

Machine Learning-Based Prediction of Two-Week Disease Risk in Children Aged 0–6 Years in Xizang

Ling Yuyun¹, Qi Jieyang¹, Zhaxi Dawa^{1, 2, *}

¹College of Medicine, Xizang University, Lhasa, China

²Lhasa Key Laboratory of Public Health Security and Health Policy Research, Xizang University, Lhasa, China

Email address:

1942289112@qq.com (Ling Yuyun), 963449251@qq.com (Qi Jieyang), zhaxi0891@qq.com (Zhaxi Dawa)

*Corresponding author

Abstract

Background: Tibet is located at high altitude and sparsely populated, and the accessibility of children's health services is poor. Harsh natural conditions and limited medical resources further exacerbate the health challenges faced by children in this region. **Objectives:** This study aims to develop and compare logistic regression, random forest, and XGBoost models for predicting two-week illness risk among children aged 0–6 years in Tibet, and to identify key influencing factors using SHAP analysis, thereby providing AI tools for precision public health interventions. **Methods:** Data were collected from the fifth (2013), sixth (2018) and seventh (2023) National Health Service surveys of Tibet Autonomous Region, and children aged 0–6 years were selected as the research subjects. A total of 4433 children were included in the three surveys, including 1889, 1547 and 997 valid samples. The included predictors covered five dimensions: demographic characteristics (age, gender), family socioeconomic status (number of registered residents, subsistence allowance households, and per capita household income), access to medical care (distance to the nearest medical institution, travel time), household sanitation (type of toilet, drinking water source), and survey year. Three prediction models, including logistic regression, random forest and XGBoost, were constructed and optimized by five-fold cross validation. The performance was evaluated by AUC and accuracy, and the explanatory analysis of the best model was performed by SHAP value. **Results:** The two-week prevalence of children was 7.69% (2013:5.61%, 2018:9.57%, 2023:8.73%). The AUC of logistic regression, random forest and XGBoost in the validation set was 0.666, 0.524 and 0.612, respectively, and the accuracy was 0.923, 0.923 and 0.707, respectively. SHAP importance analysis showed that the top five predictors were: 2018 survey year, number of registered population, distance to the nearest medical and health institution, age, and gender. **Conclusion:** Logistic regression yielded the highest AUC and was therefore used for SHAP-based interpretation. Both logistic regression and random forest achieved accuracies above 0.92, whereas XGBoost had an accuracy of 0.707. The discrepancy between the low AUC values and the high accuracies likely results from the low prevalence of the outcome. SHAP analysis indicated that survey year, family size, healthcare accessibility, and demographic characteristics were the main influencing factors, with household sanitation also contributing. This study can serve as a reference for precision health management of children and the allocation of health resources on the Tibetan Plateau.

Keywords

Machine Learning, XGBoost, Child Health, Two-week Illness, Tibet, Health Service Survey