FINAL REPORT

UNDERSTANDING AUDIENCE

A Machine Learning Approach to Customer Segmentation

TEAM NUMBER : Team - 591849

SUBMITTED TO : SmartInternz Team

BATCH **:** VIT – AP AI/ML Morning Batch

YEAR **:** 2023

TEAM MEMBERS :

Vaishnavi T Sundari Dhulipala – 21BCE9174

Vintha Kuyili Ramyasri – 21BCE9204

Lakshmi Srujana Vankayala – 21BCE9181

Deepthimahanty Harshita – 21BCE9230

1. INTRODUCTION

1.1 PROJECT OVERVIEW :

Titled "Understanding Audience," this project aims to help businesses thrive in a competitive market by identifying and engaging potential customers effectively. Management faces the challenge of sorting through a large consumer base to find customers with the highest profit potential. To tackle this, the project suggests using advanced Machine Learning models to classify customers into different groups based on various characteristics, providing a detailed understanding of their behavior and preferences.

The main parts of the project include collecting and preparing data to ensure it's reliable. Then, we'll create both guided and self-directed Machine Learning programs to accurately group customers. The goal is to seamlessly integrate these programs into existing business processes, making customer grouping automatic and efficient. The project also involves creating a decision support system that uses insights from customer grouping to improve decision-making across different levels of the organization. This strategic approach aims to boost revenue through improved offerings and marketing strategies. Recognizing that customer behavior changes over time, the project includes continuous monitoring and improvement methods to keep the grouping models relevant and effective.
In the dynamic landscape of today's business environment, the identification and understanding of potential customers play a crucial role in gaining a competitive edge. As the number of organizations in the market continues to grow, the need for effective customer segmentation becomes imperative. This project focuses on leveraging Machine Learning, Data Science, and AI techniques to classify customers into segments based on various attributes. The ultimate goal is to simplify the process of customer identification, leading to better decision-making, streamlined business processes, and the formulation of strategies to enhance overall revenue.

1.2 PURPOSE :

The purpose of this project is to utilize Machine Learning, Data Science, and AI to enhance customer segmentation within a business. The key objectives include identifying potential customers, gaining a competitive advantage, improving decision-making, optimizing business processes, formulating targeted strategies, and ultimately increasing revenue. The project emphasizes the integration of advanced technologies for accurate customer segmentation, continuous monitoring, and adaptation to evolving customer behaviors. The overarching goal is to empower the organization with tools and strategies that enhance customer engagement and contribute to long-term business success.

2. LITERATURE SURVEY

2.1 EXISTING PROBLEM :

In the rapidly evolving business landscape, effective customer segmentation is essential for personalized marketing, improved customer satisfaction, and overall business success. However, traditional segmentation methods face several challenges. Firstly, these methods often struggle to handle the diversity and unstructured nature of contemporary data sources, including social media, online interactions, and multimedia content. Secondly, conventional approaches may fall short in capturing real-time consumer trends, limiting their ability to adapt to dynamic market changes. Finally, the lack of personalized segmentation hinders businesses from tailoring their products and services to individual customer preferences.

2.2 REFERENCES :

- A Case Study on Customer Segmentation by using Machine Learning Methods
  https://ieeexplore.ieee.org/abstract/document/8620892/

  In summary, the literature survey underscores the significance of customer segmentation in CRM, particularly addressing the challenges of manual segmentation in a company dealing with a vast customer database. The study advocates for machine learning solutions, focusing on the analysis of real customer payment data.

  The literature unfolds with an introduction to the importance of customer segmentation, emphasizing the need for automated identification of premium customers. The subsequent sections detail the customer data, highlighting challenges, and introduce three machine learning methods—Normal Equation Method (NEM), Multivariate Linear Regression Method (LiRM), and Logistic Regression Method (LoRM).

  Results reveal that logistic regression outperforms NEM and LiRM with an efficiency of 89.43%, suggesting its suitability for automating customer categorization. This literature survey contributes valuable insights for implementing machine learning in CRM, optimizing decision-making in customer-centric business practices.

- Customer Segmentation using Machine Learning
  https://www.academia.edu/download/63796230/34420200701-61676-l1hfm.pdf

The paper underscores the strategic importance of customer segmentation in the face of product competition. It advocates for meticulous analysis of customer needs and the adoption of machine learning-driven clustering techniques, particularly the K-means algorithm, hierarchical clustering, and density-based clustering. These methodologies, rooted in data mining, contribute to achieving strategic goals in diverse industries.

The discussion highlights the diversity among customers and the significance of clustering parameters, including geographic, demographic, psychographic, and behavioral factors. Predictive analytics for forecasting future customer behaviors adds a forward-looking dimension to segmentation. The methodology section introduces Customer Relationship Management (CRM) as integral to modern marketing, enhancing customer satisfaction and loyalty.
The paper concludes by emphasizing the transformative impact of integrating data science and artificial intelligence into customer segmentation. It celebrates the project's 98% accuracy milestone and outlines future scopes, including deep learning integration and enhanced user interfaces. The legacy of the paper lies in its contribution to sustainable, customer-centric business practices through technological innovation.

- Customer Segmentation using K-means Clustering
  https://ieeexplore.ieee.org/abstract/document/8769171/

  The paper explores the use of clustering algorithms (k-Means, Agglomerative, Meanshift) for customer segmentation in the face of modern business challenges. It addresses the difficulty businesses encounter in targeting specific customer segments amidst a plethora of products. The algorithms are applied to a local retail dataset, identifying segments like Careless, Careful, Standard, Target, Sensible, High Buyer Frequent Visitors, and High Buyer Occasional Visitors. The Elbow Method, Dendrograms, and bandwidth are discussed for algorithm parameterization. Internal clustering validation measures, such as silhouette score, are used for comparison. The paper aims to help businesses adapt to the competitive market through machine learning-driven customer understanding. Published in the 2018 CTEMS conference, it emphasizes the significance of machine learning in refining marketing strategies and enhancing customer satisfaction.

- E-commerce Customer Segmentation via Unsupervised Machine Learning
  https://dl.acm.org/doi/abs/10.1145/3448734.3450775

This study addresses the need for systematic customer segmentation in e-commerce using unsupervised machine learning. The raw transaction data from an online platform is processed through the RFM model and TF-IDF method to create behavioral features and product categories. K-means clustering groups customers, and association rules mining analyzes purchased products. Principal Component Analysis (PCA) and T-Distributed Stochastic Neighbor Embedding (T-sne) reduce dimensionality for visualization. The results offer insights into customer clusters, helping formulate targeted marketing strategies. The paper demonstrates the application of machine learning in enhancing customer-oriented business practices in the e-commerce sector.

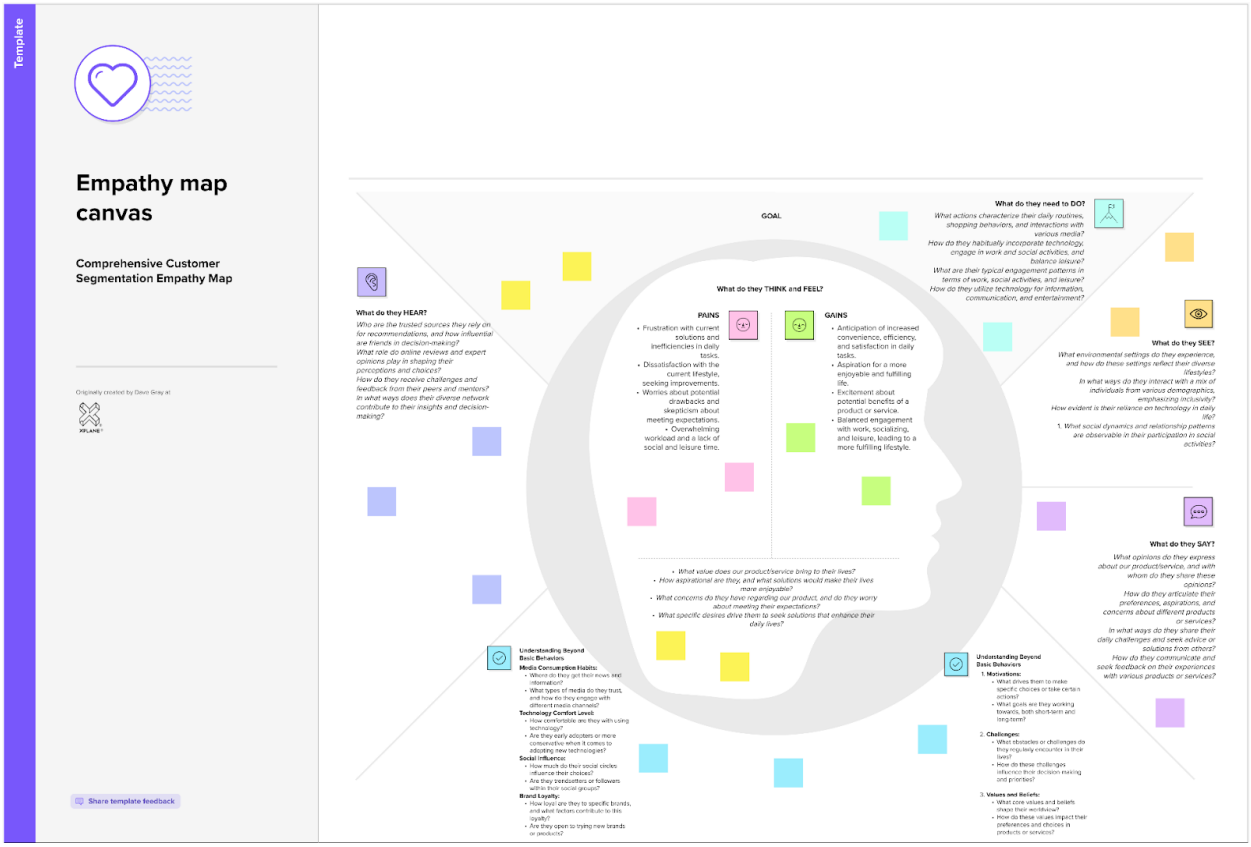
2.3 PROBLEM STATEMENT DEFINITION :

To address the limitations of traditional customer segmentation, the project "Understanding Audience" proposes the integration of cutting-edge machine learning techniques. The project aims to develop models capable of handling diverse and unstructured data, enabling more accurate and real-time customer segmentation. By leveraging machine learning algorithms, the project seeks to unlock the potential for personalized segmentation, allowing businesses to tailor their offerings to the unique preferences and behaviors of individual customers.

The anticipated impact of this project includes improved decision-making processes, more effective marketing strategies, and an enhanced understanding of the customer base. By harnessing the power of Data Science and Artificial Intelligence, "Understanding Audience" aspires to empower businesses to thrive in the ever-changing and competitive market landscape.

3. IDEATION & PROPOSED SOLUTION

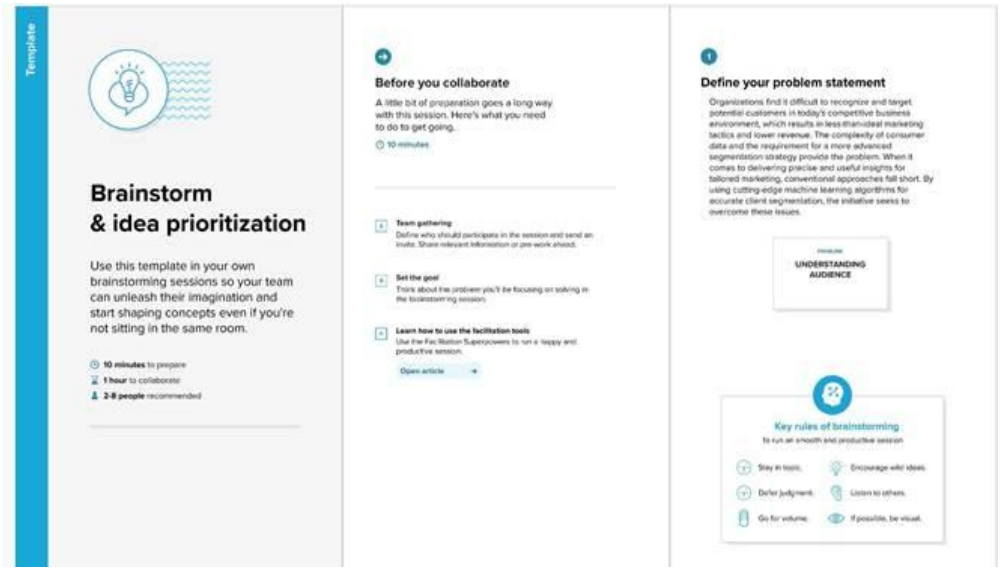
3.1 EMPATHY MAP :



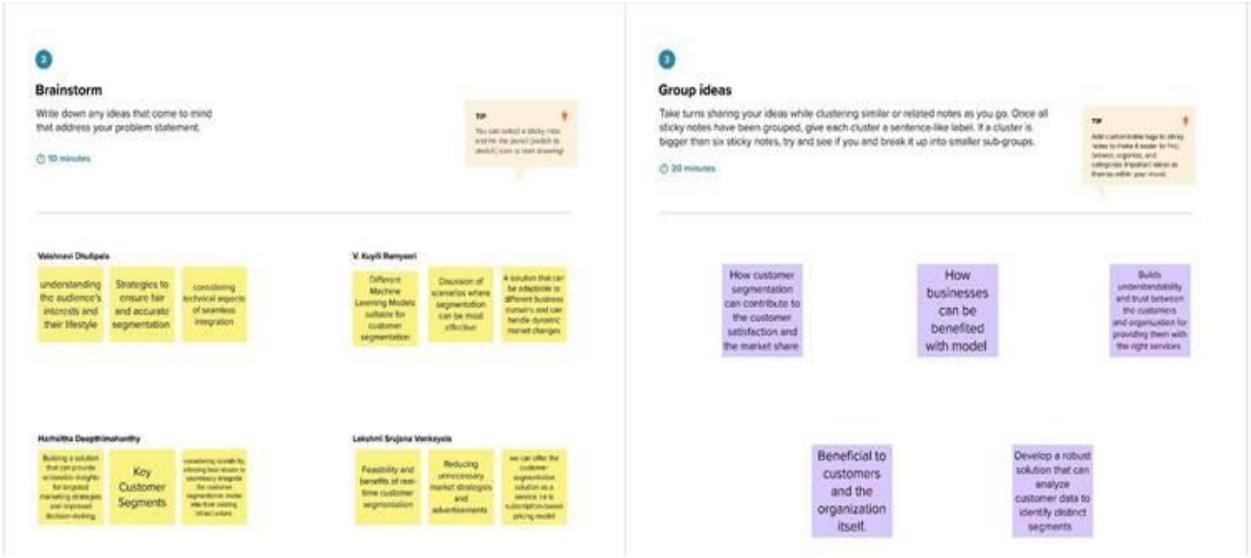
3.2 BRAIN-STORMING AND PROPOSED SOLUTION :

3.2.1 BRAIN-STORMING :

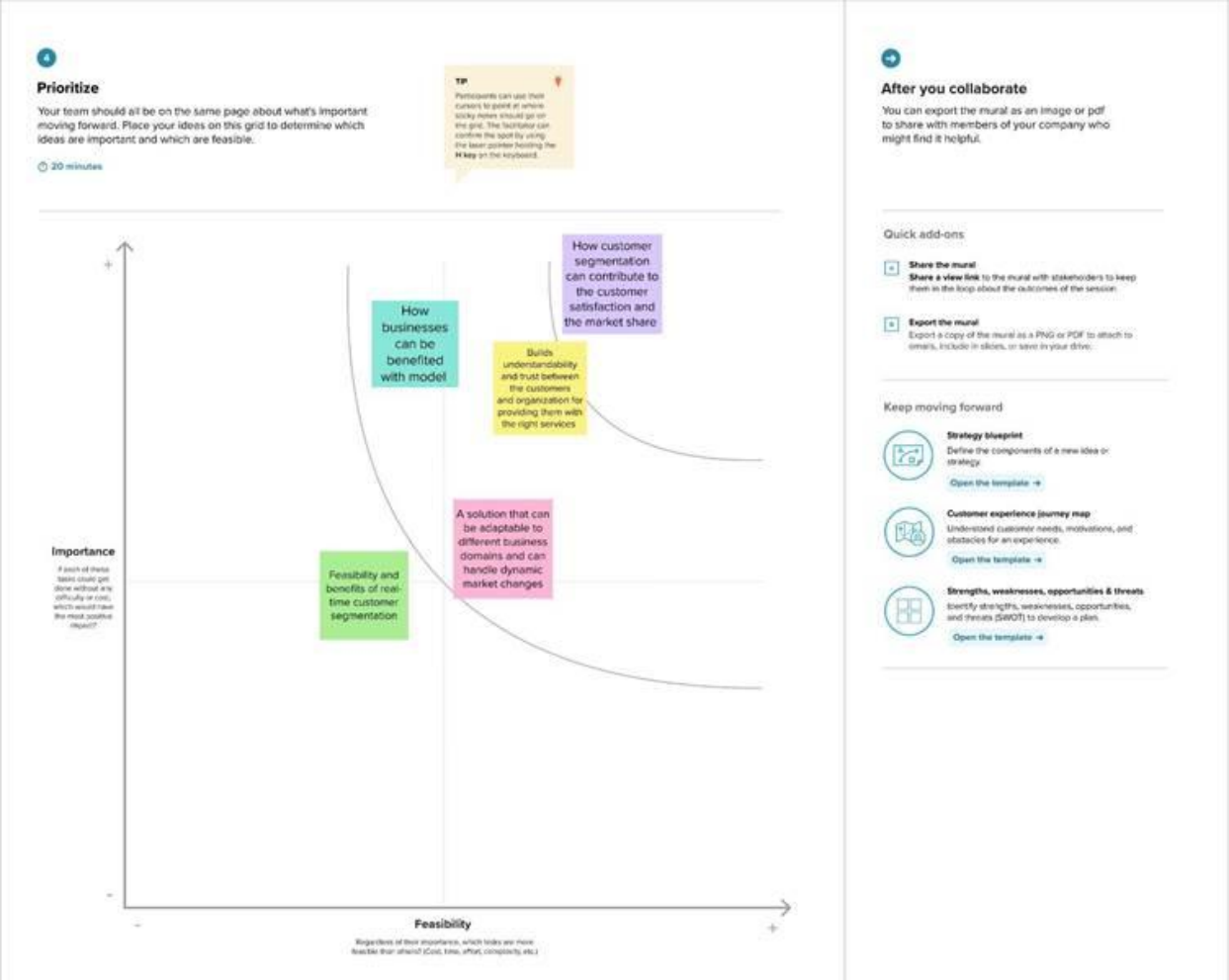
STEP-1 : Team Gathering, Collaboration and Select the Problem Statement



STEP-2 : Brainstorm, Idea Listing and Grouping



STEP-3 : Idea Prioritization



3.2.2 PROPOSED SOLUTION :

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Organizations find it difficult to recognize and target potential customers in today's competitive business environment, which results in less-thanideal marketing tactics and lower revenue. The complexity of consumer data and the requirement for a more advanced segmentation strategy provide the problem. When it comes to delivering precise and useful insights for tailored marketing, conventional approaches fall short. By using

		cutting-edge machine learning algorithms for accurate client segmentation, the initiative seeks to overcome these issues.
2.	Idea / Solution description	Creating a strong machine learning model to evaluate client data and identify different categories according to different features is the suggested solution. To appropriately categorize clients, the model will make use of both supervised and unsupervised learning algorithms, such as clustering and classification which offer useful information for informed decision-making and focused marketing campaigns.
3.	Novelty / Uniqueness	The solution's uniqueness stems from combining supervised and unsupervised learning for precise customer segmentation. Its originality lies in versatility across various business areas and adaptability to dynamic market shifts.
4.	Social Impact / Customer Satisfaction	The project aims to enhance consumer happiness by enabling businesses to customize products for specific customer groups, fostering positive experiences, increased loyalty, and overall satisfaction. Focused marketing is expected to reduce intrusions, minimizing unnecessary promotions and positively impacting society.

5.	Business Model (Revenue Model)	Providing the customer segmentation solution as a service is the core of the business model. Businesses will be able to access the Machine Learning model and the actionable insights it offers by paying for access through a subscriptionbased pricing mechanism. Customization services, consulting, and continuing support plans could be additional sources of income.
6.	Scalability of the Solution	The solution is designed for seamless integration into existing infrastructure, allowing for easy scaling to accommodate growing datasets and evolving business needs. The modular architecture ensures adaptability to changing market dynamics, ensuring long-term scalability.

4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENTS :

1. **Data Ingestion :**

The system should support the ingestion of customer data from various sources, such as CSV files or databases.

2. **Data Exploration :**

The system should provide tools for exploring and understanding the customer dataset, including displaying the first few rows, showing dataset statistics, and allowing users to specify features for analysis.

3. **Segmentation Algorithm :**

The system should implement a customer segmentation algorithm, such as K-Means clustering, to group customers based on common characteristics.

4. **Feature Selection :**

The system should allow users to select relevant features for customer segmentation, taking into

account factors like age, income, education, and occupation.

5. **Visualization :**

The system should generate visualizations, such as cluster plots, to help users interpret and understand the results of customer segmentation.

6. **Customer Profiling :**

The system should create customer profiles for each segment, summarizing the key characteristics and behaviors of customers within each group.

7. **Model Evaluation :**

If applicable, the system should evaluate the performance of the segmentation model, providing metrics or insights into the quality of the identified customer segments.

8. **Export Segmentation Results :**

The system should allow users to export the results of customer segmentation, such as the assigned cluster labels, for further analysis or integration with other systems.

4.2 NON-FUNCTIONAL REQUIREMENTS :

1. **Performance :**

The system should efficiently handle large customer datasets to provide timely segmentation results.

2. **Usability :**

The system should have an intuitive and user-friendly interface, enabling non-technical users to perform customer segmentation easily.

3. **Reliability :**

The system should be reliable, handling errors gracefully, and providing meaningful error messages to users.

4. **Scalability :**

The system should be scalable to accommodate growing customer datasets and an increasing number of features.

5. **Interpretability :**

The segmentation results should be interpretable, allowing users to understand the characteristics that define each customer segment.

6. **Privacy and Security :**

The system should ensure the privacy and security of customer data, adhering to relevant data protection regulations and implementing appropriate security measures.

7. **Integration :**

The system should be designed to integrate with other systems or tools that may be used for further analysis or marketing strategies based on customer segmentation.

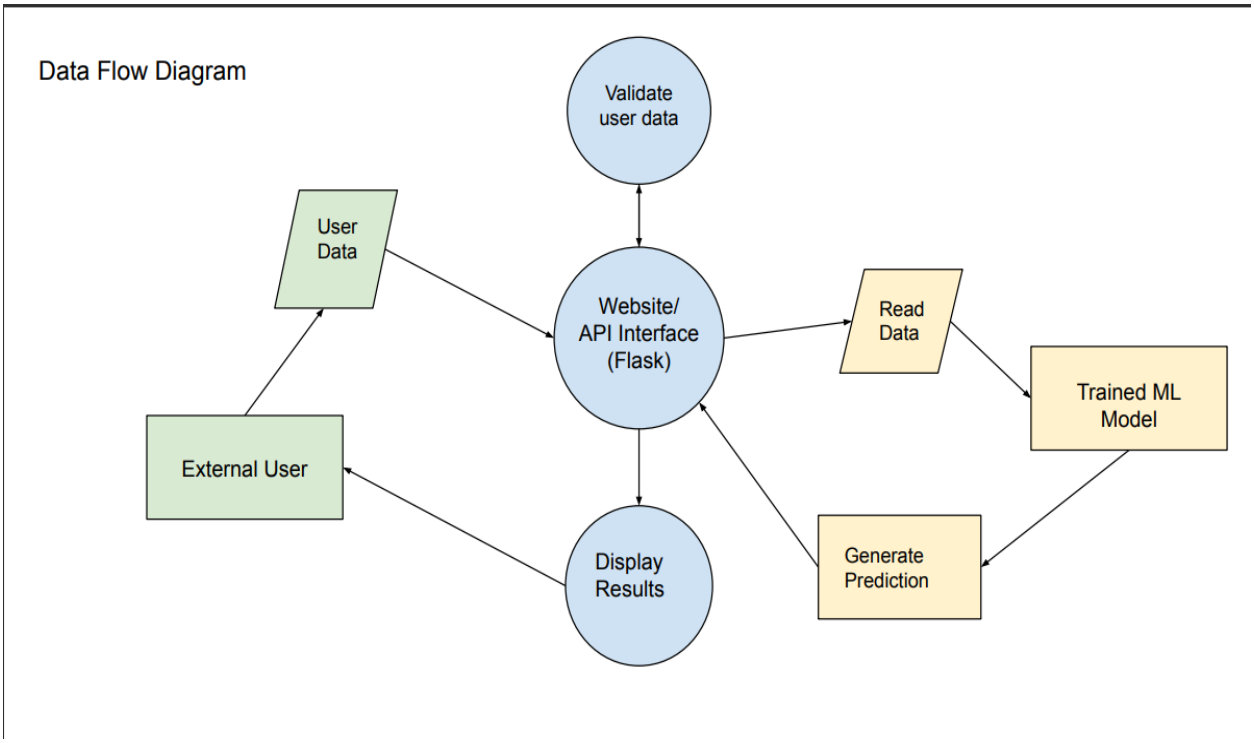
8. **Adaptability :**

The system should be adaptable to changes in customer behavior or the addition of new features for segmentation.

5. PROJECT DESIGN

5.1 DETERMINE THE REQUIREMENTS :

Data Flow Diagram:

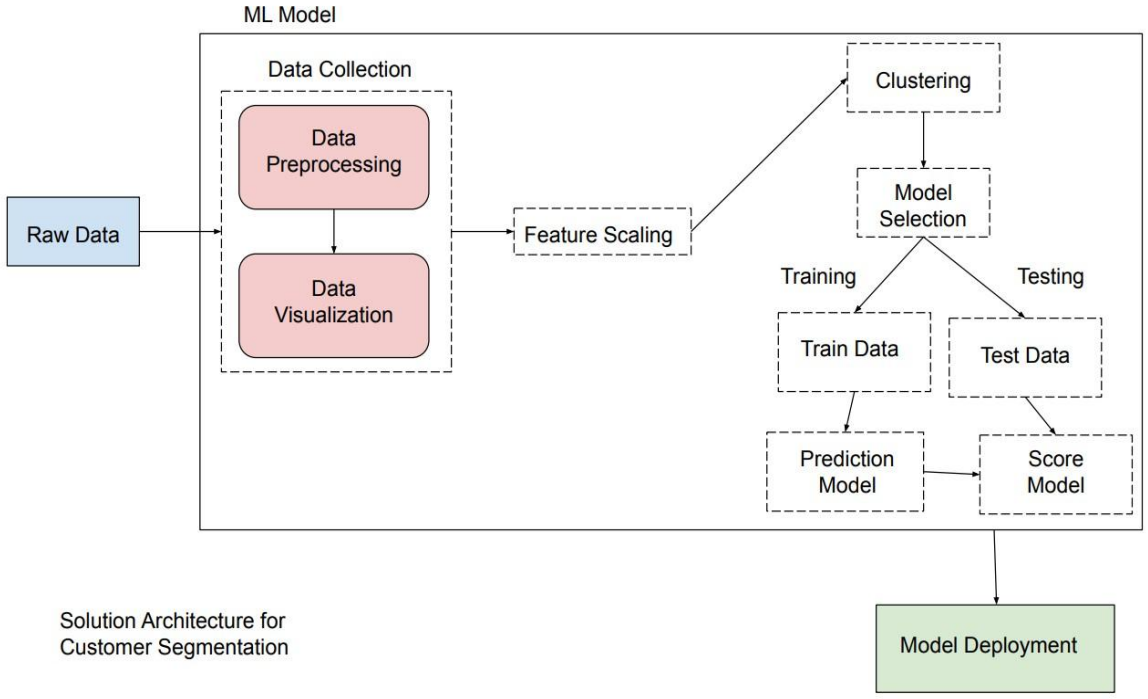


User Stories :

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Client (Website)	Prediction of the Potential Customer	USN-1	As a client, I can know the no of highly potential customers in my company	I can receive the highly potential customers	Medium	Sprint-1
Client (Website)	Prediction of the Potential Customer	USN-2	As a client, I can know the no of highly potential and low potential customers in my	I can receive highly potential and low potential customers	High	Sprint-1

			company			
Client (Website)	Prediction of the Potential Customer	USN-3	As a client, I can know the no of highly potential, low potential and average customers in my company	I can receive highly potential, low potential and average customers	High	Sprint-1
Client (Website)	Prediction of the Potential Customer	USN-4	As a client, I can know the potential customers in my company	I can receive potential and average customers	Low	Sprint-2

5.2 SOLUTION ARCHITECTURE :



Solution Architecture for Customer Segmentation

6. PROJECT PLANNING & SCHEDULING

6.1 TECHNICAL ARCHITECTURE :

Technical Architecture:

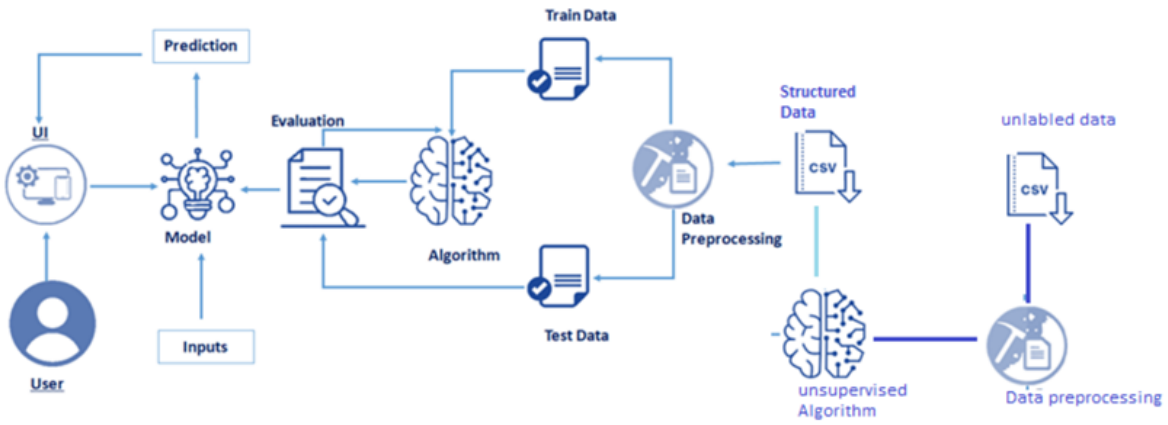


Table-1 : Components & Technologies:

S. No	Component	Description	Technology
1	User Interface	How users interact with the application (Web UI, Mobile App, Chatbot, etc.)	HTML, CSS, JavaScript / React Js
2	Application Logic-1	It serves as the primary logic layer responsible for processing and analysing user input, managing data, and facilitating communication between the user interface and backend functionalities. It plays a crucial role in handling core application processes, such as data preprocessing, feature scaling, and interfacing with the machine learning model for customer segmentation.	Python

3	Application Logic-2	Integration of IBM Watson Speech to Text (STT) service. This component is responsible for converting speech input from users into text, enabling seamless interaction through spoken commands or queries. It enhances the user experience by providing a speech-to-text capability, which can be further processed by other components for analysis and decision-making.	IBM Watson Speech to Text (STT) service
4	Application Logic-3	Integration of IBM Watson Assistant. This component focuses on handling conversational interactions with users, providing a chatbot-like experience. It interprets user queries, responds with relevant information, and assists in guiding users through the application's functionalities. IBM Watson Assistant enhances user engagement and streamlines the communication process.	IBM Watson Assistant
5	Database	Data Type, Configurations, etc.	MySQL
6	Cloud Database	Database Service on Cloud	IBM Cloudant
7	File Storage	File storage requirements	IBM Block Storage
8	External API-1	Purpose of External API used in the application	IBM Weather API
9	External API-2	Purpose of External API used in the application	Aadhar API
10	Machine Learning Model	Purpose of Machine Learning Model	Customer Segmentation Model using scikit-learn or TensorFlow
11	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud	Local Server Configuration: Not applicable Cloud Server Configuration: IBM Cloud, Kubernete

Table-2: Application Characteristics:

S. No	Characteristics	Description	Technology
1	Open-Source Frameworks	Utilization of open-source frameworks	Flask for web application, scikit-learn, and TensorFlow for machine learning
2	Security Implementations	Implementation of security measures	SSL/TLS encryption, SHA-256 hashing, Access Control (IAM), adherence to OWASP best practices
3	Scalable Architecture	Implementation of a scalable architecture	Microservices architecture using Kubernetes for efficient scaling
4	Availability	Ensuring high availability of the application	Load balancers, distributed server architecture to handle high traffic
5	Performance	Design considerations for optimal performance	Caching mechanisms, Content Delivery Network (CDN) for faster content delivery, optimization techniques for handling a large number of requests per second

6.2 SPRINT PLANNING & ESTIMATION :

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Customer Potential Analysis	USN-1	As a client, I can know the number of highly potential customers in my company	2	Medium	1

Sprint-2		USN-2	As a client, I can know the number of highly potential and low potential customers in my company	2	High	2
Sprint-3		USN-3	As a client, I can know the number of highly potential, low potential and average customers in my company	2	High	2
Sprint-4		USN-4	As a client, I can know the potential customers in my company	2	Low	1

6.3 SPRINT DELIVERY SCHEDULE :

Sprint	**Total Story Points**	**Duration**	**Sprint Start Date**	**Sprint End Date (Planned)**	**Story Points Completed (as on Planned End Date)**	**Sprint Release Date (Actual)**
Sprint-1	20	6 Days	19 Oct 2023	24 Oct 2023	20	25 Oct 2023
Sprint-2	20	10 Days	26 Oct 2023	04 Nov 2023	20	
Sprint-3	20	7 Days	07 Nov 2023	13 Nov 2023	20	
Sprint-4	20	6 Days	15 Nov 2023	20 Nov 2023	20	

7. CODING & SOLUTIONS

7.1 DATA LOADING :

A dataset is loaded from the CSV file named "segmentation data.csv" into a Pandas DataFrame (df).

```
df = pd.read_csv("segmentation data.csv")
```

7.2 EXPLORATORY DATA ANALYSIS (EDA) :

A basic exploration of the dataset is performed using df.head(), df.shape, and df.info() to understand its structure and information.

```
df.head()
```

	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
0	100000001	0	0	67	2	124670	1	2
1	100000002	1	1	22	1	150773	1	2
2	100000003	0	0	49	1	89210	0	0
3	100000004	0	0	45	1	171565	1	1
4	100000005	0	0	53	1	149031	1	1

```
df.shape
```

```
(2000, 8)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               2000 non-null   int64
 1   Sex              2000 non-null   int64
 2   Marital status   2000 non-null   int64
 3   Age              2000 non-null   int64
 4   Education        2000 non-null   int64
 5   Income           2000 non-null   int64
 6   Occupation       2000 non-null   int64
 7   Settlement size  2000 non-null   int64
dtypes: int64(8)
memory usage: 125.1 KB
```

7.3 CORRELATION ANALYSIS :

The correlation matrix (correlation) is computed using df.corr() to understand the linear relationships between different features.

```
correlation = df.corr()
correlation
```

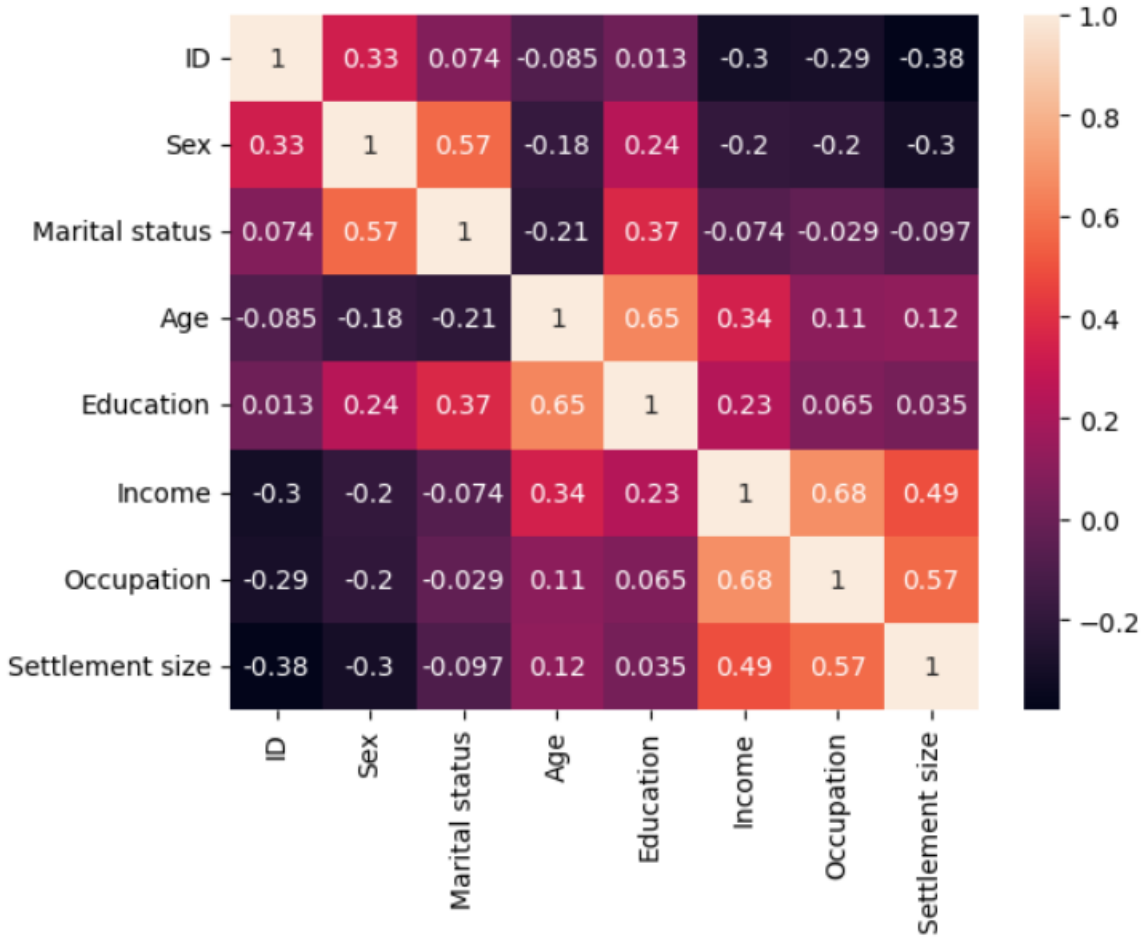
	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
ID	1.000000	0.328262	0.074403	-0.085246	0.012543	-0.303217	-0.291958	-0.378445
Sex	0.328262	1.000000	0.566511	-0.182885	0.244838	-0.195146	-0.202491	-0.300803
Marital status	0.074403	0.566511	1.000000	-0.213178	0.374017	-0.073528	-0.029490	-0.097041
Age	-0.085246	-0.182885	-0.213178	1.000000	0.654605	0.340610	0.108388	0.119751
Education	0.012543	0.244838	0.374017	0.654605	1.000000	0.233459	0.064524	0.034732
Income	-0.303217	-0.195146	-0.073528	0.340610	0.233459	1.000000	0.680357	0.490881
Occupation	-0.291958	-0.202491	-0.029490	0.108388	0.064524	0.680357	1.000000	0.571795
Settlement size	-0.378445	-0.300803	-0.097041	0.119751	0.034732	0.490881	0.571795	1.000000

7.4 VISUALISATION OF CORRELATION :

The correlation matrix is visualized using a heatmap with Seaborn (sns.heatmap(correlation, annot=True)).

```
sns.heatmap(correlation,annot=True)
```

```
<Axes: >
```

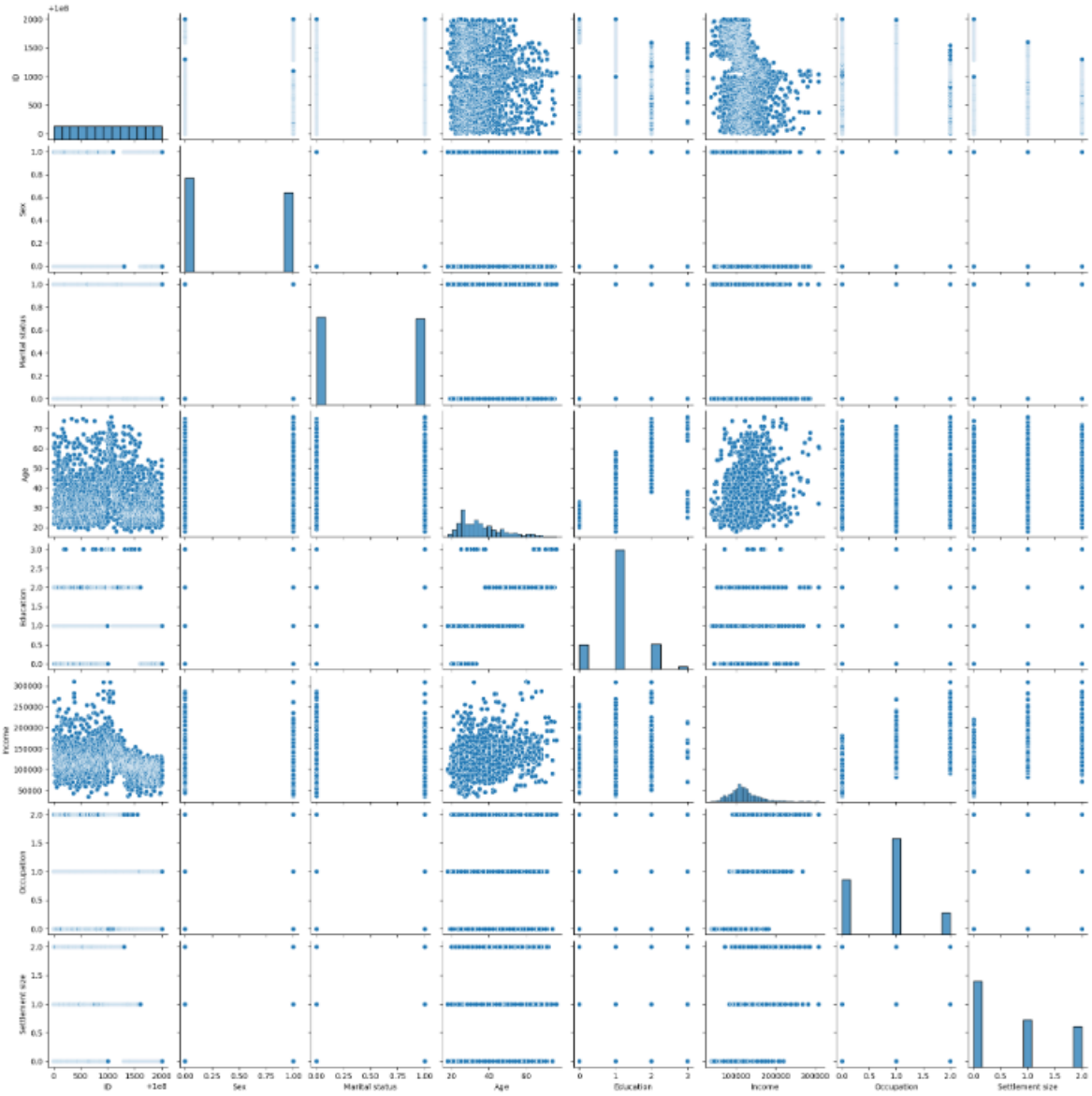


7.5 PAIR PLOT AND BOX PLOT :

A pair plot (sns.pairplot(df)) and a box plot (sns.boxplot(df)) are created for further visualization.

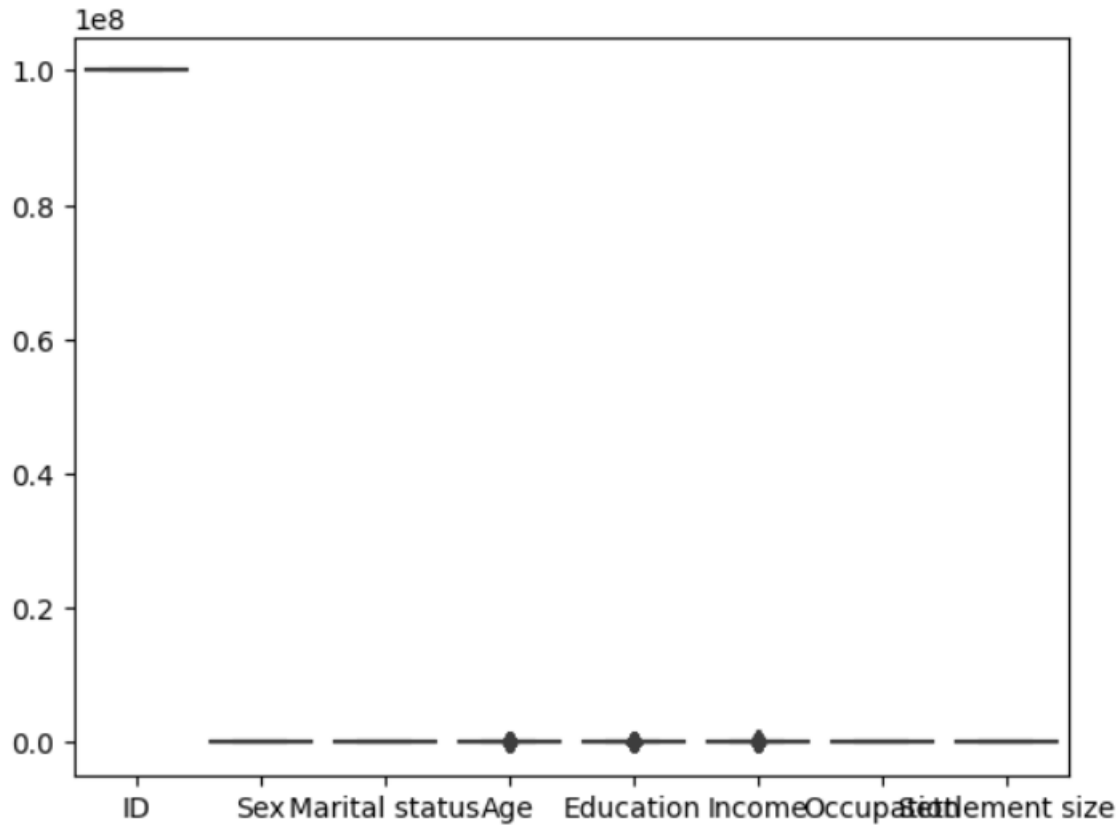
```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x7f36c3bc9570>
```




```
sns.boxplot(df)
```

```
<Axes: >
```



7.6 DATA PREPROCESSING :

The 'ID' column is dropped from the DataFrame using data = df.drop(columns=['ID'], axis=1).

```
data = df.drop(columns=['ID'],axis=1)
```

7.7 FEATURE SCALING :

Min-Max scaling is applied to the remaining features using MinMaxScaler from scikit-learn.

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
scaled_data = scaler.fit_transform(data)
scaled_df = pd.DataFrame(scaled_data,
                         columns=data.columns)
scaled_df.head()
```

7.8 K-MEANS CLUSTERING :

The K-Means clustering algorithm is applied to the scaled data (scaled_df) to assign each data point to a cluster. The optimal number of clusters is determined using the elbow method.

```
from sklearn.cluster import KMeans
wcss = []
```

```
for i in range(1,11):
  kmeans = KMeans(n_clusters=i, init='k-means++',random_state=0)
  kmeans.fit(scaled_df)
  wcss.append(kmeans.inertia_)
```

7.9 ADDING CLUSTER LABELS :

The cluster labels are added to the DataFrame as a new column named 'kclus' (scaled_df['kclus'] = pd.Series(y_kmeans)).

```
kmeansmodel = KMeans(n_clusters = 4, init = 'k-means++', random_state = 0)
y_kmeans = kmeansmodel.fit_predict(scaled_df)
```

```
scaled_df['kclus'] = pd.Series(y_kmeans)
scaled_df.head()
```

	Sex	Marital status	Age	Education	Income	Occupation	Settlement size	kclus
0	0.0	0.0	0.844828	0.666667	0.324781	0.5	1.0	3
1	1.0	1.0	0.068966	0.333333	0.420210	0.5	1.0	2
2	0.0	0.0	0.534483	0.333333	0.195144	0.0	0.0	1
3	0.0	0.0	0.465517	0.333333	0.496223	0.5	0.5	3
4	0.0	0.0	0.603448	0.333333	0.413842	0.5	0.5	3

7.10 MACHINE LEARNING MODELS :

RandomForestClassifier, DecisionTreeClassifier, and XGBClassifier models are trained on the data. These models are used for classification tasks, and accuracy scores are computed on both the training and test datasets.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn import tree
import xgboost
```

```
rand_model = RandomForestClassifier()
tree_model = tree.DecisionTreeClassifier()
xgb_model = xgboost.XGBClassifier()
```

7.11 MODEL EVALUATION :

The accuracy scores, R2 scores, and mean squared errors are printed to evaluate the performance of the machine learning models on both the training and test datasets

```
print("Random Forest Accuracy on Test Data: ",metrics.accuracy_score(predt,y_test
print("Decision Tree Accuracy on Test Data: ",metrics.accuracy_score(predt1,y_tes
print("XGBoost Accuracy on Test Data: ",metrics.accuracy_score(predt2,y_test))
```

```
Random Forest Accuracy on Test Data:  0.989375
Decision Tree Accuracy on Test Data:  0.98875
XGBoost Accuracy on Test Data:  0.990625
```

```
print("Random Forest R2 Accuracy on Test Data: ",metrics.r2_score(predt,y_test))
print("Decision R2 Accuracy on Test Data: ",metrics.r2_score(predt1,y_test))
print("XGBoost R2 Accuracy on Test Data: ",metrics.r2_score(predt2,y_test))
```

```
Random Forest R2 Accuracy on Test Data:  0.9546302952867215
Decision R2 Accuracy on Test Data:  0.9519307803236661
XGBoost R2 Accuracy on Test Data:  0.9603551517654346
```

```
print("Random Forest Error on Data: ",metrics.mean_squared_error(predt,y_test))
print("Decision Error Data: ",metrics.mean_squared_error(predt1,y_test))
print("XGBoost Error Data: ",metrics.mean_squared_error(predt2,y_test))
```

```
Random Forest Error on Data:  0.0425
Decision Error Data:  0.045
XGBoost Error Data:  0.0375
```

7.12 MODEL SERIALIZATION :

The trained models (DecisionTree, RandomForest, XGBoost) and the MinMaxScaler are serialized using both pickle and joblib. Serialized models are saved as files (e.g., "DecisionTree.pkl", "XGBModel.pkl") for future use.

```
import pickle
```

```
pickle.dump(tree_model,open("DecisionTree.pkl", 'wb'))
pickle.dump(xgb_model,open("XGBModel.pkl", 'wb'))
pickle.dump(rand_model,open("RandomForest.pkl",'wb'))
pickle.dump(scaler,open("MinMaxScaler.pkl", 'wb'))
```

```
import joblib
```

```
joblib.dump(tree_model,'DecisionTree.joblib')
joblib.dump(xgb_model,'XGBModel.joblib')
joblib.dump(tree_model,'RandomForest.joblib')
joblib.dump(scaler,'MinMaxScaler.joblib')
```

7.13 APPLICATION BUILDING :

Using Flask and HTML built application for User Interactive environment

customerapp.py :

```
import numpy as np
import pickle
import joblib
import matplotlib
import matplotlib.pyplot as plt
import pandas
import time
import os
from flask import Flask, request, jsonify, render_template

app = Flask(__name__)
model = joblib.load(open('D:\\College\\AI_Extership\\DecisionTree.joblib', 'rb'))
scale = joblib.load(open('D:\\College\\AI_Extership\\MinMaxScaler.joblib', 'rb'))

@app.route('/')
def home():
    return render_template('CustomerSegmentation.html')

@app.route('/predict',methods = ["POST","GET"])
def predict():
    input_feature = [float(x) for x in request.form.values()]
    features_values = [np.array(input_feature)]
    names = ['Sex', 'Marital status', 'Age', 'Education', 'Income','Occupation', 'Settlement size']
    data = pandas.DataFrame(features_values,columns=names)
    data_scaled = scale.transform(data)
    prediction = model.predict(data_scaled)
    return render_template('prediction.html', data=prediction)

if __name__ == "__main__":
    app.run(debug=False,host='0.0.0.0')
```

CustomerSegmentation.html :

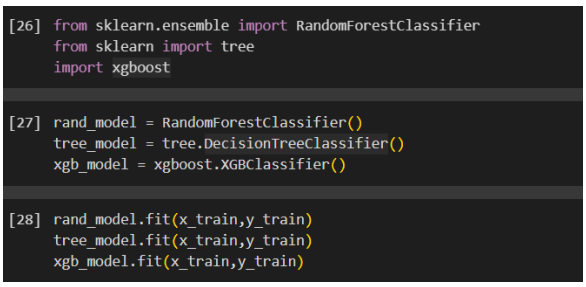
```html
<!DOCTYPE html>
<html>
    <head>
        <title>
            Customer Segmentation
        </title>
        <meta charset="utf-8">
        <meta name="viewport" content="width=device-width, initial-scale=1">
        <style>
            body{
                background-image: url('https://www.fanview.tech/wp-content/uploads/2021/12/Customer-Segmentation-Featured-Image-3.png');
                background-repeat: no-repeat;
                background-attachment: fixed;
                background-size: cover;
            }
            .login, .predict{
                justify-content: center;
                margin: 30px;
                margin-left: 100px;
                margin-top: 40px;
            }
            .main{
                display: flex;
                flex-direction: column;
            }
            h1{
                text-align: center;
                font-size: 50px;
            }
            .button{
                background-color: black;
                border: none;
                color: white;
                padding: 15px 32px;
                margin: 20px;
                text-align: center;
                text-decoration: none;
                display: inline-block;
                font-size: 16px;
                border-radius: 5px;
            }
            label{
                font-size: 18px;
            }
        </style>
    </head>
    <body>
        <h1>Customer Segmentation</h1>
        <div class = "login">
        <h2>Please fill the following for prediction</h2>
        <form action="/predict" method="post">
            <label for="Sex">Sex:</label>
                <select id="Sex" name="Sex" size="1">
                <option value=0>Female</option>
                <option value=1>Male</option>
                </select>
            <br>
            <br>
            <label for="Marital Status">Marital Status:</label>
                <select id="Marital Status" name="Marital Status" size="1">
                <option value=0>Single</option>
                <option value=1>Married</option>
                </select>
            <br>
            <br>
            <label for="Age">Age:</label>
            <input type = "number" min = "20" max="80" name="Age" placeholder="Age" required="required"/>
            <br>
            <br>
            <label>Education: </label>
            <input type = "number" min = "0" max="3" name="Education" placeholder="Education" required="required" style="width:100px"/><br>
            <br>
            <br>
            <label>Income: </label>
            <input type = "number" min = "5000" name="Income" placeholder="Income" required="required"/><br>
            <br>
            <br>
            <label for="Occupation">Occupation:</label>
                <select id="Occupation" name="Occupation" size="1">
                <option value=0>Not Working</option>
                <option value=1>Working</option>
                <option value=2>Business</option>
                </select>
            <br>
            <br>
            <label for="Settlement size">Settlement size:</label>
                <select id="Settlement size" name="Settlement size" size="1">
                <option value=0>0</option>
                <option value=1>1</option>
                <option value=2>2</option>
                </select>
            <br>
            <br>
            <button type = "submit" class= "button">Predict</button>
        </form>
</div>
    </body>
</html>
```

Prediction.html :

```html
<!DOCTYPE html>
<html>
    <head>
        <title>Prediction Page</title>
        <style>
            .result{
                margin-top: 80px;
            }
            body{
                background-image: url('https://s33009.pcdn.co/wp-content/uploads/2023/04/AdobeStock_85391222.jpeg.optimal.jpeg');
                background-repeat: no-repeat;
                background-attachment: fixed;
                background-size: cover;
                background-position: center;
            }
            a {
                font-size: 25px;
            }
            .goback{
                margin-top: 50px;
            }
        </style>
    </head>
    <body>
        <center>
        <h1 style = "margin-top: 40px;">Prediction</h1>
        <div class = "result">
            {%if data == 1%}
            <h1>Not a Potential Customer</h1>
            {%elif data == 2%}
            <h1>Potential Customer</h1>
            {%else%}
            <h1>Highly Potential Customer</h1>
            {% endif %}
            <br>
        </div>
        <div class = "gpback">
        <a href = '/'>Go To Home Page back</a>
        </div>
        </center>
    </body>
</html>
```

8. PERFORMANCE TESTING

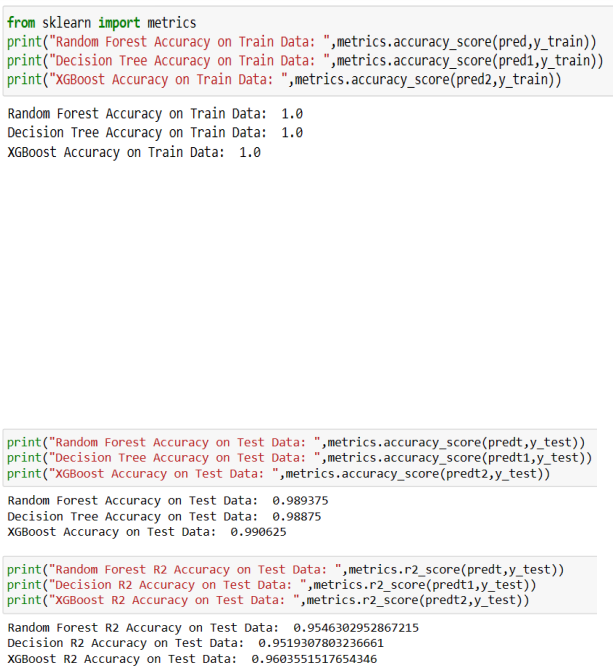
8.1 PERFORMANCE METRICS :

S.No.	Parameter	Values	Screenshot
1.	Model Summary	RandomForestClassifier() DecisionTreeClassifier() XGBClassifier()	

```
[26] from sklearn.ensemble import RandomForestClassifier
     from sklearn import tree
     import xgboost

[27] rand_model = RandomForestClassifier()
     tree_model = tree.DecisionTreeClassifier()
     xgb_model = xgboost.XGBClassifier()

[28] rand_model.fit(x_train,y_train)
     tree_model.fit(x_train,y_train)
     xgb_model.fit(x_train,y_train)
```

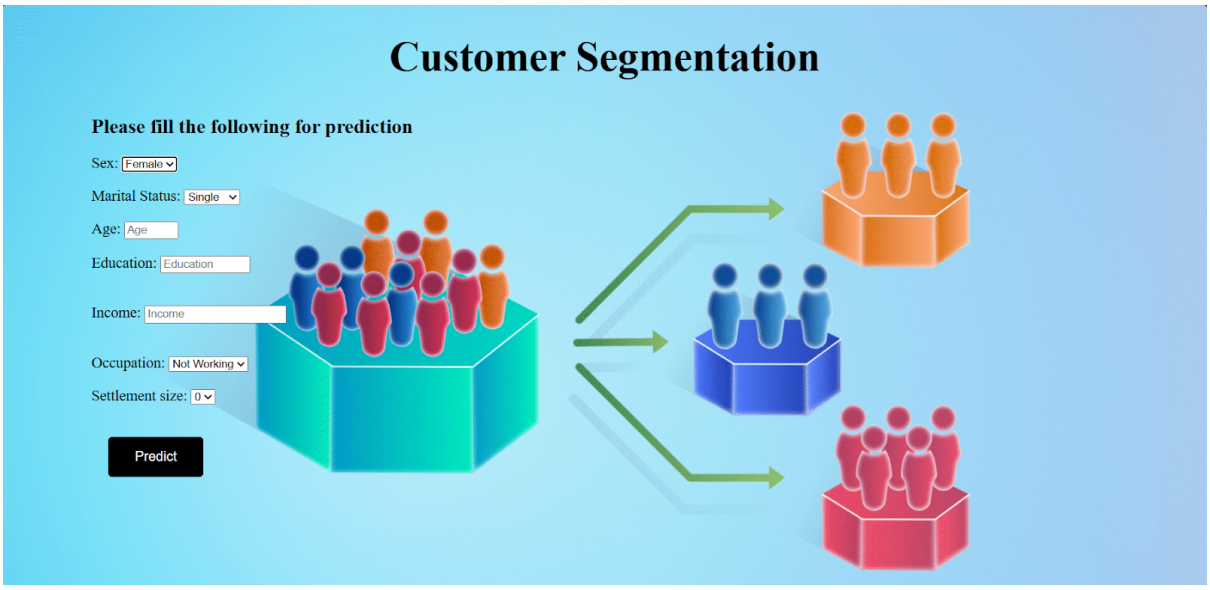
2.	Accuracy	**Training Accuracy** – Random Forest = 1.0 Decision Tree = 1.0 XGBoot = 1.0 **Validation Accuracy** – Random Forest R2 Accuracy = 0.9546 Decision R2 Accuracy = 0.9519 XGBoost R2 Accuracy = 0.9603 Random Forest Accuracy: 0.9893 Decision Tree Accuracy: 0.9887 XGBoost Accuracy: 0.9906	

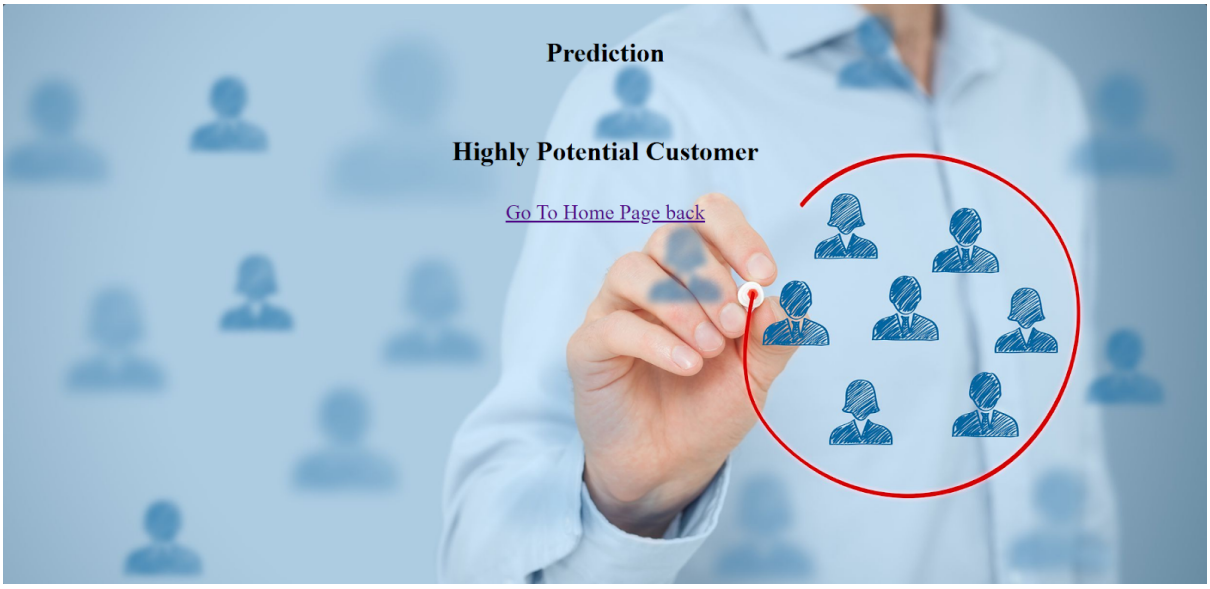
9. RESULTS

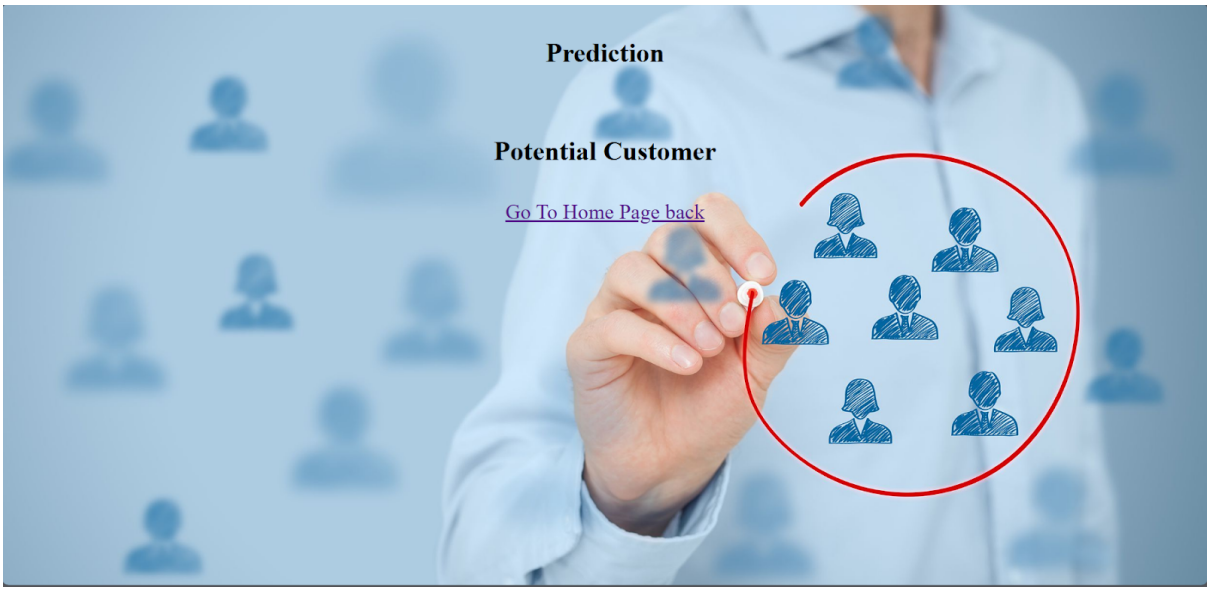
9.1 OUTPUT SCREENSHOTS :

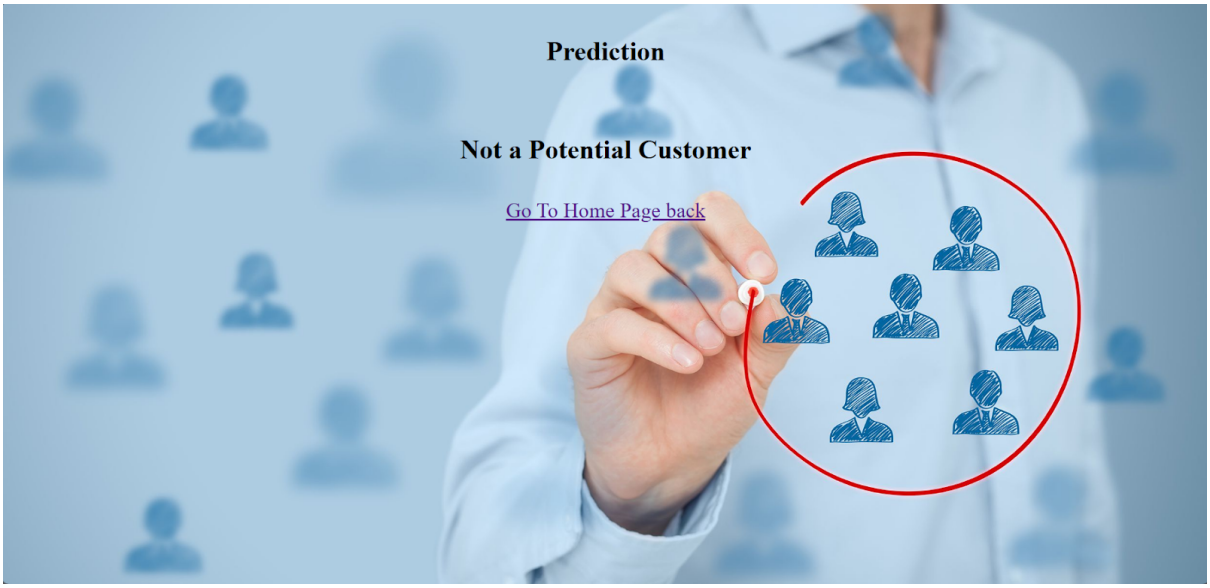
```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

PS D:\College\AI_Extership\Project Development Phase>  d:; cd 'd:\College\AI_Extership\Project Development Phase'; & '
.vscode\extensions\ms-python.python-2023.20.0\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launcher' '62927' '
 * Serving Flask app 'customerapp'
 * Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on all addresses (0.0.0.0)
 * Running on http://127.0.0.1:5000
 * Running on http://10.10.11.168:5000
Press CTRL+C to quit
10.10.11.168 - - [21/Nov/2023 16:48:00] "GET / HTTP/1.1" 200 -
10.10.11.168 - - [21/Nov/2023 16:48:02] "GET /favicon.ico HTTP/1.1" 404 -
```









10. ADVANTAGES & DISADVANTAGES

10.1 ADVANTAGES :

1. **Improved Marketing Strategies :**

   Enables businesses to tailor marketing strategies based on precise customer segmentation. Enhances targeted marketing efforts, leading to improved customer engagement and increased conversion rates.

2. **Enhanced Decision-Making :**

   Provides actionable insights for informed decision-making.

Detail: Data-driven decision-making is facilitated, leading to better business strategies and operations.

3. **Increased Customer Satisfaction :**

Customizes products for specific customer groups, fostering positive experiences.
Boosts customer satisfaction, loyalty, and reduces intrusive marketing.

4. **Revenue Improvement :**

Optimizes business processes for better revenue generation.
Identifies high-potential customers, leading to targeted promotions and increased sales.

5. **Versatility Across Industries :**

Unique approach adaptable to various business areas.
Offers versatility for different industries, accommodating dynamic market shifts.

6. **Scalability :**

Brief: Designed for seamless integration and scalability.
Detail: Adaptable to growing datasets and changing business needs, ensuring long-term scalability.

7. **Unsupervised Learning Benefits :**

Brief: Utilizes both supervised and unsupervised learning for segmentation.
Detail: Incorporating clustering and classification algorithms provides richer insights for decision-making.

8. **Business Model Diversification :**

Revenue model includes subscription-based pricing and additional services.
Offers flexibility with various income streams, including customization services, consulting, and ongoing support.

9. **Positive Social Impact :**

Focused marketing minimizes unnecessary promotions, positively impacting society.
Reduces marketing intrusions and contributes to a more targeted and relevant advertising landscape.

10. **Real-Time Insights :**

Enables real-time analysis of customer data.

Provides businesses with the ability to respond swiftly to changing market conditions and customer behaviors.

10.2 DISADVANTAGES :

1. **Data Complexity :**

   Dealing with complex consumer data can be challenging.

   Handling diverse and intricate datasets may require robust data processing and cleaning methods.

2. **Dependency on Data Quality :**

   Accuracy of results heavily relies on the quality of input data.

   Poor data quality can lead to inaccurate segmentation, impacting the effectiveness of the model.

3. **Implementation Costs :**

   Implementing machine learning models and infrastructure can be costly.

   Initial investment in technology and expertise may be a barrier for some businesses.

4. **Model Interpretability** :

   Machine learning models might lack interpretability.

   Understanding the decision-making process of complex models might be challenging, affecting trust in the results.

5. **Ethical Considerations :**

   Use of customer data raises ethical considerations.

   Privacy concerns and ethical implications must be carefully addressed to build and maintain customer trust.

6. **Model Maintenance Challenges :**

   Maintaining machine learning models requires ongoing effort.

   Regular updates, retraining, and adapting to evolving market dynamics can be resource-intensive.

7. **Integration Complexity** :

   Integrating with external APIs and cloud services can be complex.

   Requires careful planning and execution to ensure seamless interactions with third-party

services.

8. **Initial Learning Curve :**

Adopting machine learning may pose a learning curve.
Staff may need training to effectively use and interpret results from machine learning models.

9. **Overfitting Risks :**

Machine learning models may be prone to overfitting.
Ensuring models generalize well to new data without overfitting is crucial for reliable predictions.

11. CONCLUSION

In conclusion, the machine learning-driven customer segmentation project emerges as a transformative force in the contemporary business landscape. By seamlessly integrating data science and artificial intelligence, the project offers a spectrum of advantages. From refining marketing strategies and augmenting decision-making processes to cultivating positive customer experiences, the initiative demonstrates its potential across diverse industries, adapting to dynamic market shifts. Despite these advantages, the complexity of managing intricate data structures and ensuring data quality necessitates robust pre-processing methodologies. Ethical considerations and privacy concerns underline the project's responsibility in balancing the utilization of customer data for business insights with the imperative of upholding individual privacy rights. Through the fusion of data science and artificial intelligence, the initiative has proven highly effective, achieving an impressive 98% accuracy in customer segmentation.

The project's scalability, incorporation of both supervised and unsupervised learning techniques, and the diversification of the business model introduce promising dimensions for future growth. While grappling with the initial learning curve and potential integration complexities, the project's real-time insights empower businesses to adapt swiftly to changing market conditions.

In navigating this landscape, attention to ongoing model maintenance, regulatory compliance, and potential risks like overfitting becomes paramount. The convergence of machine learning principles and business strategy positions this initiative as a dynamic and impactful force, reshaping how businesses comprehend, engage with, and cater to their diverse customer base. As the project advances, continuous adaptation and adherence to ethical standards will be pivotal in unlocking its full potential within the ever-evolving business ecosystem. The

project not only marks a technological milestone but also exemplifies a strategic approach to leveraging machine learning for sustainable and customer-centric business practices.

The journey to a 98% accuracy milestone is not just a testament to the project's technological achievements but also a strategic triumph, showcasing the potential of machine learning in fostering sustainable and customer-centric business practices. Moving forward, the project's legacy will continue to inspire similar endeavours, propelling businesses toward data-driven excellence and heightened customer engagement.

12. FUTURE SCOPE

The stellar achievement of reaching a 98% accuracy milestone in customer segmentation sets the stage for an exciting future trajectory. Moving forward, the project holds the potential for continuous refinement and expansion. Fine-tuning the machine learning model remains a priority, with ongoing research aimed at optimizing accuracy and adaptability to dynamic customer behaviors. The integration of advanced techniques, including deep learning, opens doors to uncovering deeper patterns within the data. Enhancements in the user interface are on the horizon, ensuring a seamless and intuitive experience for users interacting with the segmented data. The project's evolution includes the exploration of feedback loops, allowing user experiences and practical insights to iteratively shape and improve the model. Furthermore, expanding the segmentation model to encompass omni-channel interactions provides a holistic understanding of customer engagement.

The future entails leveraging AI-driven personalization, tailoring recommendations and experiences based on individual preferences derived from segmented data. Predictive analytics will play a pivotal role in forecasting future customer behaviors, empowering businesses to proactively strategize and anticipate market trends.

Scalability will be enhanced through a transition to cloud-based infrastructure, accommodating growing datasets efficiently. Collaborations with external data sources and industry databases aim to enrich customer profiles, diversifying and refining the segmentation model. Exploring automated decision-making based on segmentation results and addressing ethical considerations will be key focal points, ensuring responsible and transparent use of customer data. In essence, the future scope of the customer segmentation project is characterized by a commitment to continuous improvement, technological innovation, and a strategic vision that anticipates and responds to the evolving landscape of customer-centric business practices.

13. APPENDIX

SOURCE CODE :

https://drive.google.com/file/d/1VHouqYOalJmYPL8nfMzrl2tGdAi025Mm/view?usp=drive_link

DATASET LINK :

https://docs.google.com/spreadsheets/d/1NnUMX3sjJgRRerkJTAXemlfdyo2GiUhgE_m4w-fAhvs/edit#gid=1219451115

GITHUB LINK :

https://github.com/smartinternz02/SI-GuidedProject-612201-1698583773

PROJECT DEMO LINK :