Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/92/20241706

Machine learning in geography education: Evaluating student performance in rural china

Wenteng Gong

Chongqing National Experimental School

1321926417@qq.com

Abstract. The disparity in the development and application of educational resources between urban and rural areas is pronounced, with rural students often encountering significant challenges in acquiring geographic knowledge and developing spatial cognitive skills due to their geographical isolation and limited access to educational resources. This study aims to investigate the effectiveness of geography education resource development in rural areas by examining the influence of various educational resource characteristics on students' scores in geography education. Through the application of data analysis and machine learning predictive methods, this research explores the impact of 13 variables, including gender, age, parental education level, family background, internet accessibility, outdoor practical activities, and family outdoor travel frequency. These factors were modeled and assessed using a range of machine learning algorithms. The findings reveal differential impacts of these characteristics on students' scores in geography education, with models such as Random Forest and XGBoost demonstrating superior performance in predicting scores in geography courses. This research provides empirical data and a scientific framework to support the optimization of geography education resources in rural areas, offering theoretical and practical insights for advancing educational equity and fostering local socioeconomic development.

Keywords: Geography education, Educational resource disparities, Machine learning.

1. Introduction

In the context of accelerating globalization, education has increasingly been recognized as a critical vehicle for enhancing students' geographic literacy and spatial cognitive abilities. The effective development of educational resources in rural areas, in particular, is of profound importance for promoting educational equity and elevating overall student quality. According to the Ministry of Education's "14th Five-Year Plan," China's basic education informatization penetration rate is projected to exceed 90% by 2025, creating substantial opportunities for the optimization of basic education resources, including those in rural regions [1].

Recent years have witnessed the rapid digitalization of education in rural areas, facilitated by the widespread adoption of information technology and the deepening reach of the Internet. Data indicate that rural internet access increased from 56.4% in 2017 to 70.7% in 2020, establishing the foundational conditions necessary for the effective distribution and utilization of educational resources [2-3]. Despite these advances, significant disparities persist in the construction and application of educational resources between rural and urban areas. Rural students continue to face considerable obstacles in accessing

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/92/20241706

geographic knowledge and developing spatial cognitive skills, largely attributable to their remote locations and the scarcity of educational resources.

In the current educational system, the significance of geographic education is increasingly evident. Geography is not merely the study of the natural environment and human societies but also a crucial means of fostering students' global perspectives and understanding complex world dynamics. Geographic education enables students to better comprehend globalization processes, environmental issues, and the regional disparities in socio-economic development. Thus, it plays an irreplaceable role in the holistic development of students and the cultivation of their social responsibility. However, despite the critical importance of geographic education in contemporary society, many rural schools continue to face significant challenges in this domain. The lack of appropriate teaching resources and qualified educators has resulted in rural students being relatively weak in acquiring geographic knowledge and developing spatial thinking skills, further exacerbating the educational imbalance between urban and rural areas.

The present study seeks to explore the effectiveness of geography educational in rural areas. Utilizing data analysis and machine learning prediction techniques, this research will investigate the effects of various educational resource characteristics on students' scores in geography courses. By collecting and analyzing data on 13 variables—including gender, age, parental education level, family background, Internet accessibility, outdoor practice activities and family outdoor travel frequency.T his study will develop a predictive model to assess the impact of these characteristics on students' scores in geography courses. The results will provide a scientific basis for optimizing educational resources in rural areas, thereby promoting educational equity and sustainable socioeconomic development in these communities.

2. Literature review

With the acceleration of globalization, the importance of geography education has become increasingly prominent in China's education system. Scholars have discussed the core concepts, educational objectives, technical applications and challenges of geography education from various angles [4]. Nick Cheung and Yuan Xiaoting (2015) put forward the screening, interpretation and characteristics of geographical core concepts, emphasizing the key role of these concepts in students' aunderstanding of the nature of geography [5]. In addition, Nick Cheung and Yuan Xiaoting (2016) pointed out that the idea of scale is the basic content in geography education, which is very important to improve students' spatial cognitive ability [6].

Although the importance of geography education has been widely recognized, there are still many challenges in actual teaching in China, especially in rural areas. In order to solve this problems, in recent years, machine learning methods have been gradually introduced into educational research to evaluate and optimize the distribution and use of educational resources. This kind of research not only helps to reveal the specific challenges faced by rural students in geography learning, but also provides a scientific basis for optimizing the allocation of educational resources and formulating more effective teaching strategies [7]. This research direction is consistent with the suggestions in existing geography education literature, such as Nick Cheung and Yuan Xiaoting's suggestion on strengthening students' spatial cognition, and provides a new perspective for future geography education research and practice.

To sum up, China's geography education has made remarkable progress in core concepts, technology application and environmental education. Future research should pay more attention to how to evaluate and optimize the effectiveness of educational resources through advanced technologies such as machine learning, especially in rural areas, so as to improve students' geographical literacy, promote educational equity, and ultimately promote the sustainable development of social economy.

3. Data and model

3.1. Data Characterization

In machine learning, features represent the variables that describe the attributes of data samples, serving as the foundational inputs for models to make predictions or classifications. These features are typically

DOI: 10.54254/2755-2721/92/20241706

utilized as independent variables within models to capture distinctions and patterns among samples. Consequently, the selection and extraction of features are pivotal to the performance of the model. High-quality features not only enhance model accuracy but also improve its generalizability. In this study, features such as gender, age, parental education level, family background, internet accessibility, outdoor practical activities, and family outdoor travel frequency, with the goal of predicting their scores in geography courses. The target column, or the dependent variable, is the variable that the model aims to predict or classify in a supervised learning task. It contains the actual values or categories that the model learns and predicts. In this study, the level of students' scores in geography courses serves as the target column, describing their ability to study geography courses. The model's task is to categorize students' studying levels based on these characteristics, thereby providing data-driven insights and recommendations for educators and policymakers.

The dataset used in this study comprises 13 features and 1 target column. These features encompass students' personal background information, educational environment, and scores. They include variables such as gender, age, parental education level, family background, internet accessibility, outdoor practical activities, family outdoor travel frequency, social media learning, participation in study groups, extracurricular tutoring, historically geographical achievements, and extracurricular reading. The target column describes the level of students' scores in geography courses.

During the data preprocessing phase, it is standard practice to convert categorical into numerical representations that are interpretable by machine learning models. This is often achieved through label encoding, which assigns a unique numerical value to each category, ensuring the model can process the data appropriately.

To allow the model sufficient data for training and validation, the dataset was divided into training and test sets. Seventy percent of the data (approximately 843 samples) was allocated to the training set for model training, while the remaining 30% (approximately 362 samples) was used as the test set to evaluate the model’s performance.

3.2. Model Introduction

In this study, five distinct machine learning algorithms were selected for the development and evaluation of predictive models: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, XGBoost, and CatBoost.

Logistic Regression, a classical algorithm primarily utilized for binary classification tasks, is rooted in linear regression principles. It applies a logistic function to transform continuous predictor variables into probability estimates, offering strengths in simplicity, interpretability, and scalability for large datasets.

The K-Nearest Neighbors (KNN) algorithm is a type of instance-based learning, where classification is determined by the labels of the K nearest neighbors within the feature space. Unlike other algorithms, KNN bypasses the traditional model training process, instead utilizing the training data directly for prediction, providing flexibility in certain applications.

Random Forest, an ensemble learning approach, enhances predictive accuracy and model robustness through the construction and aggregation of multiple decision trees. This algorithm is particularly adept at managing large numbers of input features and maintains strong performance even in the presence of missing data and outliers.

XGBoost is a gradient boosting algorithm designed to incrementally improve predictive performance. It builds a series of decision trees, with each successive tree aiming to correct the residual errors of the preceding one. XGBoost excels in handling structured data and high-dimensional features, while also offering rapid training times.

CatBoost, another gradient boosting algorithm, is specifically tailored for classification tasks and is optimized to handle categorical features and missing values automatically. It demonstrates strong robustness when working with large-scale data and sparse feature sets.

DOI: 10.54254/2755-2721/92/20241706

4. Modeling and Evaluation

4.1. Model Performance Comparison and Cross-validation

A 5-fold cross-validation was conducted to assess the stability and reliability of the selected models. The average accuracy scores on the validation set were used as the performance metric. The cross-validation results (see Figure 1) indicate that the Random Forest model achieved the highest accuracy at 0.896, followed closely by XGBoost and CatBoost, with accuracies of 0.891 and 0.886, respectively. In contrast, the Logistic Regression model demonstrated weaker performance, with an accuracy of 0.688. These findings suggest that Random Forest, XGBoost, and CatBoost are better suited to addressing the research problem, as they more effectively capture the complex relationships within the data.

Table 1. Classifiers Accuracy Score

Model	Accuracy	Precision	Recall
LogisticRgeress	0.688	0.693	0.691
KNearest	0.771	0.768	0.780
RandomForestClassifier	0.896	0.902	0.919
XGBClassifier	0.891	0.899	0.852
CatBoostClassifier	0.886	0.856	0.881

4.2. Modelling Verification

The model trained above is verified on the test set, and the results are shown in Table 2. ROC curve and PR diagram of random forest model are shown in Figure 2.

Table 2. Classifiers Accuracy Score

Model	Accuracy	Precision	Recall
LogisticRgeress	0.548	0.602	0.654
KNearest	0.663	0.682	0.680
RandomForestClassifier	0.826	0.782	0.805
XGBClassifier	0.791	0.734	0.752
CatBoostClassifier	0.802	0.761	0.789

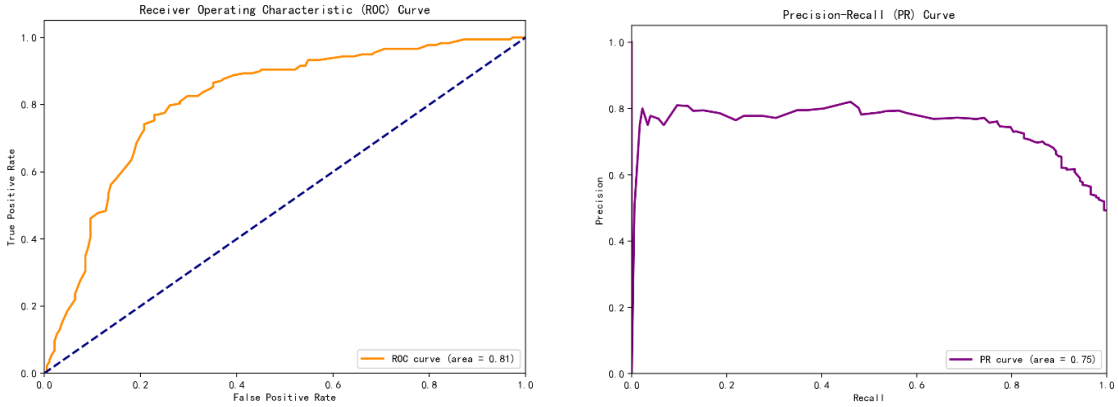

Figure 1. ROC diagram and PR diagram of RandomForestModel

The table compares the performance of five machine learning models—Logistic Regression, K-Nearest Neighbors, Random Forest, XGBoost, and CatBoost—based on their accuracy, precision, and recall metrics. Among these models, the Random Forest Classifier stands out with the highest performance across all metrics, indicating its superior ability to correctly classify both positive and negative cases, making it the most reliable model in this comparison.

DOI: 10.54254/2755-2721/92/20241706

In contrast, the Logistic Regression model exhibits the poorest performance, particularly in terms of accuracy, which is significantly lower than that of the other models. This suggests that while Logistic Regression is moderately effective at identifying true positives, it struggles with overall classification accuracy.

The K-Nearest Neighbors model shows a moderate and balanced performance with comparable precision and recall values, indicating it manages false positives and false negatives effectively, though its overall accuracy lags behind the more advanced models.

Both the XGBoost and CatBoost classifiers demonstrate strong performance, particularly in terms of recall and precision, respectively. CatBoost's slightly higher accuracy suggests it performs marginally better than XGBoost, though both models are close contenders in terms of their effectiveness.

Overall, the Random Forest Classifier is the most effective model in this analysis, with CatBoost and XGBoost also providing strong, competitive results. Logistic Regression, however, is notably less effective, particularly in terms of overall accuracy.

4.3. Eigenvector Analysis

Figure 3 illustrates the importance of different features in a Random Forest model. It shows that the most impactful factors influencing scores in geography education are "family outdoor travel", "gender", and "historically geographical achievements", which account for 15.4%, 13.2%, and 11.3% respectively. That implies that the exposure to real-world geographical settings, students' gender, and historical geographical achievements significantly affect students' scores in geography education.

Other notable contributors include "outdoor practical activities" and "age" with a feature importance of 9.5% and 8.2% respectively. This suggests that hands-on geographical activities and age are also influential to a certain extent. Additionally, the education level of parents, participation in study groups, and family background collectively contribute to around 19.8% of the model.

Furthermore, factors like "internet accessibility", "social media learning", "extracurricular reading" and "extracurricular tutoring" collectively make up around 17% of feature importance. Despite their lesser influence relative to other factors, they still play a role in shaping students' scores in geography education.

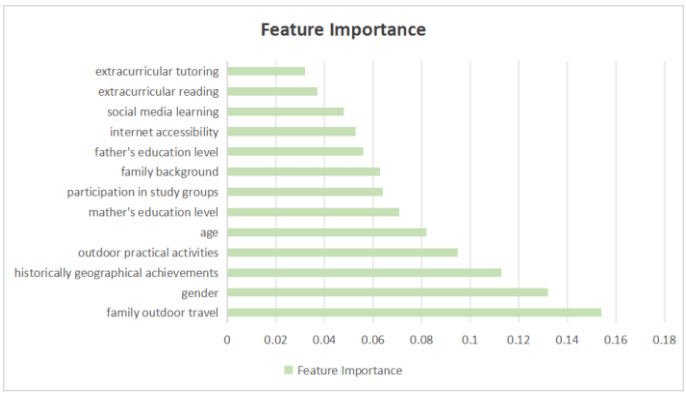

Figure 2. Feature Importance Ranking

5. Conclusions

This study aimed to explore students' perfermance in geography education, specifically within rural contexts, and to identify the factors that influence it by applying machine learning models. Through the analysis and modeling of various features in the dataset, several key conclusions were drawn.

The figure suggests that students' scores in geography education is significantly influenced by factors such as family outdoor travel, gender, and historical geographical achievements. Other factors like outdoor practical activities and age also contribute notably. Additionally, parental education level, study

DOI: 10.54254/2755-2721/92/20241706

group participation, and family background collectively constitute around 19.8% of the model. Despite lesser influence, internet accessibility, social media learning, extracurricular reading, and tutoring still play a role.

Furthermore, by comparing the performance of different machine learning models, it was found that Random Forest, XGBoost, and CatBoost models outperformed others in predicting students' scores in geography courses. The Random Forest model, in particular, demonstrated the highest average accuracy during cross-validation, indicating its effectiveness in capturing key features related to student scores and its strong generalization capability.

Based on the findings, it's suggested that educational institutions enhance outdoor geographical activities. The significant impact of factors like family outdoor travel and outdoor practical activities show that real-world exposure can improve students' geography education scores. Therefore, increasing opportunities for outdoor learning experiences could enhance their geographical perception and application skills.

References

[1] Ministry of Commerce of the People's Republic of China. (2022). "14th Five-Year" E-commerce Development Plan.

[2] General Administration of Customs of China. (2023). Import and Export Data of China's Cross-border E-commerce in January 2023.

[3] China's Cross-border E-commerce Market in 2023. iiMedia Research. (2023). Analysis Report on the Development Trends of China's Cross-border E-commerce Market in 2023.

[4] Wang Jincai, Zhang Haizhong. Re-comment on the disadvantages of psychological education in primary and secondary schools based on the compulsory education curriculum plan and curriculum standards (2022 edition) [J]. Gansu Education Research, 2024,(06):19-22.

[5] Nick Cheung, Yuan Xiaoting. The thought of scale in geography education: basic content and teaching value [J]. Curriculum, teaching materials and teaching methods, 2016,36 (06): 103-108.

[6] Nick Cheung, Yuan Xiaoting. The core concepts of geography in middle school geography curriculum: screening, interpretation and characteristics [J]. Curriculum, teaching materials, teaching methods, 2015,35 (11): 113-118.

[7] Pan Liping. The Construction of Contemporary Teaching Space: Theoretical Implication and Action Path-Reflections on the Implementation of Compulsory Education Curriculum Scheme and Curriculum Standards (2022 Edition) [J]. China Education Journal, 2023,(03):39-44.