**Customer Personality Analysis**

**using Machine Learning**

**AN INTERNSHIP REPORT**

*Submitted by*

**SOUMEN CHATTERJEE**

**[EC2432251010407]**

**Under the Guidance of**

**Dr. G. Babu**

(Assistant Professor, Directorate of Online Education)

*in partial fulfilment for the award of the degree of*

**MASTER OF COMPUTER APPLICATIONS**



DIRECTORATE OF ONLINE EDUCATION

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY KATTANKULATHUR- 603 203

**JAN 2024 MCA Batch**

**MAY-JUNE 2025 Exam (3rd SEM)**

DIRECTORATE OF ONLINE EDUCATION

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY** KATTANKULATHUR – 603 203

**BONAFIDE CERTIFICATE**

This Internship Report titled **“Customer Personality Analysis using Machine Learning”** of **“Soumen Chatterjee [EC2432251010407]”**, who carried out the Internship Project Work under the supervision of Program Coordinator of Online Education along with the company mentor.

Certified further that to the best of my knowledge, the work reported herein does not form any other internship report or dissertation based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

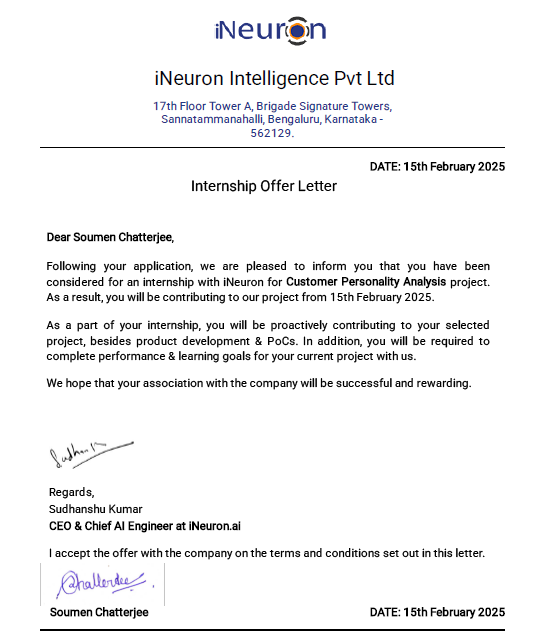
A close-up of a signature

AI-generated content may be incorrect.

Signature of the Student

(Soumen Chatterjee)

**INTERNSHIP OFFER LETTER**



**ACKNOWLEDGEMENTS**

I would like to express my sincere gratitude to **Dr. C. Muthamizhchelvan**, Vice Chancellor, SRM Institute of Science and Technology, for providing the necessary facilities and continued support that enabled the successful completion of this project.

I extend my heartfelt thanks to **Prof. Dr. Manoranjan Pon Ram**, Director, Directorate of Online Education, SRM Institute of Science and Technology, for his valuable guidance and encouragement throughout the course of this internship.

I am deeply thankful to **Dr. G. Babu**, Program Coordinator, Directorate of Online Education, for his insightful feedback during project reviews and his unwavering support. I am especially grateful to him for serving as my project guide. His mentorship, encouragement, and the freedom he offered to explore research topics aligned with my interests were instrumental in shaping this project. His dedication and passion for solving real-world problems have been truly inspiring.

I also extend my sincere appreciation to the faculty, staff, and fellow students of the **Directorate of Online Education**, SRM Institute of Science and Technology, for their cooperation and assistance during various stages of the project.

Lastly, I would like to acknowledge the unconditional love, encouragement, and support of my **parents, family members, and friends**, without whom this journey would not have been possible.

**SOUMEN CHATTERJEE**

**Table of Contents**

[1. Abstract 6](#_Toc197030155)

[2. Introduction 7](#_Toc197030156)

[3. System Analysis 8](#_Toc197030157)

[4. Analysis and Requirements 11](#_Toc197030158)

[5. Problem Description / Module Description 12](#_Toc197030159)

[6. Design 13](#_Toc197030160)

[6. a) System Design 13](#_Toc197030161)

[6. b) UML Diagrams 14](#_Toc197030162)

[6. c) Database Design 17](#_Toc197030163)

[7. Implementation 19](#_Toc197030164)

[8. Testing 34](#_Toc197030165)

[9. Output Screens 39](#_Toc197030166)

[10. Tools and Technologies 44](#_Toc197030167)

[11. Conclusion 46](#_Toc197030168)

[12. Appendices 48](#_Toc197030169)

[13. References / Bibliography 52](#_Toc197030170)

1. Abstract

Customer Personality Analysis is a data-driven approach to understanding consumer behaviour, preferences, and purchasing patterns. This project aims to analyse customer data to segment consumers into distinct groups, allowing businesses to target them effectively. The study leverages machine learning techniques, including clustering (unsupervised learning) and classification (supervised learning), to gain insights from customer demographics and purchase behaviour.

Key accomplishments of this project include:

* **Data preprocessing**, including cleaning, handling missing values, and feature engineering.
* **Unsupervised learning** using K-Means clustering to segment customers into groups.
* **Supervised learning** using Decision Tree, K-Nearest Neighbours (KNN), and Random Forest classifiers to predict customer response.
* **Visualization and analysis** of key customer attributes such as income, spending behaviour, and family structure.

My key contributions include data preparation, model selection, and result interpretation. The findings help in targeted marketing and customer relationship management strategies.

2. Introduction

* **Background**

Businesses increasingly rely on data-driven decision-making to understand their customers better. Customer Personality Analysis plays a crucial role in personalizing marketing strategies, optimizing customer engagement, and improving product recommendations.

* **Problem Statement**

Companies often struggle with segmenting their customers effectively, leading to inefficient marketing campaigns. Instead of marketing a product to the entire customer base, a company can analyse which customer segment is most likely to buy the product and target them specifically.

* **Development Process**

The project follows a structured development approach:

1. **Data Collection & Preprocessing**: Cleaning data, handling missing values, and engineering useful features.
2. **Exploratory Data Analysis (EDA)**: Understanding data distribution and identifying patterns.
3. **Unsupervised Learning (Clustering)**: Using K-Means clustering to segment customers.
4. **Supervised Learning (Classification)**: Using decision trees, KNN, and random forests to predict customer responses.
5. **Model Evaluation & Insights**: Assessing model performance and deriving business insights.

3. System Analysis

System analysis is a critical phase in any software or data science project. It involves examining the current business scenario, identifying limitations of the existing approach, and proposing an improved solution through technology and innovation. This section includes an overview of the existing system (if any), the proposed system for customer personality analysis, and the results of a feasibility study validating the project's viability.

**A. Existing System**

Traditionally, businesses rely on basic customer profiling methods using limited demographic attributes such as age, gender, and location. These approaches are typically **manual or rule-based**, lacking the depth and accuracy required to truly understand complex customer behavior. The limitations of such systems include:

* **No Personalization**: Marketing campaigns are often generalized rather than tailored to specific customer segments.
* **Inefficient Targeting**: Without deeper insights, resources are wasted on customers who are unlikely to respond.
* **Data Underutilization**: Many businesses collect extensive customer data but fail to apply advanced analytics for better decision-making.
* **No Predictive Capability**: Traditional systems cannot predict which customers are more likely to respond to a campaign or make a purchase.

**B. Proposed System**

The proposed system, *Customer Personality Analysis using Machine Learning*, is designed to overcome these limitations by leveraging modern data science techniques. It provides a **data-driven approach** to customer segmentation and behavioral prediction.

**Key Features of the Proposed System:**

* **Data-Driven Insights**: Uses historical customer data including spending patterns, demographics, and online activity.
* **Customer Segmentation**: Applies clustering algorithms (e.g., KMeans) to group customers based on similar behaviors and preferences.
* **Marketing Response Prediction**: Implements supervised learning models (e.g., Random Forest) to predict customer responses to marketing campaigns.
* **Feature Engineering**: Derives new features such as Age, Total\_Spent, and Family\_Size to improve analytical depth.
* **Visualization**: Provides graphical insights for easy interpretation of complex relationships and patterns.
* **Model Evaluation**: Measures model performance using accuracy scores, ensuring reliable results.

**Benefits:**

* Enables **personalized marketing**
* Improves **customer engagement**
* Reduces **marketing costs**
* Increases **conversion rates**
* Enhances **strategic decision-making**

**C. Feasibility Study**

To ensure the practicality and success of the project, a feasibility study was conducted in the following areas:

| **Feasibility Type** | **Assessment** |
| --- | --- |
| **Technical Feasibility** | ✅ The tools and technologies used (Python, Jupyter Notebook, scikit-learn, etc.) are well-established and supported. The project was successfully implemented on standard hardware without requiring high-end computational resources. |
| **Operational Feasibility** | ✅ The solution is user-friendly, modular, and adaptable. Business users and analysts can adopt this model for campaign optimization with minimal training. The outputs are easy to interpret via visualizations and performance metrics. |
| **Economic Feasibility** | ✅ As the project uses open-source tools, the cost of development is minimal. Implementation can lead to higher ROI by optimizing marketing strategies and improving customer targeting. |
| **Schedule Feasibility** | ✅ The project was completed within the academic timeline. A well-structured plan with clear milestones ensured timely completion of each phase—data cleaning, modeling, and evaluation. |

**Conclusion of System Analysis**

The analysis clearly shows that the proposed machine learning-based system significantly improves upon the traditional approach to customer profiling. With minimal investment and strong analytical capabilities, this system is feasible, scalable, and aligns well with modern business intelligence needs.

4. Analysis and Requirements

**Problem Analysis**

The primary challenge in this project is to segment customers based on their demographic and purchasing data. The dataset consists of 29 features, including customer age, income, education, marital status, and spending behaviour across various product categories.

**UML Analysis Model**

The analysis can be represented using the following UML models:

* **Use Case Diagram**: Represents interactions between the system and different user roles (e.g., Data Analyst, Business Manager).
* **Activity Diagram**: Shows the step-by-step process of data preprocessing, clustering, and classification.
* **Class Diagram**: Defines key entities such as Customer, Purchase History, and Segmentation Model.

**System-Level and Software-Level Requirements**

* **System Requirements**
  + Python environment (Jupyter Notebook, Anaconda)
  + Libraries: pandas, NumPy, seaborn, matplotlib, scikit-learn
  + Computational resources for machine learning processing
* **Software Requirements**
  + Data preprocessing module
  + Clustering module (K-Means)
  + Classification module (Decision Tree, KNN, Random Forest)
  + Visualization module for insights

5. Problem Description / Module Description

The project consists of the following key modules:

* 1. **Data Preprocessing**
* Importing libraries and dataset
* Handling missing values (e.g., replacing missing income values with mean)
* Feature engineering (creating new features like Age, Total\_Spent, Family\_Size)
* Encoding categorical variables
  1. **Exploratory Data Analysis (EDA)**
* Understanding data distribution using histograms and scatter plots
* Visualizing customer spending behaviour
* Analysing correlations using heatmaps
  1. **Unsupervised Learning: Clustering**
* Applying K-Means clustering to segment customers
* Finding the optimal number of clusters using the Elbow Method
* Visualizing clusters based on income and spending behaviour
  1. **Supervised Learning: Classification**
* Training machine learning models to predict customer response
* Implementing Decision Tree, KNN, and Random Forest classifiers
* Comparing model performance and accuracy
  1. **Model Evaluation and Business Insights**
* Evaluating classification models using accuracy scores
* Understanding customer segments for targeted marketing strategies

6. Design

6. a) System Design

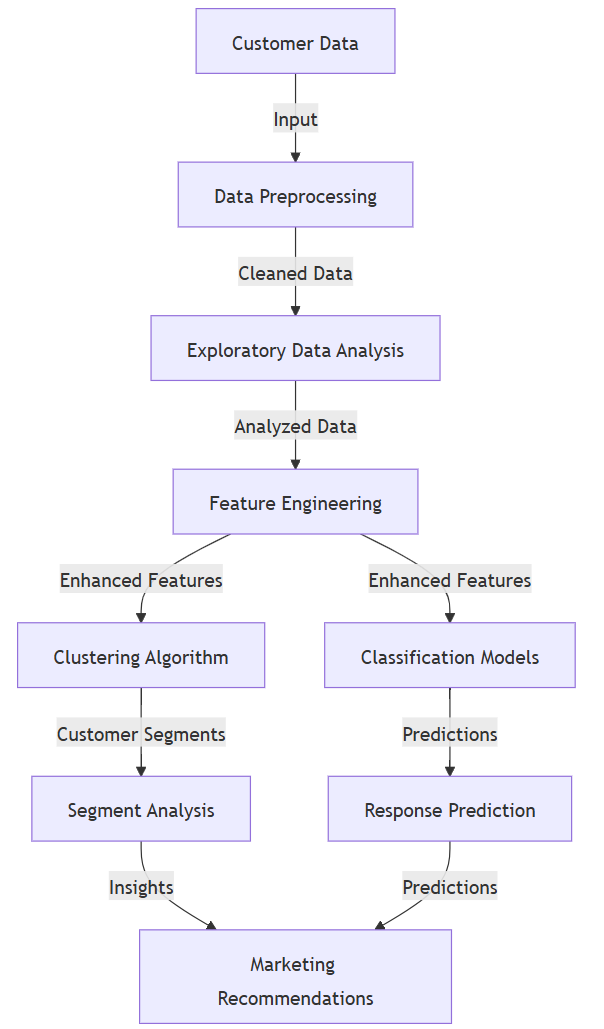
**Data Flow Diagram (DFD):**

Level 0 DFD - Customer Personality Analysis System

A diagram of customer data

AI-generated content may be incorrect.

Level 1 DFD - Customer Personality Analysis System



6. b) UML Diagrams

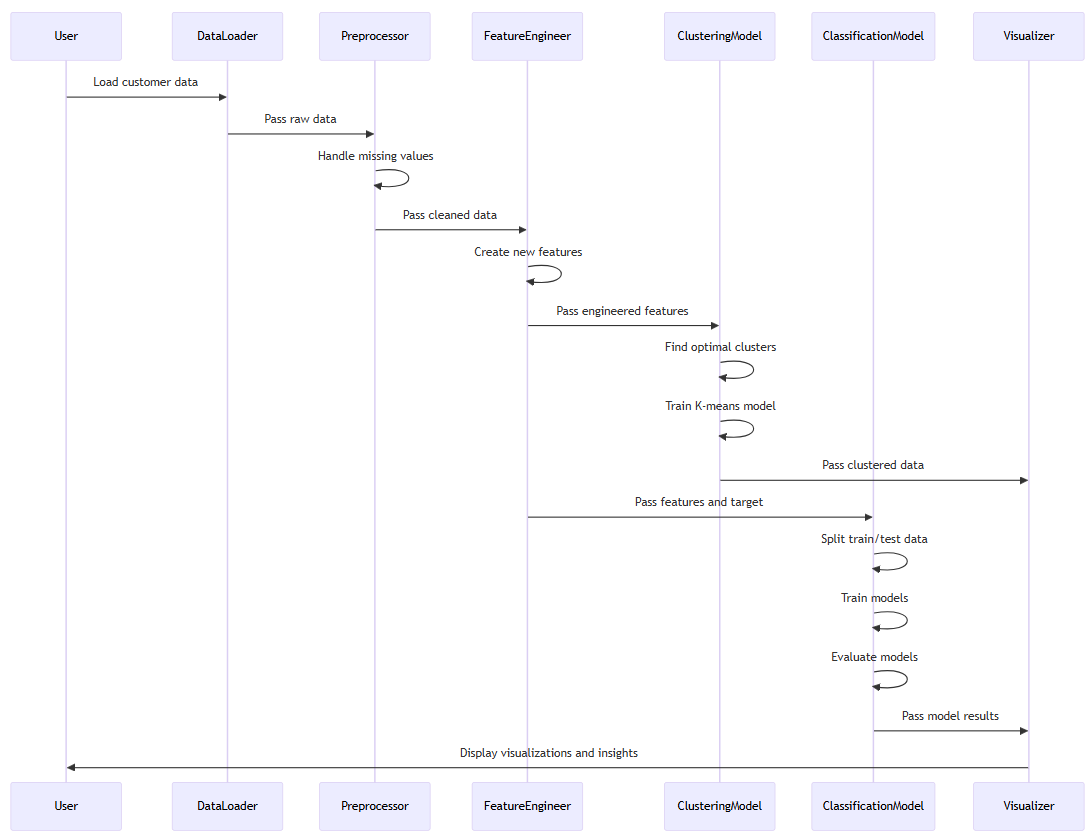
**Class Diagram** - Customer Personality Analysis

A diagram of a data processing process

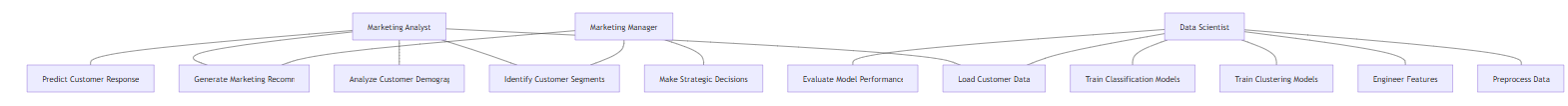
AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

**Sequence Diagram** - Customer Personality Analysis Process



**Use Case Diagram** - Customer Personality Analysis



Object Diagram

Object Diagram - Customer Segments

A screenshot of a computer screen

AI-generated content may be incorrect.

**Control Flow Diagram** - Customer Personality Analysis

A diagram of a company

AI-generated content may be incorrect.

6. c) Database Design

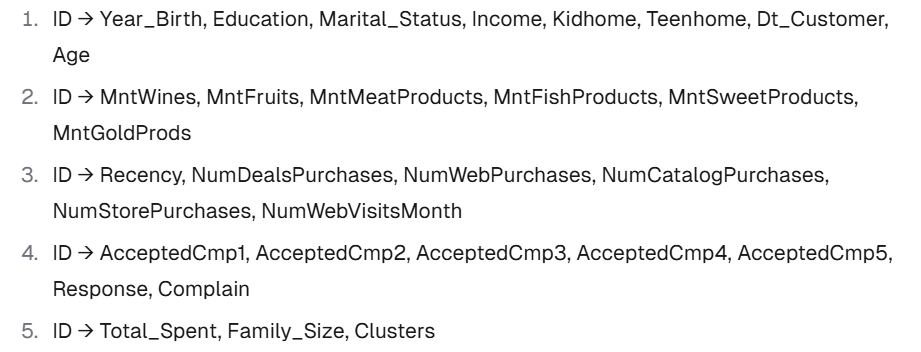
**E-R Diagram** - Customer Personality Analysis

A screenshot of a computer

AI-generated content may be incorrect.

**Functional Dependencies and Normalization**

**Functional Dependencies:**



**Normalization Process:**

**A. First Normal Form (1NF):**

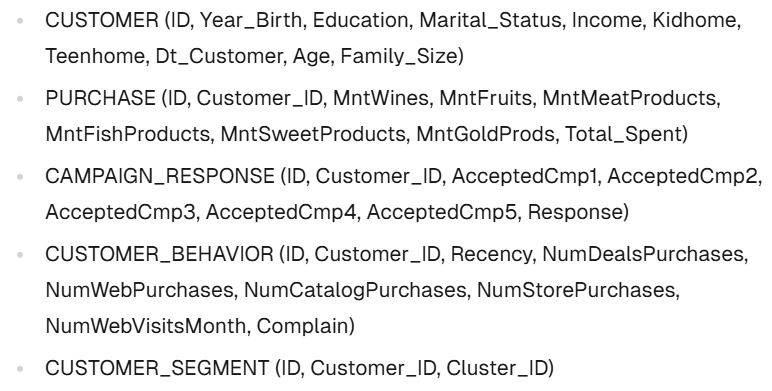
1. All attributes contain atomic values
2. No repeating groups
3. Primary key identified (ID)

**B. Second Normal Form (2NF):**

1. 1. Already in 1NF
2. 2. No partial dependencies (all attributes depend on the entire primary key)

**C. Third Normal Form (3NF):**

1. Already in 2NF
2. No transitive dependencies
3. Decomposed into:



7. Implementation

**Implementation Approach**

The implementation of this project was carried out in a modular, step-by-step manner within a Jupyter Notebook environment using Python programming language. The overall goal was to analyze customer data using both unsupervised and supervised machine learning techniques, with clear separation between data processing, model training, and evaluation phases.

**The stages of implementation included**:

* Data Ingestion
* Data Cleaning and Preprocessing
* Feature Engineering
* Data Visualization
* Modeling (Clustering and Classification)
* Evaluation and Interpretation of Results

Each of these stages was implemented using reusable and readable code blocks, allowing flexibility for changes and experimentation.

**Software Reuse and Libraries Used**

The project heavily relied on software reuse through existing open-source libraries, which significantly reduced development time and improved reliability. The following key Python libraries were reused:

| **Library** | **Purpose** |
| --- | --- |
| pandas | Data manipulation and tabular data processing |
| NumPy | Numerical operations and statistical functions |
| matplotlib | Data visualization through plots and charts |
| seaborn | Enhanced statistical data visualization |
| scikit-learn | Machine Learning algorithms for clustering, classification, and preprocessing |
| yellowbrick | Visual analysis of ML model performance (e.g., Elbow method for clustering) |

These libraries follow the industry’s best practices and are widely adopted, ensuring scalability and maintainability of the codebase.

**Special Tools Used**

| **Tool** | **Usage** |
| --- | --- |
| Jupyter Notebook | Main IDE used for developing, testing, visualizing, and documenting the project |
| Yellowbrick | Used specifically for KElbowVisualizer, which helps to find the optimal number of clusters (k) |
| scikit-learn Models | Used for KMeans Clustering, Decision Tree, KNN, and Random Forest classifiers |

The use of Jupyter Notebook provided an interactive platform to iteratively develop and debug code, visualize intermediate results, and annotate each step with Markdown documentation.

**Design Patterns and Coding Techniques**

Although the project is not object-oriented in nature, some core design principles and patterns were as follows:

* Modularity: Each step in the analysis process—such as loading data, cleaning, visualization, and modeling—was implemented in a separate code block. This aligns with the principle of *Separation of Concerns*.
* DRY (Don't Repeat Yourself): Common logic such as aggregations and visualizations were written in a reusable way with minimal repetition.
* Reusable Functions *(Optional for enhancement)*: The project could be extended further by converting repeated tasks (e.g., plot generation or model evaluation) into callable functions or class methods.
* Encapsulation of Data Transformations: Preprocessing steps such as label encoding, handling null values, and feature creation were encapsulated before feeding data to ML models.

**Data Transformation and Preprocessing Techniques**

Special coding and data preprocessing techniques included:

* Handling Missing Data: Missing Income values were replaced with the mean to preserve data without discarding rows.
* Feature Engineering:
  + Age was derived from Year\_Birth.
  + Total\_Spent was calculated as the total of all product category spendings.
  + Family\_Size was derived using the sum of Kidhome, Teenhome, and marital relationship status.
* Encoding Categorical Variables: LabelEncoder from sklearn.preprocessing was used to convert Education levels into numerical values.
* Feature Scaling *(Optional in advanced modeling)*: The use of StandardScaler was initiated to normalize feature ranges, which is beneficial for algorithms sensitive to data magnitude.

**Model Implementation Summary**

* Unsupervised Learning (KMeans Clustering):
  + The Elbow Method (via Yellowbrick’s KElbowVisualizer) was used to determine the optimal number of clusters (k=4).
  + KMeans then grouped customers into segments based on behavioral and demographic variables.
* Supervised Learning (Classification Models):
  + Three classifiers were implemented: Decision Tree, K-Nearest Neighbors (KNN), and Random Forest.
  + The Random Forest Classifier performed best with an accuracy of approximately 90.4%, indicating high prediction reliability for customer response.

**Summary**

The implementation used a structured, modular, and flexible approach that enabled:

* Efficient development using Python and industry-standard libraries
* Accurate results using tested machine learning models
* Readable, reproducible, and extensible code in a Jupyter Notebook format

This foundation allows the project to be enhanced in the future with additional models, real-time data integration, or deployment as an API or dashboard.

**Code Modules and Functionality**

**Module 1: Data Loading and Exploration**

A screenshot of a computer code

AI-generated content may be incorrect.

**Functionality**: This module imports the required libraries and loads the dataset. It then explores the basic structure of the data, including its shape, information about columns, and data types.

**Input**: Marketing campaign CSV file

**Output**: DataFrame object with loaded data and basic information about the dataset

**Module 2: Data Analysis**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Functionality**: This module performs exploratory data analysis on the dataset. It checks for null values, fills missing values, creates new features, and visualizes various aspects of the data such as age distribution, education distribution, marital status, income distribution, and family composition. It also analyzes correlations between variables and explores relationships between income and spending patterns.

**Input**: DataFrame with customer data

**Output**: Visualizations and insights about customer demographics and behaviour

**Module 3: Data Cleaning and Feature Engineering**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**Functionality**: This module cleans the data by handling missing values and performs feature engineering by creating new features such as Total\_Spent, Family\_Size, etc. It also encodes categorical variables and prepares the data for scaling.

**Input**: DataFrame with raw customer data

**Output**: Cleaned DataFrame with engineered features

**Module 4: Clustering (Unsupervised Learning)**

A screen shot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

**Functionality**: This module implements unsupervised learning using K-means clustering. It finds the optimal number of clusters using the Elbow method, applies K-means clustering with the optimal number of clusters, and visualizes the resulting clusters.

**Input**: Cleaned DataFrame with engineered features

**Output**: DataFrame with cluster assignments and visualizations of the clusters

**Module 5: Classification (Supervised Learning)**

**A screenshot of a computer code

AI-generated content may be incorrect.** **A screenshot of a computer code

AI-generated content may be incorrect.**

**Functionality**: This module implements supervised learning using various classification algorithms. It prepares the data by defining dependent and independent variables, splits the data into training and testing sets, and trains and evaluates three different classification models: Decision Tree, K-Nearest Neighbors, and Random Forest.

**Input**: DataFrame with features and target variable

**Output**: Trained classification models and their performance metrics

**Database Tables**

**Database Table Explanation -** CustomerProfile

This table contains detailed records of customer demographic information, lifestyle indicators, purchasing behavior, and responses to marketing campaigns. Each row represents one customer.

Structure of the CustomerProfile Table

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| ID | Integer | Unique identifier for each customer |
| Year\_Birth | Integer | Year the customer was born |
| Education | Categorical | Education level (e.g., Graduation, PhD, Master) |
| Marital\_Status | Categorical | Marital status (e.g., Married, Single, Divorced) |
| Income | Float | Annual income of the customer |
| Kidhome | Integer | Number of children living at home |
| Teenhome | Integer | Number of teenagers living at home |
| Dt\_Customer | Date | Date the customer enrolled with the company |
| Recency | Integer | Number of days since last purchase |
| MntWines | Integer | Amount spent on wine products |
| MntFruits | Integer | Amount spent on fruit products |
| MntMeatProducts | Integer | Amount spent on meat products |
| MntFishProducts | Integer | Amount spent on fish products |
| MntSweetProducts | Integer | Amount spent on sweet products |
| MntGoldProds | Integer | Amount spent on gold products |
| NumDealsPurchases | Integer | Number of purchases made using a discount deal |
| NumWebPurchases | Integer | Number of purchases made via the company website |
| NumCatalogPurchases | Integer | Number of purchases made using a catalog |
| NumStorePurchases | Integer | Number of purchases made in a physical store |
| NumWebVisitsMonth | Integer | Number of visits to the website in the last month |
| AcceptedCmp1 to AcceptedCmp5 | Binary | Indicates if the customer accepted each of 5 previous marketing campaigns |
| Response | Binary | Indicates if the customer accepted the last campaign |
| Complain | Binary | Indicates if the customer complained in the last 2 years |
| Z\_CostContact | Constant | Cost of customer contact (constant for all entries) |
| Z\_Revenue | Constant | Revenue from customer contact (constant for all entries) |
| Age | Integer | Derived field: Customer's age |
| Total\_Spent | Integer | Derived field: Total amount spent across product categories |
| Family\_Size | Integer | Derived field: Total number of family members (self + kids/teens + partner) |
| Education (encoded) | Integer | Label-encoded version of Education |
| Clusters | Integer | Cluster ID assigned after KMeans clustering |

**Notes on Derived Fields**

* Age = 2022 - Year\_Birth
* Total\_Spent = Sum of all product spending columns
* Family\_Size = Sum of Kidhome, Teenhome, and inferred relationship count
* Clusters are results from unsupervised KMeans clustering
* Response is the target variable used for classification in supervised learning

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a white box

AI-generated content may be incorrect.

A screenshot of a data type

AI-generated content may be incorrect. A screenshot of a phone

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a phone

AI-generated content may be incorrect.

A screenshot of a data type

AI-generated content may be incorrect. A screenshot of a phone

AI-generated content may be incorrect.

A white screen shot of a document

AI-generated content may be incorrect.

8. Testing

Testing plays a crucial role in verifying the correctness, accuracy, and performance of the implemented system. For this project, which focuses on data analysis and machine learning, the testing approach ensures that:

* The data is correctly preprocessed and transformed,
* The machine learning models behave as expected,
* Predictions are accurate and aligned with the objective,
* Code functions produce valid and interpretable outputs.

**Testing Approach**

Given the nature of the project, the following types of testing were applied:

* **Unit Testing**: To verify that individual functions such as data cleaning, feature engineering, encoding, and model training behave as expected.
* **Data Validation Testing**: Ensuring that data loading, null value handling, and transformations preserve data integrity.
* **Functional Testing**: Testing whether the pipeline—from raw data to visualization, modeling, and prediction—executes as intended.
* **Model Evaluation Testing**: Comparing predicted outcomes with actual values to determine classification model accuracy using performance metrics like accuracy score.

**Lessons Learnt from Testing**

* Early testing of preprocessing steps prevents downstream issues in model training.
* Data imbalance and feature skew can affect model performance—important to validate assumptions with plots.
* Regular checking of data types and nulls is essential in an ML pipeline.
* Model testing using different algorithms helped identify Random Forest as the most robust performer.

A screenshot of a test

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a test

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a test

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a test

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

9. Output Screens

Data Exploration and Analysis

Figure 1: Age distribution of customers

A graph of a red line

AI-generated content may be incorrect.

Figure 2: Pie chart showing education distribution

A pie chart with different colored circles

AI-generated content may be incorrect.

Figure 3: Count plot of marital status

A graph of different colored bars

AI-generated content may be incorrect.

Figure 4: Distribution of customer income

A graph of income

AI-generated content may be incorrect.

Figure 5: Correlation heatmap of numerical variables

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 6: Scatter plots showing relationship between income and spending categories

A group of blue dots

AI-generated content may be incorrect.

Figure 7: Elbow method for determining optimal number of clusters

A graph with a line and a line

AI-generated content may be incorrect.

Figure 8: Distribution of customers across clusters

A graph with different colored bars

AI-generated content may be incorrect.

Figure 9: Scatter plot showing cluster profiles based on income and total spending

A graph showing a number of colored dots

AI-generated content may be incorrect.

**Classification Results**

A screenshot of a computer

AI-generated content may be incorrect.

10. Tools and Technologies

**Programming Language:**

- Python: Primary programming language used for data analysis, preprocessing, visualization, and modeling.

**Data Analysis and Manipulation:**

- Pandas: Used for data manipulation and analysis, providing data structures and operations for manipulating numerical tables and time series.

- NumPy: Used for numerical computations, providing support for large, multi-dimensional arrays and matrices.

**Data Visualization:**

- Matplotlib: Used for creating static, animated, and interactive visualizations in Python.

- Seaborn: Built on top of Matplotlib, used for making statistical graphics more attractive and informative.

- Yellowbrick: Used specifically for machine learning visualization, particularly for the Elbow method in clustering.

**Machine Learning:**

- Scikit-learn: Used for machine learning algorithms implementation, including:

- Preprocessing tools (LabelEncoder, StandardScaler)

- Clustering algorithms (KMeans, AgglomerativeClustering)

- Classification algorithms (DecisionTreeClassifier, KNeighborsClassifier, RandomForestClassifier)

- Model evaluation metrics

- Train-test splitting functionality

**Development Environment**

- Jupyter Notebook: Interactive computing environment used for developing and documenting the analysis.

- Anaconda: Distribution of Python used for scientific computing, which includes many of the packages used in this project.

**Version Control**

- Git: Used for version control and collaboration.

**Database**

- CSV: The data was stored in CSV format, which was processed using Pandas.

**Methodologies**

- Exploratory Data Analysis (EDA): Used to analyze and investigate data sets and summarize their main characteristics.

- Feature Engineering: Process of using domain knowledge to extract features from raw data.

- Unsupervised Learning: Used K-means clustering to segment customers without labelled data.

- Supervised Learning: Used classification algorithms to predict customer response to campaigns.

- Cross-validation: Used to evaluate model performance and prevent overfitting.

This comprehensive set of tools and technologies enabled efficient data processing, insightful analysis, and effective modelling for the Customer Personality Analysis project.

11. Conclusion

**Summary and Key Achievements**

The primary objective of this project, *Customer Personality Analysis*, was to extract meaningful insights from customer data and identify patterns that can help in effective customer segmentation and targeting. Leveraging the power of **Machine Learning**, the project focused on analyzing customer demographics, behavioral attributes, and marketing response data to understand the personality traits of customers and group them accordingly.

The project was carried out in the following structured manner:

* **Data Preprocessing and Cleaning**: Handled missing values (especially in the Income column), created derived features like Age, Total\_Spent, and Family\_Size, and performed label encoding.
* **Exploratory Data Analysis (EDA)**: Visualized various relationships within the data such as age distribution, income variation, family structure, and spending behavior.
* **Unsupervised Learning – Clustering**: Implemented **K-Means Clustering** to group customers into distinct clusters based on attributes like income, spending, and family profile. The **Elbow Method** was used to determine the optimal number of clusters (k = 4).
* **Supervised Learning – Classification Models**: Built predictive models using **Decision Tree**, **K-Nearest Neighbors (KNN)**, and **Random Forest** to classify customer responses to marketing campaigns. Among these, the **Random Forest** classifier achieved the highest accuracy (~90.4%), making it a strong choice for marketing response prediction.

**Limitations and Lessons Learnt**

Despite the success of the project, a few limitations were observed:

* The dataset was relatively limited in scope, and customer behaviors may vary over time and across regions—an aspect not covered due to static data.
* Some categorical fields, such as Marital\_Status, had inconsistent values that required manual consolidation.
* No deep learning models were explored due to time and resource constraints.

Key lessons learnt include:

* The importance of **data preprocessing and feature engineering** in improving model performance.
* How combining **EDA with domain understanding** can guide meaningful feature creation.
* Practical knowledge of how different machine learning algorithms behave with real-world business data.

**Further Enhancements / Recommendations**

Future work in this area could consider the following enhancements:

* **Incorporate time-series elements** to study how customer behavior evolves over time.
* **Enrich the dataset** with external data sources (e.g., transaction history, website interactions, geographic data).
* Apply **Dimensionality Reduction techniques** like PCA or t-SNE for improved visualization and model performance.
* Explore advanced models such as **XGBoost**, **LightGBM**, or **Deep Neural Networks** for potentially better prediction accuracy.
* Develop an **interactive dashboard** using tools like Power BI or Tableau to make the insights accessible to marketing teams.

12. Appendices

This section includes supplementary information that supports the main content of the report. The materials provided here offer deeper insight into the implementation details, usage guidelines, and supporting artifacts that may otherwise interrupt the flow if placed within the main chapters.

**Appendix A: User Documentation**

**Project Title**: *Customer Personality Analysis using Machine Learning*  
**Objective**: To analyze customer profiles and behavior using Machine Learning techniques to assist marketing teams in better segmenting and targeting their customer base.

**Functionality Overview**:

* Load and clean the dataset
* Perform Exploratory Data Analysis (EDA)
* Engineer new features such as Age, Total\_Spent, and Family\_Size
* Cluster customers using K-Means
* Build classification models using Decision Tree, KNN, and Random Forest
* Visualize clusters and classifier results

**Tools Used**:

* Programming Language: Python 3.x
* Platform: Jupyter Notebook
* Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, yellowbrick

**Appendix B: Installation Instructions**

To run the project locally, follow these steps:

1. **Prerequisites**:
   * Python 3.x installed (preferably 3.8 or higher)
   * Jupyter Notebook (install via Anaconda or pip)
2. **Installation Steps**:

|  |
| --- |
| 1. pip install pandas 2. pip install numpy 3. pip install matplotlib 4. pip install seaborn 5. pip install scikit-learn 6. pip install yellowbrick |

**3. Launch the Jupyter Notebook**:

* Navigate to the project directory and open the notebook using:

|  |
| --- |
| jupyter notebook |

* Open the notebook file Customer Personality Analysis.ipynb and run all cells in order.

**Appendix C: README – How to Interact with the System**

**Step-by-step instructions to use the project:**

1. **Load the dataset**:  
   The dataset marketing\_campaign.csv is loaded using pandas.read\_csv(). Ensure the file is present in the same directory as the notebook.
2. **Execute Data Cleaning Cells**:  
   Run preprocessing cells to handle missing values, create derived columns, and prepare the data.
3. **Visualize Data**:  
   Run EDA cells to understand the dataset's structure and trends using charts and graphs.
4. **Run Machine Learning Models**:  
   Execute the clustering and classification cells to generate and view results.
5. **Understand the Output**:
   * Cluster labels will be added to the dataset
   * Classification accuracy will be printed for each model
   * Visualizations will help interpret clusters and predictions

**Appendix D: Sample Source Code**

|  |
| --- |
|  |



|  |  |
| --- | --- |
| **Appendix E: Glossary** |  |
|  |  |
|  |  |
|  |  |

13. References / Bibliography

1. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
2. Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.
3. Dataset Source: *Marketing Campaign Data* – Provided for academic and educational use.
4. Official Documentation:
   * [Scikit-learn](https://scikit-learn.org/stable/)
   * [Pandas](https://pandas.pydata.org/)
   * [Seaborn](https://seaborn.pydata.org/)
   * [Matplotlib](https://matplotlib.org/)
5. Online tutorials and resources from:
   * Kaggle: [https://www.kaggle.com](https://www.kaggle.com/)

Towards Data Science: [https://towardsdatascience.com](https://towardsdatascience.com/)

**Thank You**