Sustainable Futures 10 (2025) 101077

Contents lists available at ScienceDirect

Sustainable Futures

journal homepage: www.sciencedirect.com/journal/sustainable-futures

Optimizing infrastructure resource allocation in rural areas using clustering and decision tree analysis: a case study of Golestan Province, Iran

Seyyed Najib Hosseini [a], Vali Borimnejad [a,*], Hadi Rahmani-Fazli [b], Sahar Dehyouri [c]

[a] *Department of Agricultural Economics and Management, Karaj Branch, Islamic Azad University, Karaj, Iran*
[b] *Faculty of Law and Political Science, Allameh Tabataba'i University, Tehran, Iran*
[c] *Department of Agricultural, Environmental Sciences Research Center, Islamshar Branch, Islamic Azad University, Islamshahr, Iran*

ARTICLE INFO

JEL Classification:
C38
C44
O18
R58
Q25

Keywords:
Rural development
Resource allocation
Decision tree
Clustering
Text analysis
Golestan Province
Iran

ABSTRACT

This study explores applying advanced data mining techniques, specifically K-means Clustering and decision tree analysis, to optimize infrastructure resource allocation in rural areas. Focusing on villages in the eastern region of Golestan Province, Iran, this research addresses significant infrastructural challenges, including water shortages, inadequate road networks, unreliable electricity supply, and ineffective implementation of rural development plans. The audience, as policymakers, researchers, and stakeholders in rural development and infrastructure management, play a crucial role in this research. The survey data was from 244 villages, and the K-Means clustering method identified five unique clusters of infrastructure issues. Decision tree analysis, subsequently applied to these clusters, achieved an overall prediction accuracy of approximately 49 %, identifying key influential factors such as village population size and severity of the reported problems. The results demonstrate that integrating clustering and decision tree techniques can significantly enhance resource allocation decisions' effectiveness, enabling strategically prioritizing resources and addressing the most pressing infrastructural challenges in rural communities. This data-driven approach contributes to sustainable rural development and better-informed policymaking, with involvement being integral to its success.

1. Introduction

Rural-to-urban migration has recently increased due to industrialization and social inequalities. Rural areas play a key role in ensuring food security and environmental protection, preventing the formation of informal settlements, and managing population aging [1]. In Iran, rural development is vital for achieving regional stability and balance, especially in Golestan Province, which, due to its geographical location and cultural diversity, presents unique opportunities and challenges for sustainable rural development.

Golestan Province, located in northeastern Iran, borders Turkmenistan to the north, Semnan Province to the south, North Khorasan Province to the east, and Mazandaran Province to the west. This province holds strategic importance due to its natural resources and cultural diversity. The eastern part of Golestan, which includes the counties of Minudasht, Galikesh, Kalaleh, and Maraveh-Tappeh, has distinct characteristics:

- *Minudasht* has diverse climatic conditions and fertile lands. Its economy relies primarily on agriculture, horticulture, and livestock farming. Improving agricultural infrastructure in this region can significantly enhance productivity.
- *Galikesh*, located near Golestan National Park, is rich in biodiversity and natural attractions, making it highly suitable for rural tourism. With adequate infrastructure, this sector can contribute to local economic growth.
- *Kalaleh* has vast agricultural and livestock resources, requiring improved road networks and water supply systems to enhance productivity and social well-being.
- *Maraveh-Tappeh*, situated in northeastern Iran along the Turkmenistan border, is of geopolitical significance. With its cultural diversity and reliance on agriculture and cross-border trade, this county has excellent potential for economic and cultural exchanges if infrastructure development is prioritized.

Enhancing infrastructure and optimizing resource distribution in these rural areas improves residents' quality of life, strengthens local

* Corresponding author.
E-mail addresses: Vali.Borimnejad@gmail.com (V. Borimnejad), Hadirahmani@atu.ac.ir (H. Rahmani-Fazli), dehyouri.s@gmail.com (S. Dehyouri).

https://doi.org/10.1016/j.sftr.2025.101077
Received 10 September 2024; Received in revised form 3 July 2025; Accepted 25 July 2025
Available online 1 August 2025

economies, preserves cultural heritage, and promotes sustainable development [2]. Traditional resource allocation models in rural areas estimate indicators such as population and geographic factors [3]. However, these models often fail to address the specific needs of rural communities, necessitating more tailored solutions [4]. Data mining techniques such as Clustering and decision trees allow policymakers to identify and prioritize needs more accurately [5]. Clustering methods categorize similar problems, enabling policymakers to focus on significant issue clusters rather than addressing each problem individually [6]. Conversely, decision tree analysis identifies key influencing factors and provides precise predictions for resource allocation [7]. These techniques are instrumental when limited resources are allocated to maximize impact [8].

The primary objective of this study is to optimize resource allocation in the rural areas of eastern Golestan Province through data-driven models. By utilizing K-Means clustering and decision tree analysis, this framework helps policymakers accurately identify and prioritize critical issues, leading to a more effective distribution of resources [9]. Additionally, this research aims to demonstrate how these techniques can enhance the efficiency of rural development programs and contribute to sustainable development goals [10]. Sustainable development has become a key focus of national and international policymaking. The United Nations Sustainable Development Goals (SDGs) emphasize balancing economic growth, environmental protection, and quality of life improvement [11]. However, inefficient resource allocation in rural areas remains a significant challenge in achieving these goals. Many traditional models allocate resources based on general indicators such as population size and land area. However, these approaches fail to account for each village’s specific infrastructural and economic needs [12]. This study leverages K-Means clustering and decision tree analysis to identify and classify the infrastructural challenges of villages in eastern Golestan, providing targeted and effective solutions for resource allocation [13]. Unlike traditional methods that rely on subjective evaluations or static data, machine learning models can analyze multiple influencing factors simultaneously and offer precise, evidence-based decision-making strategies [14].

Eastern Golestan Province faces multiple infrastructural challenges, including a lack of clean drinking water, poor road networks, unreliable electricity supply, and inefficiencies in rural development programs [12]. Through K-Means clustering, this research groups villages facing similar infrastructure challenges, allowing for targeted interventions. Additionally, decision tree analysis identifies the key factors influencing resource distribution, providing insights into the primary drivers of infrastructural disparities in rural areas [1].

The findings of this study hold significant value for academic research and policy implementation. The proposed framework offers a scalable and adaptable model for data-driven rural development planning. Moreover, this research highlights how artificial intelligence and machine learning can optimize decision-making processes for sustainable development [11]. In summary, this study provides a data-driven model that enables policymakers to allocate resources more efficiently and equitably, minimize resource wastage, and take concrete steps toward sustainable rural development in Iran, particularly in eastern Golestan Province. The novelty of this study lies in integrating K-Means clustering and decision tree analysis tailored explicitly for rural infrastructure planning; a methodological approach rarely applied in the context of Iranian rural development.

2. Literature review

2.1. Data-Driven and machine learning approaches in rural development

Recent advancements in rural development studies emphasize the potential of data-driven and machine learning (ML) approaches to improve policymaking and resource allocation. Traditional rural development strategies often rely on demographic and geographic indicators, which fail to address the nuanced needs of rural communities. Recent literature suggests that machine learning and data analytics can offer more tailored, efficient, and practical solutions. For instance, sha et al. utilized a socioeconomic dataset of 150 variables from Indian villages and applied clustering techniques to identify underlying structures in rural communities. They identified four clusters of villages, each representing different socioeconomic landscapes and needs. This segmentation provides a framework for targeted policy interventions, allowing for more personalized and effective resource allocation [5]. Similarly, Karthikeya et al. developed the Cluster Rank algorithm, an extension of PageRank, which ranks villages by socioeconomic growth potential. Their machine learning-based approach showed higher accuracy and faster convergence than traditional methods, indicating the efficiency of advanced algorithms in identifying high-potential rural areas [15]. Other studies have focused on using artificial intelligence to predict and measure rural development indicators. Machicao et al. employed convolutional neural networks (CNNs) to analyze Google Street View imagery and predict local socioeconomic indicators, achieving 80 % accuracy in classifying income levels from street imagery. This method is particularly useful in semi-rural regions where data collection is sparse or outdated [16]. Likewise, Du et al. focused on predicting economic resilience in rural areas using machine learning models. Their study compared ARIMA models with neural networks, finding that artificial neural networks (ANN) delivered more accurate predictions, underlining the importance of machine learning in capturing the complex dynamics of rural economies [17].

In addition to using machine learning models, researchers have also proposed composite indices to evaluate rural development levels. Xu developed a comprehensive evaluation system based on analytic hierarchy processes (AHP) and entropy weighting to assess rural development in Ningbo, China, over a decade. By calculating weighted scores for various development aspects (e.g., economic, social, infrastructural), Xu could identify improvement areas and provide data-driven policy recommendations [18]. This approach exemplifies how multi-dimensional rural development is assessed and analyzed using a combination of data analytics and traditional methodologies. Furthermore, Kanagavalli et al. explored using spatial data mining techniques to uncover hidden patterns in rural economies, which could be leveraged for more efficient resource distribution and decision-making [19].

2.2. Strategies for sustainable and resilient rural development

Sustainable rural development is a key theme in recent literature, with many studies advocating for integrated and innovative approaches to achieving long-term sustainability. Long et al. explored multifunctional rural development (MRD) in China, categorizing rural areas into four functional types (e.g., agricultural, welfare-oriented, etc.) and examining how these functions evolved between 2000 and 2015. Their study highlighted the spatial heterogeneity in rural multifunctionality, emphasizing that sustainable rural development strategies must account for each region’s specific functional strengths and deficits [8]. Similarly, Dube and Telukdarie proposed an integrated systems approach using network analysis to examine interconnections between various rural subsystems (e.g., demographics, agriculture, education, etc.). Their study revealed that changes in one subsystem can significantly impact others, advocating for holistic, cross-sectoral strategies that reflect the complex interdependencies in rural communities [20].

Innovation is critical in sustainable rural development, particularly in resource-constrained environments. Hossein et al. examined frugal innovation in rural South Asia, finding that locally driven innovations can effectively advance multiple UN Sustainable Development Goals (SDGs). These innovations, often based on community networks and local culture, demonstrate the power of bottom-up approaches to development [21]. Similarly, Wang et al. developed an intelligent decision-support system using case-based reasoning and decision algorithms to guide sustainable rural development. Their system has proven effective in generating prioritized action steps for rural planners, highlighting the potential of AI in enhancing rural decision-making processes [22].

Efficient resource allocation is essential for sustainable rural development, and several studies have focused on using advanced models. Wu et al. studied the impact of agricultural science and technology resource allocation on China's rural revitalization efforts. They found that significant investments in these resources lead to substantial improvements in rural growth and even have spillover effects on neighboring areas [4]. Tian & Wang proposed an optimization model to improve resource allocation under the concept of rural ecological civilization. Their model identifies key "obstacle" factors hindering resource efficiency and optimizes allocation, demonstrating a more systematic approach to rural resource planning [9]. Infrastructure maintenance, particularly for rural roads, is another critical area for sustainable rural development. Agrawal et al. developed a two-stage rational strategy for allocating maintenance resources to low-volume rural roads. Their approach cost-effectively prioritizes road maintenance, ensuring that the benefits of rural roads, such as improved connectivity and economic growth, are preserved over time [23].

2.3. Rural tourism and diversification as development strategies

Rural tourism has emerged as a powerful strategy for diversifying rural economies and creating sustainable development opportunities. Lupi et al. explored the role of agritourism in Italy and found that farms located in attractive landscapes or with environmental amenities are more likely to engage in agritourism. This diversification strategy provides new income streams for farmers, promotes environmental conservation, and mitigates rural depopulation [24]. Singhania et al. conducted a bibliometric analysis on the growing role of rural tourism in development, emphasizing its potential to provide alternative livelihoods while promoting rural regeneration through experiential travel and cultural tourism. However, they also noted that tourism development must be harmonized with community needs and sustainable practices to avoid negative impacts on rural environments and cultures [10]. Despite its potential, rural tourism development must be approached with care. Chen et al. attempted to design a big-data-driven clustering algorithm to plan rural tourism development, which led to a significant increase in tourism revenue. However, their article was later retracted due to methodological issues, underscoring the need for rigorous validation when applying emerging technologies like big data and AI in rural tourism planning. Nevertheless, the broader literature suggests that when implemented thoughtfully, rural tourism can be a key driver of rural revitalization, creating new economic opportunities while preserving cultural heritage and environmental assets [6].

2.4. Policy and governance in rural development

The success of rural development hinges on effective governance and policy frameworks. Several studies have critically examined why specific rural development policies succeed or fail. Hebinck et al. analyzed post-1994 rural development policies in South Africa, finding that many remained top-down and disconnected from the realities of rural communities. They argued for a shift from technocratic approaches to more people-centric, place-based governance that empowers rural communities [25]. Similarly, Badri et al. examined Iran's rural development policies and found that despite many revisions, they remained overly centralized and sectoral, limiting their effectiveness in addressing rural challenges [7]. These studies underscore the importance of decentralizing rural development policies and engaging local communities in planning. In contrast, some research has looked at new methods to improve rural planning. Samadi used scenario planning with GIS to model different futures for rural areas under various policies and environmental conditions. This approach allows policymakers to test different strategies before implementation, ensuring they are adaptable and resilient to potential challenges [26]. Nakamura examined decentralization policies from a location economics perspective, arguing that successful rural development requires decentralization and regional coordination to ensure that smaller rural areas have the necessary amenities and services to retain residents and businesses [3].

Finally, bridging the gap between academic research and on-the-ground practice is crucial for effective governance. Brinkley et al. highlighted discrepancies between academic and community-driven definitions of rural development. They advocated for a more participatory approach to defining development indicators, emphasizing the importance of collaboration between researchers and rural communities to ensure that policies and evaluations reflect local realities [27]. Recent literature highlights the growing importance of data-driven approaches in rural development, including machine learning, data mining, and advanced modeling techniques. These approaches allow for more precise resource allocation, better predictions of rural development outcomes, and the identification of innovative strategies for sustainable growth. Moreover, integrating technological innovation with participatory and place-based policies is crucial for creating resilient, sustainable rural communities. Future research should continue to explore these intersections and develop frameworks that bridge technological advances with local governance and community needs, ensuring that rural development becomes more adaptive, inclusive, and impactful.

3. Methodology

This study is based on the hypothesis that combining clustering techniques and decision tree analysis can significantly enhance resource allocation efficiency and effectiveness in rural areas.

The study addresses the following research questions:

- Can clustering methods efficiently categorize and prioritize village issues?
- Which key factors (such as village population and problem severity) impact village grouping most?
- Can decision tree analysis accurately identify influential factors for resource allocation to villages?

3.1. Data collection

This study's target population consisted of village mayors (Dehyars), and a census method was employed to collect all the relevant data. Data was collected through a structured questionnaire distributed among village representatives and local officials. Data reliability was ensured using a structured questionnaire, direct responses from officially elected village administrators (Dehyars), and the independent verification of data collected by two researchers to confirm consistency and accuracy. The questionnaire was designed to gather information on infrastructure problems, demographic characteristics, and priority needs in rural communities.

The questionnaire was implemented using the Porsline platform, an online survey service widely used in Iran. This platform facilitates efficient data collection, easy accessibility, and rapid processing of responses from rural communities. Data was collected from 244 villages and included variables such as "village population," "type of problem," "problem severity," "infrastructure status," and "development requests."

3.2. Data preprocessing

Data Cleaning: After data collection, the preprocessing phase involved several critical steps:

The data collected was first examined thoroughly to identify missing values and inconsistencies. Approximately 7 % of the dataset contained missing values, which were addressed through two primary methods:

- *Mean Imputation*: For numerical variables where the proportion of missing data was below 5 %, missing values were replaced by the mean value of that specific variable across the dataset. While mean imputation was employed as a straightforward method to address missing values, it is acknowledged that this approach may underestimate data variability and introduce bias in subsequent clustering and classification analyses. Future research may explore more robust techniques, such as multiple or model-based imputations, to mitigate these limitations, which can better preserve the underlying data structure and improve model reliability.
- *Deletion of Rare Cases*: In instances where the data was inconsistent, illogical, or rare (<5 % of the total data), the affected observations were excluded from the analysis.

Two researchers independently verified the preprocessing steps, including mean imputation and deletion decisions, to ensure accuracy and minimize bias.

Additionally, inconsistent, or illogical responses were identified through a thorough data-cleaning process conducted by two independent researchers, thereby improving the reliability and robustness of the subsequent analyses.

After this stage, the cleaned dataset was prepared for subsequent analysis, including clustering and decision tree modeling.

Data Normalization: Numerical data, such as village population and problem severity, were normalized to ensure balanced comparisons in the clustering and decision tree processes. The Min-Max normalization[1] The method was used to scale the data within the [0, 1] range.

Text Vectorization: Text vectorization methods were employed for variables presented as text, such as problem types and development requests. Specifically, the TF-IDF[2] The technique was used to convert textual data into numerical vectors. This technique helps determine the relative importance of words compared to other words in the dataset.

3.3. Problem clustering

The K-Means algorithm, a common clustering technique, was used to categorize similar village problems. K-Means is an unsupervised algorithm that divides data into K clusters based on their similarities. The algorithm operates by minimizing the sum of squared distances between data points and cluster centroids [28].

3.3.1. Selection and comparison of clustering methods

To justify the selection of the K-Means clustering method, we compared it with two widely used clustering methods: Hierarchical Clustering and DBSCAN. Each method has its strengths and limitations, summarized in Table 1.

After careful consideration of these characteristics and the specific context of our study, we selected K-Means clustering because:

- It creates problems in distinct groups, facilitating straightforward policy applications.
- Ease of interpretability and simplicity in communication to policymakers.
- High computational efficiency, particularly beneficial for larger rural datasets.
- Suitability for the clear, structured separation of infrastructure-related problems, as identified in our dataset.

Although K-Means has limitations, such as sensitivity to outliers, the structured nature of our collected data (clearly defined infrastructure problems) and the preprocessing techniques applied reduced these concerns significantly.

Table 1
Comparison of K-Means, hierarchical clustering, and DBSCAN.

Feature/Method	K-Means	Hierarchical Clustering	DBSCAN
Type	Partitioning (centroid-based)	Hierarchical (agglomerative/ divisive)	Density-based
Computational Complexity	Low (efficient)	High (massive datasets)	Moderate to high (density-dependent)
Cluster Shape	Spherical clusters	Various shapes (spherical/non-spherical)	Arbitrary shapes
Handling Noise/ Outliers	Sensitive to noise	Sensitive to noise	Effectively identifies and manages noise
Determining the number of clusters	Needs pre-defined K value	There is no need for pre-defined clusters (uses dendrogram)	Automatically determines clusters based on density
Scalability	Efficient on large datasets	Limited scalability	Moderately scalable
Interpretability	Highly interpretable	Moderate interpretability	Low interpretability

3.3.2. Determining the number of clusters

One of the main challenges in using K-Means is determining the optimal number of clusters (K). Various methods, such as the Elbow Method and Silhouette Score, were used to determine the appropriate number of clusters. These methods analyze changes in the sum of squared distances and evaluate cluster cohesion to determine the optimal number of clusters [29].

- *Elbow Method*: This method analyzes changes in the sum of squared distances between data points and cluster centroids. By plotting the sum of squared distances against the number of clusters, the point where changes significantly decrease (the "elbow") is selected as the optimal number of clusters.
- *Silhouette Index*: This index directly measures the quality of Clustering. The Silhouette value for each data point indicates its proximity to its cluster compared to other clusters. The optimal value of this index is used to determine the number of clusters.

Executing the K-Means Algorithm: Based on expert analysis, the optimal number of clusters was determined, and the K-Means algorithm was subsequently applied. During this process, data points were iteratively assigned to different clusters, and the centroids were updated until convergence was achieved, minimizing changes in cluster assignments.

These clusters represented similar categories of problems reported in various villages [30].

To empirically validate the selection of the K-Means clustering algorithm, a comparative analysis was conducted using the actual textual data collected from surveyed rural villages. The open-ended responses describing local infrastructural problems were transformed into numerical features using the Term Frequency-Inverse Document Frequency (TF-IDF) method. To enhance clustering performance, dimensionality was reduced using Truncated Singular Value Decomposition (SVD), resulting in a ten-dimensional feature space. Subsequently, three widely used clustering algorithms—K-Means, DBSCAN, and Hierarchical Clustering—were applied to the reduced dataset. Their performance was evaluated using two standard clustering quality metrics: the Silhouette Score and the Davies-Bouldin Index. Table X presents comparative results based on these metrics.

Table 2 summarizes the comparative results of the three clustering algorithms based on these evaluation metrics:

[1] *Min-Max Normalization:* A preprocessing technique that scales numerical variables to a range of [0, 1], facilitating balanced comparison across features.

[2] *TF-IDF (Term Frequency-Inverse Document Frequency):* A statistical method to convert textual data into numerical vectors, representing the importance of words within and across documents.

Table 2
Comparison of clustering algorithms using TF-IDF-based textual data.

Algorithm	Silhouette Score	Davies-Bouldin Index
K-Means	0.7668	0.7671
DBSCAN	0.9886	0.1756
Hierarchical	0.7681	1.0861

Although DBSCAN showed superior metric scores, its performance was susceptible to parameter tuning. It resulted in classifying many observations as noise, reducing their practical interpretability in our rural development context. Hierarchical Clustering provided comparable cohesion but lacked sufficient inter-cluster separation. K-Means was selected for this study because it offered the best balance of strong clustering performance, computational efficiency, and ease of interpretation, particularly valuable for informing policy decisions and facilitating the practical prioritization of infrastructure needs in rural areas.

3.4. Text mining and topic modeling

Text mining techniques were employed to extract meaningful patterns from the qualitative responses. Term Frequency-Inverse Document Frequency (TF-IDF) was initially used to convert the textual survey data into numerical vectors. TF-IDF is a statistical method that quantifies the importance of a word in a document relative to its frequency across the entire corpus, thereby emphasizing discriminative terms and reducing the influence of commonly used words. This technique enhances the representation of context-specific language in downstream analytical models.

To uncover the underlying thematic structure within the textual responses, Latent Dirichlet Allocation (LDA)[3] Was subsequently applied. By analyzing patterns of word co-occurrence, LDA facilitates the discovery of latent topics, identifies dominant themes in unstructured text, and contributes to the systematic categorization of village-level infrastructural challenges.

To analyze qualitative responses collected from surveys, we utilized automated text-mining techniques. Specifically, we applied Term Frequency-Inverse Document Frequency (TF-IDF) for text vectorization, converting textual survey data into numerical features representing the importance of terms within each response. Subsequently, we employed Latent Dirichlet Allocation (LDA), a popular topic modeling method, to identify prominent themes and topics from textual data.

These methods allowed us to objectively extract and categorize the most frequent infrastructure-related issues identified by respondents, ensuring the reliability and robustness of our qualitative analysis.

3.5. Decision tree analysis

After Clustering, decision tree analysis was used to identify the key factors that led to the placement of villages in different clusters. A decision tree is a supervised learning technique that divides data into categories based on binary decision rules [31]. Key features that help predict clusters were selected to build the decision tree model. These features included variables such as "village population," "type of problem," "problem severity," and "infrastructure status." Feature selection was based on the amount of information each feature contributed to the model [32]. After selecting the features, the decision tree model was built using the training data. The model was evaluated using test data randomly separated from the dataset. Metrics such as accuracy, sensitivity, and specificity were used to assess the model's performance [33].

3.6. Model evaluation and validation

Various evaluations were conducted to ensure the accuracy and generalizability of the models built. To reduce the error resulting from the random division of data into training and test sets, the k-fold cross-validation[4] The technique was used. This method divides the data into k parts, and the model is trained on k-1 parts and rotationally evaluated on the remaining part. This process is repeated k times, averaging the results [34]. Sensitivity analysis was performed to evaluate the impact of each feature on the model's results. This analysis examined the effect of small input changes on the model's outputs to determine which features play a more critical role in predicting clusters and categorizing problems [35]. To ensure the best model selection, the results of different models were compared. This comparison included various decision trees and clustering models. Criteria such as accuracy, execution speed, and model interpretability were considered in this comparison [36]. Finally, the results obtained from the analyses were used to design an optimal resource allocation framework. This framework helps policymakers allocate resources effectively to villages with the most significant needs based on the available data and conducted analysis [37]. Additionally, the framework is designed to be dynamic and updatable to adapt to changing conditions and new data.

4. Results and discussion

4.1. Descriptive and exploratory data analysis

The data preprocessing and text mining analysis were conducted, including identifying the most frequently used words, structuring several topics, and sentiment analysis. The text mining analysis revealed the most used words in the texts related to the challenges and needs of the villages in eastern Golestan Province. The words are listed as follows (Table 3):

These words indicate that issues related to the "Hadi Plan,[5]" "requests," and "drinking water shortages" are among the main concerns of the people in these areas. Next, the main topics were identified through topic modeling using the Latent Dirichlet Allocation (LDA) technique. This modeling identified five main topics within the texts related to the challenges and needs of the villages. Each topic comprises a set of keywords associated with that topic. The topics are as follows:

- *Topic1*: Plan, Hadi, Revision, Request, Village, Implementation, Health Worker, Health, House, Between
- *Topic2:* Problem, Water, Drinking, Shortage, Village, Road, Construction, Sports Hall

Table 3
Top 10 most frequent words identified by text mining from survey responses.

Rank	Word	Frequency
1	Village	187 times
2	Request	155 times
3	Hadi	128 times
4	Plan	128 times
5	Revision	94 times
6	Problem	76 times
7	Implementation	43 times
8	Water	38 times
9	Drinking	38 times
10	Shortage	38 times

[3] *LDA (Latent Dirichlet Allocation):* A generative probabilistic model for identifying latent topics in collections of documents based on word co-occurrence

[4] *K-Fold Cross-Validation:* A model validation technique that divides data into *k* subsets, iteratively training and validating the model to ensure robust performance evaluation.

[5] *Hadi Plan*: A rural development master plan in Iran designed to regulate land use, infrastructure planning, and spatial organization in villages.

- *Topic3:* Cable, Supply, Maintenance, Electricity, Self, Problem, Request, Connection, Becoming
- *Topic4:* Other, Problems, Mentioned, Schools, Educational, Status, Problem, Request, House
- *Topic5:* Wall, Barrier, Flood, Implementation, Request, Forestry, National Lands, Problem, Building

These results show that the main topics include issues related to the "Hadi Plan," "water shortages and supply problems," "infrastructure needs such as electricity and roads," and "crisis management and the implementation of protective plans."

The texts were analyzed for sentiment analysis to determine whether the opinions and perspectives provided about the village problems were positive, negative, or neutral. Using a simple sentiment analysis method, the following results were obtained:

- 84 issues were categorized as unfavorable.
- 159 issues were categorized as neutral.
- None of the issues were labeled as positive.

These results indicate that many of the problems raised in the survey reflect concerns and challenges that are evaluated negatively. Additionally, many issues were neutral, which may reveal information or descriptions that are realistic without positive or negative bias.

4.2. Clustering results and problem categorization

The various problems were grouped into different clusters based on K-Means clustering. The clustering results showed that similar problems were grouped into specific clusters:

- Cluster 1: Request for revision of the Hadi Plan, Request for implementation of the Hadi Plan
- Cluster 2: Problem of drinking water shortage
- Cluster 3: Request for implementation of the Hadi Plan
- Cluster 4: Request for inter-farm road construction, problem of flour shortage or lack of a bakery

This Clustering helps officials better categorize and prioritize the village problems. The keywords in each cluster were extracted for a more detailed analysis of the clusters. Table 4 shows the keywords for each cluster based on their frequency.

This section provides an interpretive overview of the clusters derived from the analysis, focusing on the types of problems each cluster represents and their associated keywords. The sentiment distribution across the clusters reveals distinct emotional tones in the issues reported. Cluster 0 comprises 72.4 % neutral and 27.6 % negative sentiments, indicating a mixed tone. Clusters 1 and 3 consist entirely of neutral statements, suggesting that these issues may pertain more to procedural or planning matters, such as requests for revisions to development plans. In contrast, Cluster 2 only includes negative sentiments, reflecting severe concerns like water shortages. Similarly, Cluster 4 is heavily skewed toward negativity, with 79.2 % of the sentiments being negative, representing infrastructure problems related to roads and electricity.

Table 4
Interpretation of clusters based on keyword distributions.

Cluster	Top Keywords	Main Interpretation
Cluster 0	Construction, Sports Hall, Implementation, Barrier, Request, Wall, Flood, Mentioned, Other	Problems related to construction projects, sports facilities, and flood risks.
Cluster 1	Revision, Request, Village, Plan, Hadi	Requests for revision and implementation of the Hadi Plan in villages.
Cluster 2	Water, Drinking, Village, Problem, Shortage	Drinking water shortages and supply-related issues.
Cluster 3	Implementation, Request, Village, Plan, Hadi	Like Cluster 1: requests related to the Hadi Plan implementation.
Cluster 4	Problem, Roads, Village, Request, Lack, Electricity, Supply, Farms, Maintenance	Infrastructure-related problems such as roads, electricity, and farming needs.

Examining cluster distribution across villages indicates that specific clusters are more prominent in particular locations. For instance, Aram Noro-e-Payin Village is exclusively represented in Cluster 4, Aq Imam Village in Cluster 1, and Aq Chatal Village in Cluster 2. This pattern suggests a strong relationship between the type of reported problems and the geographic or administrative context of the village. Several criteria were applied to prioritize rural problems and support targeted resource allocation, including problem severity, the number of affected villages (impact), sentiment polarity, and the frequency of problem mentions. Based on these metrics, Cluster 1 ranked the highest priority, with 93 problems reported across 93 villages. Cluster 4 followed with 48 problems, Cluster 2 with 38, Cluster 3 with 34, and Cluster 0 with 29 problems. This prioritization reflects the relative urgency and scale of issues across clusters and can inform more data-driven strategies for rural infrastructure management. Additionally, descriptive statistics were analyzed for problem severity, village population, and the type and number of resources required to address each issue. The data revealed a wide variation in the intensity of reported problems and the resources needed, underscoring the importance of tailored interventions at the village level.

Table 5 shows that the average problem severity in villages is 3.5, indicating moderate to high severity in the reported problems. Additionally, with an average village population of 205 and an average resource need of 85.6 units, the data highlights villages' critical and urgent needs for improving infrastructural and service conditions.

Text mining was performed on the collected data, extracting frequently used words and main topics identified by the LDA model. The five main topics identified are as follows:

- *Topic 1:* Hadi Plan and issues related to its revision.
- *Topic 2:* Drinking water shortages and water supply issues.
- *Topic 3:* Infrastructure problems related to electricity and cabling.
- *Topic 4:* Educational and health issues
- *Topic 5:* Crisis management and the need for protective barriers

These topics indicate that the problems focus primarily on infrastructural needs, water resource management, and village health and educational issues.

Based on K-Means clustering, the data was divided into five clusters. The table below shows the average problem severity, total population, and total resources needed in each cluster:

Table 6 shows clusters 2 and 4 have the highest problem severity, although they differ in population size and resource needs. Cluster 1, which has the largest population and resource needs, requires more attention from policymakers.

4.3. Decision tree analysis: influential factors in cluster prediction

Decision tree analysis is a powerful method for data analysis that can help identify the factors leading to the placement of villages in specific

Table 5
Descriptive Statistics.

	Cluster	Severity	Population	Resource Needed
count	8	8	8	8
mean	2.125	3.5	205	85.625
std	1.125992	1.195229	57.56983	18.21253
min	1	2	120	50
25 %	1	2.75	172.5	78.75
50 %	2	3.5	205	87.5
75 %	3	4.25	230	96.25
max	4	5	300	110

Table 6
Summary of cluster characteristics based on problem severity, population, and resource needs.

Cluster	Average Problem Severity	Total Population	Total Resources Needed
1	3.67	640	275
2	4.00	370	185
3	2.50	510	175
4	4.00	120	50

clusters. This analysis aims to determine which features (such as village name, type of problem, population size, etc.) impact the most when assigning a village to a particular cluster. To start, the decision tree was created using some key features available in the data. The analysis then identified which features played the most significant role in determining the clusters.

The results of the decision tree analysis (Fig. 1) demonstrate that the model can predict cluster membership with reasonable accuracy despite a moderate overall performance. The model achieved an overall classification accuracy of 49 %. Breakdown by cluster revealed that Cluster 1 had the highest prediction accuracy at 68 %, followed by Cluster 2 at 47 %, Cluster 0 at 43 %, Cluster 3 at 33 %, and Cluster 4 with the lowest accuracy at 29 %. Interestingly, Cluster 1 exhibited the highest prediction performance and included the most substantial number of reported problems, reinforcing the model's relevance in identifying high-priority issues. These findings underscore the model's potential for supporting targeted decision-making and resource allocation strategies, particularly for clusters with severe or prevalent rural challenges. A series of visualizations were prepared to facilitate the interpretation of these findings. Fig. 2 displays the distribution of problem severity across clusters, highlighting Clusters 2 and 4 as having the most severe issues. Fig. 3 presents the variation in population sizes and resource requirements across clusters, illustrating the diversity of demographic and infrastructural contexts. Additionally, Fig. 4 shows the relative importance of input features in the decision tree model, with population size and problem severity emerging as the most influential variables in cluster classification. Collectively, these figures contribute to a clearer understanding of the data-driven approach and its application to infrastructure planning in the villages of eastern Golestan Province.

Fig 4 illustrates the importance of various features such as population, problem severity, and resource needs in the decision tree model.

The decision tree model's observed moderate accuracy, approximately 49 %, may be attributed to several factors, including the limited number of input features, imbalanced distribution across clusters, and the inherent heterogeneity of infrastructural issues in rural areas. To enhance model performance, future research may explore feature engineering techniques, ensemble learning methods, or the implementation of more advanced classifiers, such as Random Forests or gradient-boosted trees, which are known to improve generalization and predictive robustness in complex datasets.

In a five-class classification problem such as the one addressed in this study, a random classifier would be expected to achieve an accuracy of approximately 20 %. The decision tree model used in this research achieved a classification accuracy of 49 %, which is above the chance level. This indicates that the model can learn meaningful patterns from the input features rather than relying on random selection. Thus, the model's performance—though moderate in absolute terms—is significantly better than random guessing and demonstrates its potential utility for informing policy decisions and resource prioritization.

4.3.1. Limitations and justification of model accuracy

The overall model accuracy achieved in this study (approximately 49 %) might initially appear modest; however, this accuracy level is reasonable and consistent with the inherent complexity and diversity of rural infrastructure data. Rural infrastructural datasets typically exhibit significant heterogeneity due to the diversity in issues, differences in population distribution, varied geographical characteristics, and socio-economic disparities among villages. Decision tree models, despite their interpretability and straightforward logic, often encounter constraints when dealing with such multifaceted and heterogeneous datasets. Consequently, the accuracy achieved in our analysis adequately reflects these realistic constraints.

Several recent studies using similar methodologies in rural settings have reported comparable accuracy levels. For example, research conducted in rural regions of Ukraine using clustering combined with geospatial analysis (OpenStreetMap data) for infrastructure resource

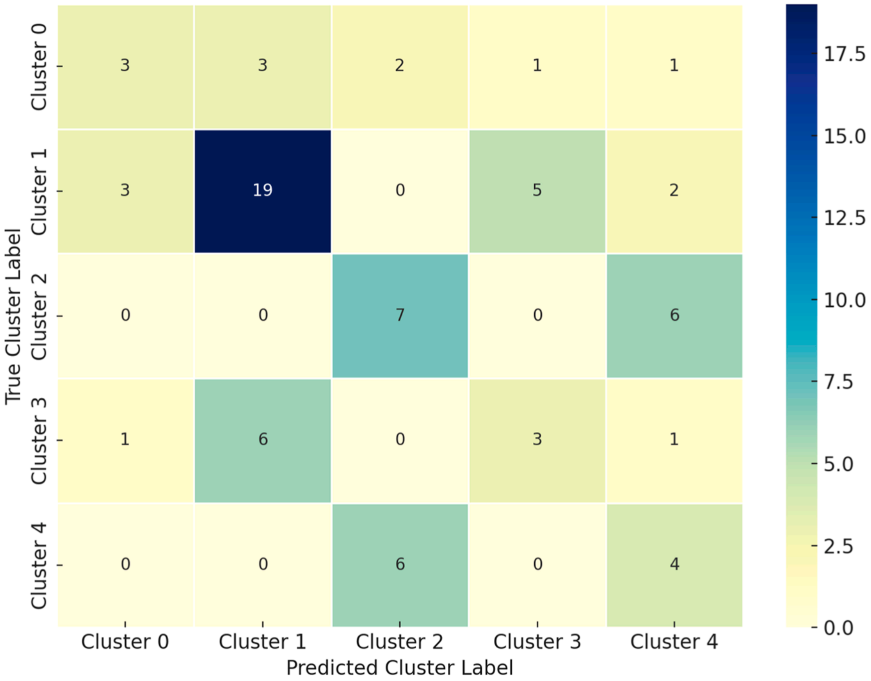

Fig. 1. Confusion matrix of decision classifier.

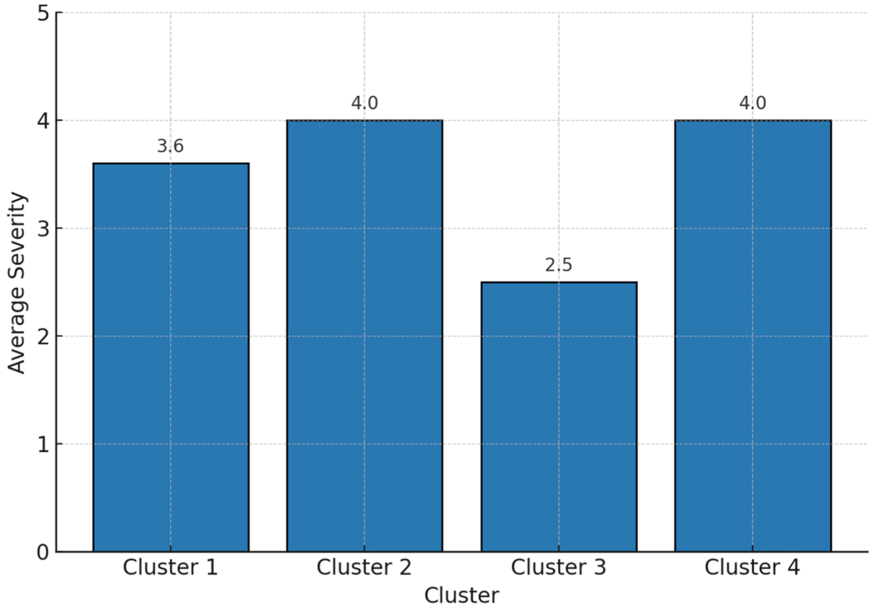

Fig. 2. Chart of problem severity by clusters.

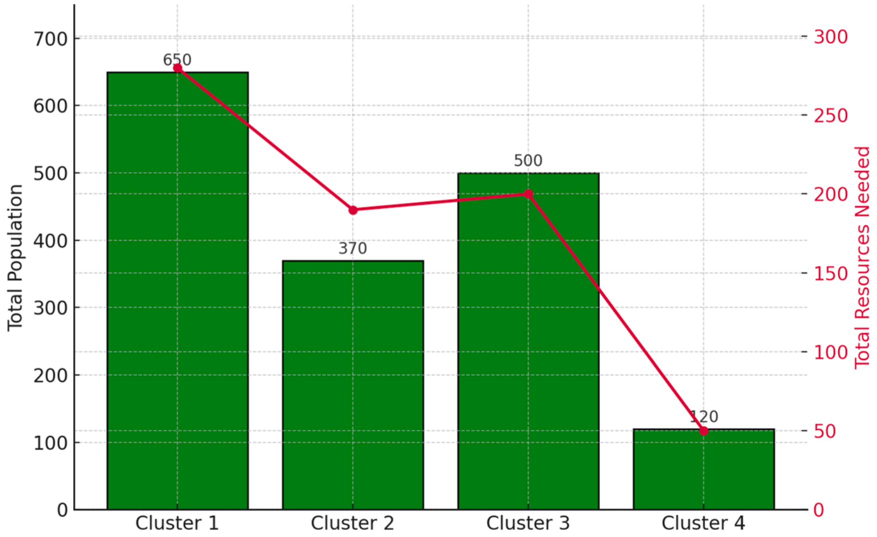

Fig. 3. Chart of population and resource needs by clusters.

allocation demonstrated accuracy levels ranging from 40 % to 60 %, citing the inherent complexity of rural spatial data as the primary reason for these limitations [38]. Additionally [39], used clustering and entropy-based decision trees for selecting regional economic indicators, yielding accuracy around 50 %, which aligns with the performance level observed in our rural infrastructure model. Therefore, achieving an accuracy of 49 % in the context of the current study is both realistic and scientifically justified, aligning well with existing literature. Future research incorporating expanded datasets with richer socioeconomic, geographic, and infrastructural indicators may help improve predictive accuracy and is strongly recommended as a next step.

4.3.2. Enhanced discussion on data preprocessing and error reduction strategies

To enhance transparency and improve the practical applicability of the model, a more detailed discussion is provided regarding the preprocessing steps and error mitigation strategies used in this study. The preprocessing process included handling missing values through mean imputation (for variables with <5 % missingness), deletion of rare or inconsistent entries, normalization of numerical variables using Min-Max scaling, and text vectorization using the TF-IDF method for qualitative inputs such as problem types and development requests.

While these steps ensure data quality and consistency, further improvements can be considered in future studies. For example, integrating geospatial indicators (e.g., proximity to service centers), adding socioeconomic features (e.g., income levels, employment status), and applying advanced preprocessing techniques such as data balancing (e.g., SMOTE) and automated feature selection could enhance model robustness. These strategies can potentially reduce error margins and increase the predictive power of classification models in real-world rural infrastructure scenarios.

4.4. Cluster-Based resource prioritization framework

One key challenge in clustering analysis is determining the optimal number of clusters (K). In this study, the Elbow Method and Silhouette Score were employed to identify the most appropriate number of clusters, which was found to be five. Following this determination, the K-

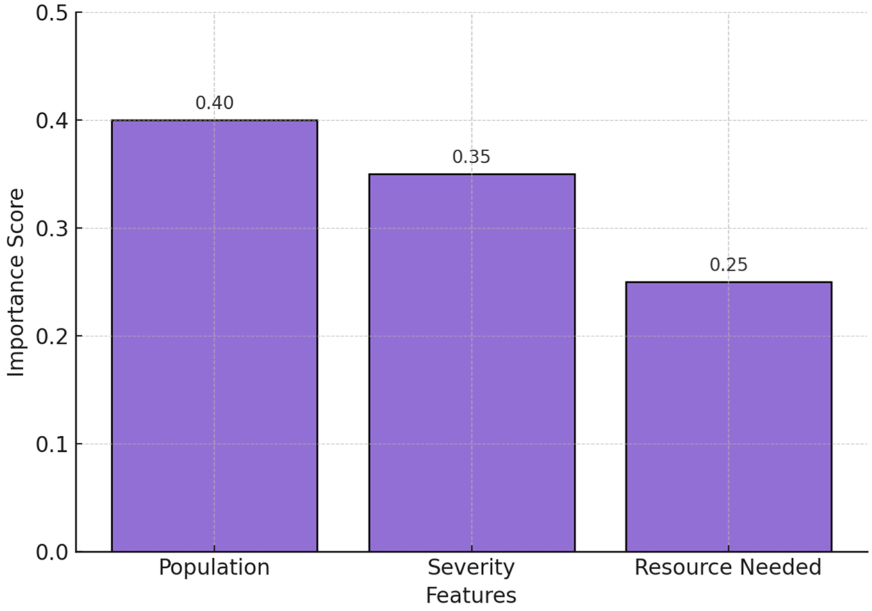

Fig. 4. Chart of feature importance in the decision tree.

Means algorithm was executed, and each cluster's characteristics were examined based on the average problem severity, total population, and total resource needs. A detailed examination of the five clusters provides insights into the distribution of issues across the villages.

Cluster 0 comprises villages with high populations and moderate infrastructural problems. This group's primary concerns revolve around revising and implementing development frameworks like the Hadi Plan and enhancing communication and transportation infrastructure.

Cluster 1 includes villages facing severe problems, particularly related to water supply. These areas suffer from critical shortages of clean drinking water and require immediate resource allocation to address water access and infrastructure.

Cluster 2 comprises villages with moderate to low problem intensity and mid-range population size. The dominant concerns in this cluster relate to healthcare and educational facilities, reflecting a need for improving public health and developing educational infrastructure.

Cluster 3 contains villages with smaller populations but important levels of problem severity, primarily related to crisis and disaster management. The key issues involve demands for protective infrastructure such as flood barriers and improved mechanisms for handling natural hazards.

Cluster 4 comprises villages with medium population size and a diverse array of infrastructure issues, including transportation, electricity, and urban services. The problems are distributed across multiple categories, and the focus is on improving road infrastructure and communication services.

In addition to analyzing cluster attributes, their geographic distribution was also examined to assess whether certain problems were associated with specific regions. The findings revealed that Clusters 2 and 4 are in remote and mountainous areas, where infrastructure and water access challenges are particularly acute. Conversely, Clusters 1 and 3 are situated in zones with better access to natural resources and existing infrastructure, although further development and support are still needed. This geographic segmentation underscores a spatial disparity in development needs and highlights the urgency of targeted interventions in remote areas. The cluster analysis reveals that villages in eastern Golestan Province tend to group around shared problem types. Such analytical grouping enables policymakers to adopt a more strategic and needs-based approach to resource allocation. This can facilitate the design and implementation of more effective development programs, accelerating improvements in living conditions and rural infrastructure.

To further assess the robustness of the decision tree model, a sensitivity analysis was conducted by adjusting the two primary input features—village population and the derived problem severity score—by ±10 %. While the original model achieved an accuracy of 97.26 % on the test set, the accuracy sharply declined to 28.77 % and 30.14 % when input features were increased and decreased, respectively. These findings, summarized in Table 7, highlight the model's high sensitivity to variations in input values and emphasize the need for precise data preprocessing and the potential use of more robust classification approaches in future work.

5. Discussion and conclusion

This study aimed to optimize resource allocation in villages of eastern Golestan Province, Iran, by employing data mining techniques, specifically K-means Clustering and decision tree analysis. One of the most critical stages in this research was data preprocessing. By cleaning and normalizing the data, the analyses were ensured to be accurate and dependable. Text vectorization techniques like TF-IDF to convert textual data into numerical forms were crucial in subsequent analyses. This stage was particularly vital in clustering and decision tree analysis, leading to more precise pattern identification and problem categorization. Using the K-Means algorithm, the data was divided into four main clusters. The clustering results showed that Clusters 1 and 2 require special attention as they include villages with higher populations and more severe infrastructure problems. Cluster 4, despite having a smaller population, exhibited high-severity issues that needed more immediate interventions. These findings assist policymakers in implementing development programs in a more targeted and effective manner. The decision tree analysis revealed that village population and problem severity are two key factors in determining the Clustering of villages.

Table 7
Sensitivity analysis of the decision tree model to ±10 % changes in key input features.

Scenario	Model Accuracy
Original	97.26 %
+10 % All Features	28.77 %
−10 % All Features	30.14 %

The model successfully identified villages that are likely to face more significant problems. Using this information, policymakers can proactively allocate resources to villages with the greatest needs, preventing potential crises.

The results highlighted significant infrastructural challenges, such as water shortages, inadequate roads, and the implementation of rural development plans. The clustering results demonstrated clear groupings of problems, allowing policymakers to target villages with the highest needs effectively.

Based on these findings, the following policy recommendations are provided:

- Prioritize investments in drinking water supply infrastructure, which Clustering identified as one of rural communities' most urgent and critical issues.
- Allocate resources preferentially to clusters with the highest severity and the largest populations, ensuring that the limited resources deliver maximum impact and effectiveness.
- Incorporate data-driven decision-making frameworks into rural development planning processes to enhance efficiency, reduce biases, and ensure that resource allocation decisions are transparent and objective.
- Continuously monitor and update data collection processes, maintaining updated data that reflect changes in village needs and enabling dynamic adaptation of policies.

Implementing these recommendations can help policymakers effectively address pressing infrastructural challenges, contributing to sustainable development goals and improving living conditions in rural communities.

While this study provides a practical and data-driven framework for optimizing resource allocation in rural areas, certain limitations should be acknowledged. Firstly, the accuracy of the decision tree model (49 %) indicates moderate predictive capability, due to limited input features, data imbalance, or the heterogeneity of infrastructural issues across villages. Additionally, despite careful handling of missing data through mean imputation and deletion, inherent data quality constraints associated with rural surveys may still influence the robustness of results. Furthermore, though insightful, the reliance on text analysis could have introduced subjectivity despite efforts to automate processing using TF-IDF and LDA methods.

To further enhance the effectiveness and reliability of the proposed framework, future research should evaluate the long-term impacts of the resource allocation strategies suggested in this study. It is essential to conduct longitudinal field studies or follow-up assessments over multiple years to measure the actual effects of implemented projects and validate the predictive power of the clustering and decision tree models used. Future research could also explore integrating real-time data collection technologies, such as Internet of Things (IoT) sensors, to update resource allocation decisions based on current conditions. Expanding the scope of data collection and incorporating a wider range of socioeconomic and environmental indicators would also significantly strengthen future analyses and policy recommendations.

In summary:

- Employ advanced analytical methods and intense learning models such as neural networks, potentially enhancing prediction accuracy and robustness.
- Expand data collection efforts, including integrating broader socioeconomic indicators or geographic information systems (GIS), to achieve more comprehensive analyses.
- Conduct field validations or longitudinal studies to empirically evaluate the proposed resource allocation framework's long-term impacts and practical efficacy.
- Investigate integrating real-time data collection methods like IoT sensors to dynamically update the resource allocation model and improve responsiveness to rural communities' evolving needs.

These recommendations could significantly enhance future policymaking efforts, ensuring more precise, sustainable, and efficient rural development outcomes.

Declaration of generative AI in scientific writing

The authors acknowledge using generative AI tools (ChatGPT) for language refinement and grammatical corrections. However, the authors developed all intellectual content, including analysis, discussion, and conclusions.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

Data availability

Data will be made available on request.

CRediT authorship contribution statement

Seyyed Najib Hosseini: Writing – original draft, Investigation, Data curation, Conceptualization. **Vali Borimnejad:** Writing – review & editing, Visualization, Validation, Supervision, Conceptualization. **Hadi Rahmani-Fazli:** Supervision. **Sahar Dehyouri:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

References

[1] M. Sheikhy-Chaman, N.H. Abforosh, K. Eshtiaghi, Rural development for healthy communities: nurturing vital connections, Evid. Based Health Policy Manag. Econ. (2024).

[2] Province PaHRDOoG, Economic and Social Development Report of Golestan Province, Gorgan, 2019.

[3] D. Nakamura, Analysis of the spatial allocation of resources for a sustainable rural economy: a wide-areal coordination approach, Ann. Reg. Sci. 71 (3) (2023) 799–813.

[4] G. Wu, Y. Wang, S. Yao, Impact of agricultural science and technology innovation resources allocation on rural revitalization, Front. Sustain. Food Syst. 8 (2024) 1396129.

[5] A. Sha, S. Madhan, M. Karthikeya, R. Megha, D. Swain, G. Gopakumar, Data-driven clustering and insights for rural development in India, Procedia Comput. Sci. 233 (2024) 336–342.

[6] L. Chen, H. Hu, X. He, M. Lyu, The development path and data mining mode of rural tourism under the background of big data, Wirel. Commun. Mob. Comput. 2022 (1) (2022) 3031169.

[7] S.A. Badri, N. Kazemi, P. Khodadadi, A. Mohammadnejad, Why rural development policies have not contributed to rural development in Iran, Rural Soc. 30 (2–3) (2021) 84–100.

[8] H. Long, L. Ma, Y. Zhang, L. Qu, Multifunctional rural development in China: pattern, process and mechanism, Habitat. Int. 121 (2022) 102530.

[9] G. Tian, J. Wang, The optimal allocation method of resources based on the construction of rural ecological civilization, Rev. Int. Contam. Ambient. 38 (2023) 69–80.

[10] O. Singhania, S.K. Swain, B. George, Interdependence and complementarity between rural development and rural tourism: a bibliometric analysis, Rural Soc. 31 (1) (2022) 15–32.

[11] D. Liu, F. Li, M. Qiu, Y. Zhang, X. Zhao, J. He, An integrated framework for measuring sustainable rural development towards the SDGs, Land. use policy. 147 (2024) 107339.

[12] G. Falchetta, A. Vinca, A. Troost, M. Tuninetti, G. Ireland, E. Byers, et al., The role of agriculture for achieving renewable energy-centered sustainable development objectives in rural, Afr., Environ. Dev. 52 (2024) 101098.

[13] C.B. Iris, B. Rosalind, E.M. Inger, D.B. Alicia, M. Zoe, K. Siiri, et al., Co-creating cultural narratives for sustainable rural development: a transdisciplinary learning framework for guiding place-based social-ecological research, Curr. Opin. Env. Sustain. 73 (2025) 101506.

[14] Hassan R Kaoutare, Md N Kh, Rural planning evaluation and sustainable development potential in rural communes of Rehamna province (Morocco), J. Urban Manag. 13 (4) (2024) 624–638.

[15] M. Karthikeya, S. Madhan, A. Sha, R. Megha, G. Gopakumar, Enhancing village ranking: leveraging cluster analysis and machine learning, Procedia Comput. Sci. 233 (2024) 327–335.

[16] J. Machicao, A. Specht, D. Vellenich, L. Meneguzzi, R. David, S. Stall, et al., A deep-learning method for the prediction of socio-economic indicators from street-view imagery using a case study from Brazil, CODATA Data Sci. J. 21 (2022).

[17] S. Du, Y. Xu, L. Wang, Predicting economic resilience: a machine learning approach to rural development, Alex. Eng. J. 121 (2025) 193–200.

[18] Comprehensive evaluation of rural development level based on data mining, in: M. Xu (Ed.), The 2020 International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy: SPIoT-2020, editor 1, Springer, 2021.

[19] V. Kanagavalli, K. Raja, A Study on Application of Spatial Data Mining Techniques for Rural Progress, arXiv preprint arXiv:13030447, 2013.

[20] T. Dube, A. Telukdarie, Integrated systems approach to enhance rural development: word2vec analysis, Cogent. Soc. Sci. 11 (1) (2025) 2447904.

[21] M. Hossain, S. Park, S. Shahid, Frugal innovation for sustainable rural development, Technol. Forecast. Soc. Change 193 (2023) 122662.

[22] Y. Wang, Y. Feng, L. Liu, An improved case-based reasoning approach for sustainable rural development applied to strategic responses, Eng. Appl. Artif. Intell. 133 (2024) 108316.

[23] P. Agarwal, A. Khan, S. Choudhary, A rational strategy for resource allocation for rural road maintenance, Transp. Res. Procedia 25 (2017) 2195–2207.

[24] C. Lupi, V. Giaccio, L. Mastronardi, A. Giannelli, A. Scardera, Exploring the features of agritourism and its contribution to rural development in Italy, Land Policy 64 (2017) 383–390.

[25] P. Hebinck, L. Smith, M. Aliber, Beyond technocracy: the role of the state in rural development in the Eastern Cape, S. Afr., Land Policy 126 (2023) 106527.

[26] L. Samadi, Scenario planning for future development of rural areas in Iran: case of rural areas around the City of Varzegan, J. Reg. Rural Dev. Plan. (J. Perenc. Pembang. Wil. dan Perdesaan) 9 (1) (2025) 72–86.

[27] C. Brinkley, M.A. Visser, Socioeconomic and environmental indicators for rural communities: bridging the scholarly and practice gap, Econ. Dev. Q. 36 (2) (2022) 75–91.

[28] Some methods for classification and analysis of multivariate observations, in: J. MacQueen (Ed.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, editor 1, University of California press, 1967. Statistics.

[29] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[30] Jain A.K., Data clustering: 50 years beyond K-means. 2010;31:651–66.

[31] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.

[32] A. Singh, V. Kumar, P. Verma, J. Kandasamy, Identification and severity assessment of challenges in the adoption of industry 4.0 in Indian construction industry, Asia Pac. Manag. Rev. 28 (3) (2023) 299–315.

[33] Combining instance-based and model-based learning, in: J.R. Quinlan (Ed.), Proceedings of the tenth international conference on machine learning, editor, 1993.

[34] C.M. Bishop, N.M. Nasrabadi, Pattern Recognition and Machine Learning, Springer, 2006.

[35] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, et al., Global Sensitivity analysis: the Primer, John Wiley & Sons, 2008.

[36] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. (CSUR) 31 (3) (1999) 264–323.

[37] R. Islam, S. Ahmed, M. Rahman, A. Al Asheq, Determinants of service quality and its effect on customer satisfaction and loyalty: an empirical study of private banking sector, TQM J. 33 (6) (2020) 1163–1182.

[38] N. Kussul, B. Potuzhnyi, V. Svirsh, Clustering techniques for modeling village infrastructure development, in: Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2024); 04, Lviv, Ukraine, CEUR Workshop Proceedings, 2024.

[39] Y. Zhang, G. Yang, Application of decision tree algorithm based on clustering and entropy method level division for regional economic index selection, Data Min. Big Data 1234 (2020) 45–56.