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

ZHAO ZHIHAN

UNIVERSITI TEKNOLOGI MALAYSIA

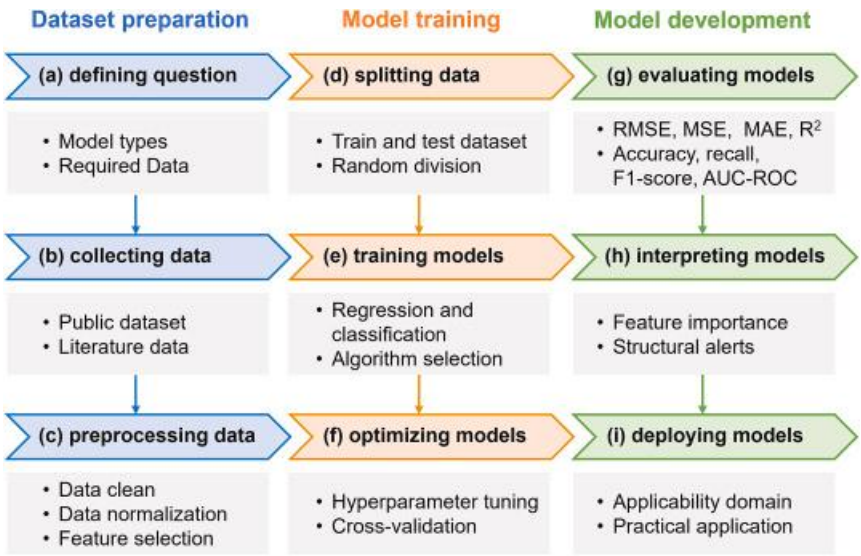
2.1 Introduction

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).



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.