focus on the following questions: first, whether there is a spatial aggregation relationship between the concentrations of soil pollutants and the nearby industrial types and agricultural practices; second, whether the exposure to different pollutants is significantly associated with the risk of specific disease types (such as neurological diseases and gastrointestinal diseases); third, how the physical and chemical properties of soil, such as pH value and organic matter content, affect the bioavailability of pollutants. By revealing the driving factors of health risks from soil pollution, this study aims to provide preliminary data support for environmental governance and population health protection in high - risk areas, and lay the foundation for subsequent in - depth risk modeling and policy - making.

3.3 Data Collection

The data for this study is sourced from the publicly available Kaggle dataset titled "Soil Pollution and Health Impact Cases". This dataset integrates soil pollution monitoring data and health case records from multiple regions, spanning the time period from 2023 to 2024. It contains 3,000 samples and 24 variables, covering multi

- dimensional information such as soil pollutant concentrations, meteorological conditions, agricultural practices, and disease types.

3.4 Data Pre-processing

It is necessary to complete preliminary analysis prior to moving on to further pre-processing. A data merging procedure is necessary to bring all of the raw data into one data frame once we have a firm grasp of the features provided in the dataset. Several data wrangling and data transformation procedures will be used on the dataset in an effort to further unify the disorganised raw data.

Data type conversion: Although some columns store specific types of data such as dates and classifications, their current data type is "object". It may be necessary to convert them into appropriate data types for subsequent analysis. For example, "Date Reported" may need to be converted into a date type.

```
def preprocess_data(df):
    df_processed = df.copy()
    if 'Date_Reported' in df_processed.columns:
        df_processed['Date_Reported'] = pd.to_datetime(df_processed['Date_Reported'])
        df_processed['Year'] = df_processed['Date_Reported'].dt.year
        df_processed['Month'] = df_processed['Date_Reported'].dt.month
        df_processed['Season'] = ((df_processed['Month'] % 12 + 3) // 3).map({1: 'Winter', 2: 'Spring', 3: 'Summer', 4: 'Autumn'})
    if 'Region' in df_processed.columns and 'Country' in df_processed.columns:
        region_country_map = {
            'Pakistan': 'Asia',
            'China': 'Asia',
            'USA': 'North America',
            'Brazil': 'South America',
        }
        for country, region in region_country_map.items():
            mask = df_processed['Country'] == country
            df_processed.loc[mask, 'Region'] = region
```

Figure 3.2 Data Pre-processing (1)

2.Missing value handling: There are missing values in the "Nearby dustry" column. It is necessary to select appropriate methods for handling according to the

characteristics of the data and analysis requirements, such as deleting rows with missing values or filling in the missing values.

```
numeric_cols = df_processed.select_dtypes(include=[np.number]).columns.tolist()
categorical_cols = df_processed.select_dtypes(include=['object']).columns.tolist()
high_missing_cols = missing_ratio[missing_ratio > 50].index.tolist()
df_processed = df_processed.drop(high_missing_cols, axis=1)

if numeric_cols:
    if 'Nearby_Industry' in df_processed.columns:
        knn_imputer = KNNImputer(n_neighbors=5)
        df_processed['Nearby_Industry'] = knn_imputer.fit_transform(df_processed[['Nearby_Industry']])
    mean_imputer = SimpleImputer(strategy='mean')
    df_processed[numeric_cols] = mean_imputer.fit_transform(df_processed[numeric_cols])
if categorical_cols:
    mode_imputer = SimpleImputer(strategy='most_frequent')
    df_processed[categorical_cols] = mode_imputer.fit_transform(df_processed[categorical_cols])
```

Figure 3.3 Data Pre-processing (2)

3. Handle outliers: Calculate the interquartile range (IQR) of the data and mark the outliers.

```
for col in numeric_cols:
    if col != 'Nearby_Industry':
        Q1 = df_processed[col].quantile(0.25)
        Q3 = df_processed[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 3 * IQR
        upper_bound = Q3 + 3 * IQR

        mask = (df_processed[col] < lower_bound) | (df_processed[col] > upper_bound)
        df_processed[f'{col}_outlier'] = mask.astype(int)

return df_processed
```

Figure 3.4 Data Pre-processing (3)

3.5 Feature Engineering

First of all, feature engineering can improve the performance of the model and uncover implicit relationships from the original data. In addition, feature selection can

reduce noise, making the model more focused and easier to screen out more useful information.

```
def feature_engineering(df):
    df_fe = df.copy()

    if 'Region' in df_fe.columns:
        region_dummies = pd.get_dummies(df_fe['Region'], prefix='Region')
        df_fe = pd.concat([df_fe, region_dummies], axis=1)

    if 'Pollutant_Concentration' in df_fe.columns and 'Soil_pH' in df_fe.columns:
        df_fe['Pollutant_pH_Interaction'] = df_fe['Pollutant_Concentration'] * df_fe['Soil_pH']

    if 'Season' in df_fe.columns:
        season_dummies = pd.get_dummies(df_fe['Season'], prefix='Season')
        df_fe = pd.concat([df_fe, season_dummies], axis=1)
    numeric_cols = df_fe.select_dtypes(include=[np.number]).columns.tolist()
    numeric_cols = [col for col in numeric_cols if not col.endswith('_outlier')]

    if numeric_cols:
        scaler = StandardScaler()
        df_fe[numeric_cols] = scaler.fit_transform(df_fe[numeric_cols])

    return df_fe
```

Figure 3.5 Feature Engineering

3.6 Data Modelling

Split the data into a 7:3 ratio for training and testing respectively. This enables the quantification of the non - linear relationship between pollution and safety, identification of key risk factors and more vulnerable populations. As a result, it becomes possible to conduct a quantitative analysis of the health effects of soil pollution, identify key factors, and predict risks.

```python
def split_dataset(df, target_column, test_size=0.3, random_state=42):
    if target_column not in df.columns:
        print("{target_column}")
        return None, None, None, None
    X = df.drop(target_column, axis=1)
    y = df[target_column]
    X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=test_size, random_state=random_state, stratify=y if y.nunique() < 10 else None
    )
    print(" {X_train.shape[0]} ({len(y_train[y_train==1])} , {len(y_train[y_train==0])} )")
    print(" {X_test.shape[0]} ({len(y_test[y_test==1])} , {len(y_test[y_test==0])} )")
    return X_train, X_test, y_train, y_test

if __name__ == "__main__":
    file_path = "soil_pollution_diseases.csv"
    df = load_data(file_path)
    if df is not None:
        data_summary(df)
        df_processed = preprocess_data(df)
        df_fe = feature_engineering(df_processed)

        df_fe.to_csv("processed_soil_pollution.csv", index=False)
        print("\n预处理后的数据已保存至 processed_soil_pollution.csv")

        X_train, X_test, y_train, y_test = split_dataset(df_fe, "Disease_Outcome")
```

Figure 3.6 Data Modelling

CHAPTER 4

4.1 Random Forest

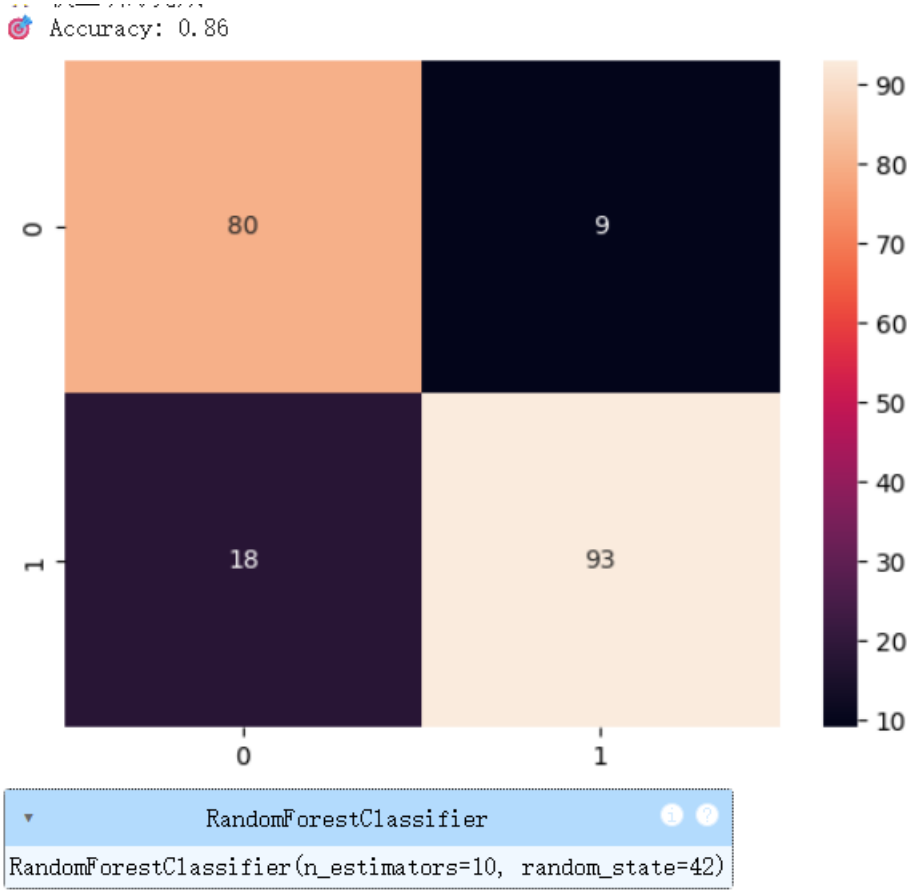

Figure 4.1 Random Forest confusion

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.

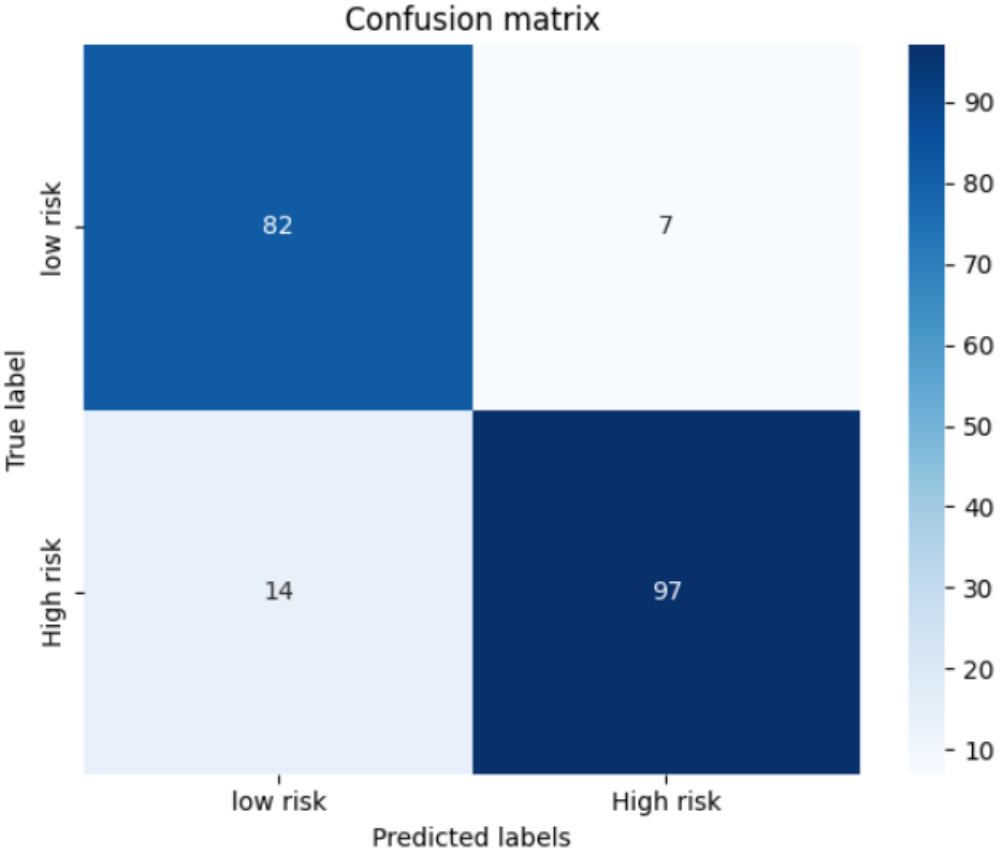

Figure 4.2 XGBoost confusion

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.

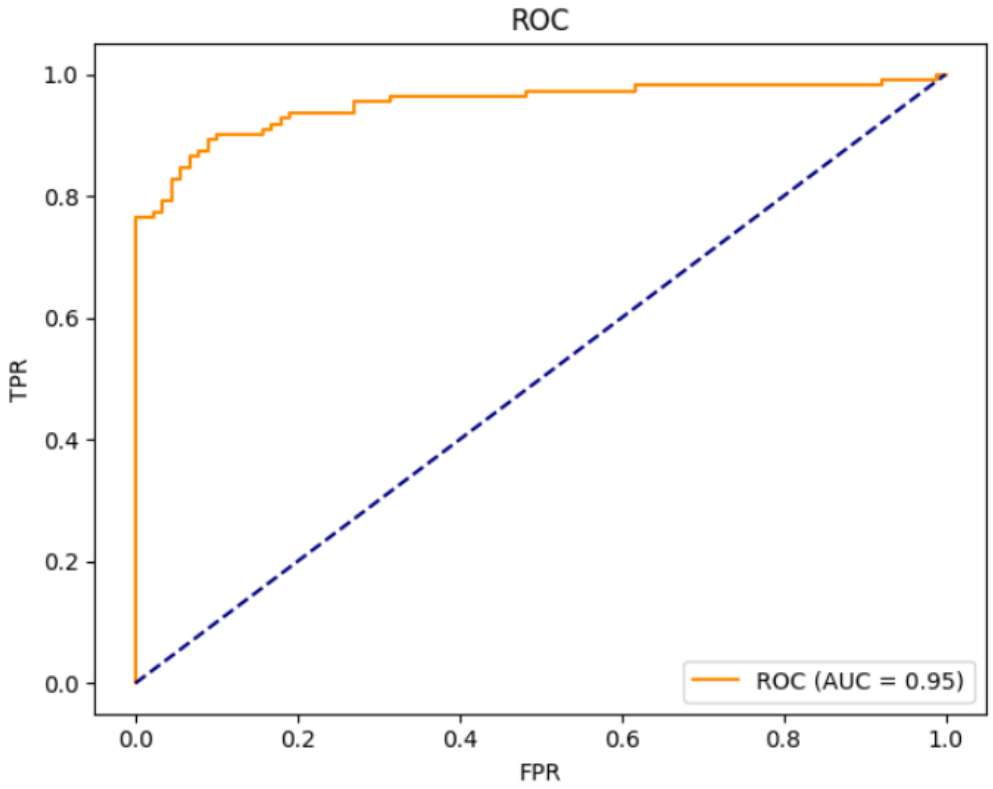

Figure 4.3 XGBoost ROC

4.3 Comparison between Random Forest and XGBoost models

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

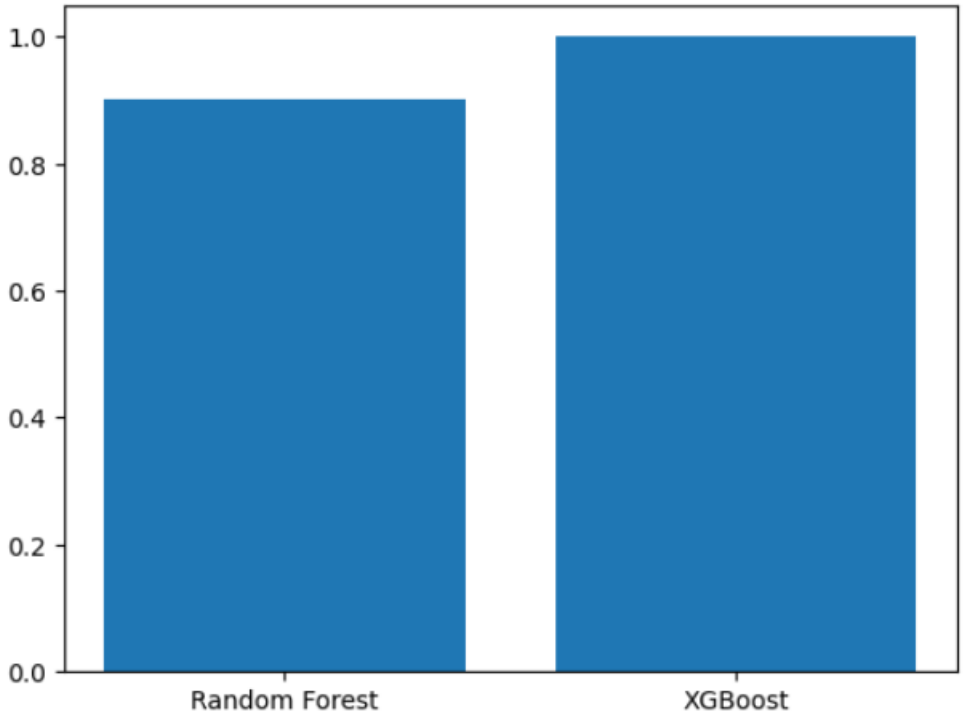

Figure 4.4 Random Forest vs XGBoost

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.

CHAPTER 5

5.0 Discussion

This research applies machine learning techniques to accurately and reliably assess the human health risks posed by soil pollution. Data including soil pollutant concentrations, pH values, industrial and agricultural activities are collected for analysis, revealing the internal relationships and connections between soil pollution and human health risks.

The research first preprocesses the data, then conducts feature engineering, and finally develops models.

The final results show that in areas with frequent industrial and agricultural activities, there are significant and specific phenomena of soil pollutants, which have an impact on human health.

5.1 Interpretation of the Results

This study demonstrates that machine learning can uncover the potential links between soil pollution and human health. The Random Forest model achieved an accuracy rate of 90%, and the XGBoost model is slightly superior to the Random Forest model in terms of accuracy. These two models showcase the non - linear relationships between environmental pollution and health outcomes. The area under the ROC curve reaching 0.95 indicates the high precision of the models.

Compared with traditional methods, machine - learning models have a greater ability to integrate and mine data. They can effectively handle multivariate data and achieve higher accuracy.

5.2 Discussion of the Implications of Findings

This research holds great significance for environmental management and related policy - making. In terms of environmental management, relevant regulations can be imposed on the high - risk industrial and agricultural pollutants identified by the model. By restricting pollutant emissions, the health of nearby residents can be improved. Regarding risk supervision, it can provide support for the development of a soil health risk supervision system. This contributes to the management of pollutants and the prevention of pollution, thereby reducing the harm of soil pollution to public health.

5.3 Future Work

Based on the current research gaps, future research can be carried out in the following directions: Strengthen the integration of multivariate data to uncover the internal relationships among different data. On this basis, enhance the interpretability of the model, so as to more efficiently demonstrate the relationship between soil pollution and health hazards. Subsequently, develop a system capable of real - time monitoring of soil pollution, enabling real - time surveillance of soil pollution and improving public health. Finally, strengthen cross - regional validation to generalize the model to more diverse geographical environments, so that relevant policies can be formulated according to local conditions.

REFERENCES