High-Risk Population Identification Based on Machine Learning

Hanbo Min #, Muyang Liu *, #, Siyu Luo #, Yihan Zhao #, Yi Ren #

College of Science, Minzu University of China, Beijing, China, 100081

* Corresponding Author Email: lmyop6331@163.com

#These authors contributed equally.

Abstract. This research centers on identifying high-risk populations with poor digital health capabilities among older adults in Beijing in the post-pandemic context. The goal is to identify high-risk populations precisely by comparing various machine learning algorithms, offering support for public health policymaking.The study adopts the social ecosystem theory to develop an indicator framework, collects data through a mix of online and offline approaches, and performs empirical analysis using SVM, Random Forest, and XGBOOST classification algorithms with L-SMOTE interpolation resampling to compare their effectiveness in identifying high-risk populations. Results indicate that the SMOTE-resampled SVM algorithm outperforms in accuracy, recall, and F1 score, while the L-SMOTE-resampled XGBOOST algorithm also demonstrates solid performance. Thus, when considering recall, precision, and F1 score comprehensively, the SMOTE-resampled SVM algorithm provides the best prediction results for high-risk populations.

Keywords: Digital Health, High-Risk Populations Identification, SVM, Random Forest, XGBoost, SMOTE.

1. Introduction

The spread of the COVID-19 pandemic and the acceleration of societal aging have made the health of older adults a growing concern. In the post-pandemic period, despite the wide use of digital technology in healthcare, older adults lack digital health literacy. Their internet usage rate is low, and they struggle to access digital health resources and apps, along with challenges related to knowledge, physical health, and privacy. Meanwhile, the government promotes age-friendly online services and “Internet + Medical Health” initiatives, stressing the need to boost older adults’ digital health capabilities. This study thus focuses on Beijing’s older adults, looking into their digital health technology anxiety and related factors to ease anxiety and improve service utilization[1]. Moreover, analyzing their digital health capabilities in the post-pandemic era, current situations, influencing factors, and seeking improvement strategies is highly significant[2].

As global aging and digital integration in healthcare continue, academic attention on older adults’ digital health capabilities has risen. Different studies have provided insights. For example, Huang aims to bridge the digital divide in elderly healthcare via electronic health records, suggesting measures like strengthening regulations for data security, building collection platforms for better service, and introducing “three-code integration” to reduce technical hurdles[3]. Si explores the internet’s impact on older adults’ lives and their coping mechanisms, advocating for a digitally inclusive society and cultural adaptation education[4]. Sun et al. spotlight the difficulties older adults had with “health codes” during the pandemic, showing the digital divide’s manifestations and causes, and stressing community, family, and intergenerational digital support[5]. However, current research mainly focuses on single aspects, lacking a comprehensive approach. In the post-pandemic era, there’s a lack of studies using the social ecosystem theory to fully explore factors and strategies for enhancing older adults’ digital health capabilities.

Considering the widespread aging population and the high prevalence of chronic diseases, the demand for managing high-risk populations is critical, yet current research is limited and struggles with data imbalance challenges.This study innovatively combines the social ecosystem theory with machine learning algorithms to create a comprehensive model for analyzing older adults’ digital health capabilities post-pandemic. It assesses the current state and factors in Beijing, identifies high-risk population, and offers targeted improvement plans. This new method fills research gaps in systematization and comprehensiveness, provides a new perspective and framework for exploring older adults' digital health, and enables more precise interventions to enhance their digital health literacy.

The study is structured as follows: Chapter 1 gives the introduction, covering the background, current status, and innovations; Chapter 2 presents relevant theories, discussing digital health capabilities and classification models; Chapter 3 centers on empirical analysis, involving data processing based on the indicator system, profile building, and algorithm comparison; Chapter 4 states the conclusions, evaluating algorithm performance and providing practical advice; Chapter 5 summarizes, highlighting the findings and suggesting the method's expansion to wider regions and populations.

2. Methods

Support Vector Machine (SVM) identifies the optimal classification hyperplane to separate samples, employs kernel functions to convert nonlinear problems into linear ones in high-dimensional space, and achieves the global optimum using quadratic programming. The model enhances classification accuracy and optimizes classifier weights, reducing the errors associated with reliance on empirical values [6].

The Random Forest classification model operates on the core principle of building multiple decision trees and aggregating their results to enhance prediction accuracy and model robustness. The key challenge in decision tree algorithms lies in selecting the best splitting attributes. During prediction, Random Forest integrates predictions from multiple decision trees, reducing generalization error, minimizing overfitting, and improving accuracy [7].

XGBoost is an optimized algorithm built on boosting trees, offering efficiency, flexibility, and robust generalization capabilities. It enhances prediction performance by combining multiple weak learners (CART decision trees) in an additive model. Its objective function is $Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$;Generating new samples around minority class data balances the data distribution, effectively addressing model bias issues arising from class imbalance [8].

SMOTE is a well-known oversampling technique designed to handle the issue of imbalanced datasets. The core concept involves analyzing minority class samples and artificially generating new ones, adding them to the original dataset to balance the ratio of positive and negative samples [9]. SMOTE is an enhanced method of random oversampling that is straightforward, efficient, and capable of mitigating the risk of overfitting [10].

3. Empirical Analysis

This empirical study is designed to assess the digital health capabilities of older adults in the post-pandemic era using the developed evaluation indicator system. By collecting and analyzing data, it creates user profiles and applies various machine learning algorithms to identify and compare high-risk groups with poor digital health capabilities. The objective is to achieve a comprehensive understanding of the current state of older adults' digital health capabilities, uncover influencing factors, and offer effective methods and strategies for improvement. The detailed process flow is shown in Figure 1:

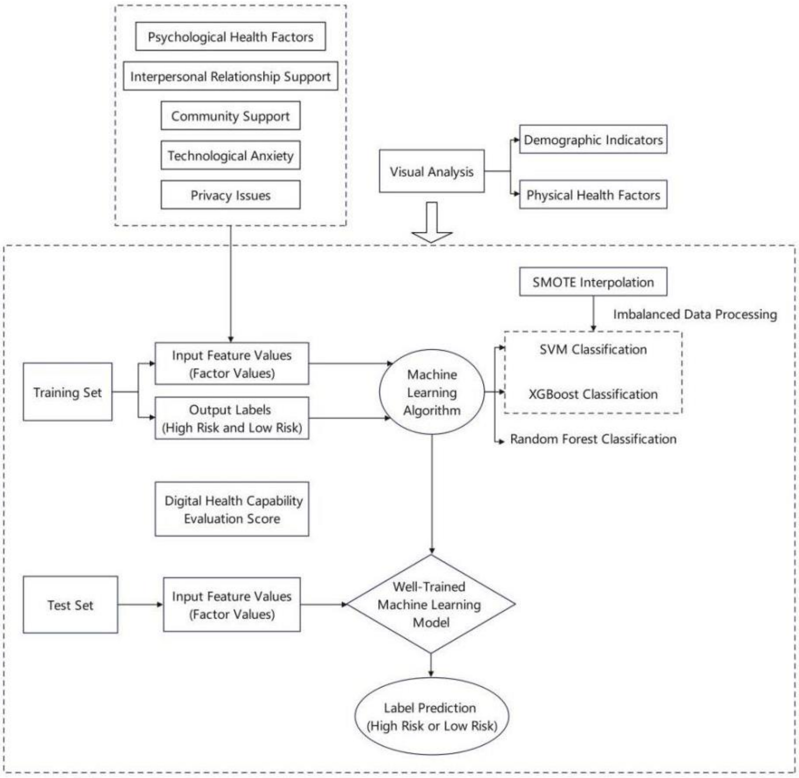


Figure 1 Empirical Study Flowchart

This research applies Charles Zastrow's social ecosystem theory, analyzing the micro, meso, and macro social ecosystems of older adults in Beijing to establish an indicator system of factors influencing digital health capability in the post-pandemic era. Moreover, the eHealth Literacy Scale designed by Norman et al. in 2006 is used as the evaluation standard for digital health capability, as detailed in Table 1-2.

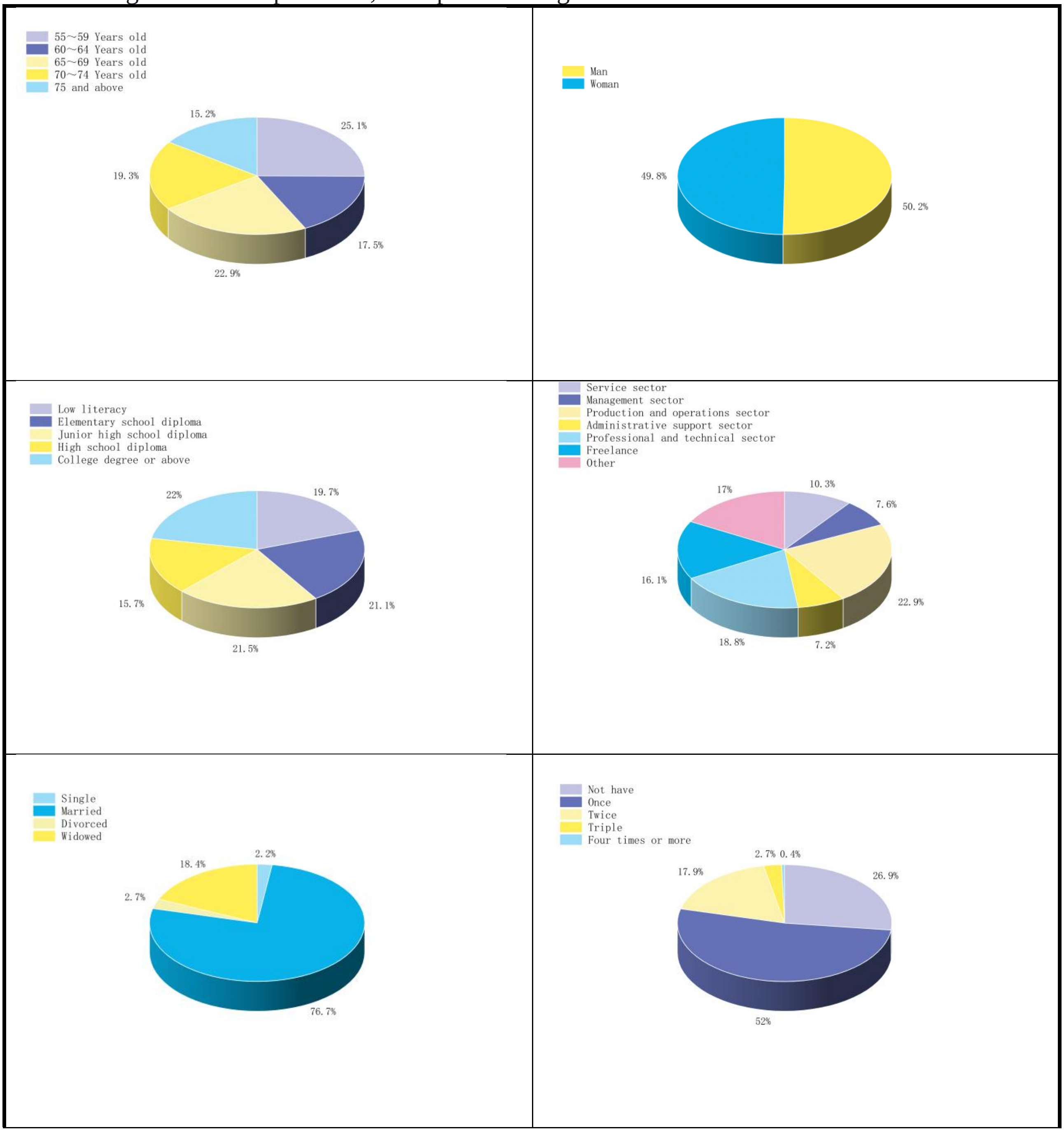
Table 1 Post-Pandemic Factors Influencing Digital Health Capability for Older Adults

Primary Dimensions	Secondary Dimensions	Investigation Content
Individual Health Dimension	X_{11}Demographic Indicators	Age, gender, profession, educational background, marital status, income, housing situation
	X_{12}Physical Health Indicators	BMI, health status, disease history, frequency of exercise, eating habits
	X_{13}Psychological Health Indicators	Alertness, adaptability, emotional regulation, resilience, decision-making, life satisfaction
Social Support Dimension	X_{21}Interpersonal Relationship Support	Family and friends' influence (e.g., health discussions, assistance with digital tools) Access to health management content (e.g., short videos, social media posts)
	X_{22}Community Support	Related posters, WeChat accounts, community noticeboards, lectures, advertisements
Era Environment Dimension	X_{31}Technology Anxiety	Recreational stress relief (intention to use), proficiency in electronic features, information anxiety
	X_{32}Privacy Concerns	Worries about data privacy breaches, etc.

Table 2 Scoring of Digital Health Abilities

Average	Median	Mode	First Quartile	Third Quartile
23.81	24.00	24.00	16.00	32.00

Utilizing the constructed indicator system, this study conducts an in-depth analysis of the elderly's characteristics and develops multidimensional user profiles to gain a thorough and detailed insight into their digital health capabilities, as depicted in Figure 2.



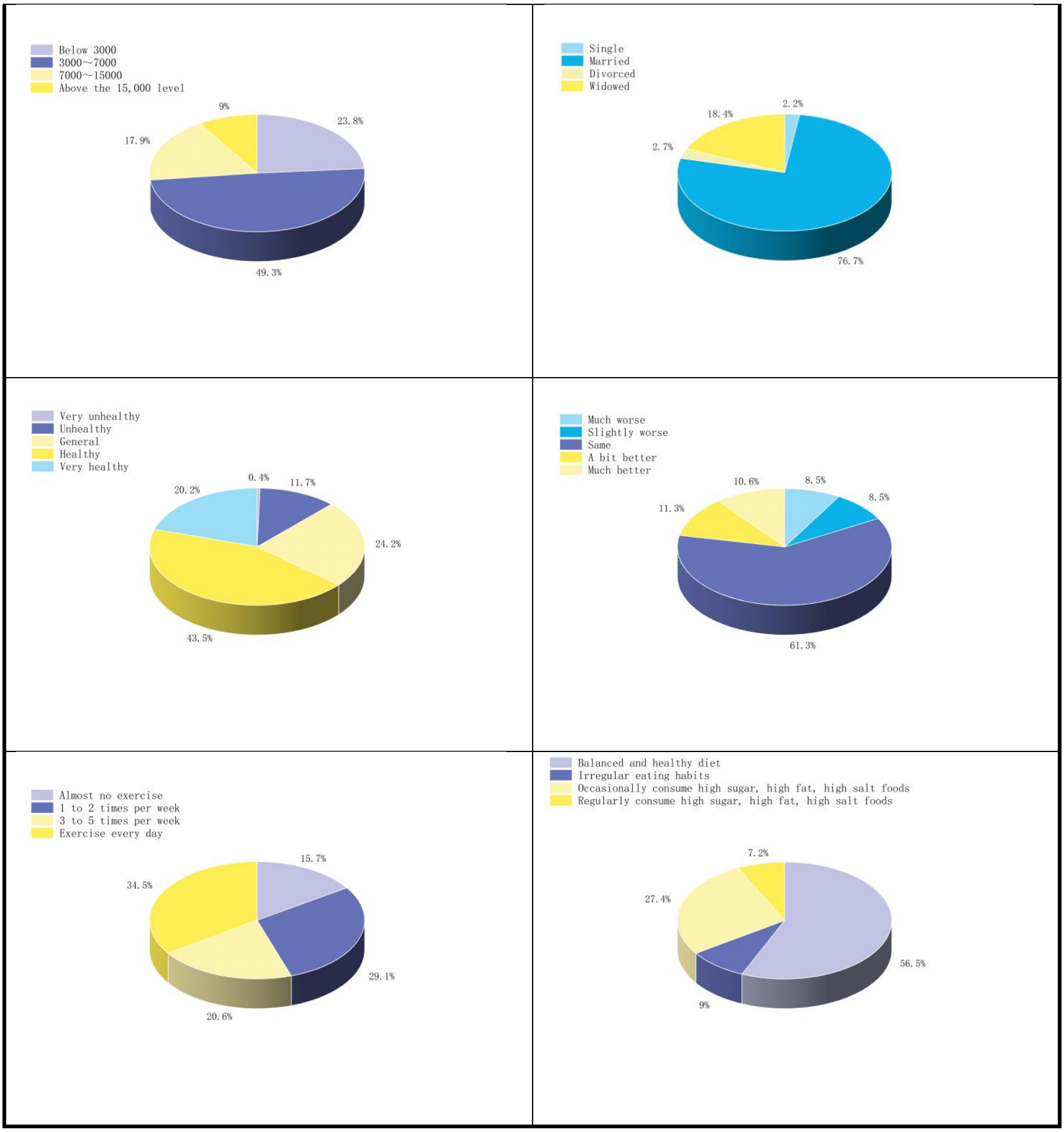


Figure 2 User Profile Feature Diagram Groups

By developing user profiles, this study gained a comprehensive understanding of the elderly's fundamental characteristics. Subsequently, various machine learning algorithms were employed to identify high-risk subgroups with poor digital health capabilities, offering a foundation for designing targeted improvement strategies.

The high-risk population identification model adopts supervised machine learning. Questionnaire section scores serve as the dependent variable (Y), with a threshold of 16 points set to categorize individuals into high-risk and non-high-risk groups. These scores are used to train the model for posterior classification. Secondary indicators in the evaluation system are used as independent variables, derived from the average scores of corresponding questionnaire sections to quantify key influencing factors.

The data in this study were all collected through questionnaires, and the respondents were the elderly in Beijing. Prior to the large - scale survey, we rigorously designed, repeatedly pre - tested, and optimized the questionnaires. After professional reliability and validity tests, both the reliability and validity of the questionnaires met the standard requirements of social science research. Thus, the questionnaires can reliably and effectively collect the data needed for the research, providing a solid data foundation for subsequent analysis.

To enhance the model further, this study defines evaluation metrics, detailed in Table 3:

Table 3 Model Assessment Indicators

Indicator	Formula
Accuracy	$Accuracy = \frac{TP + TF}{TP + TF + FP + FN}$
Recall	$Recall_{HighRisk} = \frac{TP}{TP+FN}$
Precision	$Precision_{HighRisk} = \frac{TP}{TP + FP}$
F1 Score	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

All three models have a training-to-testing ratio of 8:2. The SVM classification model employs a Gaussian kernel function with a regularization parameter of 1, and the Random Forest model includes 100 decision trees. The SMOTE-based SVM classification model utilizes a Gaussian kernel function with an influence range of 1 and a penalty parameter of 10.

The empirical analysis above provided a comprehensive understanding of the digital health capabilities of older adults, serving as a foundation for further research and practical applications. The conclusions derived from the analysis are as follows.

4. Results

This study compared the performance of four different machine learning algorithms for identifying high-risk populations, including traditional SVM and Random Forest algorithms, and SVM and XGBOOST algorithms enhanced with SMOTE and L-SMOTE interpolation resampling. Table 4 presents a summary and comparison of the performance of the four algorithms.

Table 4 Comparison of the Performance of Four Algorithms for High-Risk Population Identification (High-Risk Identification)

	SVM	**Random Forest**	**SMOTE-Resampled SVM**	**L-SMOTE-Resampled XGBOOST**
Accuracy	78.57%	76.19%	90.00%	87.00%
Recall	40.00%	50.00%	88.00%	84.00%
Precision	57.14%	50.00%	86.00%	82.00%
F1 Score	46.15%	50.00%	87.00%	83.00%

According to Table 4, firstly, the SVM algorithm excels in identifying non-high-risk individuals but performs poorly in recognizing high-risk ones, with a recall rate of only 40%, suggesting many high-risk individuals are misclassified. Secondly, the Random Forest algorithm performs marginally better than SVM in identifying high-risk individuals, achieving recall and precision rates of 50%, though its overall performance remains unsatisfactory. Thirdly, SMOTE substantially enhances the SVM algorithm's performance, particularly in recall and precision, demonstrating its effectiveness in managing imbalanced datasets. Fourthly, L-SMOTE likewise improves the XGBOOST algorithm's performance. Although its precision slightly lags behind the SMOTE-resampled SVM algorithm, its overall performance remains robust.

Thus, taking recall and precision into account, the SMOTE-resampled SVM algorithm demonstrates the best prediction performance for high-risk groups.

Table 5 Statistical Table of High-Risk Population Characteristics

Characteristic	Minimum Value	Maximum Value	Average	Standard Deviation	Kurtosis	Skewness
Mental Health Level	1.57	4.71	3.24	0.68	0.16	0.19
Interpersonal Support Level	1.00	4.50	2.83	0.77	-0.16	-0.52
Community Support Level	1.00	4.00	2.25	0.92	0.25	-0.85
Community Support Level	2.33	5.00	4.07	0.66	-0.17	-0.25
Degree of Privacy Concern	1.00	5.00	4.17	0.85	-1.16	1.93

Note: The first three features use positive scoring, while the last two use negative scoring. For positively scored features, scores exceeding the mean by 3 points indicate a healthy level, and scores below 3 points indicate an unhealthy level. The opposite applies for negatively scored features.

According to Table 5, among the positively scored features, only the mental health level is above 3 points, suggesting that despite low digital health capabilities, high-risk groups maintain a generally healthy psychological state.

Furthermore, the interpersonal and community support levels are both under 3 points, suggesting that neglect by family and friends, coupled with insufficient community attention, could be primary traits of low digital health capabilities among high-risk groups.

Regarding negatively scored features, the average levels of technology anxiety and privacy concerns exceed 4, highlighting that discomfort, fear when using digital technologies, and privacy concerns are key traits of high-risk groups and pivotal factors underlying poor digital health capabilities.

5. Conclusions

This study addresses the issue of identifying high-risk populations with poor digital health capabilities among elderly residents in Beijing in the post-pandemic context. It employs the social ecosystem theory to establish an indicator system, collect data, and conduct empirical analysis using SVM, Random Forest, L-SMOTE-resampled XGBOOST, and SMOTE-resampled SVM classification algorithms. The findings demonstrate the effectiveness of these algorithms, confirming that the SMOTE-resampled SVM algorithm performs best in terms of accuracy, recall, and F1 score, while the L-SMOTE-resampled XGBOOST algorithm also exhibits strong performance. The results of this research offer data support and decision-making guidance for related areas. In the future, the methods from this study can be expanded to additional regions and different age groups to further improve the digital health capabilities of older adults, facilitate the integration of their digital skills with the digital environment, generate governance synergies, and offer more comprehensive and precise evidence for public health policy formulation.

References

[1] Qin Chaohui, Huang Wenhao. "Digital Health Technologies in Age-Friendly Services: Current Supply and Development Strategies" [J]. Soft Science of Health, 2023, 37(10): 22-27.

[2] Peng Yanxia, Gao Yunfei, Yong Jingjing, et al. "An Analysis of Digital Health Technology Anxiety and Nursing Strategies for Community Elderly" [J]. Chinese Journal of Nursing, 2023, 58(11): 1345-1351.

[3] Huang Xinhui. "The Use of Electronic Health Records to Bridge the Digital Divide in Elderly Healthcare" [J]. Information Systems Engineering, 2023, (02): 10-13.

[4] Si Ge. "Internet Use and Its Impact on the Lives of Older Adults: Coping Strategies" [J]. Gerontology Research, 2021, 9(09): 69-78.

[5] Sun Qihu, Guan Yutong. "Research on the Digital Divide in Older Adults During the COVID-19 Pandemic: A Case Study of 'Health Code' Usage Challenges" [J]. Journal of Anhui University of Science and Technology (Social Science Edition), 2022, 24(02): 81-86.

[6] Wang Xueqing, Liu Shuang, Li Qiuyan, et al. "Classification of Tunnel Rock Mass with SVM Based on K-Fold Cross-Validation" [J]. Mining and Metallurgical Engineering, 2021, 41(06): 126-128+133.

[7] Yang Heya, Xing Wenshuo, Chen Cong, et al. "Open-Circuit Fault Detection in Modular Multilevel Converters Using a Random Forest Binary Classifier" [J]. Proceedings of the CSEE, 2023, 43(10): 3916-3928.

[8] Wang Qi, Xiong Shalina, Zhan Rou, et al. "XGBoost-Based Financial Fraud Detection for Imbalanced Datasets" [J]. Computer Era, 2023, (12): 59-63.

[9] Song Wanpeng, Yang Ming, Jin Yuanyuan. "Analysis of Prediction Models Using SMOTE Sampling and KNN Algorithm" [J]. Application of Integrated Circuits, 2023, 40(04): 76-78.

[10] Chen Hailong, Yang Chang, Du Mei, et al. "Credit Risk Prediction Using LightGBM Enhanced with Boundary-Adaptive SMOTE and FocalLoss Function" [J]. Computer Applications, 2022, 42(07): 2256-2264.