BASAVARAJESWARI GROUP OF INSTITUTIONS

BALLARI INSTITUTE OF TECHNOLOGY & MANAGEMENT

Autonomous Institute under VTU, Belagavi | Approved by AICTE,New Delhi Recognized by Govt. of Karnataka

NACC Accredited Institution*
(Recognized by Govt. of Karnataka, approved by AICTE, New Delhi & Affiliated to Visvesvaraya Technological University, Belgavi)
"JnanaGangotri" Campus, No.873/2, Ballari-Hospet Road, Allipur, Ballar1-583 104 (Karnataka) (India)
Ph: 08392 – 237100 / 237190, Fax: 08392 – 237197

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

A Machine Learning Report On

"CUSTOMER SEGMENTATION USING CLUSTERING"

Project Associates:

BHAVANA JOSHI **3BR22AI025**

Under the Guidance of

P Sahana Prasadh

R Sai Arshitha

Dept of AIML,

BITM, Ballari.

Visvesvaraya Technological University

Belagavi, Karnataka

2024-2025

BASAVARAJESWARI GROUP OF INSTITUTIONS

BALLARI INSTITUTE OF TECHNOLOGY & MANAGEMENT

Autonomous Institute under VTU, Belagavi | Approved by AICTE,New Delhi Recognized by Govt. of Karnataka

NACC Accredited Institution*

(Recognized by Govt. of Karnataka, approved by AICTE, New Delhi & Affiliated to Visvesvaraya Technological University, Belgavi)

"JnanaGangotri" Campus, No.873/2, Ballari-Hospet Road, Allipur, Ballar1-583 104 (Karnataka) (India)

Ph: 08392 – 237100 / 237190, Fax: 08392 – 237197

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

CERTIFICATE

This is to certify that the project work entitled "**Customer Segmentation Using Clustering**" is a bonafide work carried out by **BHAVANA JOSHI** bearing USN **3BR22AI025** in partial fulfillment for the award of degree of **Bachelor Degree in AIML** in the VISVESVARAYA TECHNOLOGICAL UNIVERSITY, Belagavi during the academic year 2024-2025. It is certified that all corrections and suggestions indicated for internal assessment have been incorporated in the report deposited in the library. The project has been approved as it satisfies the academic requirements in respect of mini project work prescribed for a Bachelor of Engineering Degree.

Signature of guide	Signature of guide	Signature of HOD
P Sahana Prasadh	**R Sai Arshitha**	**Dr. BM Vidyavathi**

Abstract

Customer segmentation is the process of dividing a customer base into distinct groups based on similar characteristics or behavior. This enables businesses to tailor marketing strategies, improve customer engagement, and allocate resources more effectively. This study explores a structured methodology for customer segmentation using clustering techniques,focusing on data preparation, analysis, and visualization. The process begins with organizing raw customer data, ensuring it is clean, consistent, and ready for analysis. Techniques like scaling and encoding are applied to standardize the data and make it suitable for clustering. Dimensional reduction methods, such as identifying key features, are used to simplify complex datasets while preserving the most relevant information.

Clustering methods are then applied to group customers based on patterns such as purchasing frequency, spending habits, and preferences. The optimal number of clusters is determined through methods that analyse patterns in the data, ensuring meaningful segmentation. Results are visualized using clear and intuitive plots to identify group differences and characteristics effectively.

This approach reveals actionable insights into customer behaviour, such as identifying loyal customers, infrequent shoppers, and high-value spenders. Businesses can leverage these insights to develop personalized marketing strategies, improve customer retention, and maximize profits. The study showcases a practical, step-by-step framework for achieving customer segmentation without relying heavily on automated functionalities, promoting a deeper understanding of the underlying methods.

Acknowledgement

I would like to express my sincere gratitude to everyone who contributed to the successful completion of this project, "**Customer segmentation using clustering**."I am deeply grateful to my institution **"**BALLARI INSITUTE OF TECHNOLOGY AND MANGEMENT", and the AIML, for providing the necessary resources and a conducive environment for learning and experimentation. I would also like to thank my professors and peers for their insightful discussions and encouragement. Lastly, I would like to acknowledge the immense support from my team, whose encouragement and understanding have been a constant source of motivation. This project has been a rewarding experience, enabling me to deepen my understanding of machine learning techniques, particularly the Naive Bayes algorithm, and their application in solving real-world problems like fake news detection.

Thank you all.

NAME	**USN**
Bhavana Joshi	**3BR22AI025**

Table of Contents

1. Introduction

In today's competitive business landscape, understanding customer behaviour is crucial for developing effective marketing strategies and improving customer satisfaction. One of the most powerful tools for achieving this is customer segmentation, which involves grouping customers based on similarities in their behaviours, demographics, or purchasing habits. By categorizing customers into meaningful segments, businesses can personalize their offerings, target the right audience with the right message, and optimize resource allocation.

Customer segmentation using clustering techniques, such as K-Means, has gained popularity due to its ability to uncover hidden patterns in data. Clustering, an unsupervised machine learning method, helps to automatically group customers based on data without predefined labels, making it ideal for exploring complex customer datasets. This process allows businesses to create customer profiles that represent distinct needs and preferences, ultimately improving customer engagement and loyalty.

In this project, we aim to design and develop a customer segmentation system using clustering algorithms. This project focuses on predicting customer segments using customer data, specifically their annual income and spending score. This system will utilize tools such as KMeans, to analyse customer behaviour patterns and develop actionable insights for businesses.

1.1 Objectives

1. To **train** a model to classify customers into different segments to allow targeted marketing.
2. To predict the appropriate cluster based on Annual income and Spending score of a particular customer.

2. System Analysis:

- The primary functions of the Customer Segmentation System are:

1. Input Collection:
    - The user should be able to input data on two key parameters:
        - Annual Income (k$).
        - Spending Score (1-100).
    - The data will be collected via an interactive web-based user interface.
2. Predict Customer Segment:
    - The system uses the input data to predict which customer cluster the user belongs to.
    - The output is a cluster label (e.g., Cluster 1, Cluster 2, etc.).
3. Output Presentation:
    - Once the prediction is made, the predicted segment should be displayed to the user in an easily understandable format.
4. Model Training and Prediction:
    - Training: A machine learning model (K-Means Clustering) will be pre-trained on historical data to recognize patterns in customer segments.
    - Prediction: When the user inputs data, the trained model will be used to predict the customer's segment based on their income and spending score.
5. Scalability:
    - The system should be designed to handle an increasing number of users, ensuring low-latency responses, especially when integrated with large-scale data.

3. Flow Chart

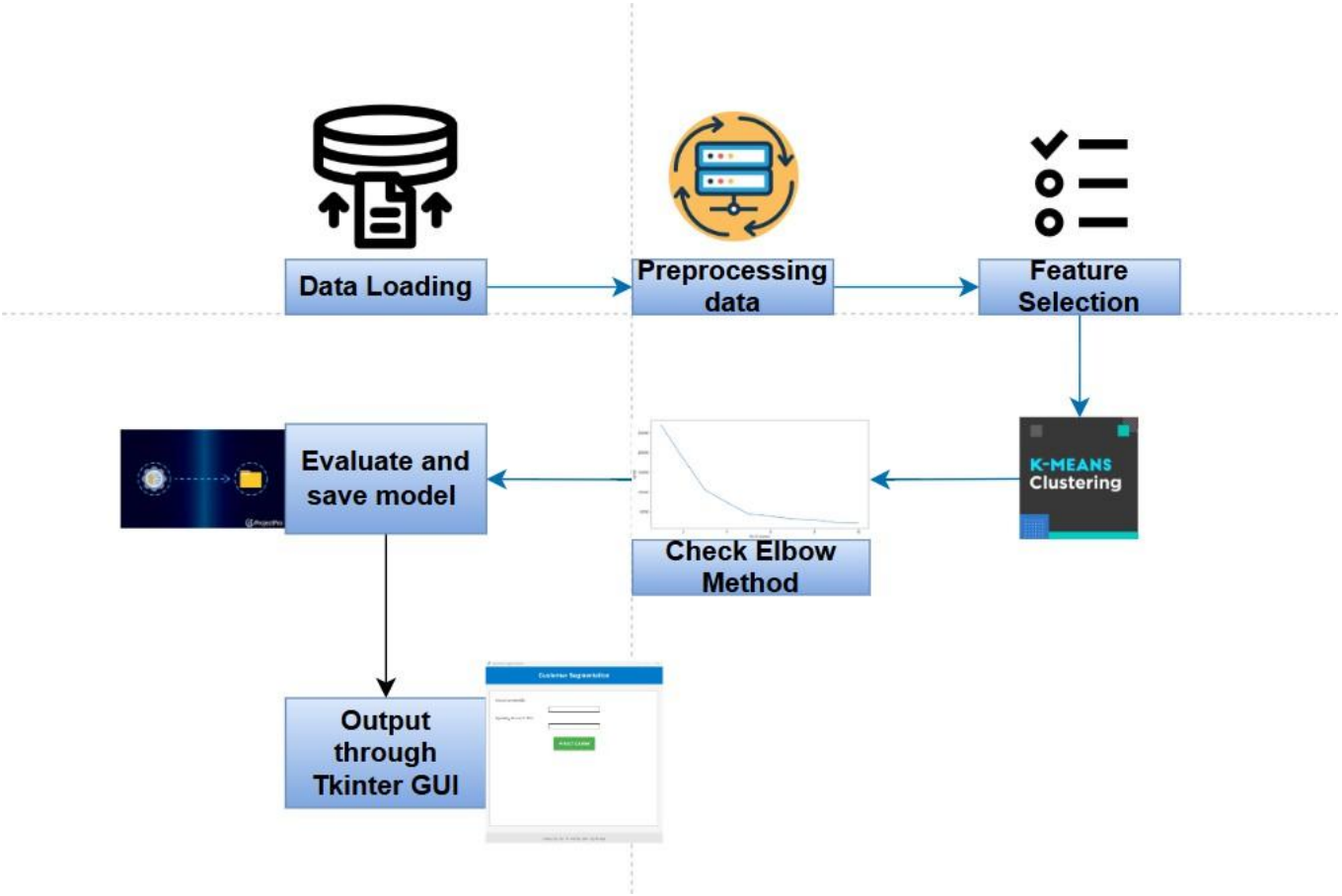

Figure 1 Flow-chart

4. Implementation

4.1 Dataset

- We have imported our dataset from Kaggle.
- The dataset consists of
    - Customer ID
    - Gender
    - Age
    - Annual Income
    - Spending Score

4.2 Preprocessing the dataset

- We have checked for null values and performed one hot encoding on the categorical data in the dataset.
- Described the data.

4.3 Feature Selection

- From the dataset to classify into clusters we have chosen only particular columns.
- We have selected two features from the dataset i.e. Annual income and Spending score.
- Further training will be performed on this features.

4.4 K-Means Clustering

- KMeans is a popular unsupervised machine learning algorithm used for clustering data into groups (clusters) based on their similarities.
- The algorithm divides a dataset into **K** distinct, non-overlapping clusters.
- Here we are going to group the similar groups of customers into k distinct clusters using KMeans Algorithm.

4.5 Elbow Method

- The **Elbow Method** is a technique used to determine the optimal number of clusters (**K**) in KMeans clustering.
- It involves plotting the **Within-Cluster Sum of Squares (WCSS)** against different values of **K**.
- The "elbow" point, where the rate of decrease in WCSS slows down, indicates the optimal value of **K**, balancing between model performance and complexity.
- In this case when we plotted elbow graph we got 5 as optimal no.of clusters (k).

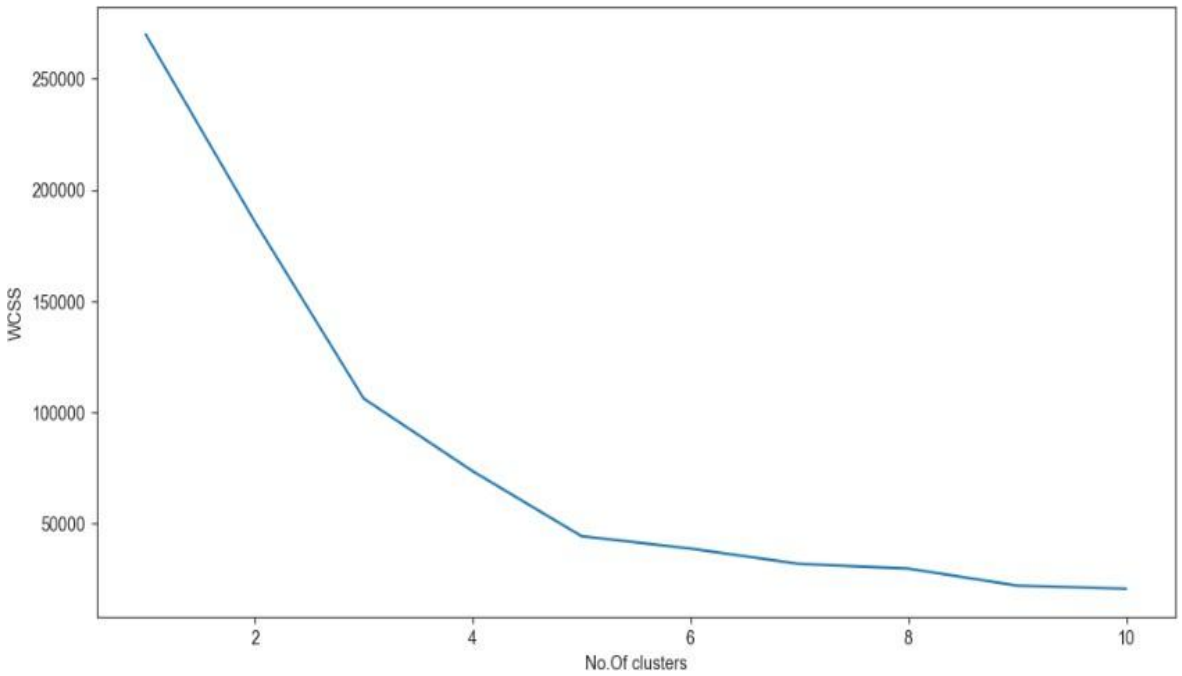

Figure 2 Elbow Graph

- The 5 clusters are:
    1. Cluster 0: Typically includes customers with low annual income and low spending score.
    2. Cluster 1: Contains customers with moderate annual income and moderate spending score.
    3. Cluster 2: Represents customers with high annual income but a low spending score.
    4. Cluster 3: Consists of customers with low annual income and a high spending score.
    5. Cluster 4: Includes customers with high annual income and a high spending score.

4.6 Evaluate and Save the model

- After evaluating the model save the model using joblib library.

4.7 Output

- Output the Model using Tkinter GUI.

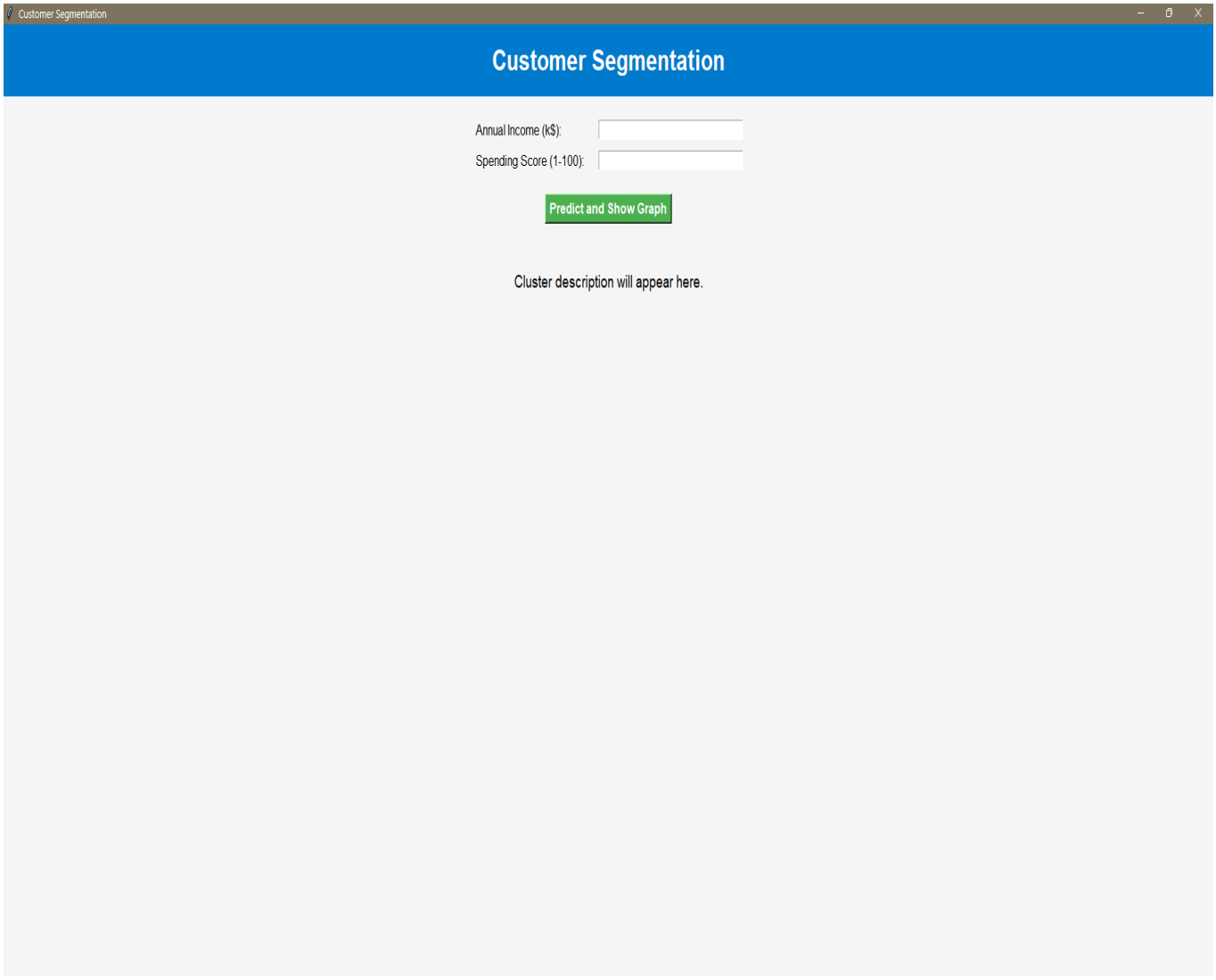

Figure 3 Tkinter GU

5. Testing

SL NO.	Step	Objective	Expected Output	Test Outcome
01	Library Imports	Ensure Required Libraries Are Imported Successfully.	Libraries Imported Without Error.	**Pass**
02	Dataset Loading	Load The Dataset And Inspect Its Structure.	Data Should Load And Display First Few Rows Correctly.	**Pass**
03	Missing Data Handling	Verify That Missing Data Is Handled Correctly.	Checking the shape and it shld have all same rows of the data.	**Pass**
04	Feature Selection	Verify That required features are selected.	Features like Annual income and Spending score are selected to train the model.	**Pass**
05	Predicting the correct clusters in GUI.	Giving Annual Income and Spending Score and checking for the correct cluster.	Displaying the respective cluster for the inputs.	**Pass**

6. Results

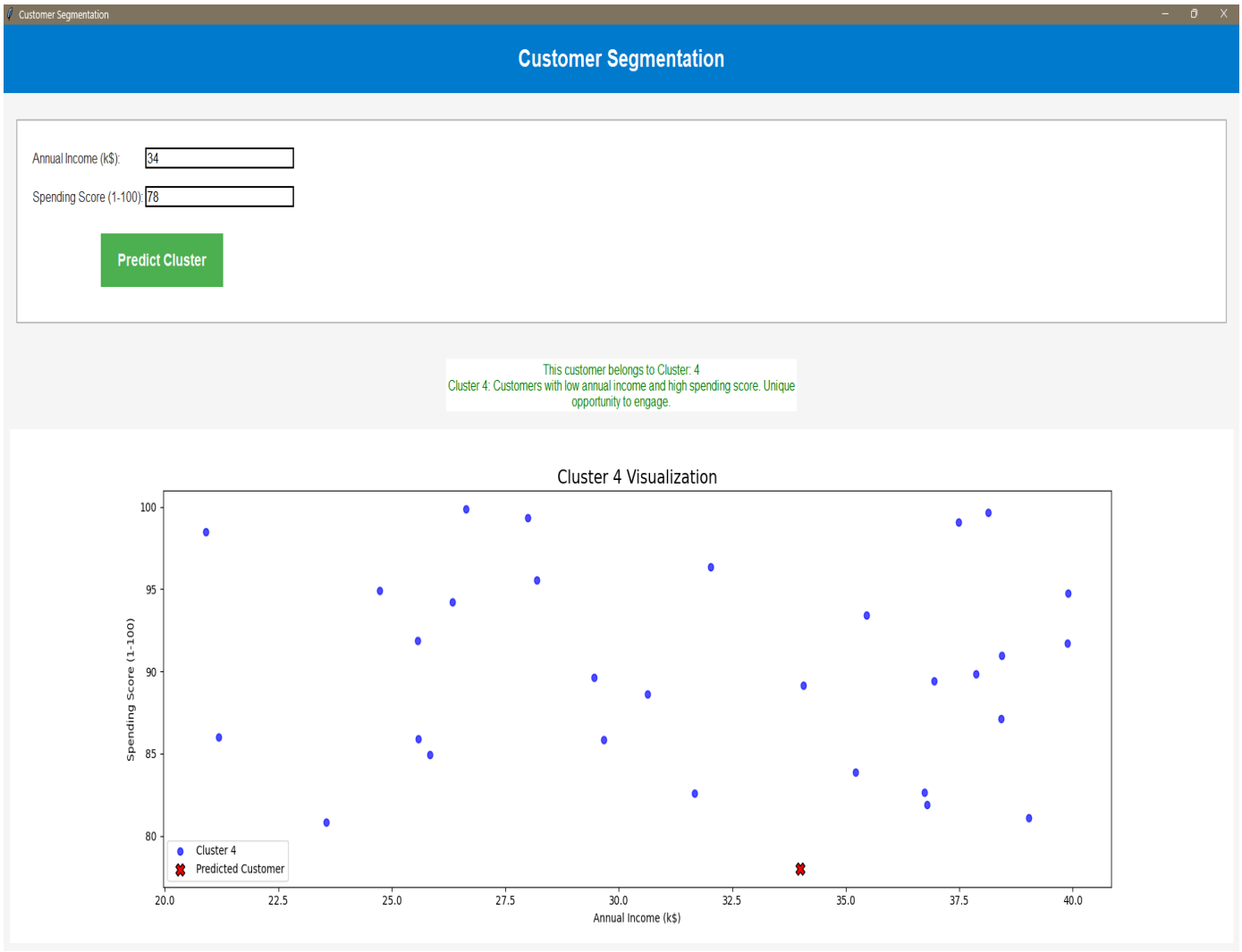

Figure 4 Output-1

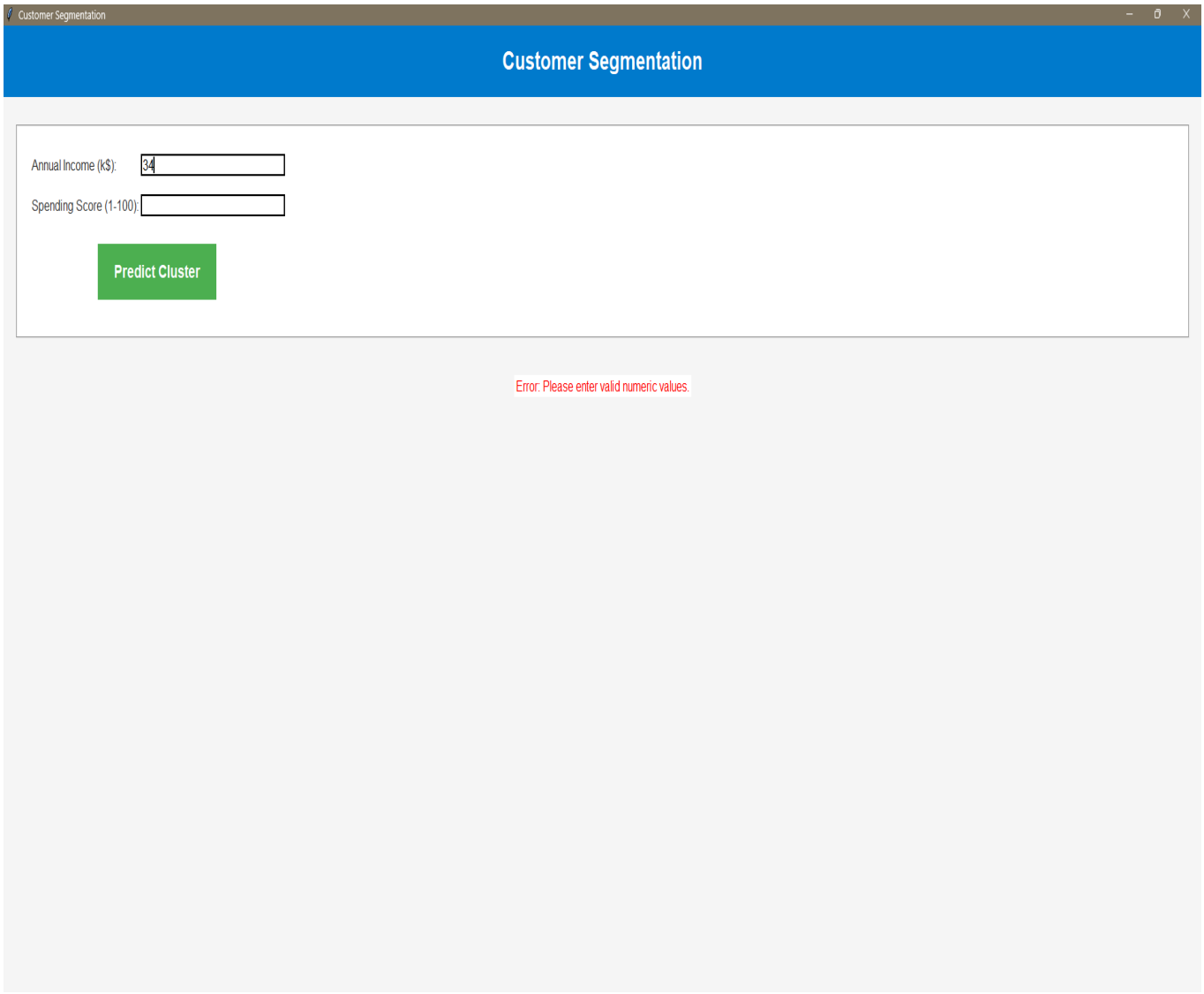

Figure 5 Output-2

7. Conclusion

The code effectively preprocesses and analyzes a marketing dataset, starting with loading and inspecting the data, followed by cleaning missing values. It creates several new features, such as "Customer_For," "Age," and "Spent," which enhance the understanding of customer behavior. Data visualization, including pair plots and heatmaps, is used to explore relationships between various features. The code handles outliers by applying caps to "Income" and "Age" to ensure data integrity. Categorical features are encoded using LabelEncoder, and numeric features are scaled for consistency. Principal Component Analysis (PCA) reduces dimensionality, allowing for easier visualization. Clustering techniques like KMeans and Agglomerative Clustering are applied to identify patterns in the data. The overall workflow is well-structured, ensuring that the dataset is transformed into a usable format for machine learning models.

8. References

[1] Kansal, Tushar, et al. "Customer segmentation using K-means clustering." *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)*. IEEE, 2018.

[2] Tabianan, Kayalvily, Shubashini Velu, and Vinayakumar Ravi. "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data." *Sustainability* 14.12 (2022): 7243.

[3] Wu, Jing, and Zheng Lin. "Research on customer segmentation model by clustering." *Proceedings of the 7th international conference on Electronic commerce*. 2005.