KARNATAKA LAW SOCIETY'S

GOGTE INSTITUTE OF TECHNOLOGY

UDYAMBAG, BELAGAVI-590008

(An Autonomous Institution under Visvesvaraya Technological University, Belagavi)

(APPROVED BY AICTE, NEW DELHI)

Course Activity Report on

" Customer Segregation Using K-means Algorithm"

Submitted in the partial fulfillment for the academic requirement

of ***5th Semester B.E.***

in

Data Warehousing and Data Mining - 21CS543

Submitted by

Supreet Tembadmani 2GI21CS168

Aditya Muchandikar 2GI21CS014

Omkar Patil 2GI21CS106

Brahmaraj kashetti 2GI21CS041

Rohit Bolgundi 2GI21CS129

GUIDE

Prof. Prashant Niranjan

Computer Science Department, Gogte Institute of Technology 2023 – 2024

KARNATAKA LAW SOCIETY'S

GOGTE INSTITUTE OF TECHNOLOGY

UDYAMBAG, BELAGAVI-590008

(An Autonomous Institution under Visvesvaraya Technological University, Belagavi)

(APPROVED BY AICTE, NEW DELHI)

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that Supreet Tembadmani, Aditya Muchandikar, Omkar Patil, Brahmaraj Kashetti, Rohit Bolgundi of 5th semesterand bearing USN 2GI21CS142, 2GI21CS186 , 2GI21CS136, 2GI21CS118 has satisfactorily completed the course activity (Seminar/Project) in Data Warehousing and Data Mining ***(Course code: 21CS543)*** . It can be considered as a bonafide work carried out in partial fulfillment for the academic requirement of 5th Semester B.E. Computer Science & Engineering prescribed by KLS Gogte Institute of Technology,Belagavi during the academic year **2023- 2024**

The report has been approved as it satisfies the academic requirements in respect of Assignment (Course activity) prescribed for the said Degree.

Signature of the Faculty Member

Signature of the HOD

Marks allocation:

	Batch No. :						
1.	Project Title:	Marks Range	USN				
			2GI21CS168	2GI21CS014	2GI21CS106	2GI21CS041	2GI21CS129
2.	Problem statement (PO2)	0-1					
3.	Objectives of Defined Problem statement(PO1,PO2)	0-2					
4.	Design / Algorithm/Flowchart/ Methodology(PO3)	0-3					
5.	Implementation details/Function/ Procedures/Classes and Objects (Language/Tools) (PO1,PO3,PO4,PO5)	0-4					
6.	Working model of the final solution(PO3,PO12)	0-5					
7.	Report and Oral presentation skill (PO9,PO10)	0-5					
	Total	20					
	Reduced to	10					

*** 20 marks is converted to 10 marks for CGPA calculation**

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals and an engineering specialization to the solution of complex engineering problems.

2. Problem Analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences and Engineering sciences.

3. Design/Development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the publichealth and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding ofthe limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need
for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of theengineering practice.

9. Individual and team work: Function effectively as an individual and as a member or leader in diverse teams, and in multidisciplinary settings.

10. Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of the engineering management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. Life-long learning: Recognise the need for and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological channel

Customer Segregation Using K-means Algorithm

Objective

The primary objective of this report is to explore and implement the K-means algorithm for customer segregation. By leveraging the power of clustering, businesses can gain valuable insights into their customer base, enabling more targeted marketing strategies, personalized product offerings, and enhanced customer experiences.

Customer Segmentation

Traditional customer segmentation methods, such as demographic or geographic segmentation, have been widely employed. However, these approaches often oversimplify customer diversity, hindering the ability to tailor strategies effectively.

Introduction to K-means Algorithm

The K-means algorithm is a popular clustering technique that partitions a dataset into K clusters based on similarity. Its simplicity and efficiency make it well-suited for customer segmentation tasks.

Previous Studies

Several studies have successfully applied the K-means algorithm in various domains, demonstrating its effectiveness in uncovering hidden patterns within datasets. By reviewing these studies, we can draw insights into best practices and potential challenges.

Algorithm

Input: Raw dataset (in CSV format)

Output: Model evaluation metrics (accuracy, confusion matrix, classification report)

1. Choose the Number of Clusters (k): Decide on the number of clusters you want to create in your dataset.This is a critical step, as the choice of k can impact the quality of the clustering results.

2. Feature Selection/Extraction (Optional): Depending on the nature of your dataset, you may choose to select relevant features or perform feature extraction before applying k-means. This step can help improve the quality of clustering.

3. Normalize/Standardize Data (Optional):It's often a good practice to normalize or standardize the data before applying k-means, especially if the features have different scales.

4. Apply K-Means Algorithm: Use the k-means algorithm to cluster the data points into k clusters. The algorithm iteratively assigns data points to clusters based on the mean of the cluster and updates the clustercentroids.

5. Assign Clusters to Data Points: After the k-means algorithm converges, each data point will be assigned to respective cluster groups

Methodology:

1. Load the Dataset: Use a library like pandas to load the dataset into a DataFrame.

2. Handle Missing Values: Identify columns with missing values. Decide on a strategy to handle missing values (e.g., fill with mean, median, or use more advanced imputation techniques). Implement the chosenstrategy to fill or drop missing values.

3. Encode Categorical Variables :Identify categorical columns in the dataset. Choose an encoding method(e.g., Label Encoding or One-Hot Encoding). Apply the selected encoding method to convert categorical variables into numerical representations.

4. Select Relevant Features for Clustering: Decide on the features that are relevant for k-means clustering.Create a new Data Frame containing only the selected features.

5. Apply K-Means Clustering: Choose the number of clusters (k) based on the characteristics of the data. Apply the k-means clustering algorithm to the selected features. Assign cluster labels to each data point.

6. Visualize Clusters Using PCA: Use PCA (Principal Component Analysis) for dimensionality reduction.Fit PCA to the clustered data. Plot the data points in a 2D or 3D space using the first few principal components. Visualize clusters using different colours for each cluster. Optionally, plot centroids of the clusters.

7. Generate Pre-processed Features: Add the cluster labels as a new feature to the original dataset.Ifneeded, create additional features based on the clustering results.

8. Further Analysis: Depending on the goals of your analysis, explore the impact of the generated features on other tasks. Consider using the pre-processed dataset for machine learning tasks, such as classification,regression, or anomaly detection.

Results and Analysis

Cluster Interpretation

The results of the K-means algorithm are personalized distinct customer clusters. Each cluster was analyzed to understand the characteristics and behaviours that defined it. Visualizations, such as cluster centroids and silhouette plots, aided in the interpretation.

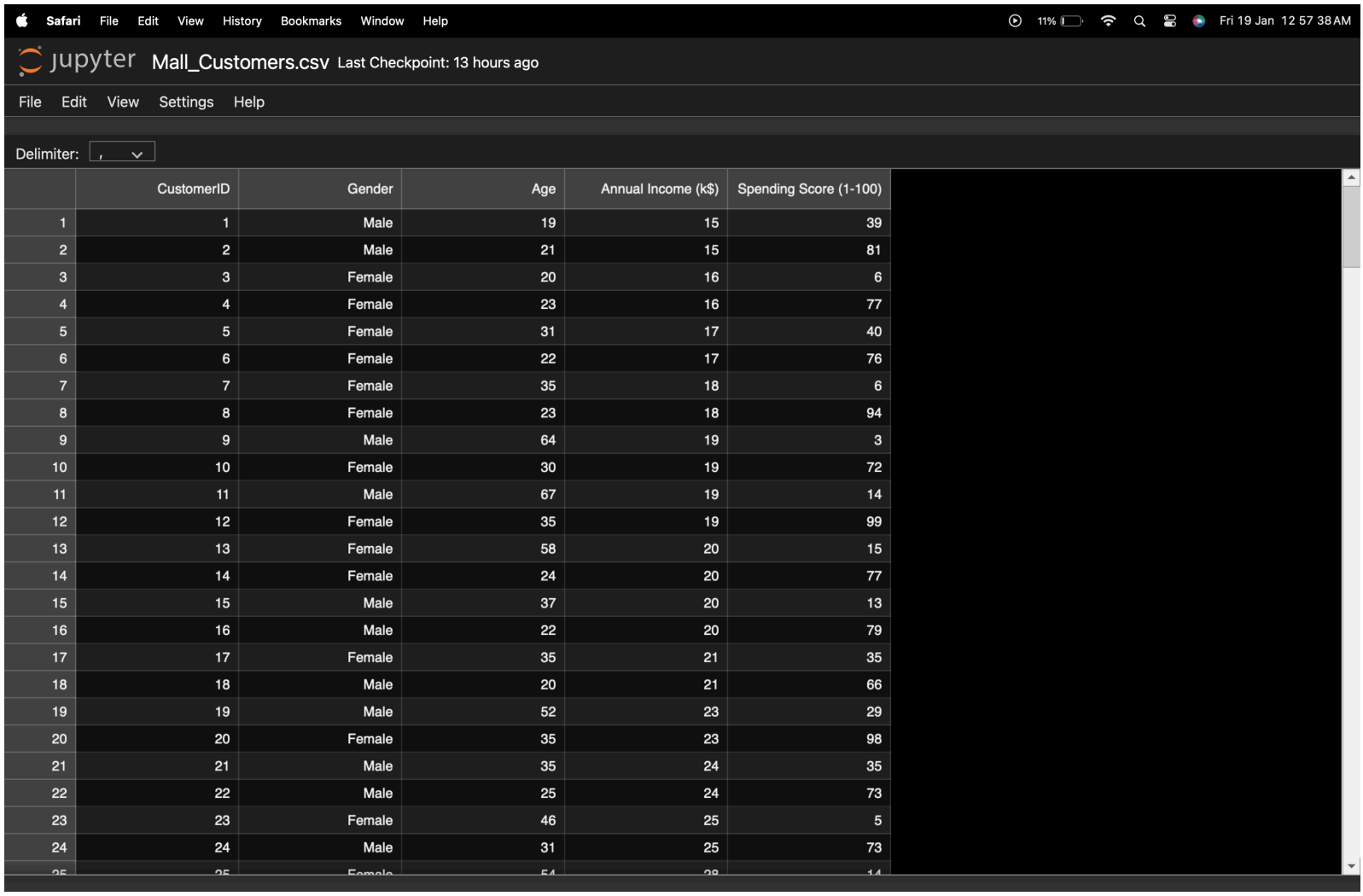
Validation Metrics

To assess the quality of the clustering, validation metrics such as the silhouette score and Davies-Bouldin index were employed. These metrics provided a quantitative measure of how well-defined and separated the clusters were.

Insights for Business

The identified clusters yielded valuable insights for businesses. For instance, Cluster 1 may represent price-sensitive customers, while Cluster 2 may consist of high-value, loyal customers. These insights can inform targeted marketing campaigns and personalized services.

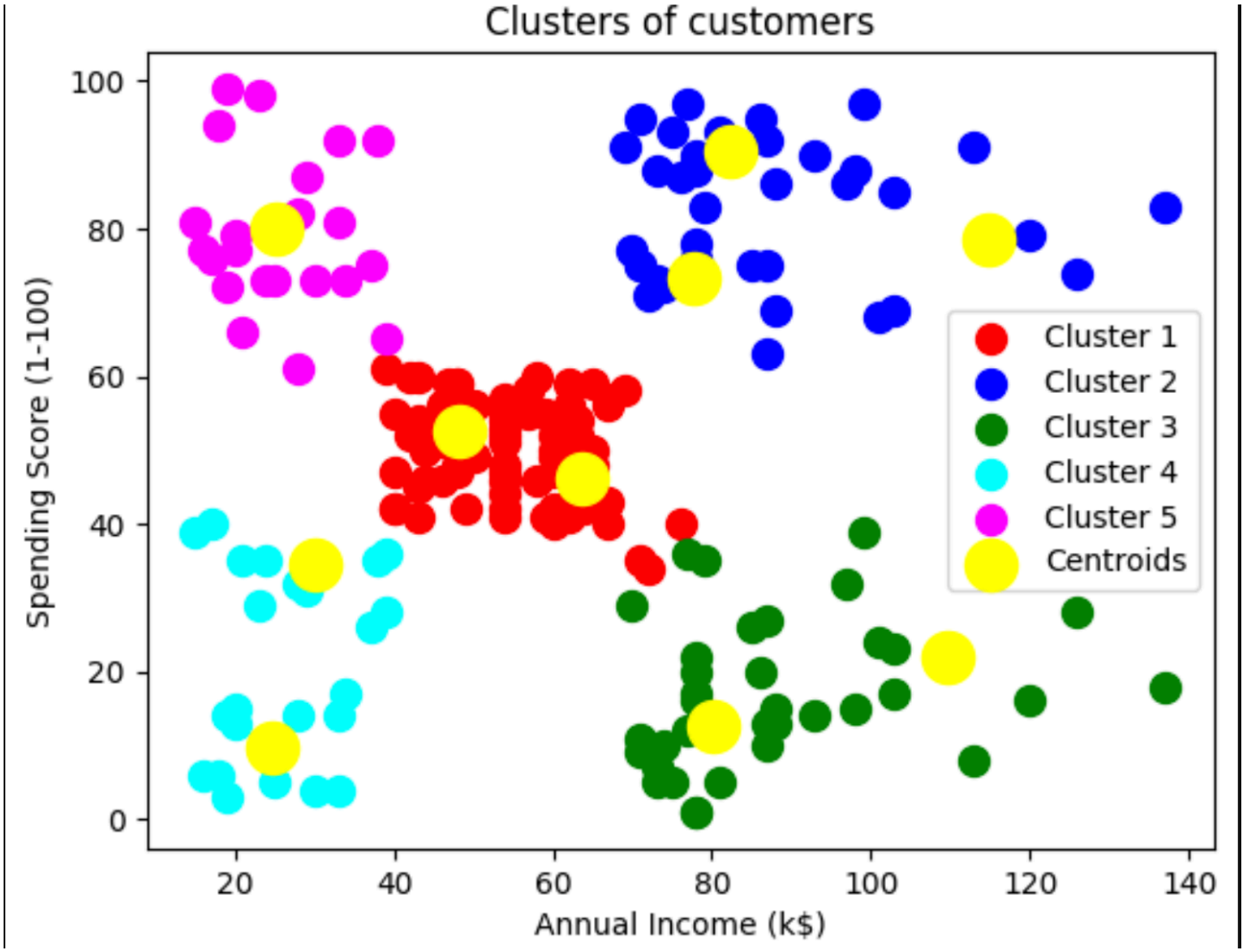
Dataset:



	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	1	Male	19	15	39
2	2	Male	21	15	81
3	3	Female	20	16	6
4	4	Female	23	16	77
5	5	Female	31	17	40
6	6	Female	22	17	76
7	7	Female	35	18	6
8	8	Female	23	18	94
9	9	Male	64	19	3
10	10	Female	30	19	72
11	11	Male	67	19	14
12	12	Female	35	19	99
13	13	Female	58	20	15
14	14	Female	24	20	77
15	15	Male	37	20	13
16	16	Male	22	20	79
17	17	Female	35	21	35
18	18	Male	20	21	66
19	19	Male	52	23	29
20	20	Female	35	23	98
21	21	Male	35	24	35
22	22	Male	25	24	73
23	23	Female	46	25	5
24	24	Male	31	25	73

Customers Data

Output:


Clusters of customers
Spending Score (1-100)
Annual Income (k$)
100
80
60
40
20
0
20
40
60
80
100
120
140
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Centroids

Summary of Findings

In summary, the application of the K-means algorithm successfully segmented the customer base into meaningful clusters. The insights gained from this segmentation have the potential to revolutionize how businesses approach customer engagement and marketing strategies.

Limitations

While the K-means algorithm is powerful, it is not without limitations. Sensitivity to initial cluster centers and the assumption of spherical clusters are factors that may impact results. Additionally, the quality of segmentation heavily relies on the features chosen.

Future Work

Future research could explore more advanced clustering algorithms or hybrid models that combine multiple techniques. Additionally, incorporating more diverse data sources and refining feature engineering could enhance the accuracy and robustness of customer segmentation.

Conclusion

In summary, the application of k-means clustering to preprocess a given dataset facilitates the identificationof natural patterns and the segmentation of data into distinct groups. This clustering approach enhances our understanding of the inherent structure within the dataset, offering valuable insights for decision-making and subsequent analysis. In conclusion, the utilization of the K-means algorithm offers a promising avenue for customer segregation. Businesses adopting such data-driven approaches can gain a competitive edge by better understanding their customer base and tailoring their strategies to meet evolving consumer needs.