Research on Relevant Factors of Helping Rural Revitalization Based on Machine Learning Analysis - Taking Kengwei Village in Shantou City as an Example

Chun Zhou[1a], Yisi Chen[2b], Jingyi Zhao[1c], Changzhi Niu[1d], Mengyao Wang[1e*], Xin Ma[3f], Ying Feng[2g], Yi Zhang[2h]

[1]Big Data College, Zhuhai College of Science and Technology, ZHuhai, China 519000
[2]School of Internet Finance and Information Engineering, Guangdong University of Finance, Guangzhou, China 510000
[3]School of Pharmaceutical Business, Guangdong Pharmaceutical University, Zhongshan, China 528400

[a]zhouchun1231@163.com, [b] 13425304131@163.com, [c]18927466176@163.com, [d]18346700563@163.com, [f]13433855815@163.com, [g]2761864305@qq.com, [h]1937410563@qq.com

[e*] Corresponding author: wangmengyao_o@qq.com

Abstract. Nowadays, the development of rural areas still faces difficulties and challenges, such as unbalanced economic development, insufficient infrastructure and public services, and backwardness of education level due to brain drain still exist. In order to promote the modernization of agriculture and industrial upgrading, and to support the sustainable development of the national economy, the article takes Kengwei Village in Shantou City as an example, and carries out a specific analysis of the factors that promote the revitalization of the countryside based on the relevant application of machine learning using the algorithms of K-nearest neighbour (KNN), decision tree, logistic regression, and random forest. Through data visualization, a more intuitive understanding of the cognitive village, enhance the understanding of the three rural issues, and actively contribute to rural revitalization.

Keywords*:* data analysis; rural revitalization; decision tree; K-nearest neighbour (KNN); logistic regression

1 INTRODUCTION

Rural revitalization is an inevitable requirement for building a socialist modernization in China. While modern agriculture has developed rapidly, rural areas have lagged behind, facing economic disparities, environmental degradation, and inadequate infrastructure compared to urban counterparts[1]. This study employs algorithms such as knn, decision trees, logistic regression, and random forests to build models aimed at promoting agricultural modernization and industrial upgrading. Evaluation metrics

V. Vasilev et al. (eds.), *Proceedings of the 2024 5th International Conference on Management Science and Engineering Management (ICMSEM 2024)*, Advances in Economics, Business and Management Research 306,
https://doi.org/10.2991/978-94-6463-570-6_146

such as confusion matrices, accuracy, ROC curves, AUC, precision, recall, and F1 score assess model performance. Detailed analysis of factors contributing to rural revitalization is conducted, enhancing understanding through visualization to provide a scientific basis for rural revitalization strategies.

2 MATERIALS AND METHODS

2.1 Overview of the Study Area

Kengwei Village, located at the southeast end of Chaoyang District in Shantou City, borders the Leian River estuary on three sides, enjoying a unique geographical advantage and scenic beauty. The village features 50 acres of rapeseed and mulberry flower fields that complement each other, showcasing not only beautiful surroundings but also profound historical and cultural significance. As an old revolutionary area, Kengwei Village preserves remnants of anti-Japanese militia sites, embodying the spirit of the revolution that continues to inspire villagers in their perseverance. The village has effectively connected its achievements in poverty alleviation with rural revitalization efforts, continually leading local residents toward prosperity.

2.2 Research Indicators and Content

The study combines factors such as villagers' education levels, presence of returning university graduates in households, influx of external population, public safety, infrastructure development, and overall quality of life for analysis. The influencing factors and meanings are shown in Table 1. Due to the multitude of factors influencing rural revitalization, this research primarily focuses on villagers' sense of well-being and includes field interviews[2].

Table 1. Rural Revitalization Influencing Factors

Influencing factors	**explanation**
ducational attainment	Literacy of villagers, education in villages.
College students returning home to work	Whether the college students in the village return to work, participate in rural construction, and promote rural revitalization.
Foreign population inflows	Influx of foreign population in the village and whether it is conducive to village development and construction.
Law and order situation	The security situation in the village, order in public places and illegal and criminal behaviors.
Infrastructure development	Infrastructure construction in the village, such as houses, roads, recreational areas, etc.
Overall quality of life	Quality of life and sense of well-being of villagers and outsiders in the village.

2.3 Research Ideas

To gain deeper insights into the current status and issues of rural revitalization in China, it is crucial to conduct research and analyze relevant data. Initially, demographic variables such as gender, education level, presence of returning university graduates, awareness of external population inflow, assessment of local public safety, and evaluation of infrastructure development should be treated as control variables in the model. The dependent variable will be villagers' assessments of their quality of life[3].

2.4 Research Theory and Algorithm Model

Data Preprocessing.

Data Collection: This study employed a questionnaire survey method to randomly survey residents and migrants in Kengwei Village, Haimen Town, Chaoyang District, Shantou City, Guangdong Province. A total of 500 questionnaires were distributed, with 462 collected. The survey collected data on residents' gender, education level, presence of returning college students in their families, awareness of migrant influx into the village, public safety in Kengwei Village, infrastructure development, and overall quality of life.

Control of sample selection bias: Random sampling was used to select residents for the questionnaire survey, ensuring the sample's representativeness and randomness to minimize potential sample selection bias.

Quality control measures: Measures included defining the theme of rural revitalization factors, simplifying questionnaire structure and content, ensuring sample diversity and representativeness, and focusing on usability and accessibility to improve survey quality and provide a solid data foundation.

Data Cleaning: Data collected were cleaned and organized, involving deletion of blank rows, handling duplicate entries, and addressing outliers.

Feature Selection: For rural revitalization research, select features related to rural quality of life aspects and exclude invalid features that do not affect the prediction.

Data Standardization: Standardization involved transforming feature values to have a mean of 0 and a standard deviation of 1, eliminating scale differences among features.

Data Splitting: Split into training set and test set.

K-nearest Neighbour (KNN) Algorithm. KNN is a basic and intuitive classification and regression algorithm. It classifies or predicts by finding the K closest data points in the training set to a new data point[4].

Model Instantiation: Instantiate a KNN classifier.

KNN = KNeighborsClassifier(n_neighbors=5)

Model Training:.

Using training data to fit the KNN model.
KNN.fit(X_train, y_train)
# X_train is the feature data, y_train is the corresponding class labels

Model Prediction: Use the trained model to predict the testing set. Predict the impact of various factors on rural revitalization.

KNN_pred = KNN.predict(X_test)

Logistic Regression Model. Logistic Regression is a linear model for dichotomous or multicategorical tasks that predicts category probabilities by learning the weights of features in the data[5].

Model Establishment: LR = LogisticRegression().

Model Training: LR.*fit(X_train, y_train)*

Model Prediction: LR_pred = LR.*predict(X_test)*

Decision Tree Model. Decision tree is a common machine learning algorithm for classification and regression problems. A decision tree model builds a tree-like structure by recursively partitioning features in a dataset for the purpose of classifying or predicting data[6] .

Model Instantiation: reg = DecisionTreeRegressor().

Model Training: DT.*fit(X_train, y_train)*

Model Prediction: DT_pred = DT.*predict(X_test)*

Random Forest Model. Random forests improve the performance and stability of the model by combining multiple decision tree-based classifiers for tasks such as classification and regression[7].

Model Instantiation: RF = RandomForestClassifier().

Model Training: RF.*fit(X_train, y_train)*

Model Prediction: RF_pred = RF.*predict(X_test)*

Model Evaluation.

Confusion Matrix: Confusion matrix is a tool used to assess the performance of a classification model. It measures the accuracy of the model by comparing actual classes with predicted classes[8].

In a scenario where the predicted classes are divided into Positive and Negative, and the accuracy of predictions is classified as True or False.

There are four relationships between the prediction labels and the true labels of the data as follows:

TP(TruePositive,correctly predicted positive overdue result)

TN(TrueNegative,correctly predicted negative normal result)

FP(FalsePositive,false predicted positive overdue result)

FN(FalseNegative,false predicted negative normal result)

Model Accuracy: Accuracy is the most intuitive evaluation metric,the number of correctly classified samples divided by the total number of samples, as in Equation 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

Figure 1 shows a visualization of the accuracy of the four models, comparing the accuracy of the four models.

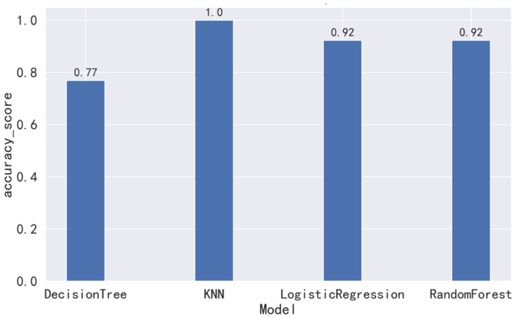

Fig. 1. Model Accuracy

ROC And AUC: Receiver Operating Characteristic (ROC) Curve and AUC (Area Under the Curve) are important indicators for evaluating the performance of binary classification models.

ROC Curve: Describes the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) at different thresholds. TPR denotes the proportion of true cases and FPR denotes the proportion of false positive cases.

AUC Indicator: The AUC metric represents the area under the ROC curve. It provides a comprehensive evaluation of a model's performance across all possible thresholds. A higher AUC value indicates better model performance.

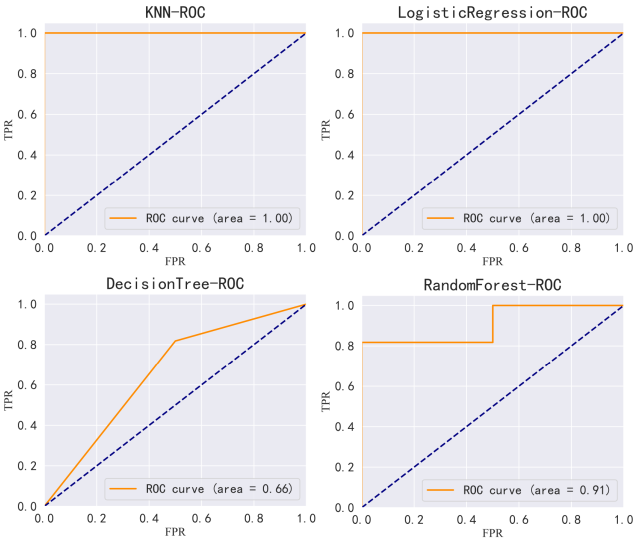

Fig. 2. ROC

Figure 2 shows the results of KNN, logistic regression, decision tree, and random forest, four models trained and evaluated using ROC curves with AUC as evaluation metrics[9].

*Comparison of performance indicators:*The performance of four different classification algorithms was evaluated using Accuracy, Precision, Recall, F1_score, AUC, five metrics respectively, with the following formulas.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\_Score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{4}$$

The performance of four algorithms, decision tree, logistic regression, KNN and random forest, were evaluated using five indexes respectively, as shown in Figure 3.

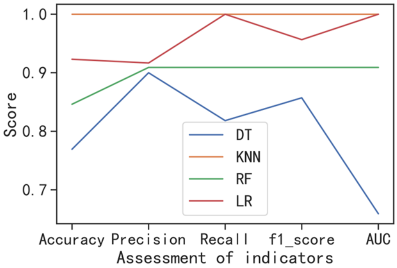

Fig. 3. Algorithm Performance Comparison

3 RESULTS AND ANALYSIS

3.1 Survey Results

This study conducted a random survey among villagers and outsiders in Kengwei Village, Haimen Town, Chaoyang District, Shantou City, Guangdong Province. Out of 500 questionnaires distributed, 462 were returned, achieving an effective response rate of 92.4%. Given ongoing improvements in local industrial development and the implementation of various pro-people policies, there has been an influx of outsiders, prompting their inclusion as subjects to ensure objective data and conclusions.[10].

According to Table 2, among the surveyed villagers and outsiders, males account for 53.9% and females for 46.1%, with a nearly equal distribution. The majority have attained primary (39.17%) or secondary education (31.82%). Only 5.41% of households have university graduates who have returned to work in the village. Regarding public safety, most respondents rate it as fair or poor, comprising 29.65% and 43.51%, respectively. Similarly, opinions on infrastructure development in the village are predominantly fair or poor, at 43.07% and 30.09% respectively.

Table 2. Factors related to rural revitalization in Kengwei Village, Haimen Town

project	category	frequency	Proportion（%）
Gender	masculine	249	53.90
	woman	213	46.10
Educational level	primary school	181	39.17
	junior high school	147	31.82
	high school	98	21.21
	College or undergraduate degree	36	7.79
	Graduate or above	0	0
Are there any college students returning to work at home	No college students	178	38.53
	There are college students who have no plans to return to their hometowns for work	259	56.06
	There are college students who have returned to their hometowns to work	25	5.41
The inflow of foreign population into the village as understood	Very rare	56	12.12
	Less	121	26.19
	commonly	158	34.20
	More	89	19.26
	A lot	38	8.23
The security situation in the village	Very good	26	5.63
	good	98	21.21
	commonly	137	29.65
	Poor	201	43.51
Infrastructure construction in the village	Very good	32	6.93
	good	92	19.91
	commonly	199	43.07

project	**category**	**frequency**	**Proportion（%）**
	Poor	139	30.09
The overall quality of life of villagers	Very good	27	5.84
	good	128	27.71
	commonly	221	47.84
	Poor	86	18.61

3.2 Analysis Results

Data visualization and analysis. Through sample analysis in Kengwei Village, villagers' evaluations of their quality of life show that 5.8% rate it as excellent, 27.7% as good, 47.8% as average, and 18.6% as poor. Nearly half of the surveyed villagers perceive their quality of life as average, as depicted in Figure 4.

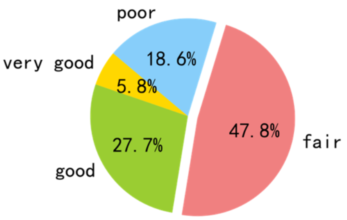

Fig. 4. Overall quality of life in Kengwei Village.

Calculate the correlation coefficients between quality of life and various factors, then plot a histogram like Figure 5. This will help identify which factors have a significant impact on the quality of life in Kengwei Village, enabling targeted strategies or policies to enhance living standards there.

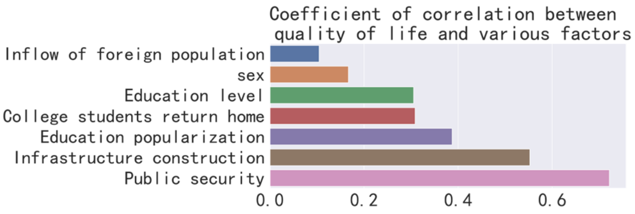

Fig. 5. Coefficient of correlation between quality of life and various factors

The figure indicates a strong correlation between public safety, infrastructure development, and residents' perception of quality of life, whereas external population and gender show lower correlation. Improvements should focus on enhancing public safety management and optimizing village infrastructure to make daily life more convenient for residents.

Analysis of modeling results. Decision Tree: Visualize and output the importance of each feature in the decision tree model. Train the decision tree model on the training set and predict on the test set. Obtain the feature importance from the trained decision tree model, sort the feature importance, and present it in a bar chart[11].

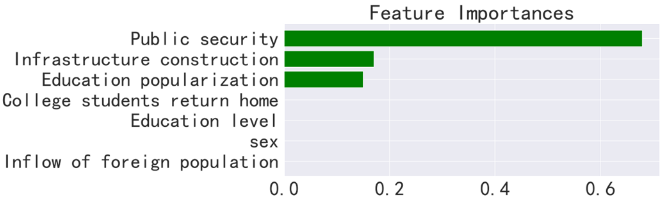

Fig. 6. Feature importances

From Figure 6, it is evident that public safety and infrastructure development in Kengwei Village have a significant impact on the quality of life for its residents.

Trained K-Nearest Neighbors (KNN) classification and Logistic Regression models using StandardScaler to standardize feature data for both training and testing sets. Created a KNeighborsClassifier object, fitted it with training data, and predicted using the test set to assess KNN model performance. Similarly, created a LogisticRegression model object, fitted it with standardized training data, and evaluated its performance. Visualized feature importance of the logistic regression model's coefficients using model.coef_ in a bar chart as shown in Figure 7[12].

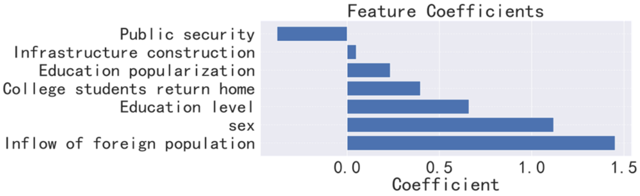

Fig. 7. Feature Coefficients

In general, public safety significantly impacts the quality of life satisfaction among residents of Kengwei Village, followed by infrastructure development. Rural revitalization efforts should prioritize improving these two aspects to effectively enhance the quality of life for villagers[13].

4 OPTIMIZATION SUGGESTIONS FOR RURAL REVITALIZATION IN KENGWEI VILLAGE BASED ON ALGORITHM ANALYSIS

Our study focuses on distributing questionnaires, organizing data, and analyzing rural revitalization factors through model algorithm and visualization analysis. We identified public safety and infrastructure development as crucial factors affecting villagers' quality of life. Based on this analysis, we propose the following optimization suggestions for promoting rural revitalization in Kengwei Village:

4.1 Improve Public Safety

Enhance police presence and patrol frequency, strengthen village security monitoring and prevention measures. Enforce public security laws and regulations, with law enforcement actively serving the community. Maintaining rural social stability is pivotal, including cracking down on illegal activities, a key strategy for rural revitalization.

4.2 Enhance Infrastructure Development

Prioritize investments in improving village roads, water supply, drainage systems, and other infrastructure to enhance residents' convenience and comfort. Promote public service facilities such as healthcare centers, cultural venues, and sports facilities to meet diverse community needs. Improve communication network coverage and internet access to support the spread of information technology in villages, fostering rural e-commerce and remote work.

4.3 Enhance Villagers' Education and Cultural Literacy

Strengthen education and training for villagers, particularly enhancing cultural literacy and vocational skills among rural youth and adults. Support and attract university graduates to return to or stay in villages for entrepreneurship or employment through policy and project support. Establish diverse cultural and educational resources like libraries, art exhibitions, and cultural festivals to enrich villagers' cultural and spiritual life.

By implementing these measures, we can effectively enhance villagers' quality of life and satisfaction. Additionally, initiatives such as establishing job platforms, enhancing policy support, and creating entrepreneurship subsidies can transform the current agricultural industry model in rural areas, further advancing comprehensive development for Kengwei Village's rural revitalization.

5 CONCLUSION

Through research, a detailed analysis of factors promoting rural revitalization is conducted. By comparing the importance of these factors, it provides a scientific basis for formulating rural revitalization strategies. Improvements in public safety, infrastructure, cultural education, and social security will enhance the quality of life for villagers and achieve comprehensive well-off status for rural populations. Rural revitalization strategies are crucial for building a modern economic system. Measures include upgrading rural industries, modernizing agriculture, and developing distinctive rural industries to adjust and upgrade the rural economic structure.

REFERENCES

1. Li X. Rural revitalization strategy writes an epoch-making stroke[J]. Agricultural Market Weekly,2017(48):46-47.
2. SUNYONG EOM, MINYOUNG JANG, NAM-SEOK JI. Human Mobility Change Pattern and Influencing Factors during COVID-19, from the Outbreak to the Deceleration Stage: A Study of Seoul Metropolitan City[J]. The Professional Geographer,2022,74(1):1-15. DOI:10.1080/00330124.2021.1949729.
3. VAN PARYS HANNA, PROVOOST VEERLE, DE SUTTER PETRA, et al. Multi family member interview studies: a focus on data analysis[J]. Journal of family therapy,2017,39(3):386-401. DOI:10.1111/1467-6427.12169.
4. Niu Chunyan, . Application of Feature Weighted KNN Classification Algorithm in Professional Curriculum Talent Training. 2024, :271-284.
5. Zaloa Sanchez-Varela. Prediction Analysis Based on Logistic Regression Modelling[M]. IntechOpen, 2022
6. Sharaf AlKheder, . Experimental road safety study of the actual driver reaction to the street ads using eye tracking, multiple linear regression and decision trees methods[J]. Expert Systems With Applications, 2024, 252:124222-.
7. Gonzalez Ricardo, Saha Ashirbani, Campbell Clinton J.V., et al. Seeing the random forest through the decision trees. Supporting learning health systems from histopathology with machine learning models: Challenges and opportunities[J]. Journal of Pathology Informatics, 2024, 15:100347-.
8. Thais Berger Barbosa da Silva, José Elievam Bessa, Leise Kelli de Oliveira, et al. Evaluation of models for estimating free-flow speed on two-lane rural highways in Brazil[J]. Latin American Transport Studies, 2024, 2:100011-.
9. Das Akhil Kumar, Biswas Saroj Kr., Mandal Ardhendu, et al. Machine Learning based Intelligent System for Breast Cancer Prediction (MLISBCP)[J]. Expert Systems With Applications, 2024, 242:122673-.
10. P. R. HANLON, R. BANDYOPADHYAY, G. P. BRORBY. Evaluating the Applicability of a Risk-based Approach (Decision Tree) to Mycotoxins Mitigation[J]. Food Protection Trends,2019,39(5):406-416.
11. Wenbang Niu, Yi Feng, Shicun Xu, et al. Revealing suicide risk of young adults based on comprehensive measurements using decision tree classification[J]. Computers in Human Behavior, 2024, 158:108272-.
12. Seyed Behnam Jazayeri, Seyed Farzad Maroufi, Shaya Akbarinejad, et al. Development of a regional-based predictive model of incidence of traumatic spinal cord injury using machine learning algorithms[J]. World Neurosurgery: X, 2024, 23:100280-.
13. Zhichun Fang, Jia Cheng, Chao Xu, et al. Comparison of machine learning and statistical approaches to estimate rock tensile strength[J]. Case Studies in Construction Materials, 2024, 20:e02890-.