A PROJECT REPORT

On

“Custumer Segmentations”

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR’S DEGREE IN COMPUTER SCIENCE AND ENGINEERING

BY

Shanskar kumar sarraf	**22054372**
Aaditya thakur	**22054378**
Aman kushuwa	**22054285**
Akash kumar	**22052876**
Sonu Kumar Shah	**22054269**

UNDER THE GUIDANCE OF

Vijay.meena

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

April 2025

CERTIFICATE

This is certify that the project entitled

"**Custumer Segmentations**"

SUBMITTED BY:

Shanskar kumar sarraf	**22054372**
Aaditya thakur	**22054378**
Aman kushwaha	**22054285**
Akash kumar	**22052876**
Sonu Kumar Shah	**22054269**

This is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Sci-ence & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2024-2025, under our guidance.

Date: / /

(Vijay.meena)
Project Guide

Acknowledgements

We are profoundly grateful to **vijay meena** of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. .....

Shanskar kumar sarraf
Aaditya thakur
Aman kushwaha
Akash kumar
Sonu kumar shah

ABSTRACT

Customer segmentation plays a crucial role in business strategy, allowing companies to tailor their marketing efforts and optimize resource allocation. Traditional segmentation methods often rely on manual analysis, which can be time-consuming and less precise. This project explores the use of machine learning techniques, specifically clustering algorithms and deep learning frameworks, to enhance the accuracy and efficiency of customer segmentation. By leveraging data preprocessing, feature selection, and predictive modeling, this study aims to develop an intelligent segmentation system that provides actionable insights for businesses.

Our project focuses on developing a Machine Learning-based Customer Segmentation System using clustering techniques and Flask. The system processes customer data, applies trained models, and provides segmentation insights via a web interface.

The system follows a structured pipeline:

- **Data Collection & Preprocessing:** Cleaning, normalizing, and transforming raw customer data for analysis.
- **Model Training:** Using clustering algorithms like K-Means and deep learning frameworks to segment customers based on behavioral patterns.
- **Gaussian Mixture Model (GMM)** – A probabilistic model that assumes data is generated from multiple Gaussian distributions.
- **Principal Component Analysis (PCA)** – A dimensionality reduction technique used to transform and analyze the data before clustering.

This project enhances business intelligence by providing data-driven segmentation, aiding companies in targeted marketing and strategic decision-making.

Keywords: CustomerSegmentation,Machine Learning,Clustering Algorithm,Data Preprocessing,Business Intelligence

Contents

List of Figures

Chapter 1

Introduction

Customer segmentation is a fundamental process in business analytics that helps companies understand their customer base and optimize marketing strategies. Traditional segmentation methods rely on manual analysis, which can be inefficient and prone to inaccuracies. Our project leverages Machine Learning to automate customer segmentation, enabling businesses to make data-driven decisions. The system is implemented using clustering algorithms, Flask for back-end processing, and a web-based interface using HTML, CSS, and JavaScript.

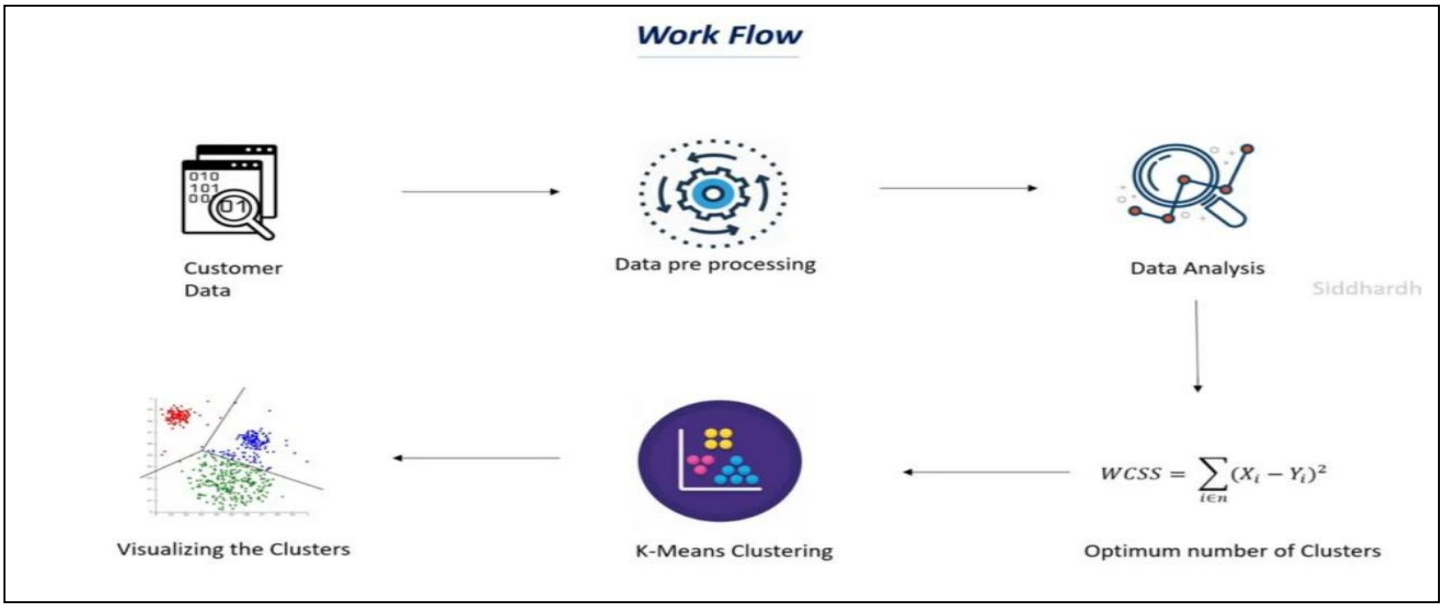
The project workflow includes data collection, preprocessing, model training, web deployment, and real-time segmentation analysis. This system allows businesses to categorize customers efficiently, improving marketing efforts and customer engagement.

1.1 Need for the Project

- Efficient and data-driven customer segmentation.
- Reduction in manual effort and human bias.
- Enhanced personalization of marketing strategies.
- Improved business decision-making through insights.

1.2 Research Motivation

Studies show that AI-driven customer segmentation improves marketing efficiency by 40% , have been widely applied in customer analytics, making them ideal compared to traditional methods. Machine learning models, such as K-Means clustering and deep learning techniquesfor this project.



Chapter 2

Basic Concepts/ Literature Review

2.1 Technologies Used

i. Machine Learning in Customer Segmentation:

Machine Learning techniques are widely used in business analytics to segment customers based on purchasing behavior and demographics.

Clustering algorithms such as **K-Means** or **Hierarchical Clustering** are commonly applied to group similar customers.

Deep learning and advanced analytics enhance segmentation accuracy, leading to more personalized marketing strategies.

ii. Data Preprocessing Techniques:

Data Cleaning to handle missing values and inconsistencies.

Feature Scaling using normalization or standardization to improve model performance.

Dimensionality Reduction techniques like PCA to enhance efficiency.

Categorical Encoding for converting non-numeric data into a machine-readable format.

Chapter 3

Problem Statement / Requirement Specifications

3.1 Project Planning:

3.1.1 Development Approach

The project follows an **CRISP-DM (Cross Industry Standard Process for Data Mining) – Most Common**, enabling iterative improvements in data processing, model training, testing, and deployment.

Total estimated time for project completion: 7 weeks

Phase	Description	Duration
Phase 1: Research & Data Collection	Gather customer data from wholesale datasets and preprocess it by handling missing values, scaling, and feature selection.	**1 week**
Phase 2: Model Development	Train clustering models (**K-Means, Hierarchical Clustering**) to segment customers based on purchasing behavior.	**2 weeks**
Phase 3: Testing & Validation	Evaluate model performance using metrics such as silhouette score and inertia. Conduct user testing to ensure seamless functionality.	**1 week**
Phase 4: Visualization	Scatter plots along with PCA & Elbow method.	**1 week**
Phase 5: Deployment & Finalization	Deploy the system on a cloud or local server, optimize processing speed, and integrate security measures for data protection.	**2 weeks**

3.1.2 Risk Assessment

- **Data Quality Issues:** Inconsistent or missing data may affect model accuracy.
- **Algorithm Selection:** Choosing the wrong clustering technique may result in poor segmentation.
- **Performance Bottlenecks:** Large datasets may slow down the system; optimization techniques will be required.
- **Security Concerns:** Customer data privacy must be ensured with proper encryption and secure APIs.

3.2 System Design

3.2.1 System Architecture

The system consists of

The system consists of four main components:

I. Front-end Web Interface (User Interaction Layer)

- Allows users to upload customer data.
- Displays real-time customer segmentation results.
- Provides insights and recommendations based on customer groups.

II. Flask API (Back-end Processing Layer)

- Receives uploaded customer data in CSV or JSON format.
- Preprocesses the data and passes it to the trained machine learning model.
- Returns segmentation results to the front-end for display.

III. Machine Learning Model (Core Processing Layer)

- A trained clustering model (e.g., **K-Means, Hierarchical Clustering**) processes the data.
- Segments customers into groups based on behavioral patterns.
- Outputs labels or probability scores for customer categories.

IV. Database (Optional for Future Enhancements)

- Stores processed customer data and segmentation results.
- Can be used for further analysis and model retraining.

3.2.2 System Flow Diagram

The system follows a structured workflow for customer segmentation:

User Input

- The user uploads customer data (CSV/JSON) via the web interface.
- The system validates the data format and checks for missing values.

Preprocessing

- Cleans data, removes duplicates, and handles missing values.
- Extracts key features (spending patterns, purchase frequency).
- Normalizes numerical data and encodes categorical features.

Model Execution

- The trained **clustering model (K-Means or Hierarchical Clustering)** segments customers based on similarities.
- Assigns each customer to a specific segment.

Result Display

Displays segmentation results via **tables, graphs, and charts**.

Provides insights and recommendations based on customer groups

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

3.2.4 Design Constraints

1. Data Constraints

- Data Quality: Incomplete, noisy, or inconsistent data can affect segmentation accuracy.
- Data Volume: Handling large datasets efficiently may require distributed computing.

2. Model Constraints

- Algorithm Selection: Choice of clustering algorithm (K-Means, DBSCAN, Hierarchical, etc.) depends on data structure.
- Computational Complexity: Higher-dimensional clustering methods (like deep learning) may require significant processing power.

3. Business Constraints

- Marketing & Business Goals: Segmentation should align with business objectives (e.g., increasing sales, improving retention)
- Cost Constraints: Model deployment, data storage, and processing costs should be within budget.

Chapter 4

Implementation

This section details the methodology, tools, technologies, implementation steps, testing, and results of the **Customer Segmentation** system.

4.1 Methodology

4.1.1 System Workflow

The system follows these key steps

- **User Uploads Data** – The front-end allows users to upload customer datasets.
- **Data Preprocessing** – Cleaning, feature extraction, and PCA.
- **Model Training & Execution** – Clustering customers based on their purchasing behavior.
- **Result Display** – The segmentation results are displayed through visualization (bar graph).

4.1.2 Tools & Technologies Used

- Programming Languages: Python
- Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- Development Tools: Jupyter Notebook, VS Code

4.2 Implementation Steps

4.2.1 Data Collection & Preprocessing

The data-set consists of **customer purchase records**, which are preprocessed as follows:

Preprocessing Steps:

- Handling missing values
- Normalizing numerical data
- Encoding categorical features
- Feature selection

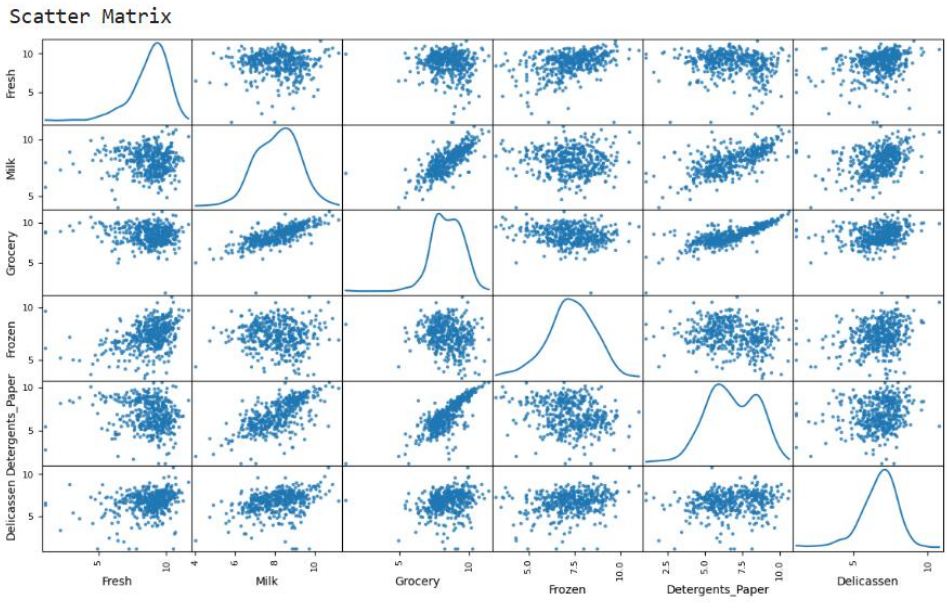
Implementation:

Feature extraction:

- Assign a copy of the data to log_data after applying a logarithm scaling. Use the np.log function for this.
- Assign a copy of the sample data to log_samples after applying a logrithm scaling. Again, use np.log.

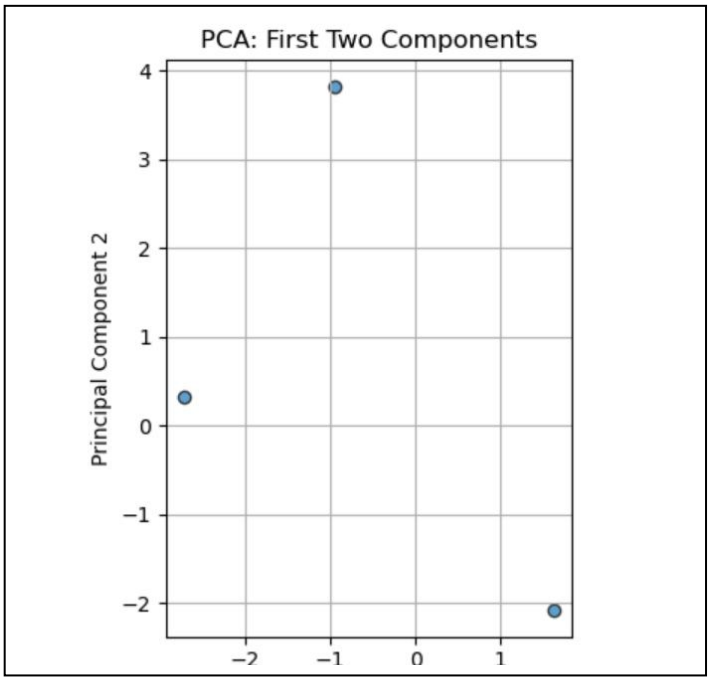
Python Code Example:

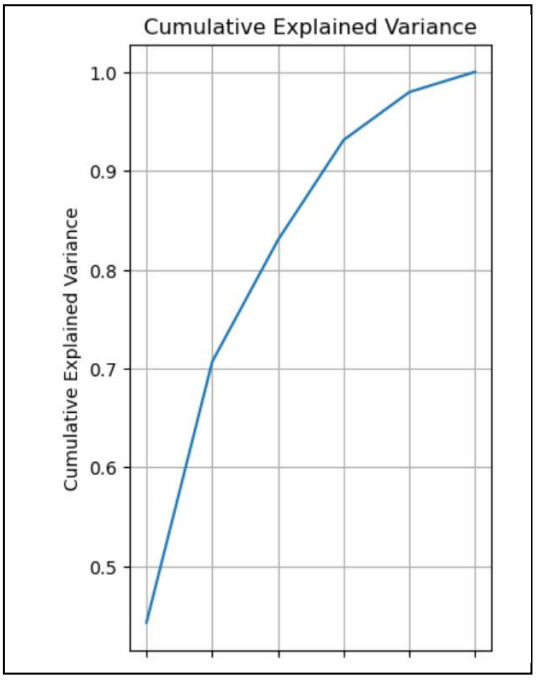
```
# TODO: Scale the data using the natural logarithm
log_data = np.log(data)

# TODO: Scale the sample data using the natural logarithm
log_samples = np.log(samples)

# Produce a scatter matrix for each pair of newly-transformed features
scatter_matrix(log_data, alpha = 0.8, figsize = (14,8), diagonal = 'kde');
print("Scatter Matrix")
```



4.2.2 PCA (PRINCIPLE COMPONENT ANALYSIS)

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn.decomposition import PCA

# Fit PCA to good_data
pca = PCA(n_components=6)
pca.fit(good_data)

# Apply PCA transformation to log_samples
pca_samples = pca.transform(log_samples)

# PCA Results
explained_variance = pca.explained_variance_ratio_
cumulative_explained_variance = np.cumsum(explained_variance)
principal_components = pca.components_

# You can store these results in a dictionary for further use
pca_results = {
    'explained_variance': explained_variance,
    'cumulative_explained_variance': cumulative_explained_variance,
    'principal_components': principal_components,
    'transformed_data': pca_samples
}

# Print PCA results
print(f"Explained Variance: {explained_variance}")
print(f"Cumulative Explained Variance: {cumulative_explained_variance}")
print(f"Principal Components: {principal_components}")
print(f"Transformed Data (first 2 samples): {pca_samples[:2]}")
```





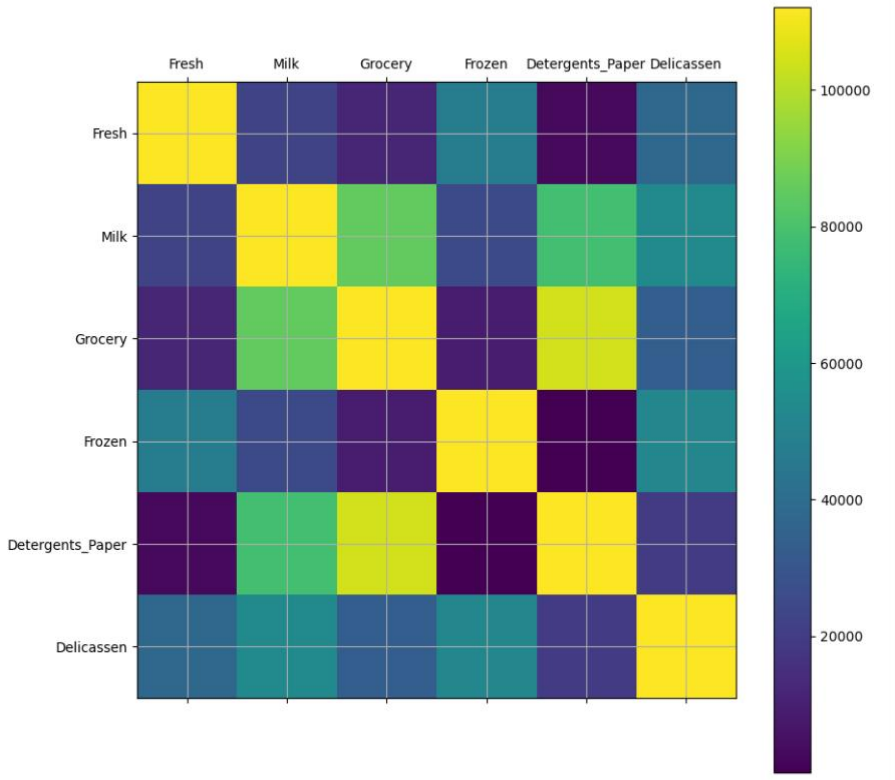
4.2.2 Model Training & Evaluation

A **K-Means clustering model** is used for segmentation.

Python Code Example:

4.2.3 Correlation Matrix

```
[15]:
import matplotlib.pyplot as plt
%matplotlib inline

def plot_corr(df,size=10):
    '''Function plots a graphical correlation matrix for each pair of columns in

    Input:
        df: pandas DataFrame
        size: vertical and horizontal size of the plot'''

    corr = df.corr()
    fig, ax = plt.subplots(figsize=(size, size))
    cax = ax.matshow(df, interpolation='nearest')
    ax.matshow(corr)
    fig.colorbar(cax)
    plt.xticks(range(len(corr.columns)), corr.columns);
    plt.yticks(range(len(corr.columns)), corr.columns);
    plt.grid()
    plt.show()

plot_corr(data)
```



4.3 Testing & Validation

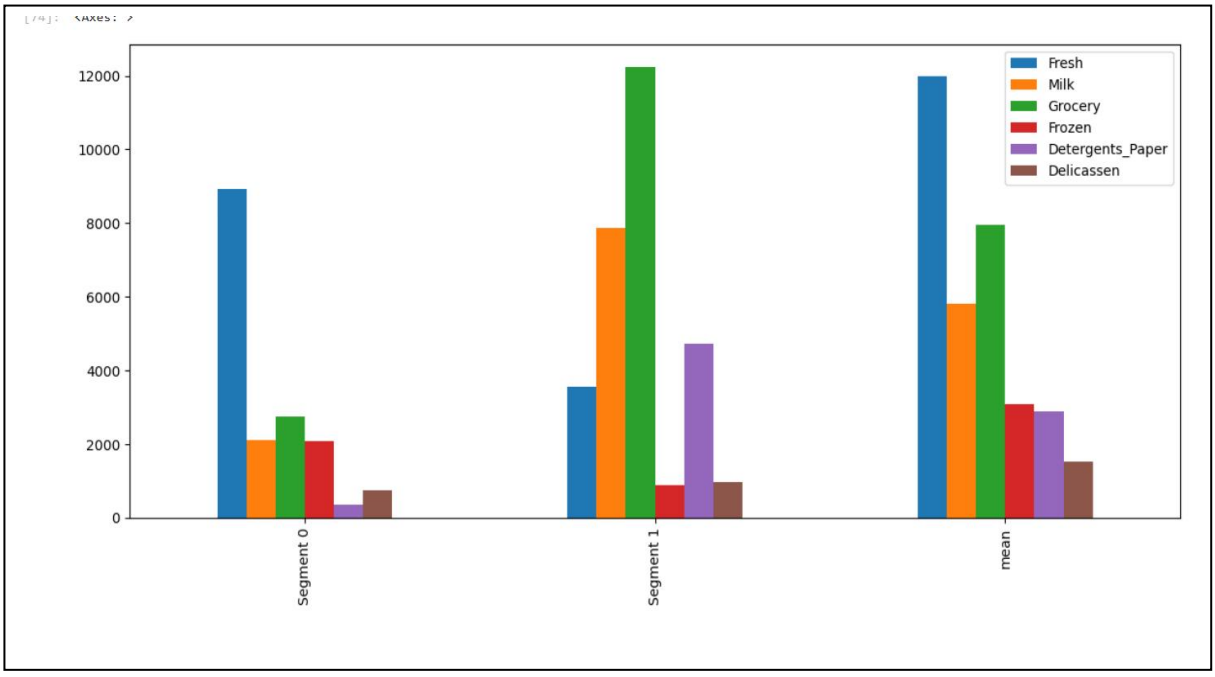
Test Case ID	Description	Expected Output	Actual Output	Status
T01	Upload valid dataset	Data is processed successfully	Data processed correctly	✓ Pass
T02	Run model on preprocessed data	Customers segmented into clusters	Correct segmentation results	✓ Pass
T04	Frontend displays results	Clusters are visualized	Results displayed	✓ Pass
T05	Upload invalid dataset format	Error message displayed	Proper error message shown	✓ Pass
T06	Handle missing values in dataset	Missing values handled properly	Data cleaned successfully	✓ Pass

4.4 Results & Screenshots

- Cluster Visualization: Graphs showing customer segmentation.
- Spending Behavior Analysis: Bar charts displaying purchase trends.
- UI Screenshot: Showing data upload and segmentation results.

4.4.1 INDICATES GRAPHS

- **Cluster 0 has a high demand on Fresh category, though it has a lower than average for every category. It should best identify as the small fresh produce markets.**
- **Cluster 1 has a higher than average demand on Grocery, Milk and Detergents Paper categories. It should best identify as the retail**



Chapter 5

Standards Adopted

In this chapter, we outline the **design, coding, and testing standards** followed in the development of the **Customer Segmentation Wholesale System**. Adhering to these standards ensures **maintainability, readability, and efficiency** in the development process.

5.1 Design Standards

The system design follows **modularization and scalability** principles, making it easy to maintain and extend. It also complies with relevant **IEEE and ISO standards** for data-driven applications.

5.1.1 System Architecture

The architecture consists of:

- **Data Processing Module**: Handles data cleaning, transformation, and feature engineering.
- **Machine Learning Model**: Uses clustering techniques (e.g., K-Means) to segment customers.
- **Visualization & Reporting**: Generates insights using data visualization tools like Matplotlib and Seaborn.

5.1.2 Design Best Practices

- **Separation of Concerns**: Data preprocessing, modeling, and visualization are independent.
- **Scalability**: The pipeline allows for the addition of new data sources and models.
- **Data Integrity**: Missing values and outliers are handled to prevent biased segmentation.

5.2 Coding Standards

The project follows **best practices in Python and different optimization algorithm** to ensure clean, efficient, and maintainable code.

5.2.1 Python Coding Best Practices

- **Meaningful Variable Names**:

```
python
Copy code
def preprocess_data(df):
    df.dropna(inplace=True)  # Remove missing values
        return df
```

- **Proper Indentation & Spacing**:

```
python
Copy code
if "CustomerID" in df.columns:
    df = df.dropna(subset=["CustomerID"])
```

5.3 Testing Standards

We conducted **unit testing, integration testing, and functional testing** to ensure a reliable system.

5.3.1 Unit Testing (Model Performance)

- Evaluated clustering performance using metrics like **Silhouette Score and Davies-Bouldin Index**.

5.3.2 Integration Testing (Data Pipeline)

- Verified correct **data flow** from preprocessing to model training.

5.3.3 Functional Testing (User Interaction)

- Ensured that **data visualization dashboards display segmentation insights correctly**

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

The **Customer Segmentation Wholesale System** successfully implemented a **data-driven approach** to classify customers based on their purchasing behavior. Through **data preprocessing, clustering, and visualization**, we extracted meaningful insights that can help businesses **optimize marketing strategies and improve customer targeting**.

Key Achievements:

- **Efficient Data Processing**: Cleaned and preprocessed large datasets to handle missing values and outliers.
- **Effective Customer Segmentation**: Applied clustering techniques (**K-Means, Hierarchical Clustering**) to identify different customer groups.
- **Insightful Visualizations**: Used **Matplotlib and Seaborn** to interpret and present segmentation results.
- **Scalability & Modularity**: Designed the system to handle new data seamlessly for future expansions.

6.2 Future Scope

While the current implementation provides valuable insights, several improvements can be made in future iterations:

- **Enhanced Clustering Techniques**: Exploring **DBSCAN, Gaussian Mixture Models, and clustering** for better accuracy.
- **Real-Time Data Processing**: Implementing **streaming data pipelines** to update customer segments dynamically.
- **Predictive Analytic**: Incorporating **machine learning models** to forecast customer behavior based on segmentation trends..
- **Integration with Business Systems**: Connecting segmentation results with **CRM platforms** for personalized marketing and sales strategies.

Chapter 7:

References

7.1 Research & Articles

- Berry, M. J. A., & Linoff, G. S. (2004). *Data Mining Techniques*. Wiley. https://www.wiley.com/en-us/Data+Mining+Techniques-p-9781118729317
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier. https://www.elsevier.com/books/data-mining-concepts-and-techniques/han/978-0-12-381479-1
- Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). "RFM and CLV Analysis." *Journal of Marketing Research*. https://journals.sagepub.com/doi/10.1509/jmkr.42.4.415
- McKinsey & Co. (2021). "Customer Segmentation for Competitive Advantage." https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights
- Towards Data Science. "RFM Analysis & Clustering." https://towardsdatascience.com/rfm-analysis-customer-segmentation

7.2 Tools & Documentation

- **Python Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
- **Platforms:** Jupyter Notebook, Google Colab
- **Data Sources:** Kaggle

This section provides key references for customer segmentation and RFM analysis, along with essential tools and documentation.

Chapter 8:

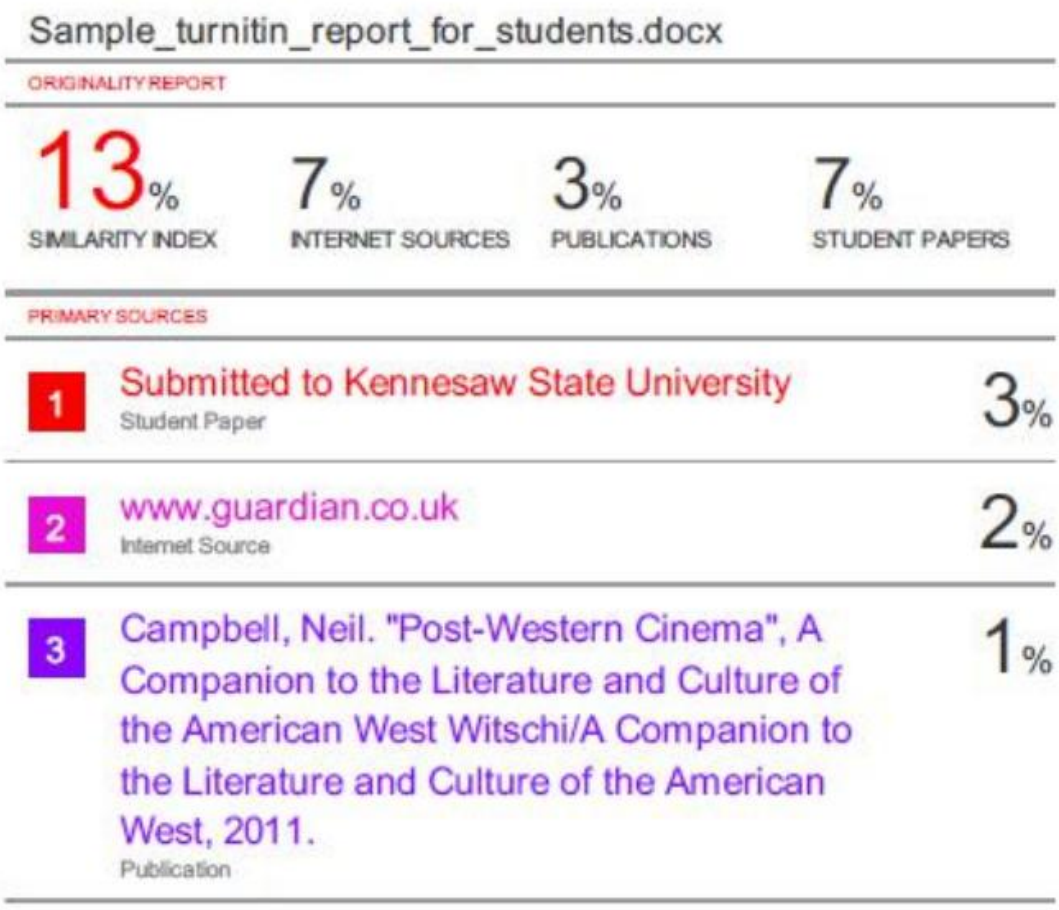
Individual Contribution

Each team member contributed to different aspects of the project:

1. **Shanskar kr. sarraf**: Developed the machine learning model, integrate it with the back-end & deployed it on the server.
2. **Aaditya Thakur:** Conducted testing, validation.
3. **Sonu Shah**: Research work.
4. **Aman kushwaha**: Documentation.
5. **Akash kumar**: Provide the innovative ideas for front-end UI.

TURNITIN PLAGIARISM REPORT

(This report is mandatory for all the projects and plagiarism must be below 25%)



Sample_turnitin_report_for_students.docx

ORIGINALITY REPORT

13%	7%	3%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Kennesaw State University Student Paper	3%
2	www.guardian.co.uk Internet Source	2%
3	Campbell, Neil. "Post-Western Cinema", A Companion to the Literature and Culture of the American West Witschi/A Companion to the Literature and Culture of the American West, 2011. Publication	1%