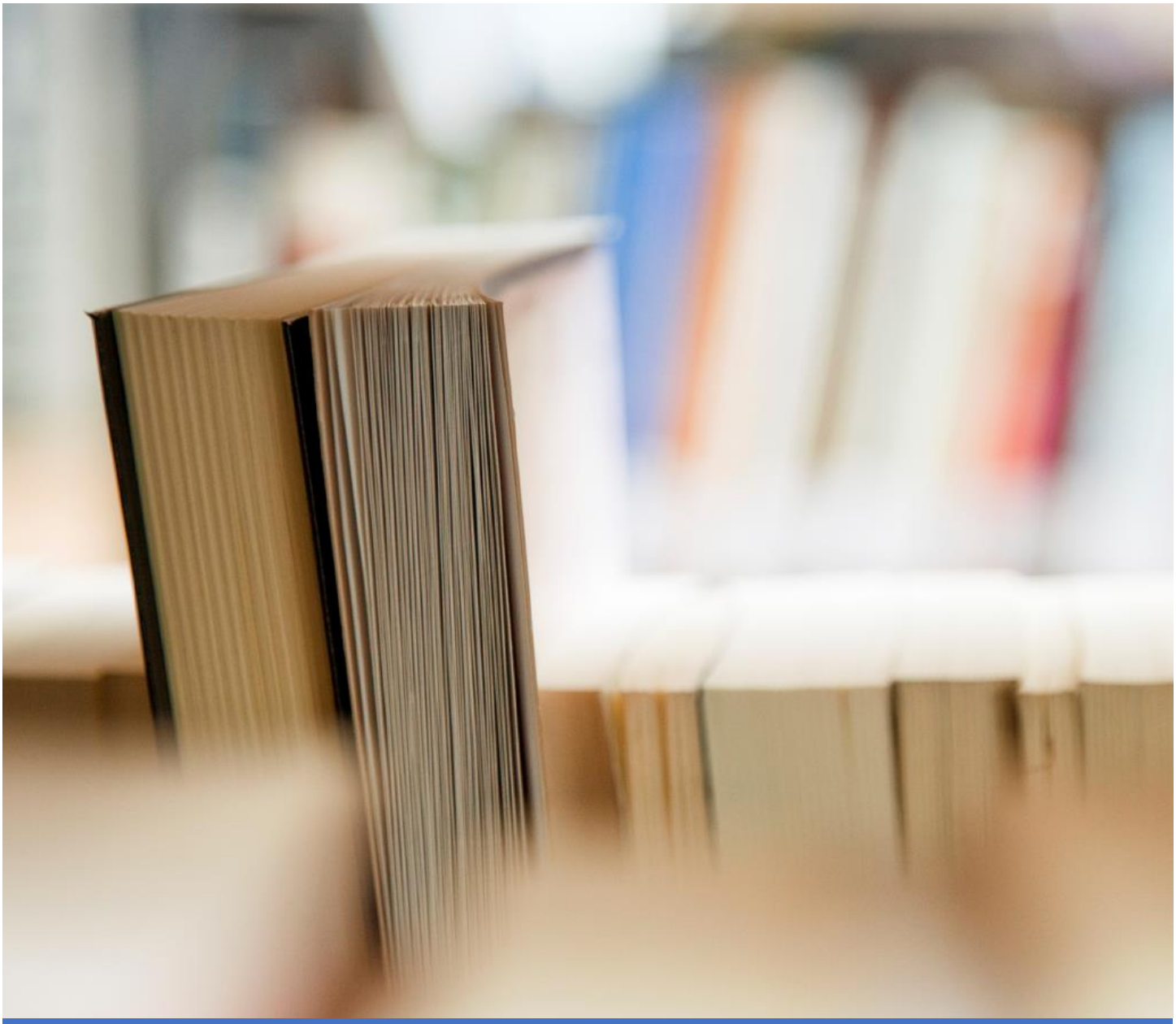


Wholesale Customer Segmentation Analysis

PROJECT REPORT

TEAM-592110

11/22/23

1. INTRODUCTION

1.1 Project Overview:

This project focuses on analyzing the spending behaviour of wholesale customers with the goal of uncovering patterns and identifying growth opportunities. The dataset contains annual spending data (in monetary units) across various product categories, such as fresh, milk, grocery, frozen, detergents and paper, and delicatessen products. Additionally, it includes information on the customer's channel (hotel/restaurant/cafe or retail) and region (Lisbon, Oporto, or other).

The primary objectives of this project are:

1. **Segmentation of Customers: Identify distinct customer segments based on their spending behaviours. By doing so, the project aims to reveal unique patterns and characteristics within the dataset.

2. **Insight Generation: Provide insights into the identified customer segments. Understanding the specific needs and preferences of each segment can guide wholesale businesses in tailoring their marketing strategies and product offerings.

3. **Predictive Modeling: Develop predictive models to categorize new customers into the identified segments. This can assist businesses in real-time decision-making and customization of services based on predicted spending behaviours.

Potential applications of the project outcomes include:

- Targeted Marketing Strategies: Tailor marketing campaigns to specific customer segments, optimizing the effectiveness of promotional efforts.

- Product Offerings: Adjust product offerings based on the preferences of different customer segments, ensuring a more personalized and appealing selection.

- Resource Allocation: Efficiently allocate resources by understanding the varying needs of different customer groups.

By undertaking this analysis, wholesale businesses can gain actionable insights, enhancing their ability to adapt and cater to the diverse spending behaviours of their customers.

1.2 Purpose:

Wholesale customer segmentation is a strategic business practice aimed at dividing a wholesale customer base into distinct groups based on shared characteristics, behaviors, and needs. The primary purpose of this segmentation is to enable businesses to tailor their marketing, sales, and service efforts more effectively to meet the unique requirements of each customer segment. By understanding the diverse needs and preferences of different customer groups, wholesalers can optimize their product offerings, pricing strategies, and communication channels. This approach allows wholesalers to enhance customer satisfaction, build stronger relationships, and ultimately drive increased sales and profitability. Through segmentation, wholesalers can identify high-value customers, allocate resources more efficiently, and develop targeted strategies that resonate with specific market segments, contributing to long-term business success in a competitive marketplace.

2. LITERATURE SURVEY

2.1 Existing problem:

One existing problem in wholesale customer segmentation is the dynamic and evolving nature of customer preferences and behaviors. Market conditions, consumer trends, and economic factors are constantly changing, making it challenging for wholesalers to maintain accurate and up-to-date customer segments. Static segmentation models may become outdated quickly, leading to ineffective targeting and missed opportunities. Additionally, some wholesalers face difficulties in accessing comprehensive and real-time data to inform their segmentation strategies. Incomplete or inaccurate data can result in misidentified customer segments, leading to suboptimal marketing efforts and resource allocation. Moreover, there is a risk of overlooking emerging customer segments or failing to adapt to shifting market dynamics. Overcoming these challenges requires a continuous and data-driven approach to customer segmentation, incorporating

advanced analytics, machine learning, and regular reassessment to ensure the segmentation strategy remains aligned with the evolving needs and behaviors of wholesale customers.

2.2 References:

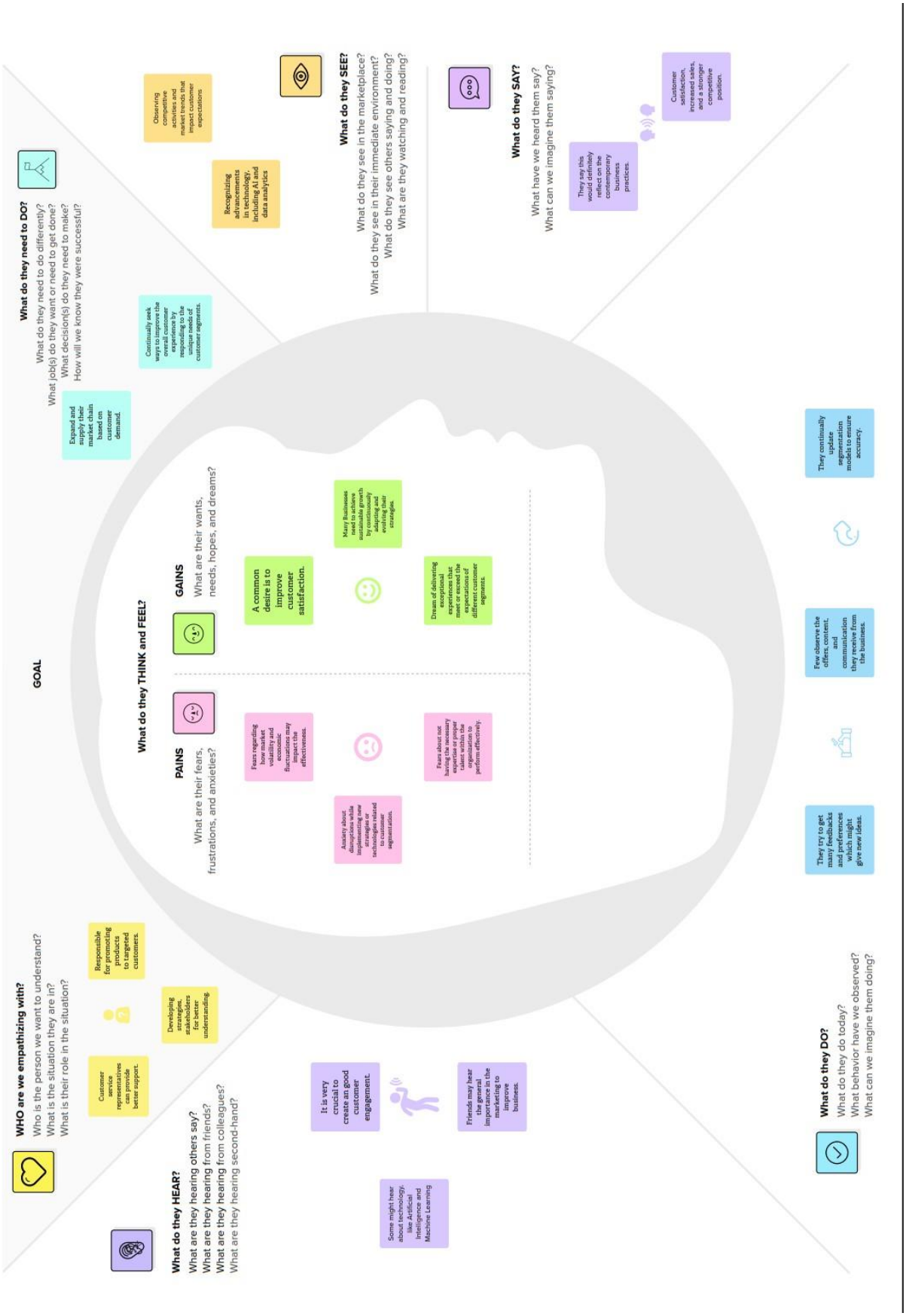
- https://ijsrst.com/paper/8152.pdf
- https://www.kaggle.com/datasets/aggle6666/wholesale-customer-segmentation-dataset/discussion
- https://rpubs.com/NikRoy/778968
- https://www.researchgate.net/profile/MehnazTabassum/publication/323771551_Comparative_Performance_Of_Using_PCA_With_KMeans_And_Fuzzy_C_Means_Clustering_For_Customer_Segmentation/links/5aa9f8680f7e9b88266f6e35/Comparative-Performance-Of-Using-PCA-With-K-Means-And-Fuzzy-C-Means-Clustering-For-Customer-Segmentation.pdf

2.3 Problem Statement Definition:

The problem statement for wholesale customer segmentation revolves around the difficulty in effectively categorizing and understanding the diverse needs and behaviors of wholesale customers. Wholesalers encounter challenges in developing and maintaining accurate customer segments due to the dynamic nature of market conditions, evolving consumer preferences, and the lack of real-time, comprehensive data. This dynamic landscape makes it challenging to create static segmentation models that remain relevant over time. Incomplete or inaccurate data can lead to misidentified customer segments, resulting in suboptimal marketing strategies and resource allocation. Additionally, there's a risk of overlooking emerging customer segments and failing to adapt to shifting market dynamics. The overarching issue is the need for a more agile and data-driven approach to wholesale customer segmentation to ensure that strategies align with the ever-changing demands of the market and customer base.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas



What do they THINK and FEEL?

PAINS

What are their fears, frustrations, and anxieties?

Fears regarding how market volatility and economic fluctuations may impact the effectiveness.

Anxiety about disruptions while implementing new strategies or technologies related to customer segmentation.

Fears about not having the necessary expertise or proper talent within the organization to perform effectively.

GAINS

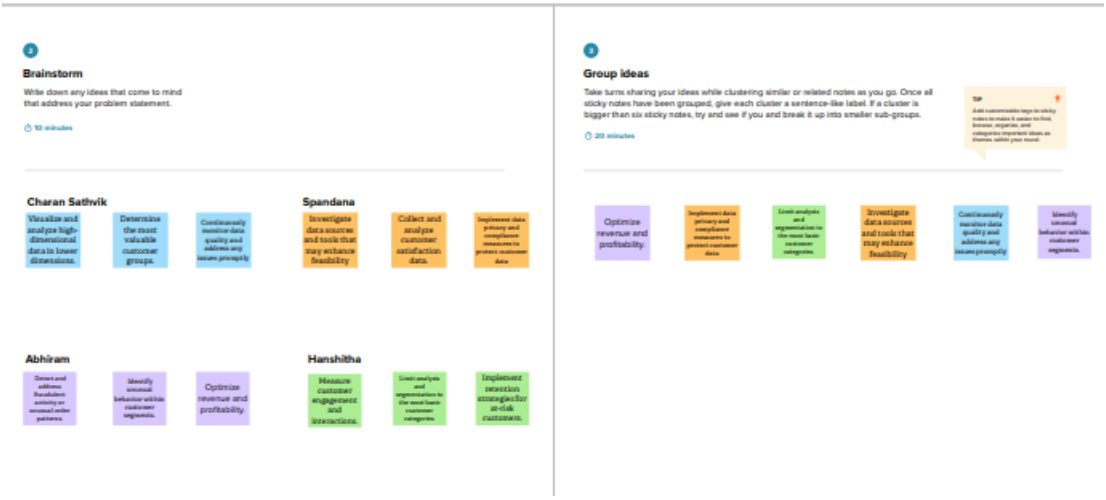
What are their wants, needs, hopes, and dreams?

A common desire is to improve customer satisfaction.

Many Businesses need to achieve sustainable growth by continuously adapting and evolving their strategies.

Dream of delivering exceptional experiences that meet or exceed the expectations of different customer segments.

3.2 Ideation & Brainstorming



4. REQUIREMENT ANALYSIS

4.1 Functional requirement:

Developing an effective wholesale customer segmentation system necessitates a clear set of functional requirements to ensure its capability to meet the diverse needs of the business and its clientele. Firstly, the system must excel in customer data management, enabling the collection, storage, and comprehensive management of customer information, including purchase history and behavioral patterns. Advanced segmentation algorithms form another critical requirement, allowing the system to categorize customers based on criteria such as order frequency, geographical location, and demographics.

Real-time data updates are imperative to keep customer segments current, ensuring that the system remains responsive to evolving market conditions and consumer behaviors. Customizable segmentation criteria are essential, providing flexibility to tailor the system to the specific goals and preferences of the wholesale business. Integration with CRM systems is crucial for maintaining consistency across different business applications and facilitating seamless data synchronization.

To safeguard sensitive customer data, the system must incorporate robust user access controls and security measures, employing role-based permissions. Reporting and analytics capabilities are essential for generating insights into the performance of different customer segments, facilitating data-driven decision-making. Scalability is a key requirement to accommodate the growing volume of customer data and adapt to the expanding needs of the wholesale business.

4.2 Non-Functional requirements:

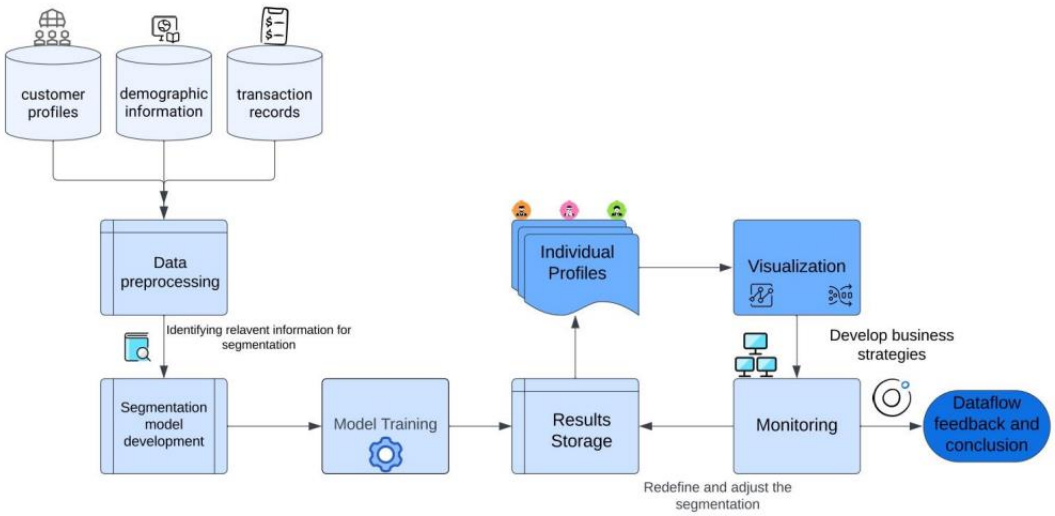
The successful development and deployment of a wholesale customer segmentation system hinge not only on its functional capabilities but also on a set of critical non-functional requirements that dictate how the system should perform and behave. Firstly, performance is paramount, demanding quick response times for data queries and updates to ensure a seamless and efficient user experience. Scalability follows closely, necessitating the system's ability to handle growing volumes of customer data and increased user demands as the business expands.

Reliability and availability are crucial non-functional requirements, minimizing downtime and ensuring continuous access to the system's segmentation capabilities. Security measures must be robust, safeguarding sensitive customer data against unauthorized access and data breaches, while also ensuring compliance with relevant data protection regulations. Usability is another key consideration, requiring an intuitive and user-friendly interface that accommodates users with varying levels of technical expertise.

Compatibility is vital, necessitating the system's ability to work seamlessly across different devices and platforms, promoting accessibility and facilitating integration with other existing business applications. Maintainability is equally important, with a need for clear documentation and a modular design that simplifies future enhancements and updates. Interoperability ensures smooth integration with other systems and technologies, fostering data exchange and synergy between different business applications.

5. PROJECT DESIGN

5.1 Data Flow Diagrams & User Stories:



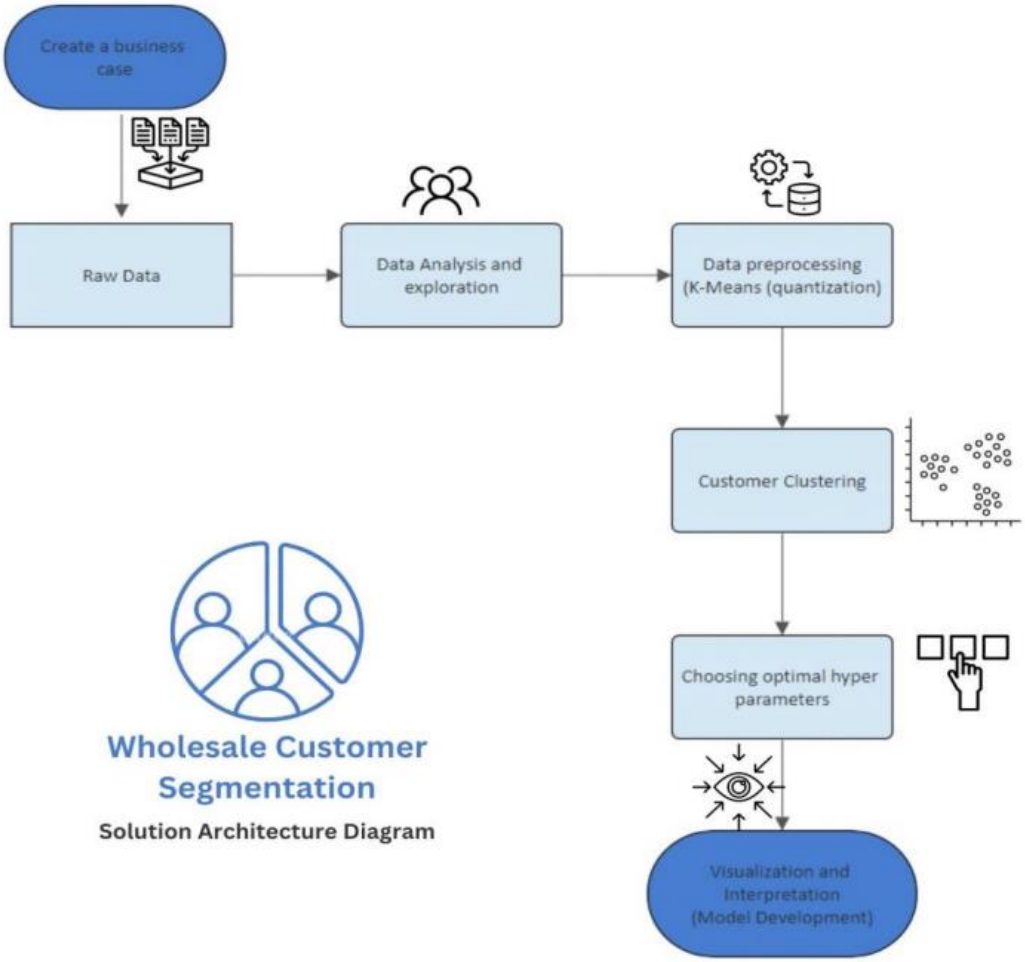
User Stories

Use the below template to list all the user stories for the product.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Governmen t Agencies	Project setup &Infrastructure	USN-1	By defining the objectives and using relevant customer data we setup the infrastructure for wholesale customer segmentation	Initalized the all the necessary aspects that required	High	Sprint-1
Consultants	Developing environment	USN-2	The main focus is on data collection, we will be able to classify the groups. It will be able to customize the current marketing strategies.	Understood the need for development of new marketing strategies	High	Sprint-1
customers	Data Collection	USN-3	Gather data on customer purchases, such as the types and quantities of products they buy over a period of time	Collected the dataset of customers in a particular region.	High	Sprint-1
E-commerce Retailers	Data Preprocessing:	USN-4	Clean and preprocess the data by handling missing values, scaling features if necessary, and removing outliers	Preprocessed the data for reducing overfitting problems.	Medium	Sprint-2
Management and decision makers	Feature Selection:	USN-5	Identify the relevant features for segmentation. In this case, it could be the product categories, purchase frequency, or total spending.	Created a meaningful segment to specific business goals.	Medium	Sprint-2
Healthcare Providers	Model Building or selection	USN-6	By selecting the number of clusters (K) through methods like the elbow method or silhouette score. Then, initialize K centroids, assign each customer to the nearest centroid based on their features, and recalculate centroids by taking the mean	Selected and performed the required algorithm that required	High	Sprint -3
Retailers	Model evaluation	USN-7	Analyze these clusters to understand the purchasing patterns and characteristics of each segment.	We could test the scalability.	Medium	Sprint -3
	Marketing strategies	USN-8	Develop specific marketing strategies for each customer segment based on their preferences and behavior	We could implement many more business and marketing strategies based on the group.	Low	Sprint -4

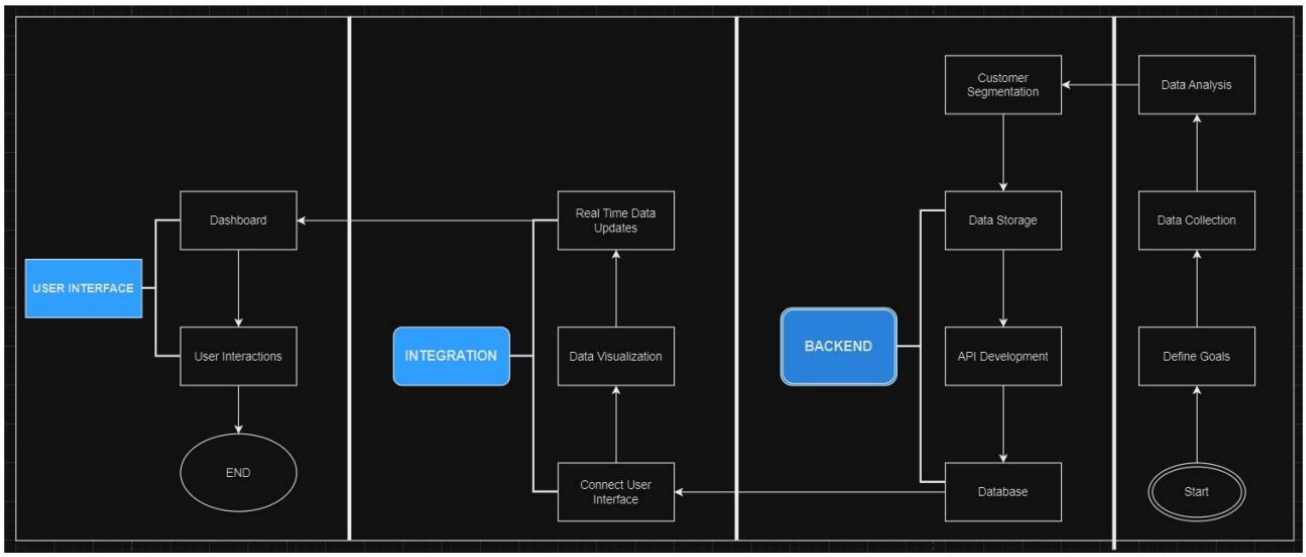
5.2 Solution Architecture:

Solution Architecture Diagram for Wholesale Customer Segmentation



6. PROJECT PLANNING & SCHEDULING

6.1 Technical Architecture:

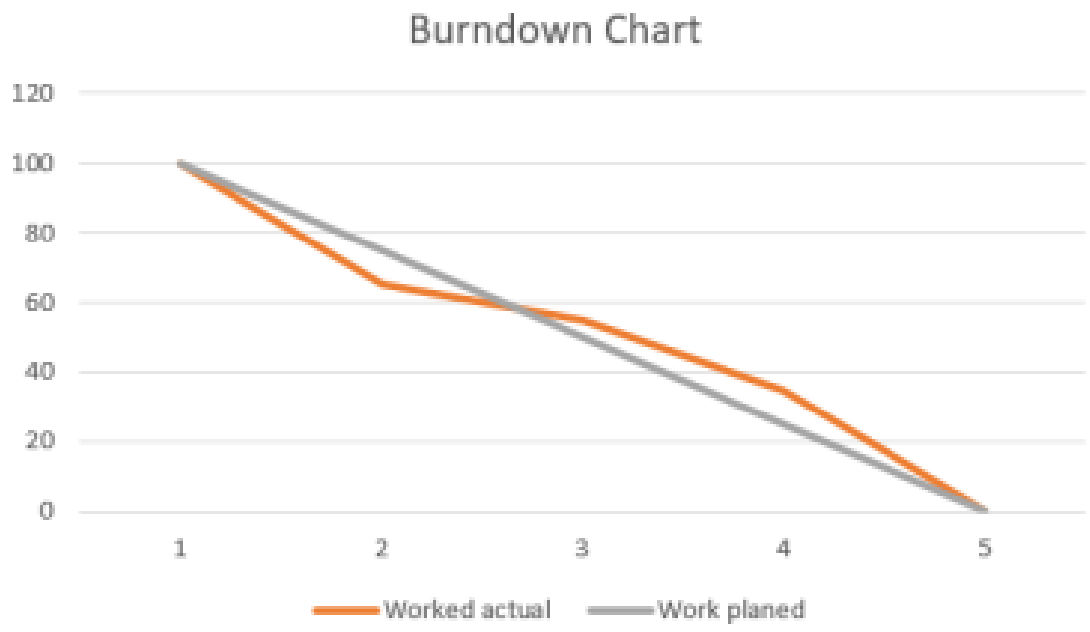


6.2 Sprint Planning & Estimation:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Project setup & Infrastructure	USN-1	By defining the objectives and using relevant customer data we setup the infrastructure for wholesale customer segmentation	1	High	Charan Sathvik
Sprint-1	Development environment	USN-2	The main focus is on data collection, we will be able to classify the groups. It will be able to customize the current marketing strategies.	2	High	Abhiram
Sprint-1	Data collection	USN-3	Gather data on customer purchases, suchas the types and quantities of products they buy over a period of time	2	High	Hanshitha

Sprint-2	data preprocessing	USN-4	Clean and preprocess the data by handlingmissing values, scaling features if necessary, and removing outliers	3	Medium	Spandana
Sprint-3	feature selection:	USN-5	Identify the relevant features for segmentation. In this case, it could be the product categories, purchase frequency, or total spending.	4	Medium	Spandana
Sprint-3	Model Building or selection	USN-6	By selecting the number of clusters (K) through methods like the elbow method or silhouette score. Then, initialize K centroids,assign each customer to the nearest centroid based on their features, and recalculate centroids by taking the mean	6	High	Abhiram
Sprint-3	Model evaluation	USN-7	Analyze these clusters to understand the purchasing patterns and characteristics of each segment.	1	Medium	Charan Sathvik
Sprint-4	Marketing strategies	USN-8	Develop specific marketing strategies for each customer segment based on their preferences and behavior	1	Low	Hanshitha

6.3 Sprint Delivery Schedule:



7. CODING & SOLUTIONING

7.1 Feature 1:

We train our data ican train our data on different algorithms. For this project we are applying Three classification algorithms KNN, Logistic Regression and Random Forest Classifier. The best model is saved based on its performance.

KNN

```
alpha = [5, 11, 15, 21, 31, 41, 51, 99]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = KNeighborsClassifier(n_neighbors=i)
    clf.fit(train_df, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_df, y_train)
    sig_clf_probs = sig_clf.predict_proba(cv_df)
    cv_log_error_array.append(log_loss(y_cv, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(y_cv, sig_clf_probs))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()
```

```
best_alpha = np.argmin(cv_log_error_array)
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_df, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_df, y_train)

predict_y = sig_clf.predict_proba(train_df)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_df)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_df)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

LOGISTIC REGRESSION

```
alpha = [10 ** x for x in range(-6, 3)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = SGDClassifier(class_weight='balanced', alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_df,y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_df, y_train)
    sig_clf_probs = sig_clf.predict_proba(cv_df)
    cv_log_error_array.append(log_loss(y_cv, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(y_cv, sig_clf_probs))
```

```
fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_df, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_df, y_train)

predict_y = sig_clf.predict_proba(train_df)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_df)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_df)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

RANDOM FOREST CLASSIFIER

```
alpha = [100,200,500,1000,2000]
max_depth = [5, 10]
cv_log_error_array = []
for i in alpha:
    for j in max_depth:
        print("for n_estimators =", i,"and max depth = ", j)
        clf = RandomForestClassifier(n_estimators=i, criterion='gini', max_depth=j, random_state=42, n_jobs=-1)
        clf.fit(train_df,y_train)
        sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
        sig_clf.fit(train_df,y_train)
        sig_clf_probs = sig_clf.predict_proba(cv_df)
        cv_log_error_array.append(log_loss(y_cv, sig_clf_probs, labels=clf.classes_, eps=1e-15))
        print("Log Loss :",log_loss(y_cv, sig_clf_probs))

'''fig, ax = plt.subplots()
features = np.dot(np.array(alpha)[:,None],np.array(max_depth)[None]).ravel()
ax.plot(features, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[int(i/2)],max_depth[int(i%2)],str(txt)), (features[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()
'''
```

```
best_alpha = np.argmin(cv_log_error_array)
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max_depth[int(best_alpha%2)], random_state=42, n_jobs=-1)
clf.fit(train_df,y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_df,y_train)

predict_y = sig_clf.predict_proba(train_df)
print('For values of best estimator = ', alpha[int(best_alpha/2)], "The train log loss is:",log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_df)
print('For values of best estimator = ', alpha[int(best_alpha/2)], "The cross validation log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_df)
print('For values of best estimator = ', alpha[int(best_alpha/2)], "The test log loss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

7.2 Feature 2:

We found the confusion matrix

KNN

```
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
predict_and_plot_confusion_matrix(train_df.values, y_train.values, cv_df.values, y_cv.values, clf)
```

LOGISTIC REGRESSION

```
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_df.values, y_train.values, cv_df.values, y_cv.values, clf)
```

RANDOM FOREST CLASSIFIER

```
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max_depth[int(best_alpha%2)], random_state=42, n_jobs=-1)
# predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y,cv_x_onehotCoding,cv_y, clf)
predict_and_plot_confusion_matrix(train_df.values, y_train.values, cv_df.values, y_cv.values, clf)
```

7.3 Database Schema (if Applicable):

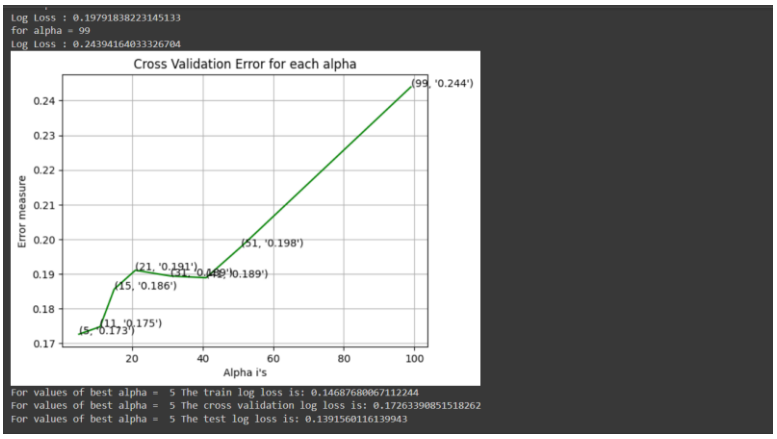
Channel

	A	B	C	D	E	F	G	H
1	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergent:	Delicasse
2	2	3	12669	9656	7561	214	2674	133
3	2	3	7057	9810	9568	1762	3293	177
4	2	3	6353	8808	7684	2405	3516	784
5	1	3	13265	1196	4221	6404	507	178
6	2	3	22615	5410	7198	3915	1777	518
7	2	3	9413	8259	5126	666	1795	145
8	2	3	12126	3199	6975	480	3140	54
9	2	3	7579	4956	9426	1669	3321	256
10	1	3	5963	3648	6192	425	1716	75
11	2	3	6006	11093	18881	1159	7425	209
12	2	3	3366	5403	12974	4400	5977	174
13	2	3	13146	1124	4523	1420	549	49
14	2	3	31714	12319	11757	287	3881	293
15	2	3	21217	6208	14982	3095	6707	60
16	2	3	24653	9465	12091	294	5058	216
17	1	3	10253	1114	3821	397	964	41
18	2	3	1020	8816	12121	134	4508	108
19	1	3	5876	6157	2933	839	370	447
20	2	3	18601	6327	10099	2205	2767	318
21	1	3	7780	2495	9464	669	2518	50
22	2	3	17546	4519	4602	1066	2259	212
23	1	3	5567	871	2010	3383	375	56
24	1	3	31276	1917	4469	9408	2381	433
25	2	3	26373	36423	22019	5154	4337	1652
26	2	3	22647	9776	13792	2915	4482	577

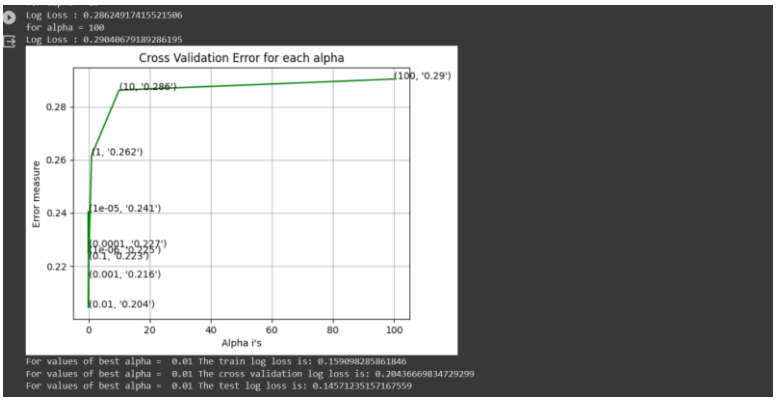
8. PERFORMANCE TESTING

8.1 Performace Metrics:

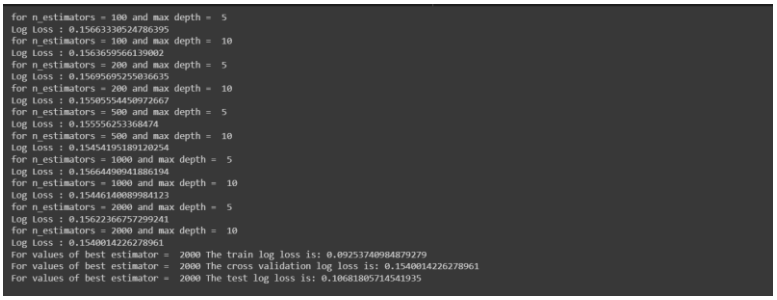
KNN



LOGISTIC REGRESSION



RANDOM FOREST

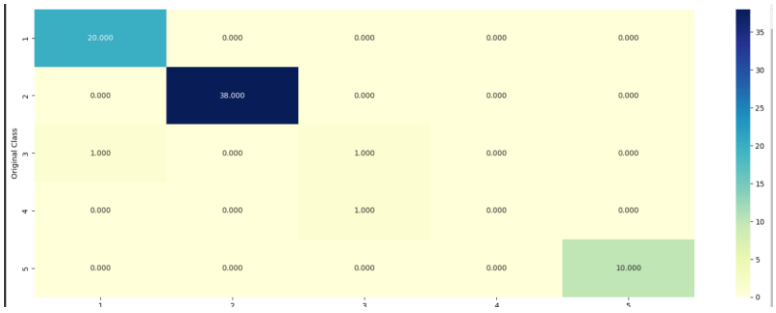


9. RESULTS

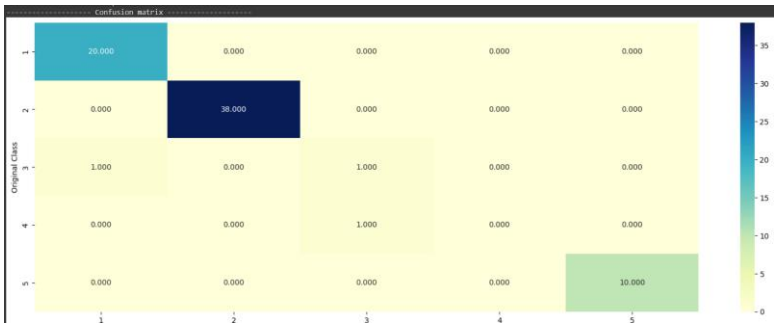
9.1 Output Screenshots

Confusion Matrix:

KNN



Logistic Regression



Random Forest Classifier

10. ADVANTAGES & DISADVANTAGES

Implementing a wholesale customer segmentation system comes with a range of advantages that can significantly impact a business's overall performance. Firstly, targeted marketing stands out as a key benefit. By understanding the unique characteristics and behaviors of different customer segments, wholesalers can tailor their marketing strategies to resonate more effectively with specific groups. This targeted approach often results in higher customer engagement and satisfaction.

Furthermore, improved customer relationships are a direct outcome of effective segmentation. When wholesalers can offer products and services that align closely with the preferences of distinct customer segments, it fosters a sense of understanding and responsiveness, leading to increased customer loyalty and repeat business.

The optimization of resource allocation is another notable advantage. With insights gained from segmentation, businesses can channel their resources more efficiently. This involves concentrating efforts on high-value customer segments, fine-tuning product offerings, pricing strategies, and marketing initiatives to maximize their impact.

In terms of financial outcomes, the implementation of a wholesale customer segmentation system has the potential to boost sales and revenue. The tailored approach to customer needs can result in a higher conversion rate as businesses capture a larger share of their target market and capitalize on growth opportunities.

Additionally, the system facilitates data-driven decision-making. Through comprehensive data analytics, businesses gain valuable insights into customer behaviors and market trends, enabling them to make informed decisions that align with real-time and accurate information, thereby improving overall operational efficiency.

11. CONCLUSION

In conclusion, the implementation of a wholesale customer segmentation system offers significant advantages, including targeted marketing, improved customer relationships, optimized resource allocation, increased sales, and data-driven decision-making. However, businesses must navigate challenges such as implementation costs, system complexity, data security concerns, and resistance to change. Striking a balance between leveraging the benefits of segmentation and addressing these challenges is essential for a successful integration that positively impacts overall business performance.

12. FUTURE SCOPE

Looking ahead, the future scope of wholesale customer segmentation is marked by the integration of advanced technologies and a strategic shift toward achieving more nuanced personalization. The incorporation of advanced analytics and artificial intelligence is poised to revolutionize segmentation practices, enabling businesses to derive deeper insights from vast datasets and offer more targeted experiences. The emphasis will not only be on segmenting customers broadly but on achieving personalization at scale, tailoring interactions to individual preferences.

Moreover, the integration of customer segmentation with emerging technologies, such as the Internet of Things (IoT) and blockchain, is expected to open new avenues for understanding customer behaviors. Real-time data from connected devices and secure, transparent data practices will contribute to a more accurate and ethical approach to segmentation.

Cross-channel integration will be a key focus, ensuring a consistent and personalized experience for customers across various platforms. Predictive analytics will play a crucial role in anticipating trends and adapting segmentation strategies proactively. Additionally, as concerns about data privacy continue to rise, businesses will need to prioritize ethical data practices, ensuring compliance with evolving regulations.

13. APPENDIX

SOURCE CODE:

https://colab.research.google.com/drive/1NTwkc6lWyTRUVhu8cF0Ql8657_LTFT96?usp=sharing

GITHUB LINK:

https://github.com/smartinternz02/SI-GuidedProject-613533-1699035540

PROJECT DEMO LINK:

https://drive.google.com/file/d/1tyAx8ILJB_sIFDXYNqT6Mv6pmMs4ny1c/view?usp=drive_link