International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

# CUSTOMER SEGMENTATION DATA PREPROCESSING

[1]R.Shirisha,[2]P.Shivakeshav,[3]P.Abhitha,[4]Y.Manoj kumar

[1]Assistant Professor, [234]Students

Department of CSE(Data Science)

Siddhartha institute of technology & sciences,narapally

shirisharangu.cse@siddhartha.co.in, 23TQ1A6707@siddhartha.co.in, 23TQ1A6728@siddhartha.co.in, 23TQ1A6744@siddhartha.co.in

## ABSTRACT

Customer segmentation is an important technique used by organizations to understand their customers better and develop effective marketing strategies. The main objective of this project is to analyze customer data and divide customers into meaningful groups based on their characteristics, behavior, and purchasing patterns. By identifying different customer segments, businesses can target specific groups with personalized products, services, and promotional campaigns. In real-world scenarios, customer data collected from various sources such as online transactions, surveys, and customer relationship management systems often contains missing values, duplicate records, and inconsistent formats. Therefore, data preprocessing plays a crucial role in preparing the dataset for analysis. This project focuses on applying various preprocessing techniques such as data cleaning, handling missing values, normalization, feature selection, and data transformation to improve the quality of the dataset before performing segmentation. After preprocessing the dataset, machine learning techniques such as clustering algorithms (especially K-Means clustering) are applied to group customers based on similarities in their attributes like age, income, spending score, and purchasing behavior. These clusters help identify different types of customers such as high-value customers, regular customers, and low-engagement customers. Visualizations and analytical methods are used to interpret the clustering results and evaluate the effectiveness of the segmentation process. The results of this project demonstrate that proper data preprocessing combined with clustering algorithms can significantly improve the accuracy and usefulness of customer segmentation. The insights obtained from the segmentation process can help businesses design better marketing strategies, improve customer satisfaction, increase sales, and enhance overall customer relationship management. Overall, this project highlights the importance of data preprocessing and machine learning techniques in customer

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

analytics, and it shows how businesses can use data-driven approaches to make informed decisions and gain a competitive advantage in the market.

## I INTRODUCTION

Customer segmentation is the process of dividing customers into different groups based on similar characteristics such as demographics, purchasing behavior, and spending patterns. It helps organizations understand their customers better and develop effective marketing strategies. Businesses collect large amounts of customer data from various sources such as transactions, websites, and customer management systems. However, raw data often contains missing values, duplicate records, and inconsistencies. Therefore, data preprocessing is required to clean and prepare the data before analysis. Data preprocessing includes steps such as data cleaning, transformation, normalization, and feature selection. These steps improve the quality of the dataset and make it suitable for analysis. In this project, customer data is preprocessed and then analyzed using clustering techniques to group customers with similar characteristics. These groups help businesses identify different types of customers and understand their behavior. Customer segmentation can be applied in various fields such as retail, banking, and e commerce to improve marketing strategies and customer satisfaction.

## II LITERATURE SURVEY

Customer segmentation is an important technique used to divide customers into different groups based on similar characteristics such as purchasing behavior, income, and preferences. It helps businesses understand their customers and develop effective marketing strategies. Many researchers have applied data mining and machine learning techniques to perform customer segmentation. Among these techniques, clustering algorithms are widely used because they group customers based on similarities in the dataset. Several studies have shown that the K-Means clustering algorithm is one of the most commonly used methods for customer segmentation due to its simplicity and efficiency. It helps identify patterns in customer data and classify customers into meaningful groups. Researchers have also highlighted the importance of data preprocessing before applying clustering algorithms. Data preprocessing techniques such as data cleaning, handling missing values, normalization, and feature selection improve the quality of the dataset and increase the accuracy of segmentation results. Overall, previous studies indicate that combining data preprocessing with clustering techniques is an effective approach for analyzing customer data and identifying useful customer segments that can help businesses improve marketing strategies and customer management.

## III SYSTEM ANALYSIS

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

The existing customer data management system in many organizations is largely manual and relies on spreadsheets or basic scripts to collect, clean, and organize data from multiple sources such as CRM systems, e-commerce platforms, and social media channels. This approach is time-consuming, prone to errors, and often results in inconsistencies due to varying data formats and missing values. Moreover, manual segmentation typically uses simple rules based on demographics or purchase history, which limits the ability to uncover deeper patterns in customer behavior and reduces the effectiveness of marketing strategies. To address these issues, the proposed system focuses on automating the preprocessing of customer data to make it accurate, standardized, and ready for advanced segmentation techniques. It integrates data from multiple sources, handles missing values and duplicates, normalizes numerical features, encodes categorical variables, and selects relevant features for clustering algorithms such as K-Means or hierarchical clustering. The system also includes visualization and reporting modules to present insights in dashboards, enabling faster and data-driven decision-making. By automating these processes, the proposed system reduces manual effort, enhances the quality of customer segments, and scales efficiently with growing data, thereby providing businesses with more precise and actionable insights for targeting and engagement strategies.

**Existing system**

The existing customer management system in most organizations is largely manual and relies on spreadsheets or simple scripts to collect and process customer data from various sources such as CRM systems, sales records, and social media. Data cleaning, handling missing values, and standardizing formats are performed manually, which is time-consuming and prone to errors. Customer segmentation is usually done using basic rules based on demographics or purchase history, limiting the ability to identify deeper patterns. The system also struggles with scalability, as large datasets are difficult to manage efficiently, and real-time updates or automated insights are not possible. Overall, the current system is inefficient, inconsistent, and unable to provide advanced, data-driven segmentation.

**Disadvantages of existing system**

- Manual data collection and preprocessing is time-consuming.
- High chance of errors and inconsistencies due to human intervention.
- Handling missing values, duplicates, and different formats is inefficient.
- Segmentation is based on simple rules, limiting insights into complex customer behavior.
- Not scalable for large datasets; performance decreases as data grows.

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

**Proposed system**

The proposed customer segmentation preprocessing system offers several advantages over the existing manual approach. It automates data collection, cleaning, and transformation, which significantly reduces human effort and errors. Standardization of data across multiple sources ensures consistency and accuracy, while feature selection and preprocessing prepare the dataset for advanced clustering algorithms, enabling deeper insights into customer behavior. The system is scalable, capable of handling large datasets efficiently, and can support real-time updates and automated reporting. These improvements allow businesses to quickly generate actionable insights, create precise customer segments, and make data-driven marketing and strategic decisions more effectively.

**Advantages of proposed system**

- Automates data collection and preprocessing, reducing manual effort and errors.
- Ensures data consistency and standardization across multiple sources.
- Prepares data for advanced segmentation using clustering algorithms, capturing complex customer patterns.
- Scalable for large datasets, handling growing data efficiently.
- Supports real-time updates and automated reporting, enabling faster decision-making.

## IV METHODOLOGY

The methodology describes the step-by-step process followed in this project to perform customer segmentation using data preprocessing and clustering techniques. It explains how the data is collected, processed, analyzed, and how the final results are obtained.

1. Data Collection The first step in the methodology is collecting the customer dataset. The dataset contains information related to customers such as age, income, gender, and spending score. This data serves as the input for the analysis process.

2. Data Preprocessing After collecting the dataset, preprocessing techniques are applied to improve the quality of the data. This step includes removing duplicate records, handling missing values, and correcting inconsistent data. Data preprocessing ensures that the dataset becomes clean and suitable for further analysis.

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

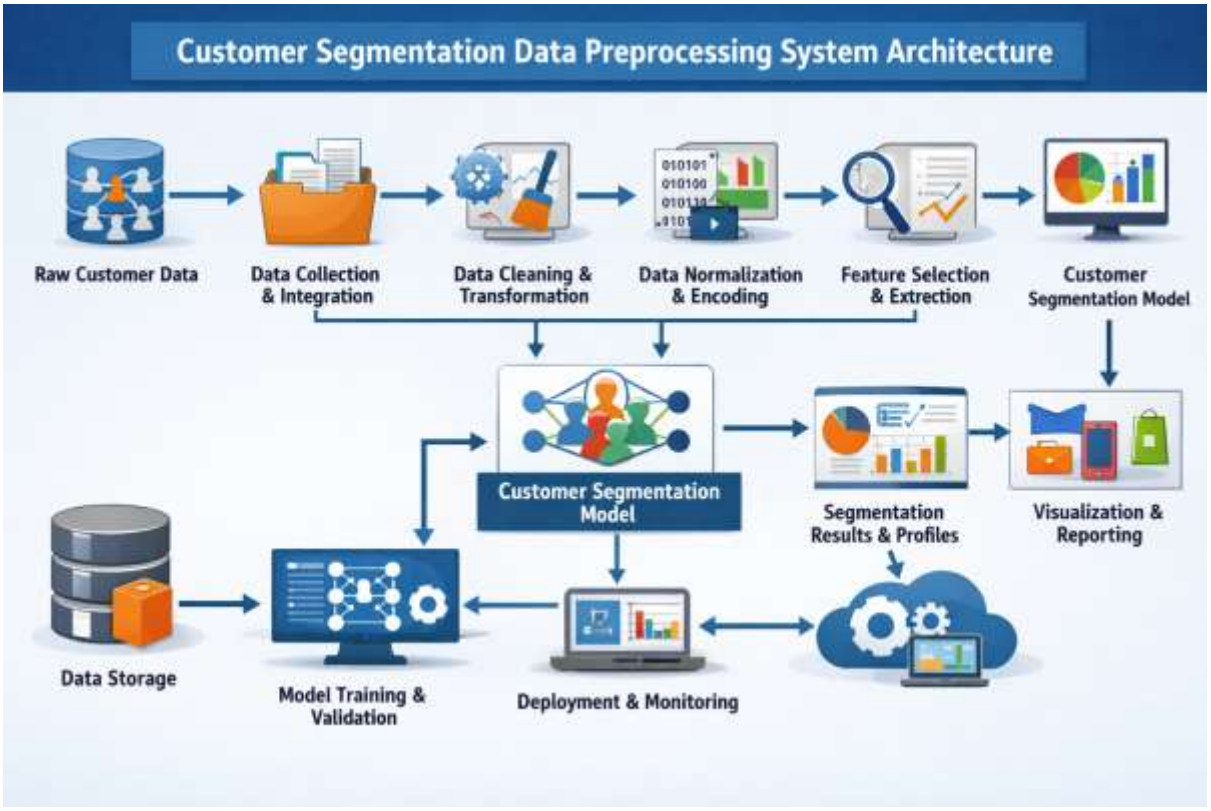
ISSN: 3068-272X ww.ijdim.com Original Research Paper

3. Data Transformation In this step, the data is converted into a proper format for analysis. Numerical and categorical values are transformed where necessary, and the dataset is prepared for machine learning algorithms.

4. Feature Selection Important attributes that influence customer behavior are selected from the dataset. Features such as age, income, and spending score are chosen because they help identify meaningful customer segments.

5. Clustering Algorithm After preprocessing and feature selection, a clustering algorithm such as K-Means is applied. This algorithm groups customers into different clusters based on similarities in their attributes.

**System Architecture**



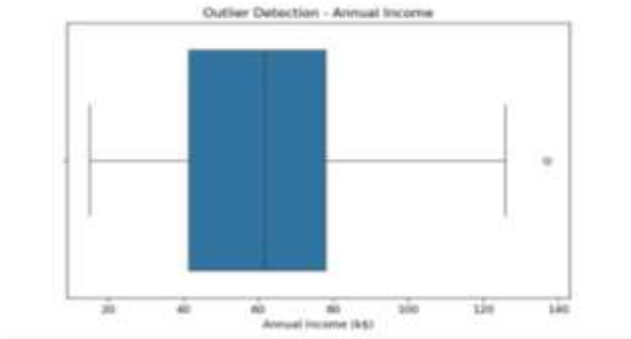

## V RESULTS&OUTPUT

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper




## VI CONCLUSION

Customer segmentation is an important technique that helps organizations understand customer behavior and improve their marketing strategies. In this project, customer data was analyzed using data preprocessing techniques and clustering algorithms to divide customers into meaningful groups. The preprocessing process helped improve the quality of the dataset by cleaning the data, handling missing values, and selecting important features. After preprocessing, clustering techniques were applied to group customers based on similarities in their characteristics such as age, income, and spending patterns. The results obtained from the segmentation process help businesses identify different types of customers and understand their needs and preferences. This information can support organizations in making better marketing decisions, improving customer satisfaction, and increasing business performance. Overall, the project demonstrates that data preprocessing combined with machine learning techniques can effectively analyze customer data and provide useful insights for business decision-making.

## REFERENCE

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.