QUANTITATIVE STUDY OF FACTORS AFFECTING THE INCIDENCE OF TUBERCULOSIS

Zhuoyang Wang[1] *Zhiling Li*[2] *Jinda Dong*[2]

[1] Department of Electronic and Electrical Engineering, SUSTech, China
[2] School of System Design and Intelligent Manufacturing, SUSTech, China

ABSTRACT

This study investigates the impact of various factors, such as smoking prevalence and air pollution, on the incidence of tuberculosis using statistical learning methods to improve the prevention and treatment of tuberculosis. The research utilizes data from the World Bank, selecting multiple potential influencing factors and conducting correlation analysis. Various models, including linear regression, decision trees, and random forests, were trained and tested, with random forest and decision tree models showing the best performance. The results indicate that population density and safe drinking water coverage are the primary factors affecting tuberculosis incidence. The study also found that, while smoking and gender had minimal impact on tuberculosis incidence, other factors such as health expenditure, urban population rate, alcohol consumption, and PM2.5 pollution had significant contributions. Future work could involve incorporating time factors to predict tuberculosis incidence at given points in time.

***Index Terms*—** Tuberculosis, quantitative study, statistical learning, regression

1. INTRODUCTION

Tuberculosis is a chronic infectious disease caused by a complex of Mycobacterium tuberculosis. The disease can cause tuberculosis in the lungs as well as invade other organs such as the liver, kidneys, brain, lymph nodes, etc. Tuberculosis continues to pose a threat to human health and is considered a major public health issue [1].

The aim of this study was to analyze the effect of various factors such as smoking prevalence and air pollution on the incidence of tuberculosis by using statistical learning methods for the prevention and treatment of tuberculosis.

2. RELATED WORK

Several studies have explored the relationship between various risk factors and the incidence of tuberculosis. For instance, Bates et al. (2007) conducted a comprehensive review of the literature and identified smoking as a significant risk factor for tuberculosis infection and disease progression [2]. Additionally, Lin et al. (2007) demonstrated that air pollution, particularly fine particulate matter (PM2.5), is associated with an increased risk of tuberculosis [3]. More recent research by Yates et al. (2016) employed machine learning techniques to analyze large-scale epidemiological data, revealing complex interactions between multiple environmental and sociodemographic factors that contribute to the incidence of tuberculosis [4].

These studies highlight the importance of integrating multiple risk factors and advanced statistical methods to improve our understanding and control of tuberculosis. The current study builds on this body of work by focusing specifically on the combined effects of smoking prevalence and air pollution on tuberculosis incidence.

3. DATA AND ASSUMPTIONS

3.1. Data Acquisition

This study employs various statistical learning techniques to analyze data obtained from the World Bank [5].

The following indicators were selected for study from the data available, including the dependent variable, TB incidence, and a range of possible influencing factors (Table 1). The intuitive ideas for the selection of the indicators is shown in Table 2.

3.2. Assumptions

Our data come from WorldBank's year-by-year statistics for the world's countries, which makes some indicators for the same country highly covariant with each other due to covariation with national development, and the amount of data is small. Using data from more than one country would provide a different perspective, but at the same time, data from countries with large gaps may contain more contextual information about the country.

So we specifically screened a group of developed European countries (GBR, FRA, DEU, ITA, NLD, NOR, SWE, FIN, DNK, ISL, CHE, BEL, LUX, IRL, ESP, PRT, AUT, CZE, SVK, HUN, GRC, SVN, POL, EST, LVA, LTU, MLT, CYP) , assuming that background indicators such as the general health situation in these countries are relatively similar and do not cause the aforementioned problems. In addition,

Name	Abbreviation	Unit
Total alcohol consumption pc*, male	alcohol_m	Liters of pure alcohol pc
Total alcohol consumption pc, female	alcohol_f	Liters of pure alcohol pc
Prevalence of current tobacco use, males	tobacco_m	% of male adults
Prevalence of current tobacco use, females	tobacco_f	% of female adults
PM2.5 air pollution, mean annual exposure	pm25	mg/m^2
People using safely managed drinking water services	water	% of population
People using safely managed sanitation services	sanitation	% of population
Domestic general government health expenditure pc	health_expenditure	Current international $ (PPP*)
Population density	population_density	People per km^2 of land area
Urban population	urban_population	% of total population
Incidence of tuberculosis	tuberculosis	Per 100,000 people

Table 1. Selected indicator names, units, and their subsequent abbreviations. pc stands for per capita and PPP stands for purchasing power parity.

Abbreviation	Intuitive Ideas
alcohol_m, alcohol_f	Alcohol consumption impairs liver function and may lower immunity
tobacco_m, tobacco_f	Tobacco damages the respiratory tract and may make you more susceptible to respiratory diseases
pm25	PM2.5 impairs respiratory function, may lower resistance
water	Unsafe drinking water sources may contain pathogenic bacteria that induce tuberculosis
sanitation	Well-established, independent sanitation infrastructure reduces the likelihood of disease transmission
health_expenditure	Health expenditures have a multifaceted impact and may be representative of the country's overall health situation
population_density	High population density increases the probability of disease transmission
urban_population	Concentration of water and sewerage in urban areas may make disease transmission more likely
tuberculosis	Tuberculosis develops from Mycobacterium tuberculosis (bacterial infection)

Table 2. The intuitive ideas for the selection of the indicators.

we selected data for the period 2010 to 2023, which ensures that all selected indicators have data for this period, with no missing values and no need for interpolation.

4. METHODOLOGY

4.1. Dataset Splitting

First, all the acquired data are fixedly partitioned into a training set and a test set, with the latter accounting for about 20%, after which all the fitting operations are performed on the training set and tested with the test set.

4.2. Data Understanding

Based on the normalized data for each indicator, correlation heat maps are drawn to characterize the data at a macro level (Figure 1).

Observation of the figure yields some findings as follows:

1. In the red box of the figure we can find that sex-disaggregated raw data (alcohol) has high inter-sex correlation, which means alcohol_m and alcohol_f are highly covariant, perhaps one can be rounded up or down, and the difference is only reflected in the scale. This feature is weaker on tobacco.
2. At the green box in the figure we can find that correlations between response and gender-specific tobacco use data vary widely. We can merge the original _m and _f data and separate sex as a new feature.

Accordingly, the sex-segregated data were merged and dummy variable sex was added to indicate sex, with 0 representing MALE and 1 representing FEMALE. A new correlation heat map was plotted after obtaining the new dataset (Figure 2).

Observation of the new heat map yields some findings as follows:

1. The part in red box shows that there is higher correlation between sex, alcohol and tobacco, i.e., Males generally drink more alcohol than females. The relationship is relatively less strong for tobacco use. Subsequent consideration may need to be given to downscaling or adding

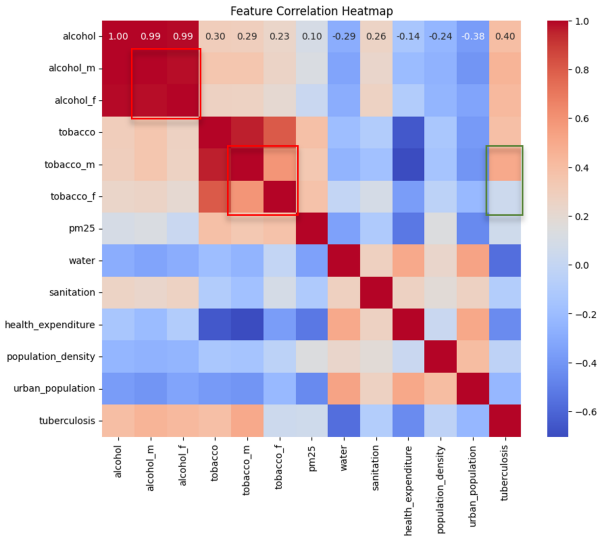

Fig. 1. The correlation heat map for the raw data.

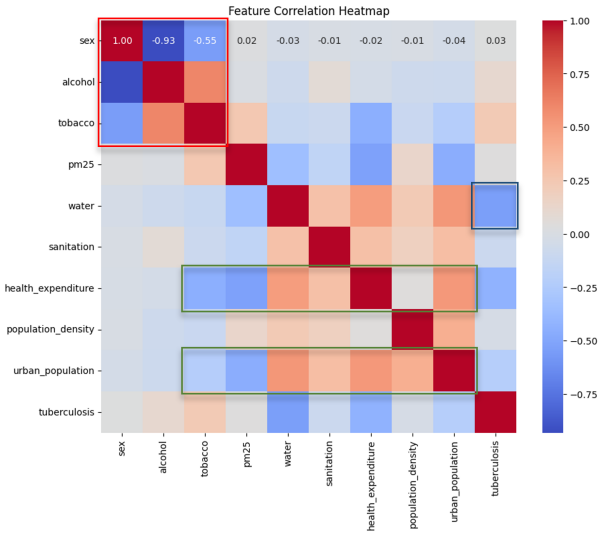

Fig. 2. The correlation heat map for the processed data.

non-linear terms.

2. The part in dark box shows that the predictor water may contribute more to the response.
3. The part in green box shows that both indicators affect more aspects and are likely to have indirect links between them and other characteristics. Subsequent downscaling may be considered.

Plot pair plots to facilitate the observation of the relationship between the quantities (Figure 3).

4.3. Modeling

4.3.1. Cross Validation

Next some improvement attempts are made for some of the models, which need to be verified whether the performance of the improved models is really improved.

Cross-validation is needed here, where different training and test sets are separated from the dataset each time, and the obtained test results are averaged as a basis for judgment to reduce chance.

4.3.2. Basic Methods

Use a variety of common models to train on the training set (after cross-validation) and give their performance on the test set (Table 3). All predictors are unprocessed, each as a variable.

Model	**Test MSE**	**Test R^2**
Multiple Linear Regression	40.223	0.6273
Decision Tree	2.3059	0.9674
Random Forest	1.014	0.9901
Support Vector Regression	104.6991	0.0298
Multi-Layer Perceptron	40.4337	0.6253

Table 3. Results on test dataset for several basic methods. The models are all trained under cross validation. Sort by test results from best to worst.

4.4. Model Improvement

4.4.1. Non-linear Terms

Observing the pairplots, we have reason to believe that there is a more significant nonlinearity between the reponse and some of the predictors, so we consider adding nonlinear terms as new predictors.

Experiments are conducted on the basis of the multiple linear regression, random forest and decision tree. The nonlinear terms mainly include two kinds, one is the interaction term and the other is the high order term. The training data is relatively small, adding the high order term is very easy to lead to overfitting, and the pre-experiment also shows that adding the high order term will lead to worse results in most cases, so the main test is the performance after adding the interaction term.

The main consideration is, for example, the effect of smoking on the incidence of tuberculosis may be moderated by the gender factor. Add interaction terms between variables with high correlation: $sex \times alcohol$, $sex \times tobacco$, $health\_expenditure \times tobacco$, $health\_expenditure \times pm25$, $health\_expenditure \times water$, $health\_expenditure \times sanitation$.

The results of the experiment are summarized later.

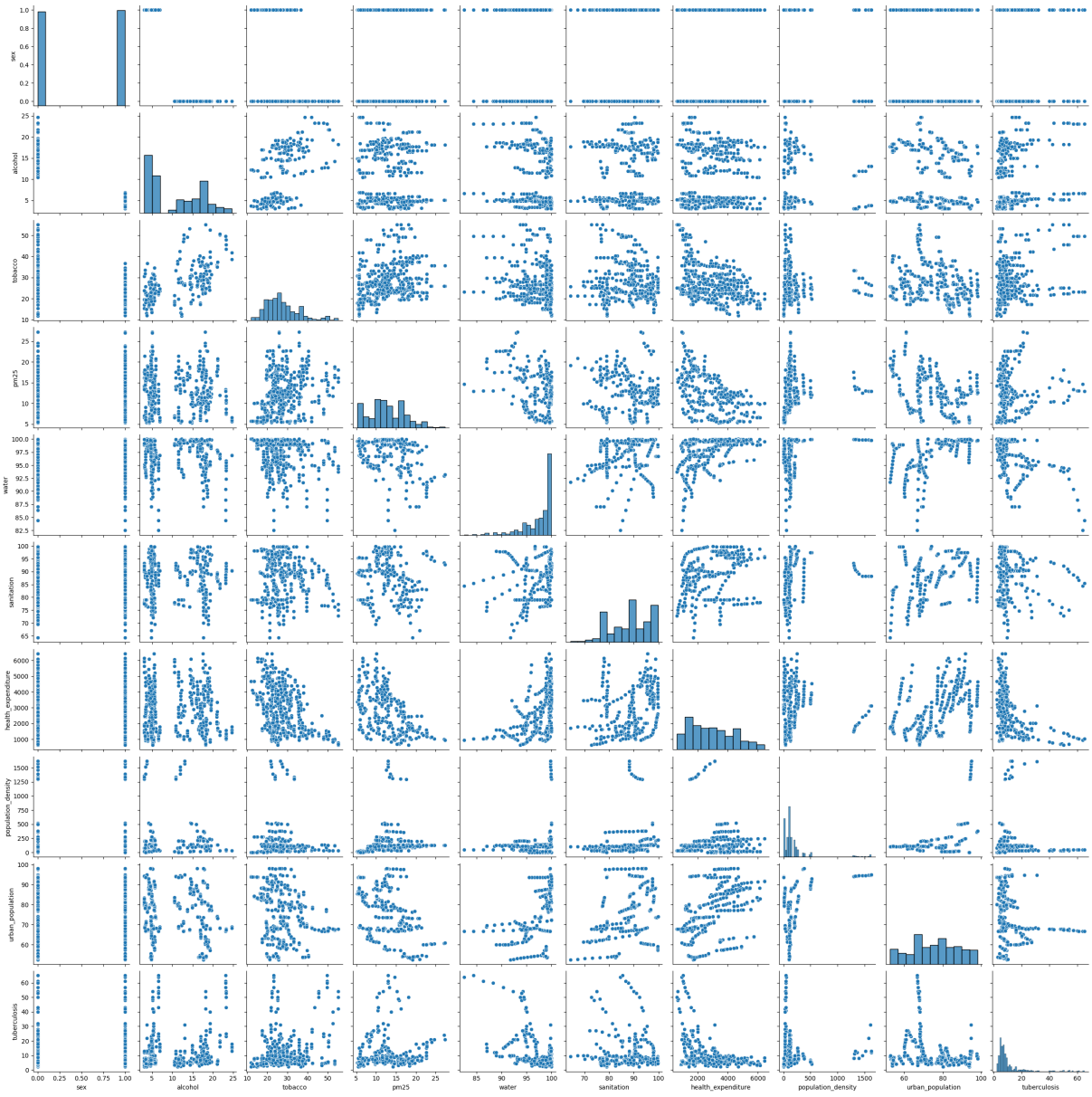

Fig. 3. The pair plots for the processed data.

4.4.2. Stepwise Selection

If the interpretability of the model is disregarded for the time being, experiments can reveal that adding any interaction term can cause the Test MSE to decrease. Too many interaction terms cause the model to be too complex, and stepwise selection method is used to select the main parameters.

After adding all possible cross terms to the dataset, feature selection is performed using Forward Selection and Mixed Selection, respectively, to retain the features that are more significant and make the model better.

The results of the experiment are summarized later.

4.4.3. Merging Terms (PCA)

Variables can be combined using principal component analysis (PCA) to eliminate covariances between variables and streamline model parameters.

Referring to the correlation heat map, combine sex, $alcohol$, and $tobacco$ as A_0 and A_1, and $pm25$, $water$, $sanitation$, $health\_expenditure$, $urban\_population$, and $population\_density$ as B_0 through B_3, and plot a new correlation map for the resulting new data (Figure 4).

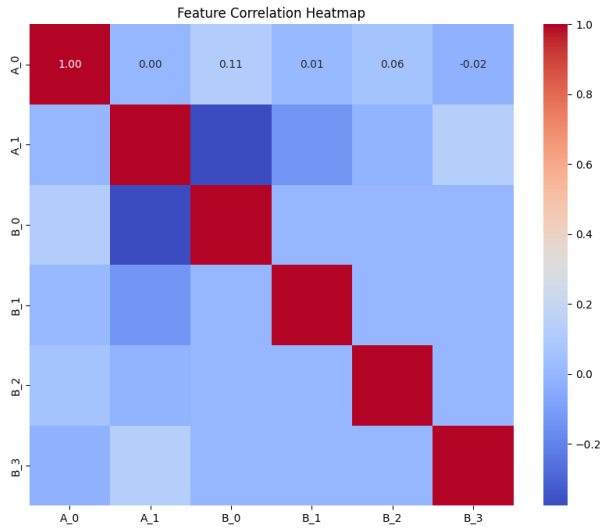

Fig. 4. The correlation heat map for the data after PCA.

5. EXPERIMENTS AND DISCUSSION

5.0.1. Test Results

The experimental results of the basic method as well as the improved modeling experiments mentioned earlier are listed in sorted order as in Table 4.

Random forests and decision trees fitted best, which may mean that there are more nonlinear relationships in the problem; In the average case, the further artificial inclusion of the cross terms resulted in worse results.

Multiple linear regression is significantly more effective with the addition of the cross terms; What is not shown here is that the average effect is still boosted after all possible cross terms are added.

Support vector regression has the worst results with different kernel functions.

5.0.2. Residuals Analysis

The distribution of residuals from the above modal test results was plotted as a box-and-line plot (Figure 5) and sorted according to Table 4.

According to the results of the comparison between the models, the fitting effect of decision tree and random forest models is much better than the other models, which are suitable for dealing with nonlinear relationships. Therefore, there may be more nonlinear relationships between the features and the response of this problem, and the R2 value of MLR cannot be greatly improved, and similarly the SVR effect is also poor.

5.0.3. Features Importance

Further analyze the importance of internal features after training of decision tree and random forest models (Figure 6 and Figure 7).

It was observed that the most significant factors affecting response within both models were $population\_density$, $water$, and $health\_expenditure$. this suggests that the incidence of tuberculosis stems mainly from inter-individual infectious factors.

Similarly, chart the importance of features after adding cross terms for both methods (Figure 8 and Figure 9).

It can be seen that the main influencing factors are $health\_expenditure \times pm25$, $water$, $population\_density$ and $health\_expenditure \times water$. The first factor is somewhat difficult to interpret and may be related to segmentation within health expenditures, with little explanatory significance; the remaining main factors are in line with the results of the two original models, mainly with regard to population and safe drinking water coverage.

6. CONCLUSION AND FUTURE WORK

Based on the results of the original Random Forest and Decision Tree, the characteristics considered that have the highest impact on TB incidence are population density and safe drinking water coverage.

Among other characteristics, health expenditure, urban population rate, alcohol consumption, PM2.5, and sanitation infrastructure coverage had more significant contributions, while gender and smoking prevalence had little effect on TB incidence.

In contrast to initial speculation, it is possible that the damage to the liver and immune system caused by alcohol abuse is more severe in targeting tuberculosis than the damage to the respiratory tract caused by smoking.

Based on the results of the optimized MLR, the highest impact of the considered characteristics on TB incidence was also safe drinking water coverage, gender, sanitation infrastructure had insignificant impact on incidence, and all other individual characteristics had high significant impact on the results. The cross-tabs related to urban population rate and health expenditure were marginally insignificant, while the cross-tab between health expenditure and drinking water coverage had a significant effect on the results.

In contrast to inference and cause analysis, in the follow-up work, we can extract the time factor and use a similar model to predict the incidence of TB at a given point in time.

7. ACKNOWLEDGMENT

This research was conducted as part of the course project for ”Statistical Learning in Data Science” under the guidance of

Model	Test MSE	Test R2
Random Forest	1.014	0.9901
Random Forest (Interaction Terms)	1.8464	0.9816
Decision Tree	2.3059	0.9674
Decision Tree (Interaction Terms)	3.2844	0.9631
Multiple Linear Regression (Interaction Terms)	25.676	0.7621
Multiple Linear Regression (Mixed Selection)	33.3072	0.6914
Multiple Linear Regression (Forward Selection)	35.446	0.6715
Multiple Linear Regression	40.223	0.6273
Multi-Layer Perceptron	40.4337	0.6253
Multiple Linear Regression (PCA)	42.3802	0.6073
Support Vector Regression	104.6991	0.0298

Table 4. Results on test dataset for all mentioned methods. The models are all trained under cross validation.

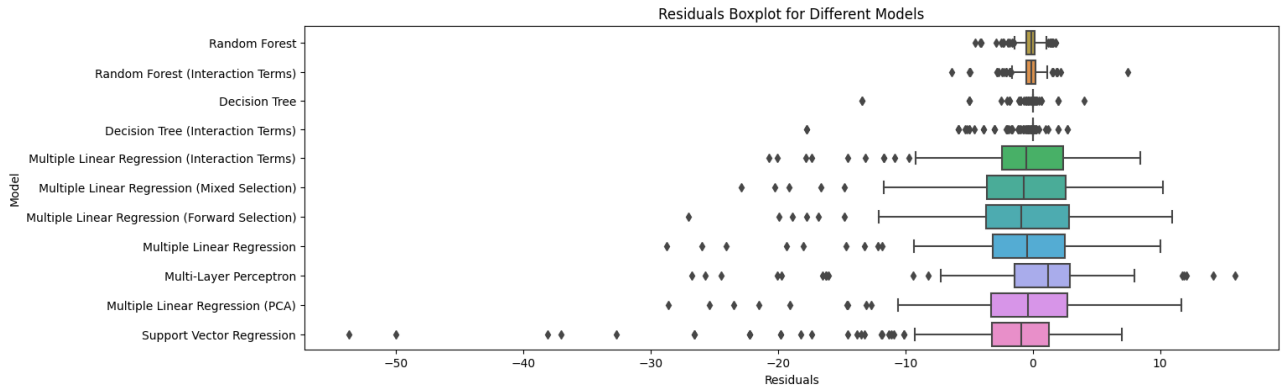

Fig. 5. The residual distributions for models.

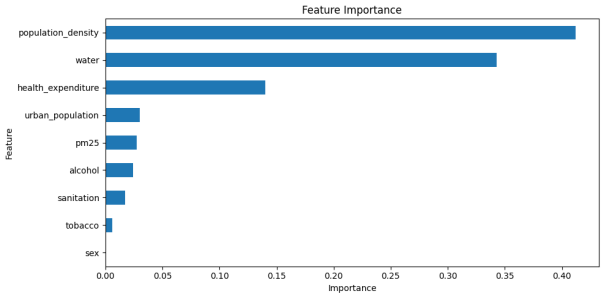

Fig. 6. The feature importance ranking for random forest method.

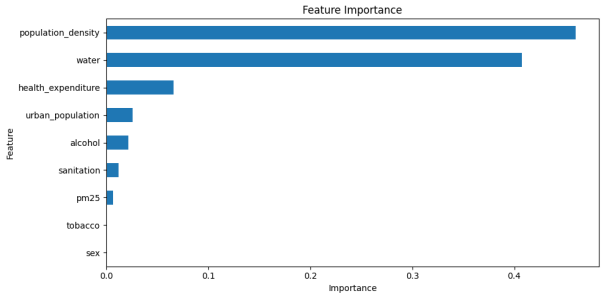

Fig. 7. The feature importance ranking for decision tree method.

Professor Tang. The authors would like to express their gratitude to Professor Tang for his insightful lectures and invaluable guidance throughout the project. We also extend our appreciation to the teaching assistants for their continuous support and constructive feedback, which significantly contributed to the success of this study.

8. REFERENCES

[1] World Health Organization, "Tuberculosis," 2024, Accessed: 2024-06-11.

[2] Matthew N Bates, Abhay Khalakdina, Madhukar Pai, Lu-Yu Chang, Fernanda Lessa, and Kenneth R Smith, "Risk of tuberculosis from exposure to tobacco smoke: a systematic review and meta-analysis," *Archives of Internal Medicine*, vol. 167, no. 4, pp. 335–342, 2007.

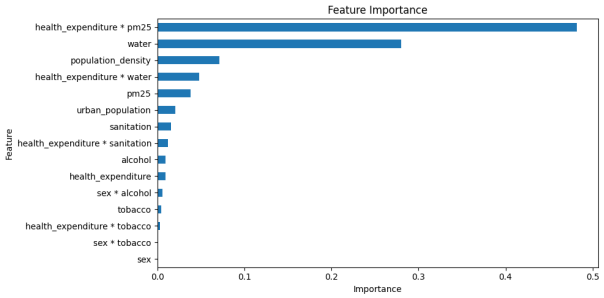

Fig. 8. The feature importance ranking for random forest method with interaction terms.

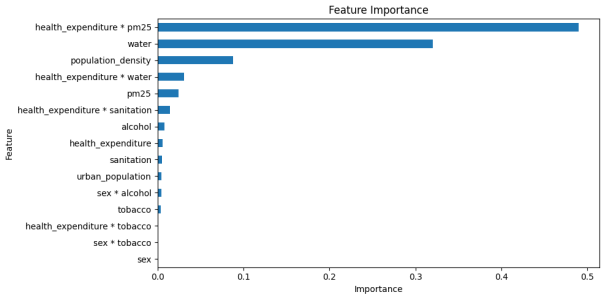

Fig. 9. The feature importance ranking for decision tree method with interaction terms.

[3] Hsien-Ho Lin, Majid Ezzati, and Megan Murray, "Tobacco smoke, indoor air pollution and tuberculosis: A systematic review and meta-analysis: e20," *PLoS Medicine*, vol. 4, no. 1, pp. e20, 2007.

[4] Tom A. Yates, Palwasha Y. Khan, Gwenan M. Knight, Jonathon G. Taylor, Timothy D. McHugh, Marc Lipman, Richard G. White, Ted Cohen, Frank G. Cobelens, Robin Wood, David A. J. Moore, and Ibrahim Abubakar, "The transmission of mycobacterium tuberculosis in high burden settings," *The Lancet Infectious Diseases*, vol. 16, no. 2, pp. 227–238, 2016.

[5] World Bank, "World bank open data," 2024, Accessed: 2024-06-11.