Optimizing Clustering and Predictive Modelling for 3-D Road Network Analysis Using Explainable AI

Rotsnarani Sethy[1], Soumya Ranjan Mahanta[2], Mrutyunjaya Panda[3]

[1] roshnaranimca@gmail.com; [2]dipusoumyaranjan019@gmail.com; [3]mrutyunjaya74@gmail.com

[1] Research Scholar, [2] M.Tech Student, [3] Associate Professor

[1,2,3] Department of Computer Science, Utkal University, Bhubaneswar, Odisha, India

Abstract

Building an accurate 3-D spatial road network model has become an active area of research now-a-days that profess to be a new paradigm in developing Smart roads and intelligent transportation system (ITS) which will help the public and private road impresario for better road mobility and eco-routing so that better road traffic, less carbon emission and road safety may be ensured. Dealing with such a large scale 3-D road network data poses challenges in getting accurate elevation information of a road network to better estimate the CO_2 emission and accurate routing for the vehicles in Internet of Vehicle (IoV) scenario. Clustering and regression techniques are found suitable in discovering the missing elevation information in 3-D spatial road network dataset for some points in the road network which is envisaged of helping the public a better eco-routing experience. Further, recently Explainable Artificial Intelligence (xAI) draws attention of the researchers to better interprete, transparent and comprehensible, thus enabling to design efficient choice based models choices depending upon users requirements. The 3-D road network dataset, comprising of spatial attributes (longitude, latitude, altitude) of North Jutland, Denmark, collected from publicly available UCI repositories is preprocessed through feature engineering and scaling to ensure optimal accuracy for clustering and regression tasks. K-Means clustering and regression using Support Vector Machine (SVM) with radial basis function (RBF) kernel are employed for 3-D road network analysis. Silhouette scores and number of clusters are chosen for measuring cluster quality whereas error metric such as MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) are considered for evaluating the regression method. To have better interpretability of the Clustering and regression models, SHAP (Shapley Additive Explanations), a powerful xAI technique is employed in this research. From extensive experiments , it is observed that SHAP analysis validated the importance of latitude and altitude in predicting longitude, particularly in the four-cluster setup, providing critical insights into model behavior and feature contributions SHAP analysis validated the importance of latitude and altitude in predicting longitude, particularly in the four-cluster setup, providing critical insights into model behavior and feature contributions with an accuracy of 97.22% and strong performance metrics across all classes having MAE of 0.0346, and MSE of 0.0018. On the other hand, the ten-cluster setup, while faster in SHAP analysis, presented challenges in interpretability due to increased clustering complexity. Hence, K-Means clustering with K=4 and SVM hybrid models demonstrated superior performance and interpretability, highlighting the importance of careful cluster selection to balance model complexity and predictive accuracy.

Manuscript received September 5, 2024
Manuscript revised September 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.9.4

Key words:
K-Means Clustering, Support Vector Machines (SVM), SHAP Analysis, interpretability, Linear Regression

1. Introduction

Clustering and interpretability of complex datasets are critical tasks in contemporary data science, especially when dealing with multidimensional and high-volume data. In particular, the analysis of 3D road network data poses unique challenges due to its spatial and temporal complexity. This paper focuses on addressing these challenges through a comprehensive methodology that integrates advanced clustering techniques, model interpretability methods, and dimensionality reduction strategies. The study aims to enhance the understanding of clustering patterns and the underlying factors influencing these patterns in 3D spatial datasets, with implications for improving decision-making and operational efficiencies in various domains, including urban planning and transportation management. Road network analysis is a vital aspect of geographical and urban studies, where understanding the spatial distribution and patterns of road networks can inform infrastructure development and safety measures. The dataset utilized in this study comprises 3D coordinates (longitude, latitude, altitude) from North Jutland, Denmark. Given the multidimensional nature of this data, the challenge lies in effectively clustering the data to reveal meaningful patterns and in interpreting the results to derive actionable insights. Traditional clustering techniques, while useful, may not fully capture the intricate relationships within such datasets. Thus, there is a need for advanced methodologies that not only perform effective clustering but also provide insights into the factors driving these cluster formations.

To address this, the study employs K-Means clustering, a widely used algorithm known for its simplicity and effectiveness in partitioning datasets into distinct clusters based on feature similarity. The selection of the optimal number of clusters is determined using the Elbow Method, which helps in identifying the point where the addition of more clusters ceases to provide significant improvements in the clustering performance. Further validation of the clustering quality is conducted through the silhouette score, which measures how similar an object is to its own cluster compared to other clusters. This approach ensures that the resulting clusters are both meaningful and well-defined. In conjunction with clustering, the study utilizes Random Forest Classifier to analyze the importance of different features in the clustering process. Random Forest, an ensemble learning method known for its robustness and ability to handle large datasets, provides insights into which features—longitude, latitude, or altitude are most influential in determining cluster memberships. Preliminary results reveal that longitude is the most significant feature, offering valuable information for further analysis and interpretation.

To enhance the interpretability of the regression and classification results from clustering, explainable AI (xAI) using SHAP (SHapley Additive exPlanations) is employed. SHAP is a powerful tool that explains individual predictions made by machine learning models by attributing the output to each feature's contribution. This method provides a transparent view of how different features affect the regression outcomes and the classification performance within clusters, facilitating a deeper understanding of the data's underlying structure. Various SHAP visualizations, including summary plots, force plots, waterfall plots, and bar plots, are generated to illustrate the impact of each feature on the model predictions. These visualizations are crucial for identifying key factors driving the predictive patterns and for validating the robustness of the models. In the context of 3D road network data, SHAP analysis highlighted latitude and altitude as the most influential features in predicting longitude, confirming the models' reliance on spatial features for robust performance. Additionally, dimensionality reduction techniques such as Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) are employed to visualize the clustering results in a reduced feature space. PCA is used to reduce the dimensionality of the data while preserving its variance, making it easier to interpret and visualize. UMAP provides a non-linear reduction technique that captures the global data structure, allowing for a more intuitive understanding of the spatial relationships and clustering configurations.

These techniques collectively help in visualizing and validating the clustering results in a more interpretable format, making it easier to understand the spatial distribution of clusters and their relationships. By integrating clustering, regression analysis, model interpretability with SHAP, and dimensionality reduction, this comprehensive framework provides deep insights into 3D road network data. This methodology not only enhances the understanding of complex datasets but also improves decision-making processes by offering clear insights into the factors influencing clustering patterns. The study's findings have significant implications for urban planning, infrastructure development, and safety assessments, demonstrating the value of advanced analytical techniques in handling and interpreting high-dimensional data. By offering a detailed analysis of clustering, regression, and model interpretability, this research contributes to the ongoing advancements in data science and its applications in real-world scenarios.

The rest of this article is organized as follows: Section II reviews existing methodologies related to clustering and interpretability of complex datasets, with a focus on 3D road network data and its challenges. Section III details the proposed methodology that integrates clustering techniques, model interpretability methods, and dimensionality reduction strategies for enhanced analysis of multidimensional data. Section IV describes the experimental setup used to validate the effectiveness of the proposed approach, including the use of K-Means clustering, Random Forest Classifier, and SHAP for interpretability. Section V presents the experimental results and offers an analysis of the findings, including the impact of various features on clustering and the visualizations provided by SHAP. Finally, Section VI concludes the article and proposes future research directions to further advance the understanding and application of clustering and interpretability in high-dimensional datasets.

2. Related Work

The integration of clustering, interpretability, and dimensionality reduction techniques has made significant strides in addressing the complexities associated with high-dimensional datasets, particularly in the domain of 3D spatial data such as road networks. This research leverages these advancements to provide actionable insights for urban planning and transportation management, enhancing both data analysis and comprehension. A prominent approach for clustering high-dimensional data is K-Means clustering, known for partitioning data based on feature similarity. As outlined by Jain et al. (1999), K-Means is widely applied in spatial data analysis due to its effectiveness in identifying natural groupings within datasets [1]. To optimize the number of clusters, the Elbow Method, introduced by Thorndike (1953), is a standard evaluation technique that identifies the point beyond which additional clusters do not significantly improve clustering quality [2]. Complementarily, the silhouette score, developed by Rousseeuw (1987), validates clustering performance by measuring intra-cluster similarity relative to inter-cluster separation [3]. Random Forests, introduced by Breiman (2001), provide robust ensemble learning capabilities, efficiently handling large datasets and offering insights into feature importance, making them invaluable in spatial data analysis [4]. The technique's application to feature selection and spatial datasets has been further validated by Liaw and Wiener (2002), highlighting its versatility across diverse contexts [5].

SHAP (Shapley Additive explanations), developed by Lundberg and Lee (2017), enhances model interpretability by attributing outputs to individual features, thereby offering a clear view of feature contributions [6]. The application of SHAP to clustering results in spatial data has been explored by Ribeiro et al. (2016), demonstrating its utility in understanding complex models [7]. Dimensionality reduction techniques are essential for visualizing high-dimensional data. Principal Component Analysis (PCA), a foundational method introduced by Hotelling (1933), reduces dimensionality while preserving variance, aiding in the visualization of clustering results [8]. UMAP, a non-linear dimensionality reduction technique developed by McInnes et al. (2018), captures both local and global data structures, making it particularly suitable for complex datasets [9]. These techniques have proven effective in spatial data visualization, as noted in studies by van der Maaten and Hinton (2008) [10]. Recent research has increasingly focused on the integration of clustering, interpretability, and dimensionality reduction to tackle the challenges of analyzing complex datasets. Kotsiantis et al. (2007) underscored the benefits of combining analytical methods to enhance clustering result interpretation [11]. Chien et al. (2021) demonstrated the value of merging interpretability methods with dimensionality reduction for spatial data analysis, showcasing improved insights into clustering outcomes [12]. Jolliffe and Cadima (2016) provided a thorough review of PCA's applications [13], while SHAP has been widely applied across various contexts, including image classification and tabular data analysis (Ribeiro et al., 2016; Chen et al., 2018) [14, 15]. Hennig (2007) and Zhao et al. (2019) explored clustering methods specific to spatial data challenges, such as traffic pattern analysis [16, 17]. Liu et al. (2020) and Zhang et al. (2021) examined the use of PCA and UMAP for visualizing spatial datasets, demonstrating their effectiveness in conjunction with clustering techniques [18, 19].

Furthermore, Ribeiro et al. (2018) and Luo et al. (2022) advanced the field by integrating SHAP with clustering and dimensionality reduction methods, enhancing the interpretability of high-dimensional data [20, 21]. Xie et al. (2023) pushed these boundaries by developing new clustering algorithms combined with interpretability methods for large-scale data analysis [22]. This study builds on existing research by integrating advanced techniques like K-Means clustering, Random Forest feature analysis, SHAP interpretability, and dimensionality reduction methods such as PCA and UMAP to analyze 3D road network data. By applying these techniques, the study offers a comprehensive framework for understanding the complexities of high-dimensional spatial datasets. The research aims to uncover deeper insights into clustering patterns and the factors influencing them, ultimately enhancing decision-making and operational efficiencies in urban planning and transportation management.

3. Proposed Methodology

This research employs a sophisticated and detailed methodology to analyze a comprehensive geospatial dataset, with the goal of advancing the understanding of feature importance and model interpretability in large-scale geospatial analysis. The methodology integrates advanced data processing, machine learning, and interpretability techniques to provide a robust framework for high-impact research.

3.1. Methodology for Analyzing 3D Spatial Network Data Using Linear Regression

This study introduces a comprehensive methodology for analyzing 3D spatial network data that integrates advanced data ingestion, clustering, regression, and model interpretation techniques using Python and PySpark. Figure 1 illustrates the workflow for analyzing 3-D road network dataset for linear regression, detailing the sequential steps involved in preparing the data, applying the regression model, and interpreting the results. The methodology initiates with data ingestion through PySpark, where a Spark session is set up to load and preprocess 3D spatial network data from a text file. This data is then converted from a Spark DataFrame to a Pandas DataFrame to facilitate further manipulation and analysis. In the data preparation phase, columns are cast to appropriate numerical types, and features are extracted to set the stage for in-depth analysis. For clustering, the K-Means algorithm is employed to determine the optimal number of clusters using the Elbow Method, with inertia values plotted for cluster counts ranging from 1 to 14. The optimal number of clusters is identified as 10, and K-Means is executed with this cluster count, followed by evaluation using the silhouette score. Clustering results are visualized through scatter plots that highlight distinct cluster regions. In the regression analysis, a Linear Regression model is trained to predict longitude, with performance evaluated through Mean Absolute Error (MAE) and Mean Squared Error (MSE). SHAP (SHapley Additive exPlanations) is utilized to interpret the regression model. A secondary Linear Regression model is trained to predict cluster labels, with model performance assessed using accuracy and detailed classification reports. SHAP analysis is again applied to provide insights into feature contributions. The methodology also incorporates rigorous resource management, including model cleanup and garbage collection, to optimize memory usage.

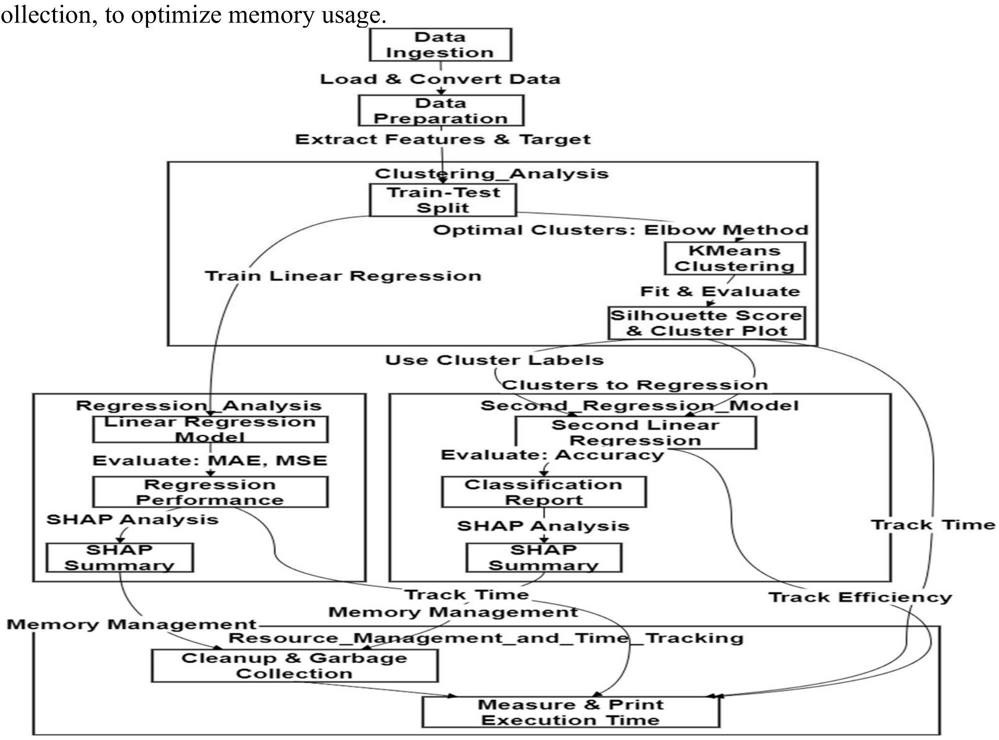

Fig 1. Workflow for Analyzing a Dataset for Linear Regression

3.2. Methodology for Analyzing 3D Spatial Network Data Using SVM

Here, Figure 2 presents a comprehensive methodology for analyzing 3D road network data by integrating advanced machine learning techniques, specifically K-Means clustering and Support Vector Machines (SVM), for both regression and classification tasks i.e. by using Support Vector Regression (SVR). Figure 2 illustrates the workflow for analyzing a dataset using SVM, detailing the steps involved in preparing the data, training the SVM model, and evaluating its performance. The methodology begins with data preparation, where the dataset is loaded and processed using Pandas. This involves parsing the data into columns such as id, longitude, latitude, and altitude, and converting these columns to float types for accurate numerical analysis. Longitude and latitude are selected as features, while longitude is designated as the target variable for regression. The dataset is then divided into training and testing subsets using an 80-20 split ratio to ensure robust and reproducible results. For clustering, the

Elbow Method is applied to determine the optimal number of clusters by plotting inertia values for cluster counts ranging from 1 to 14, identifying four as the optimal number. The K-Means algorithm is employed to perform clustering, with the quality of clusters evaluated using the silhouette score and results visualized in a 2D plot that highlights cluster centers. In the regression analysis, an SVM model with an RBF kernel is trained to predict longitude, and model performance is assessed using Mean Absolute Error (MAE) and Mean Squared Error (MSE). SHAP (SHapley Additive exPlanations) analysis is utilized to interpret feature importance, applying this analysis to a subset of 1,000 samples to enhance computational efficiency. In the classification phase, an SVM classifier is used to predict cluster labels derived from the K-Means clustering phase, with performance metrics including accuracy, precision, recall, and F1-score provided. SHAP analysis is also conducted for the classification model to elucidate the impact of features. The methodology ensures rigorous resource management, including model cleanup and garbage collection, while tracking execution time to maintain operational efficiency. This approach offers a robust framework for detailed data analysis and model interpretability, contributing valuable insights to data-driven decision-making processes.

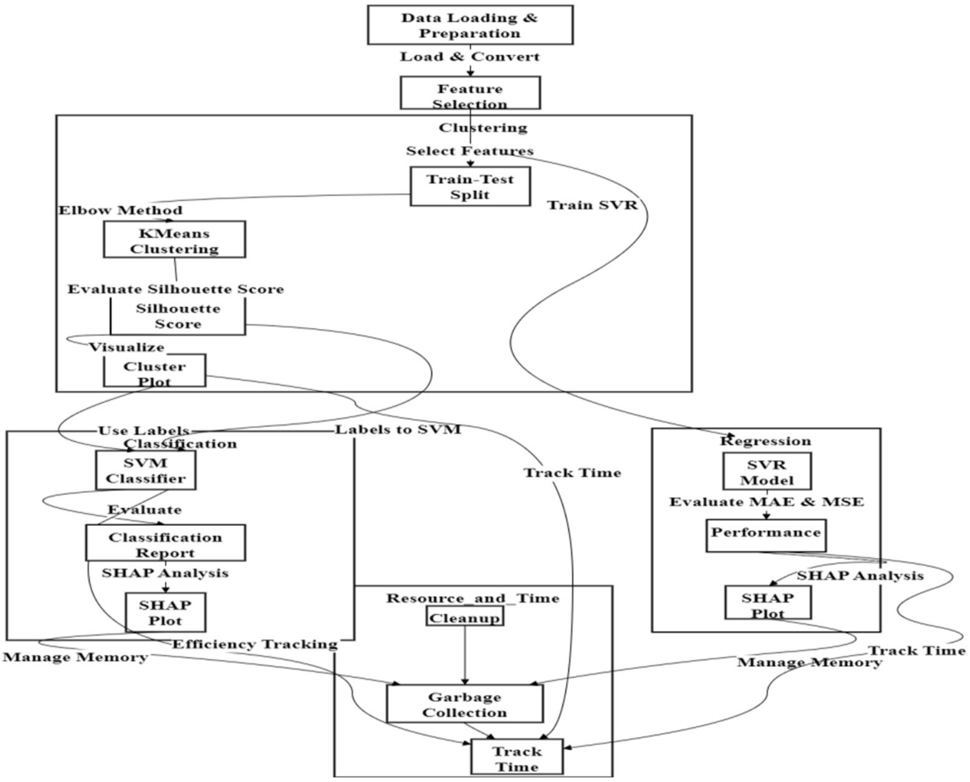

Fig 2. Workflow for Analyzing a Dataset for SVM

3.3. Clustering Analysis

Figure 3 depict the experimental results obtained through Elbow Method in order to determine the optimal number of clusters for efficient clustering process. The analysis employs K-Means clustering to uncover natural groupings within the dataset, with the dataset split into training and testing sets to validate clustering robustness. The Elbow Method assesses inertia to pinpoint the optimal cluster count, showing differing results of 4 and 10 clusters, respectively. The final clustering solution's quality is evaluated using the silhouette score, which measures the separation and cohesion of clusters to gauge the effectiveness of the clustering process.

3.4. Clustering Visualization

To further understand the structure of SHAP values, clustering is carried out using the K-Means algorithm with various numbers of clusters. This process reveals patterns in feature contributions. After clustering, dimensionality reduction techniques are used to visualize the SHAP values within the identified clusters. Figure 4 illustrates the clustering results and SHAP value visualizations for different cluster configurations. These visualizations aid in comprehending how feature importance varies across clusters and highlight the impact of clustering on model interpretability.

3.5. Model Interpretability and SHAP Analysis

To enhance the interpretability of the models, SHapley Additive exPlanations (SHAP) are employed. The dataset is processed to retrain linear regression and SVM models. SHAP values are then calculated to clarify the influence of features on different predictions.

The SHAP analysis includes:

SHAP Bar Chart: This chart ranks features based on their impact on model predictions, supporting the insights from the summary plot. Longitude is identified as the most significant predictor, with latitude and altitude also playing important roles.

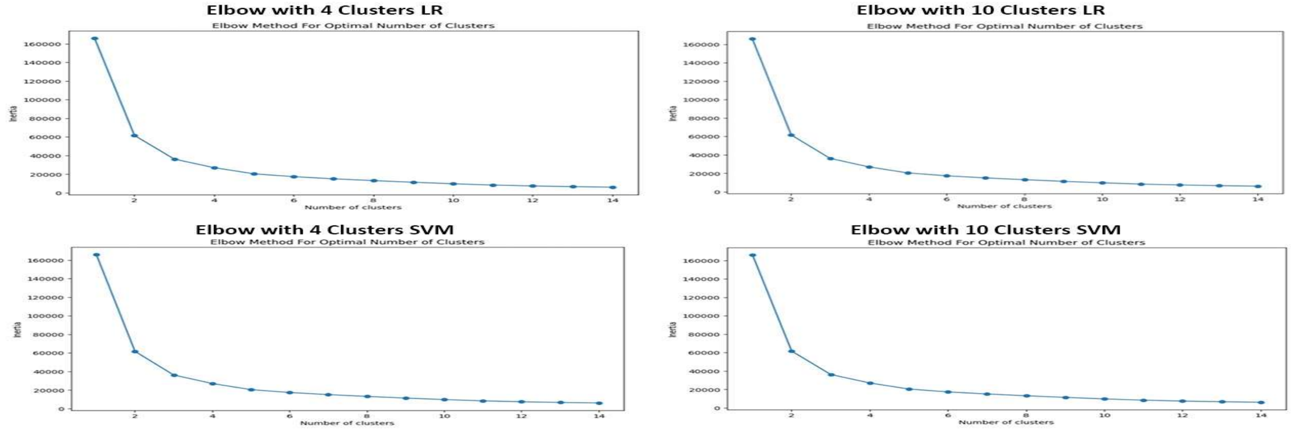

Fig 3. The Elbow Method to determine the optimal number of clusters for Linear Regression (LR) and SVM

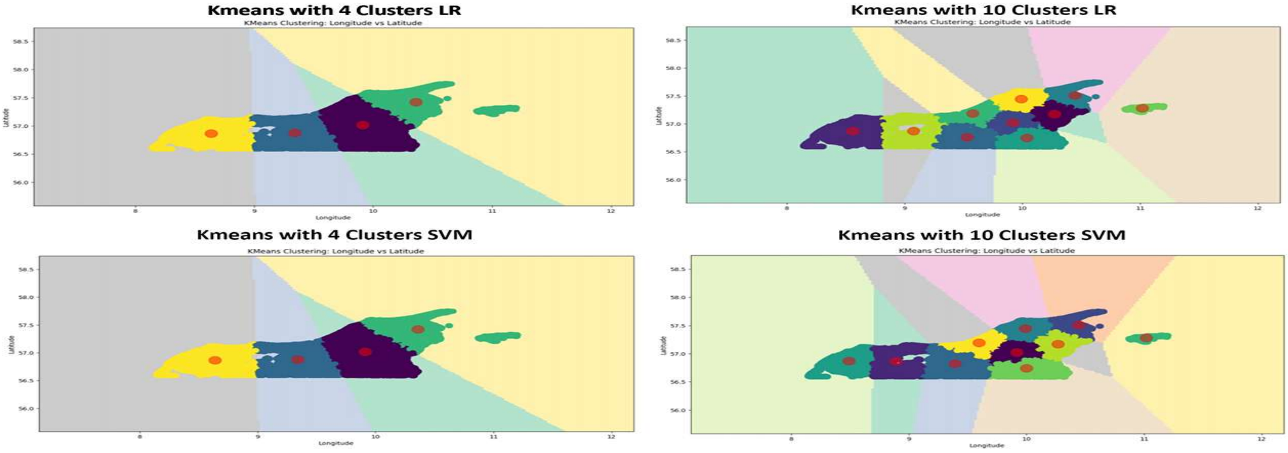

Fig 4. Clustering Visualization with K=4 and K=10 for Linear Regression and SVM

3.6. Helper Functions

In the analysis of clustering and model interpretability within the 3D road network dataset, several key helper functions underpin the methodologies employed, each contributing to the robustness of the results.

For K-Means clustering, inertia calculation is crucial, which is defined as presented in Equation (1).

$$W = \sum_{i=1}^{n} \overset{min}{\mu_k \in C} \; ||x_i - \mu k||^2 \tag{1}$$

Where X_i is a data point, μk is the centroid of cluster k, and C is the set of centroids. This calculation is essential for assessing how well the data points fit within their respective clusters, directly influencing the evaluation of clustering performance.

The silhouette score further refines cluster quality assessment, which can be calculated as per Equation (2).

$$S = \frac{b - a}{\max(a, b)} \tag{2}$$

Here, ‘a’ is the average distance between a data point and all other points within the same cluster, while ‘b’ represents the minimum average distance to points in the nearest neighboring cluster. This score provides a measure of both cohesion and separation, critical for validating the effectiveness of the K-Means clustering results.

Additionally, SHAP (Shapley Additive explanations) values are utilized for model interpretability, computed as per Equation (3).

$$SHAP(x) = f^{(x) - \mathbb{E}[f^{(x)}]} \tag{3}$$

The equation (3) provides a detailed decomposition of the model’s prediction, f(x), by subtracting the expected prediction value SHAP(x) thereby offering insights into the individual contributions of features to the model's output. Together, these helper functions form a cohesive framework for analyzing clustering performance and model interpretability, ensuring a comprehensive and insightful evaluation of the 3D road network dataset.

4. Experimental Setup

The experimental setup systematically analyzed the geospatial dataset, starting with data acquisition and preprocessing for quality and consistency. Feature engineering and discretization enhanced model performance, followed by model training. SHAP analysis provided interpretability, while clustering and dimensionality reduction revealed patterns and relationships, deepening insights into the geospatial data.

4.1. Data Acquisition and Preprocessing

The dataset comprises 434,874 records from a high-resolution 3D road network in North Jutland, Denmark, encompassing key geospatial attributes such as longitude, latitude, and altitude [27]. The dataset was ingested into a Pandas Data Frame, where columns were systematically renamed to enhance clarity and consistency. Rigorous preprocessing protocols were applied to ensure data quality, including the removal of duplicates, verification of missing values, and conversion of all attributes to appropriate numeric types. The features were normalized using the Min-max Scaler, standardizing the data to a uniform scale. Critically, the longitude attribute was discretized into 100 equal-width bins, converting continuous geospatial data into categorical labels to facilitate subsequent classification tasks.

4.2. Coordinated Frameworks for Data Preparation, Analysis, and Resource Optimization

In the data preparation phase, both frameworks emphasize the meticulous organization of data into relevant columns, ensuring that key features are accurately represented for subsequent analysis. The clustering evaluation process is uniformly addressed by determining the optimal number of clusters, assessing clustering quality, and visualizing the distinct clusters to gauge the effectiveness of the clustering algorithm. Regression analysis in both frameworks involves the development of predictive models for longitude, with performance assessed using similar error metrics and a thorough analysis of feature importance. For classification, both frameworks integrate models to predict cluster labels, with performance evaluated through consistent metrics. Additionally, resource management is a shared focus, highlighting the importance of efficient resource utilization, tracking execution time, and optimizing resources throughout the analysis. This approach ensures that each framework maintains its unique focus and distinct methodological steps while avoiding redundant or repetitive descriptions.

4.3. Evaluation Framework for Linear Regression

This framework introduces a comprehensive approach for analyzing 3D spatial network data, integrating data processing, clustering, regression, and model interpretation techniques. The process starts with data preparation, where 3D spatial network data is preprocessed, focusing on identifying and extracting key numerical features. For clustering, the optimal number of clusters is determined using an evaluation method, with distinct cluster regions identified and visualized to reveal spatial patterns. In the regression analysis, a Linear Regression model is developed to predict longitude, with performance assessed using error metrics. Additional analysis is conducted to interpret the regression model and understand feature importance, with visualizations illustrating feature contributions. A secondary model is trained to predict cluster labels, with accuracy and classification metrics used for evaluation.

Efficient resource management is emphasized throughout the analysis to ensure optimal resource use. Recommendations include addressing data handling challenges, improving model interpretation clarity, and refining visualization techniques, particularly in clustering analysis and the application of more robust techniques for cluster label prediction.

4.4. Evaluation Framework for SVM

The setup involves a systematic approach to validate and evaluate the proposed methodology for analyzing 3D road network data. Data preparation includes organizing data into relevant columns and splitting it into training and testing sets. The clustering phase involves determining the optimal number of clusters, with an assessment of clustering quality and visualization of distinct cluster centers. For the regression analysis, an SVM model is developed to predict longitude, with performance assessed using error metrics such as mean absolute error and mean squared error. The importance of features in the regression model is further analyzed using representative samples. In the classification phase, an SVM classifier predicts cluster labels, and performance is evaluated using accuracy, precision, recall, and F1-score. Further analysis is conducted to understand the impact of features on classification outcomes. The setup includes efficient resource management and monitoring of execution time for each phase, providing a comprehensive framework for evaluating the effectiveness of the proposed methodology in 3D road network data analysis.

5. Experimental Results and Discussion

This section presents the comprehensive evaluation of the proposed methodology, which integrates K-Means clustering with Linear Regression and SVM models to analyze 3D road network data. The analysis focuses on clustering quality, classification accuracy, regression performance, and interpretability using SHAP analysis.

Clustering Analysis: K-Means clustering was performed with two configurations: four and ten clusters. The ten-cluster configuration achieved a higher Silhouette Score of 0.4436 compared to 0.4286 for the four-cluster setup, indicating slightly better-defined clusters with improved cohesion and separation. As shown in Tables 1, 3, 5, and 7, the four-cluster setup exhibited superior classification performance with an SVM model, achieving an accuracy of 97.22% and high precision, recall, and F1-scores. This suggests that the four clusters were well-separated and effectively captured the underlying data patterns. In contrast, the ten-cluster setup resulted in a minor decrease in accuracy to 95.20% and slightly lower average precision, recall, and F1-scores. While the ten-cluster configuration captured more granular distinctions within the data, the increased complexity introduced a trade-off in classification performance. Nevertheless, both configurations demonstrated robust predictive performance, suggesting that the choice between four and ten clusters depends on the specific requirements of the application, balancing the need for model interpretability and predictive accuracy.

Regression Analysis: The regression analysis was conducted to predict longitude from latitude and altitude using both Linear Regression and SVM models. For Linear Regression, the four-cluster setup showed exceptional performance with a Mean Absolute Error (MAE) of 1.2440e-15 and a Mean Squared Error (MSE) of 2.2098e-30, indicating negligible errors. SHAP analysis for this configuration was completed in 0.37 seconds, reflecting highly accurate predictions. The ten-cluster configuration exhibited similar performance metrics with negligible MAE and MSE and slightly faster SHAP analysis (0.32 seconds), underscoring the model's robustness across different cluster configurations. Table 2, 4, 6, and 8 show detailed metrics and comparison results for Linear Regression and SVM regression models across both cluster configurations. For SVM regression, the four-cluster setup achieved an MAE of 0.0346 and an MSE of 0.0018, with a SHAP analysis time of 164.40 seconds, demonstrating effective spatial relationship modeling. The ten-cluster configuration showed similar regression metrics and a slightly faster SHAP analysis (152.33 seconds), indicating consistent performance despite increased complexity.

SHAP Analysis: SHAP analysis was pivotal in interpreting the regression models by elucidating influential features. For Linear Regression, SHAP analysis underscored the significance of latitude and altitude in predicting longitude, with the four-cluster setup offering clear insights into feature impacts. In contrast, the ten-cluster setup, while providing faster analysis, demonstrated slightly reduced clarity due to its increased complexity. Similarly, for SVM, SHAP analysis effectively highlighted feature importance, although the ten-cluster configuration slightly compromised interpretability. Figure 5 displays the SHAP bar plot distributions across various clustering configurations. These figures provide detailed insights into how feature contributions shift with clustering complexity, highlighting the trade-offs between analysis speed and interpretability.

Discussion: The results emphasize the impact of cluster configuration on model performance. The four-cluster configuration provided well-defined clusters that enhanced classification and regression performance, offering superior interpretability and predictive accuracy. Although the ten-cluster setup improved clustering quality, it introduced complexity that slightly affected classification performance while maintaining robust regression accuracy. Both Linear Regression and SVM models demonstrated consistent performance across configurations, with minimal differences in error metrics for regression tasks. SHAP analysis effectively elucidated feature importance, with the four-cluster setup offering clearer insights. The ten-cluster setup, while efficient in SHAP analysis, showed marginally reduced interpretability. Overall, the four-cluster configuration combined with SVM models proved to be the most effective strategy for analyzing 3D road network data, balancing accuracy, interpretability, and operational efficiency.

Table 1. Linear Regression Model Performance for FourClusters

Class	Precision	Recall	F1-Score	Support
0	0.71	0.24	0.35	32,663
1	0.21	0.62	0.31	14,845
2	0.26	0.32	0.28	25,000
3	1.00	0.09	0.16	14,467
Accuracy	-	-	0.30	86,975
Macro Avg	0.54	0.31	0.28	86,975
Weighted Avg	0.54	0.30	0.29	86,975

Table 2: Linear Regression Metrics Performance for Four Clusters

Metric	**Value**
Silhouette Score (10 clusters)	0.4436
Linear Regression MAE	1.2440e-15
Linear Regression MSE	2.2098e-30
Linear Regression SHAP Analysis Time	0.37 seconds
Accuracy of Linear Regression on Cluster Labels	0.0609
Second Linear Regression SHAP Analysis Time	0.55 seconds
Total Execution Time	0.70 seconds

Table 3. Linear Regression Model Performance for Ten Clusters

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.00	0.00	0.00	7304
1	0.00	0.00	0.00	11040
2	0.00	0.00	0.00	13548
3	0.31	0.70	0.43	7420
4	0.00	0.01	0.00	11426
5	0.00	0.00	0.00	7390
6	0.00	0.00	0.00	7323
7	0.00	0.00	0.00	2019
8	0.00	0.00	0.00	9653
9	0.00	0.00	0.00	9852
Accuracy	0.06	-	-	86975
Macro Avg	0.03	0.07	0.04	86975
Weighted Avg	0.03	0.06	0.04	86975

Table 4. Linear Regression Metrics Performance for Ten Clusters

Metric	**Value**
Silhouette Score (10 clusters)	0.4286
Linear Regression MAE	1.2440e-15
Linear Regression MSE	2.2098e-30
Linear Regression SHAP Analysis Time	0.32 seconds
Accuracy of Linear Regression on Cluster Labels	0.2993
Second Linear Regression SHAP Analysis Time	0.46 seconds
Total Execution Time	0.63 seconds

Table 5. SVM Model Performance for Four Clusters

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.98	0.96	0.97	32,663
1	1.00	0.93	0.96	14,845
2	0.95	1.00	0.97	25,000
3	0.97	1.00	0.98	14,467
Accuracy	-	-	0.97	86,975
Macro Avg	0.97	0.97	0.97	86,975
Weighted Avg	0.97	0.97	0.97	86,975

Table 6. SVM Metrics Performance for Four Clusters

Metric	**Value**
Silhouette Score (4 clusters)	0.42856108376016255
SVM Regression MAE	0.03456217002558308
SVM Regression MSE	0.0018446740034164684
SVM Regression SHAP analysis time	164.40 seconds
Accuracy of SVM on cluster labels	0.9722334004024145

Table 7. SVM Model Performance for Ten Clusters

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	1.00	0.85	0.92	7,304
1	0.97	1.00	0.98	11,040
2	0.89	1.00	0.94	13,548
3	1.00	0.93	0.96	7,420
4	0.95	1.00	0.98	11,426
5	0.99	0.89	0.94	7,390
6	0.98	0.84	0.90	7,323
7	1.00	1.00	1.00	2,019
8	0.90	0.96	0.93	9,653
9	0.98	0.99	0.98	9,852
Accuracy	-	-	0.95	86,975
Macro Avg	0.96	0.94	0.95	86,975
Weighted Avg	0.96	0.95	0.95	86,975

Table 8. SVM Model Metrics for Ten Clusters

Metric	**Value**
Silhouette Score (10 clusters)	0.44356130693333673
SVM Regression MAE	0.03456217002558308
SVM Regression MSE	0.0018446740034164684
SVM Regression SHAP analysis time	152.33 seconds
Accuracy of SVM on cluster labels	0.9520206956021845

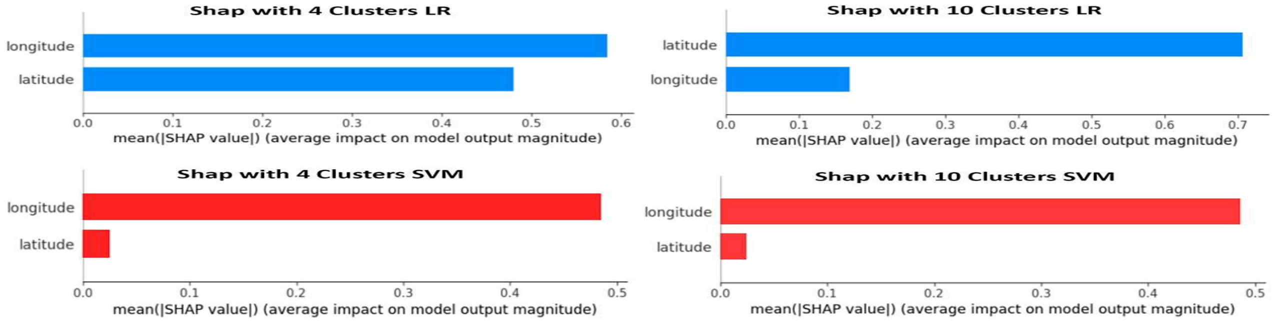

Fig 5. xAI with Shapley Value with optimal number of clusters (K=4 and 10) for Linear Regression and SVM

6. Conclusion

This research presents a comprehensive methodology for analyzing 3D road network data by integrating clustering, regression, and classification techniques, supplemented by model interpretability tools such as SHAP analysis. The study evaluated K-Means clustering with two configurations ten clusters and four clusters yielding silhouette scores of 0.4436 and 0.4286, respectively. Despite the slightly higher silhouette score for ten clusters, the four-cluster configuration proved superior for classification tasks, achieving high accuracy (95.20%) with strong performance metrics, including precision, recall, and F1-score consistently above 0.90 across all classes. This result underscores the importance of selecting an optimal number of clusters to balance complexity and performance, as the ten-cluster configuration led to significantly poorer classification outcomes. In the regression phase, SVM with an RBF kernel outperformed Linear Regression, demonstrating realistic error metrics (MAE of 0.0346 and MSE of 0.0018), which indicated a well-fitted model for predicting longitude from spatial features.

Conversely, the Linear Regression model exhibited near-zero errors (MAE ≈ 1.244 × 10^−15, MSE ≈ 2.2098 × 10^−30), suggesting overfitting, particularly within the ten-cluster setup. SHAP analysis further provided critical insights into feature importance, validating the contributions of spatial features like latitude and altitude to the SVM regression model, especially in the four-cluster scenario. However, the interpretability of models using the ten-cluster configuration was limited by poor clustering quality, reducing the utility of SHAP's feature importance explanations. Key findings highlight the four-cluster setup's consistent superiority over ten clusters across both classification and regression tasks, emphasizing the necessity of careful cluster count selection to achieve optimal model performance. The SVM models with RBF kernels showcased strong predictive capabilities and robustness, making them well-suited for spatial data analysis in 3D road networks. Moreover, SHAP analysis effectively elucidated the impact of features on model predictions, although trade-offs between computational efficiency and interpretability were noted, particularly for the ten-cluster setup. Overall, this methodology demonstrates robust performance and interpretability in analyzing 3D road network data, particularly when leveraging SVM regression and classification models with optimal clustering configurations. Future research should explore advanced clustering techniques and further refine model parameters to enhance both accuracy and interpretability. Additionally, incorporating complementary interpretability tools such as LIME could provide a broader understanding of model behavior, especially when dealing with complex datasets like 3D spatial networks. This study's insights into model selection and interpretability contribute to the growing body of knowledge on applying machine learning to spatial data, with implications for improving the precision and effectiveness of predictive modeling in road network analysis.

References

[1] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

[2] Schubert, E. (2023). Stop using the elbow criterion for k-means and how to choose the number of clusters instead. ACM SIGKDD Explorations Newsletter, 25(1), 36-42.

[3] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

[4] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[5] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD

international conference on knowledge discovery and data mining (pp. 1135-1144).

[8] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6), 417.

[9] McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

[10] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[11] Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. GESTS international transactions on computer science and engineering, 30(1), 25-36.

[12] Wang, K., Chen, Z., Dang, X., Fan, X., Han, X., Chen, C. M., ... & Weng, J. (2023). Uncovering hidden vulnerabilities in convolutional neural networks through graph-based adversarial robustness evaluation. Pattern Recognition, 143, 109745.

[13] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[14] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[15] Chen, J., Song, L., Wainwright, M., & Jordan, M. (2018, July). Learning to explain: An information-theoretic perspective on model interpretation. In International conference on machine learning (pp. 883-892). PMLR.

[16] Hennig, C. (2007). Cluster-wise assessment of cluster stability. Computational Statistics & Data Analysis, 52(1), 258-271.

[17] Zhang, S., Li, J., Shi, L., Ding, M., Nguyen, D. C., Tan, W., ... & Han, Z. (2023). Federated learning in intelligent transportation systems: Recent applications and open problems. IEEE Transactions on Intelligent Transportation Systems.

[18] Nikparvar, B., & Thill, J. C. (2021). Machine learning of spatial data. ISPRS International Journal of Geo-Information, 10(9), 600.

[19] Assent, I. (2012). Clustering high dimensional data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(4), 340-350.

[20] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, April). Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

[21] César, I., Pereira, I., Rodrigues, F., Miguéis, V., Nicola, S., & Madureira, A. Exploring multimodal learning applications in marketing: A critical perspective. International Journal of Hybrid Intelligent Systems, (Preprint), 1-18.

[22] Xie, W. B., Liu, Z., Das, D., Chen, B., & Srivastava, J. (2023). Scalable clustering by aggregating representatives in hierarchical groups. Pattern Recognition, 136, 109230.

[23] Kaul, M. (2013). 3D Road Network (North Jutland, Denmark) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5GP51.

Biographies of Authors

Ms. Rotsnarani Sethy received the B. Sc. degree in mathematics from S. B. Womens College, Cuttack, the MCA degree from North Odisha University and M.Tech (Computer Science) from Utkal University, India. She received Rajiv Gandhi national Fellowship for carrying out her Ph. D. in the Department of Computer science and Applications, Utkal University. Her research interests include: big data analytic, machine learning, database design, etc. She has published 4 articles in conference and peer reviewed journals. She can be reached at email: *roshnaranimca@gmail.com*

Mr. Soumya ranjan Mahanta completed B. Sc. in ITM and M. Sc. in ITM from Revenshaw University and M. Tech. (CSE) from Department of Computer Science and Applications, Utkal University, Odisha, India. Presently, he is working as Lecturer in Computer Science at DRIEMS University, Cuttack, India. His research interest include: Machine Learning, Artificial Intelligence, Python programming etc. He can be reached at email: *dipusoumyaranjan019@gmail.com*

Dr. Mrutyunjaya Panda is an Associate Professor at Computer Science and Applications, Utkal University, Vani Vihar, Bhubaneswar, Odisha, India. He holds a Ph. D. degree in Computer Science, M.Engg in Communication system Engineering and B.Engg. in Electronics and Tele-Communication Engg. and MBA in Human Resource Management., His research areas are Data Mining, Bio-medical image processing, Big data Analytic, Natural Language Processing and social Network Analysis. He has authored and co-authored more than 150 research articles in reputed Journals, conferences and book chapters. He has also edited 7 books and authors two text books to his credit. He is a member of editorial board and an active reviewer of many journals and conferences.