CUSTOMER SEGMENTATION AND REVIEW SENTIMENT ANALYSIS USING MACHINE LEARNING MODELS

A PROJECT REPORT

In partial fulfilment of the Requirements for the award of the degree of

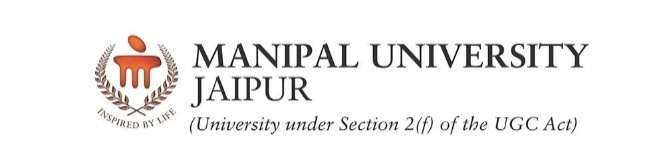
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING USING PYTHON

Under the guidance of

MAHENDRA DUTTA

By

DEBDATTA BHOUMIK

ONLINE MANIPAL

UNIVERSITY JAIPUR

In association with

(ISO 9001:2015)

Declaration

I hereby declare that the project work being presented in the project proposal entitled “Customer Segmentation and Review Sentiment Analysis using Machine Learning Models” in partial fulfilment of the requirements for the award of the degree of **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING USING PYTHON** at **ONLINE MANIPAL UNIVERSITY** **JAIPUR** is an authentic work carried out under the guidance of Mahendra Dutta. The matter embodied in this project work has not been submitted elsewhere for the award of any degree of our knowledge and belief.

Date: 04/04/2025

Name of the Student: DEBDATTA BHOUMIK



Ardent Computech Pvt. Ltd (An ISO 9001:2015 Certified)

CERTIFICATE

This is to certify that this proposal of the minor project entitled “**Customer Segmentation and Review Sentiment Analysis using Machine Learning Models**” is a record of bona fide work, carried out by **DEBDATTA BHOUMIK** under my guidance at Ardent Computech Pvt. Ltd. In my opinion, the report in its present form is in partial fulfilment of the requirements for the award of the degree of **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING USING PYTHON** and as per regulations of the Ardent Computech Pvt. Ltd. To the best of my knowledge, the results embodied in this report, are original and worthy of incorporation in the present version of the report.

Guide / Supervisor

------------------------------------------------

**MR. MAHENDRA DUTTA**

Project Engineer Ardent Computech Pvt. Ltd (An ISO 9001:2015 Certified)

Acknowledgement

The success of any project depends largely on the encouragement and guidelines of many others. I take this sincere opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project work. I want to show our most significant appreciation to **MAHENDRA DUTTA**, Project Engineer at Ardent Computech Pvt. Ltd. I always feel motivated and encouraged every time by his valuable advice and constant inspiration; without his encouragement and guidance, this project would not have materialized. Words are inadequate in offering our thanks to the other trainees, project assistants and other members at Ardent Computech Pvt. Ltd. for their encouragement and cooperation in carrying out this project work. The guidance and support received from all the members who are contributing to this project were vital for the success of this project.

**Abstract**

This project aims to revolutionize marketing strategies by leveraging AI and machine learning techniques to harness customer data effectively. It focuses on two critical components: **Customer Segmentation Using ML Models** and **Customer Review Sentiment Analysis Using NLP**.

The **Customer Segmentation** component applies machine learning algorithms to analyse structured data and group customers into distinct segments based on their purchasing behaviour, demographics, and engagement patterns. By uncovering these natural groupings, businesses can personalize marketing campaigns, optimize resource allocation, and enhance customer engagement.

The **Sentiment Analysis** component uses natural language processing (NLP) to analyse customer reviews, transforming unstructured feedback into valuable insights. It determines the sentiment behind customer opinions—positive, negative, or neutral—by employing advanced techniques such as transformer-based models. This enables businesses to detect emerging trends, assess product performance, and improve customer experience in real-time.

Together, these methodologies create a synergistic approach that offers a holistic view of customer behaviour and sentiment. The integration of segmentation and sentiment analysis drives precise marketing strategies, enhances customer satisfaction, and ensures sustainable growth in competitive markets. This project bridges the gap between raw data and actionable insights, positioning businesses for success in the modern marketing landscape.

In this research, a machine learning based customer segmentation model was proposed. The model was implemented using four classifier algorithms, which include **K-Means Clustering, Random Forest Classifier, Principal Component Analysis (PCA), and Grid Search CV.**



Chapters page no.

Abstract 5

Chapter 1: Introduction 7- 8

Chapter 2: Related Work 9

Chapter 3: Literature Survey 10- 11

Chapter 4: Methodology 12- 25

* Proposed Model 12
* Data acquisition and pre- processing 13- 14
* Data Visualization 15- 16
* Analysis 17- 19

Chapter 5: Evaluation and Result 26- 32

Chapter 6: Result Analysis 33

Chapter 7: Conclusion and Future Work 34- 35

Chapter 8: References 36- 37



**Chapter- 1: Introduction**

In today's hyperconnected and data-driven world, marketing has transformed into a sophisticated discipline where insight and precision are paramount. The project, titled Customer Segmentation Using ML Models and Customer Review Sentiment Analysis Using NLP, leverages cutting-edge artificial intelligence (AI), machine learning (ML), and natural language processing (NLP) techniques to convert vast amounts of data into actionable strategies that drive business growth.

**Customer Segmentation Using ML Models**

At the heart of any successful marketing strategy is the ability to understand customers deeply. By applying ML models, this project dissects structured transactional and behavioural data to identify natural groupings or segments within the customer base. Techniques such as K-means clustering, hierarchical clustering, or even advanced methods like density-based clustering are employed to reveal hidden patterns in purchasing behaviours, demographics, and interactions. These clusters allow businesses to create personalized marketing campaigns, optimize advertising budgets, and build robust customer profiles that guide strategic decisions. The segmentation process not only fine-tunes targeted marketing efforts but also helps in predicting customer lifetime value and enhancing overall customer engagement.

**Customer Review Sentiment Analysis Using NLP**

Customer reviews are potent indicators of market sentiment and provide nuanced insights into the consumer experience. This component of the project focuses on mining unstructured data from reviews to gauge customer sentiment accurately. Employing state-of-the-art NLP algorithms, the model classifies reviews into positive, negative, or neutral sentiments—analysing context, tone, and specific word patterns. Such sentiment analysis transcends the simple aggregate rating, revealing subtle shifts in customer perception and emerging trends. The extracted insights can drive improvements in product development, customer service, and overall brand management, ensuring that the voice of the customer is at the forefront of business strategy.

**Synergizing ML and NLP for Holistic Marketing Intelligence**

Together, customer segmentation and sentiment analysis forge a powerful toolkit for modern marketers. While ML-based segmentation answers the critical question of "who" the customers are by grouping them into actionable segments, NLP-based sentiment analysis delves into "what" customers are saying about their experiences. This combined approach provides a 360-degree view of the market, from understanding distinct customer groups to analysing the emotional tone of their feedback. As businesses continue to navigate a competitive landscape, this holistic strategy empowers decision-makers to implement precisely targeted marketing campaigns, ultimately leading to enhanced customer satisfaction and sustainable revenue growth.

By integrating these methodologies, the project not only harnesses the potential of advanced algorithms but also bridges the gap between raw data and strategic marketing insights. This innovative fusion is poised to redefine how businesses understand consumer behaviour and craft market campaigns, making it an essential tool in the modern marketing toolkit.

The objectives are as follows:

1. **Customer Segmentation Using ML Models**

* Identify distinct customer groups based on behaviour, demographics, and preferences.
* Create actionable profiles for personalized marketing and engagement strategies.
* Enhance resource allocation by prioritizing high-value customer segments.
* Improve customer retention through data-driven targeting and loyalty programs.

1. **Customer Review Sentiment Analysis Using NLP**

* Analyse customer feedback to classify sentiments as positive, negative, or neutral.
* Monitor sentiment trends to detect emerging issues or successes in real-time.
* Use sentiment insights to refine products and services for better customer satisfaction.
* Strengthen brand reputation by addressing customer concerns proactively.



Chapter 2: **Related Work**

In today's data-driven world, artificial intelligence plays a key role in our society, Researchers have successfully applied several algorithms and models to predict market demand by using various demographic and psychographic features, attributes and data generated from various sources.

**Customer Segmentation Using ML Models**

* **Evolution from Traditional Methods:**  
  Early segmentation efforts relied on rule-based and statistical methods (like RFM analysis) to manually group customers. However, with the advent of machine learning, researchers and practitioners have moved toward data-driven approaches that can uncover hidden patterns beyond obvious demographics or transactional thresholds.
* **Machine Learning Techniques:**  
  Several studies and practical implementations have explored the use of unsupervised learning algorithms. For example, **K-Means clustering** and **hierarchical clustering** have been widely adopted to segment customers based on behavioural and demographic attributes. Work highlighted in tutorials on platforms like PythonGuides and Analytics Vidhya demonstrates how these models can effectively identify multiple customer groups from large datasets.
* **Industry Best Practices:**  
  Blogs from industry experts, such as those from SoftTeco, have detailed the challenges of traditional segmentation and showcased how integrating machine learning leads to more agile and precise customer targeting strategies. These works emphasize the benefits of dynamic segmentation, which adapts as customer behaviour evolves.

**Customer Review Sentiment Analysis Using NLP**

* **Lexicon-Based to Deep Learning Approaches:**  
  Early sentiment analysis work typically used lexicon-based approaches (like using SentiWordNet or VADER) to classify customer opinions. This method provided a foundational understanding of sentiment by mapping words to predefined sentiment scores.
* **Advancements with Modern NLP:**  
  More recent work has shifted towards leveraging sophisticated NLP techniques, with transformer-based models such as BERT. These models capture context and subtle nuances in language, improving the accuracy of sentiment classification. Tutorials and real-world examples on platforms like Codezup illustrate how to build and fine-tune these models for practical applications.
* **Application and Impact:**  
  Studies have shown how sentiment analysis can be directly applied to monitor real-time customer feedback. By analysing review text, these systems help capture trends in customer sentiment, offer early warnings of potential issues, and fuel improvements in product development and customer service.

**Summary**

Overall, past work in both domains has paved the way for integrating advanced machine learning and NLP techniques into marketing analytics. On the segmentation side, moving from static, manually curated groups to dynamic clusters driven by ML have significantly enhanced targeted marketing strategies. Meanwhile, in sentiment analysis, the progression from simple lexicon-based methods to deep learning models has made it possible to derive richer, contextual insights from customer feedback.

Chapter 3: Literature Survey

Below is a literature survey that summarizes key advances and major studies in both customer segmentation using ML models and customer review sentiment analysis using NLP.

**Customer Segmentation Using ML Models**

* **Clustering-Based Approaches:**  
  Several studies have centred on applying clustering techniques to segment customers. For example, research published in the *International Journal of Contemporary Research in Technology (IJCRT)* employed the k-means algorithm to partition retail customers based on behavioural data like purchasing frequency and transaction values. This work demonstrated that ML-based segmentation can uncover latent customer groups that traditional RFM approaches may overlook.
* **Comprehensive Literature Reviews:**  
  A literature review featured in the *AIP Conference Proceedings* provided an extensive survey of customer segmentation techniques using ML. The authors discussed various clustering methods (K-Means, hierarchical clustering, DBSCAN) and highlighted the importance of dynamic segmentation to adjust to evolving customer behaviours. Their work also pinpointed challenges in data quality and feature engineering when applying these techniques.
* **Behavioural Segmentation Studies:**  
  Another significant contribution is seen in research presented in the *International Journal of Novel Research in Data (IJNRD)*. This study focused on segmenting customers based on behavioural characteristics (such as spending patterns and engagement levels) using k-means clustering. The findings demonstrated that targeted segmentation not only enhances marketing strategies but also optimizes resource allocation for customer retention.

These studies collectively underscore the evolution from traditional, rule-based segmentation to sophisticated, data-driven approaches that leverage multi-dimensional customer data and advanced clustering algorithms.

**Customer Review Sentiment Analysis Using NLP**

* **Comparative Studies with Machine Learning and Deep Learning:**  
  A comparative study published in *Computers* examined various sentiment analysis techniques on customer reviews. This research compared traditional classifiers (like Logistic Regression, Naïve Bayes) and deep learning models (such as CNNs and RNNs), highlighting that deep learning methods could better capture complex linguistic nuances in text. Their experiments on Amazon product reviews showcased significant improvements in sentiment classification accuracy with deep learning approaches.
* **Multilingual and Context-Aware Analysis:**  
  Research presented in the *IJRASET Journal for Research in Applied Science and Engineering Technology* discussed sentiment analysis on e-commerce product reviews. The study addressed challenges like language diversity by incorporating language identification and translation before performing sentiment analysis using tools such as VADER. This approach enabled a more robust analysis of reviews across different languages, thereby enhancing the overall understanding of customer feedback.
* **Application in Location-Based Reviews:**  
  Another intriguing study focused on sentiment analysis of Google Map reviews, combining traditional NLP techniques with machine learning. This work tackled the issue of noisy, informal language typical of user-generated content on maps and location-based services. The study demonstrated that even in such challenging contexts, integrating NLP with ML techniques can yield reliable sentiment insights that inform business decisions.

These investigations illustrate the progression from early lexicon-based approaches to advanced models employing deep neural networks and transformer architectures. They emphasize the importance of context and domain adaptation in accurately interpreting customer sentiments.

**Summary**

Past research in customer segmentation has moved from simple statistical models to advanced clustering algorithms that can dynamically adapt to customer behaviour using ML. Meanwhile, sentiment analysis research has evolved from basic rule-based methods to sophisticated deep learning and transformer-based models capable of handling complex, multilingual, and noisy text data. These studies provide the foundation for today’s integrated approaches in marketing analytics, aiming to tailor marketing strategies and improve customer satisfaction based on nuanced data insights.

Chapter 4: Methodology

* 1. For Customer Segmentation

In this section, I present the proposed model for Customer Segmentation using machine learning. It also contains a description of the dataset, its acquisition and pre-processing and the analysis of the Algorithm used.

**3.1.a) Proposed Model**

Input Data

Data pre- processing

Attribute Selection

Training data

Testing Data

Machine Learning Algorithm

Output: Predicted Result

**3.2.b) Data acquisition and pre-processing**

There are two datasets on which I have performed Customer Segmentation using Machine Learning Models- one dataset is of Mall Customers and another is of Online Retail. The 1st dataset has a CSV file format, i.e., Comma Separated Values consisting of 200 rows and 5 columns, the columns predominantly being named as follows:

* Customer ID
* Gender
* Age
* Annual Income (k$)
* Spending Score (1- 100)

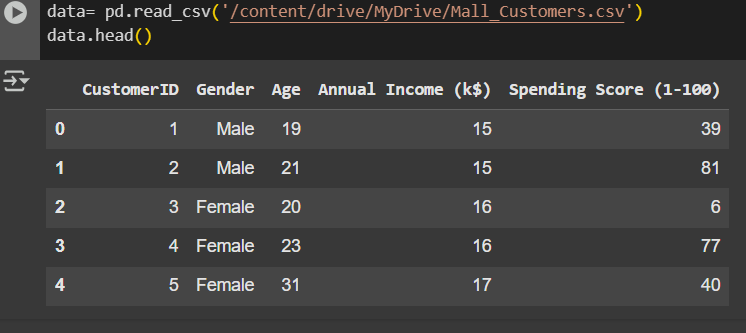


Fig 3.1 Data Head Overview

The 2nd dataset, namely Online Retail, has an XLSV file format, consisting of 541909 rows and 8 columns, the columns are as follows:

* Invoice No
* Stock Code
* Description
* Quantity
* Invoice Date
* Unit Price
* Customer ID
* Country



Fig 3.2 Data Head Overview

To achieve accurate segmentation and ensure high performance of the algorithm, the data must be pre-processed.

Data pre-processing consists of transforming the acquired data into an understandable format, removing duplicate or null values, and dropping undesired attributes. For this analysis, the data was pre-processed by removing the column date as it was deemed irrelevant; the dataset was complete and had no null values.

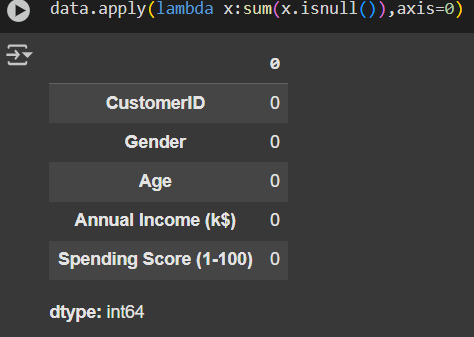
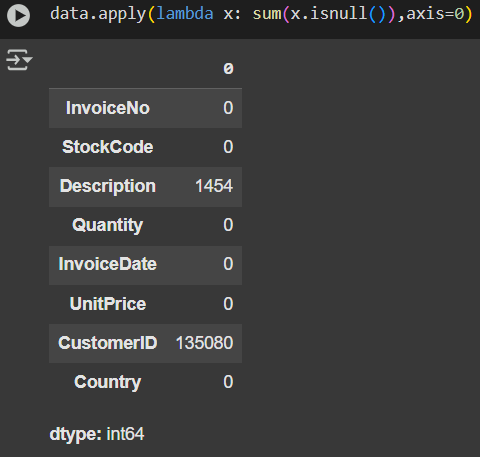
 

Fig 3.3 Checking Null Values

Since in the 1st dataset, there are no null values, they do not need to be treated. But in the 2nd dataset, the null values need to be replaced.



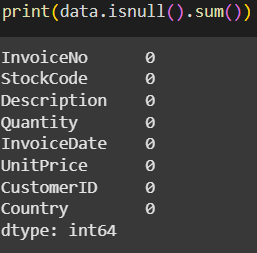


Fig 3.4 Checking data completeness

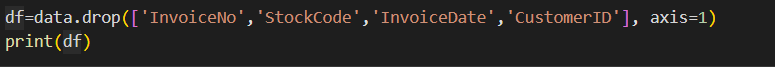
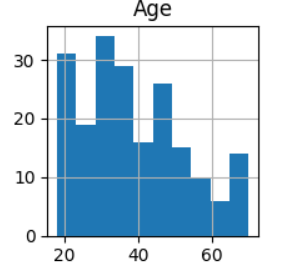
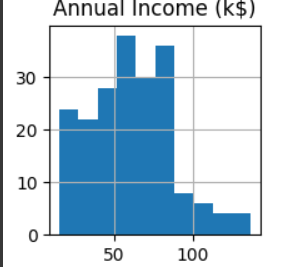


Fig 3.5 Removing the undesired columns

Following the pre-processing of the dataset, to better understand the attributes and their relationship, a graph count of the attributes was generated.

**3.3.a) Data Visualization**

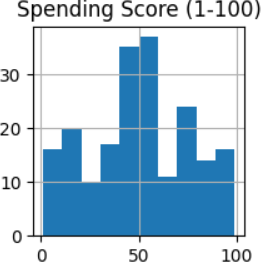
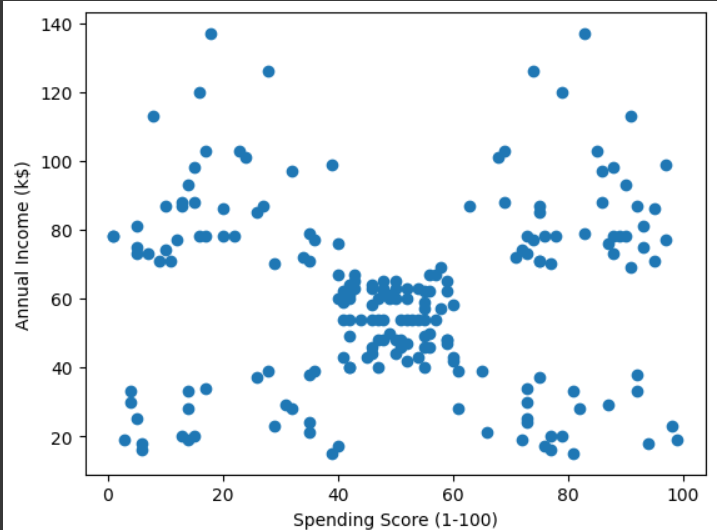
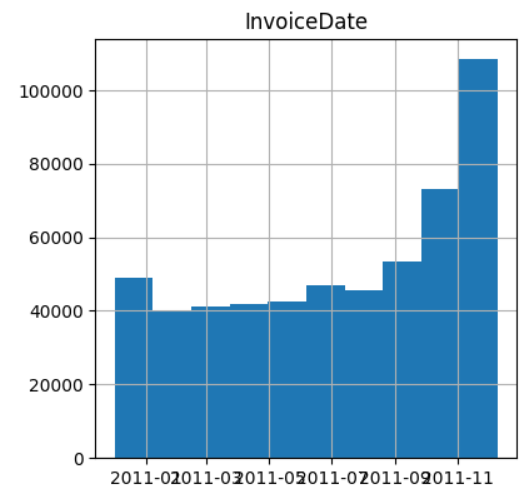
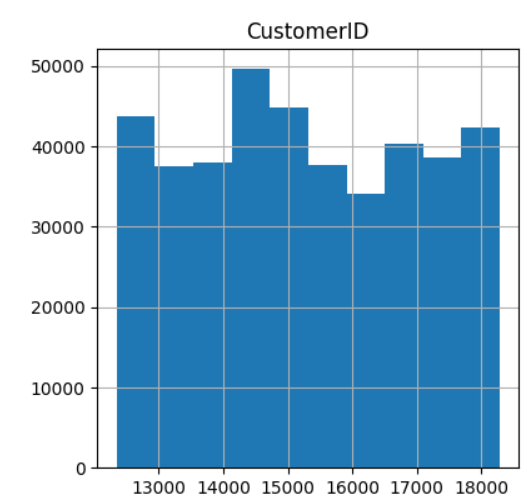
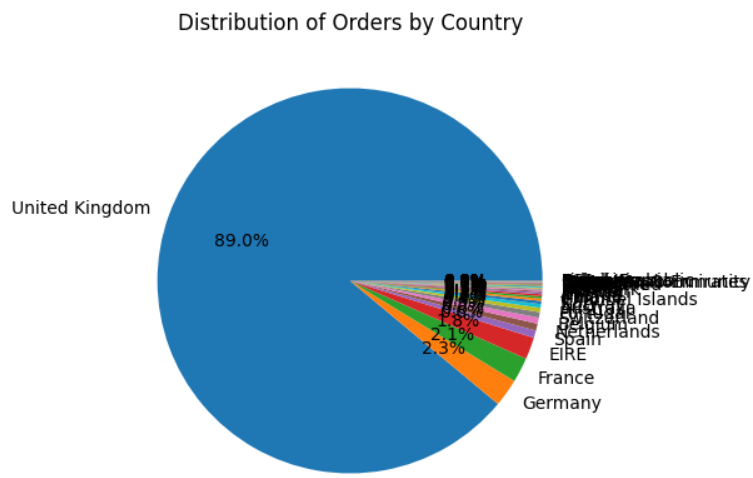
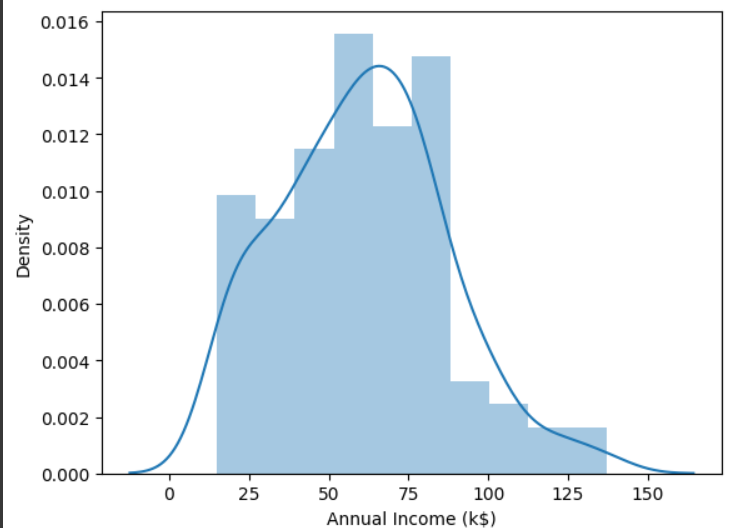
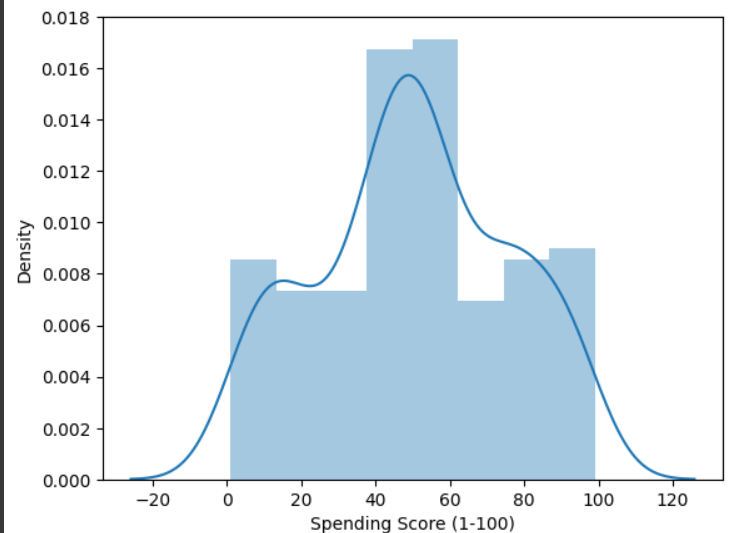
 

Fig 3.5 Data Visualisation for MALL CUSTOMERS

 Fig 3.6 Data Visualisation for ONLINE RETAIL

**3.4.a) Analysis**

** **

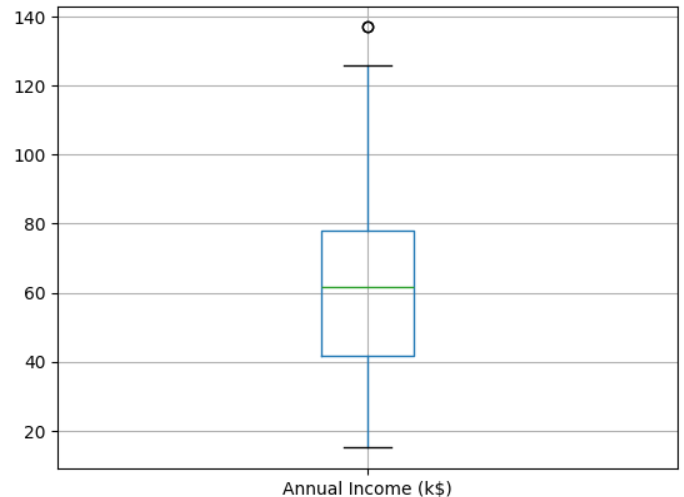
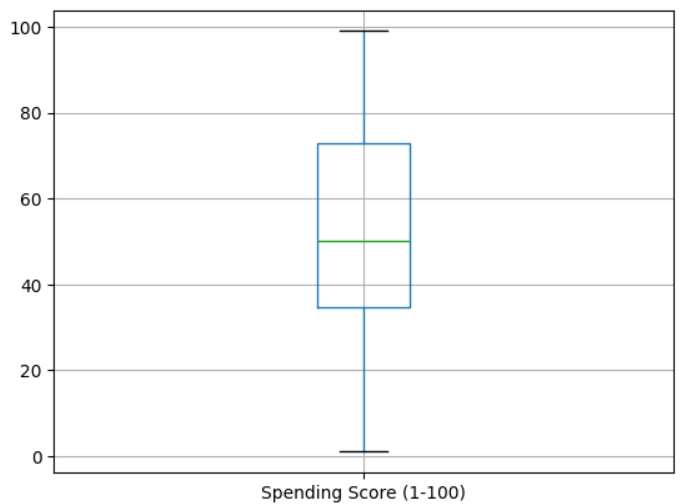
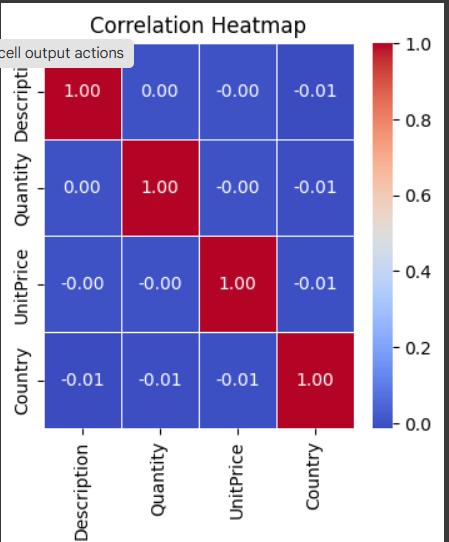
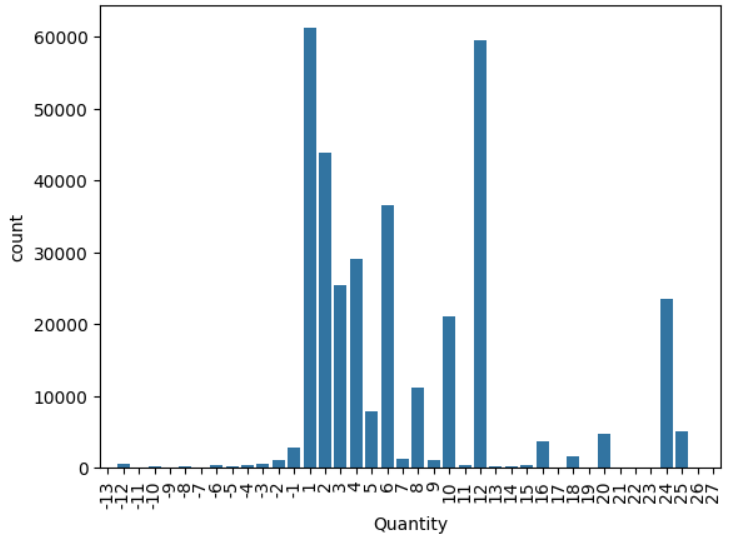
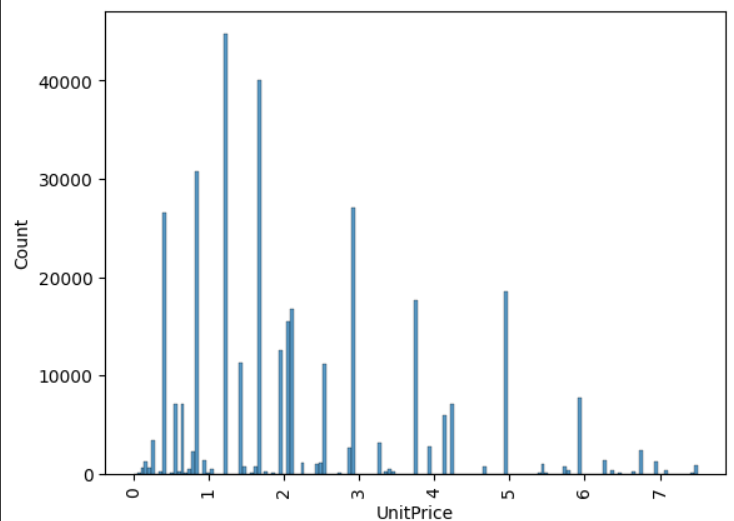
** **

Figure 3.7 Analysis for MALL CUSTOMERS

 Fig 3.8 Analysis for ONLINE RETAIL

For the analysis of our data, the data was split into a test and a Training dataset, using 20% for testing. Four machine learning classification algorithms were implemented, they are:

* K-Means Clustering
* Random. Forest classifier
* Principal Component Analysis (PCA)
* Grid Search CV

**K-Means Clustering**

It is a popular unsupervised clustering algorithm used in data analysis and machine learning. It partitions a dataset into K distinct clusters, based on their feature similarities. The algorithm iteratively assigns data points to clusters by minimizing the distance between points and their corresponding cluster centroids. The centroids are then recalculated until convergence or when changes stabilize. Its simplicity and efficiency make it ideal for discovering patterns in unlabelled data.

**Random Forest Classifier**

A random forest classifier is a collection of tree-structured classifiers whose results are compounded into one result; it is an ensemble machine learning algorithm which can be implemented for both classification and regression tasks and is made up of a set of classifiers known as a decision tree, random forest classifier is known to produce accurate predictions, provides flexibility and reduced the risk of overfitting.

**Principal Component Analysis**

PCA is a dimensionality reduction technique used in statistics and machine learning. It transforms a high-dimensional dataset into a smaller number of dimensions (called principal components) while preserving as much variance as possible. This method simplifies data visualization and analysis by identifying the most important patterns, reducing noise, and minimizing information loss. It's widely used in areas like data preprocessing, compression, and feature extraction.

**Grid Search CV**

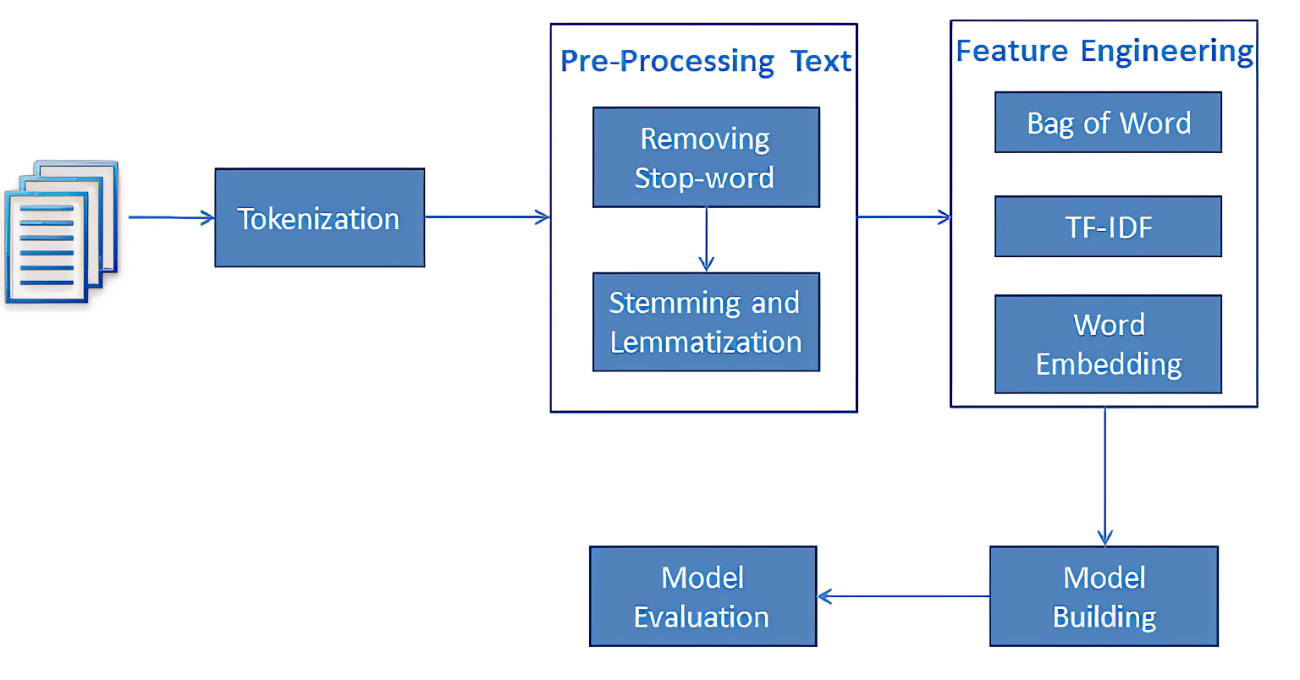
It stands for **Grid Search Cross-Validation**. It's a technique used in machine learning to optimize the hyperparameters of a model. It systematically searches through a specified hyperparameter grid, evaluates different combinations using cross-validation, and selects the combination that provides the best performance. It ensures the model is tuned for better accuracy and reliability.



* 1. ***For Sentiment Analysis***

In this section, I present the proposed model for Customer Review Sentiment Analysis using Natural Language Processing (NLP), it also contains a description of the dataset, its acquisition and pre-processing and the analysis of the Algorithm used.

**3.1.b) Proposed Model**



**3.2.b) Data acquisition and pre-processing**

The data was for the e-commerce company named Amazon, enlisting its customer reviews over a wide range of products sold on its website. The file format was a Comma Separated Value (CSV), which contains 413840 rows and 6 columns, the columns of which were identified as follows:

* Product Name
* Brand Name
* Price
* Rating
* Reviews
* Review Votes

****

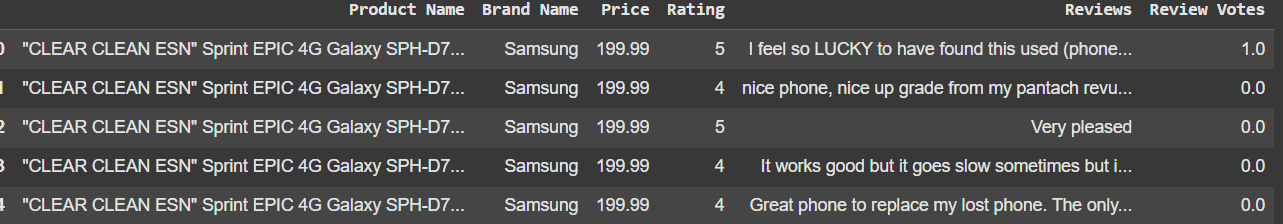
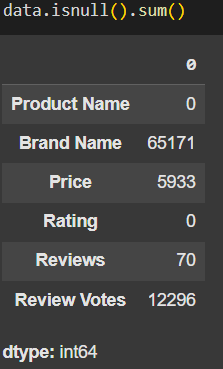
****

Fig 3.9 Data Head Overview

The dataset contains 3 emotional classifications- they are Positive (happy), Negative (sad) and Neutral. To achieve accurate forecasting and ensure high performance of the algorithm, the data must be pre-processed.

Data pre-processing consists of transforming the acquired data into an understandable format, removing duplicate or null values, and dropping undesired attributes, for this analysis the data was pre-processing by removing the column date as it was deemed irrelevant. The dataset was complete and had no null values.

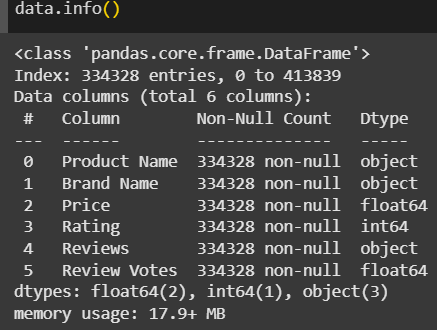


Since the dataset contains null values, they need to be treated.

Fig 3.10 Checking Null values

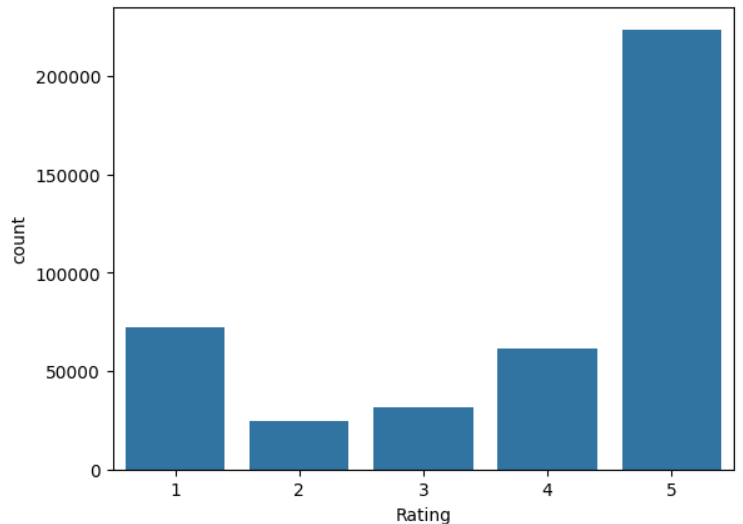


Fig 3.11 Treating Null Values



Following the pre-processing of the dataset, to better understand the attributes and their relationship, a graph count of the attributes was generated.

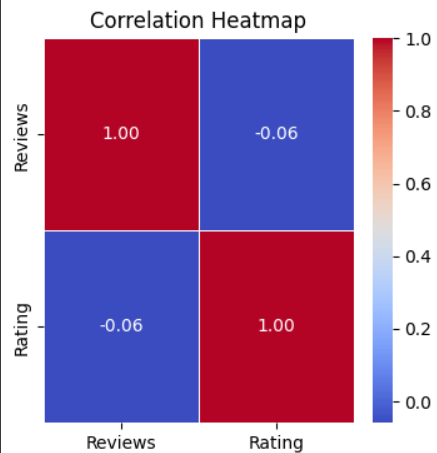
**3.3.b) Data Visualization**

****

**3.4.b) Analysis**

****

Fig 3.12 Generating Word Cloud

 Fig 3.12 Generating Correlation Heatmap

For the analysis of our data, the data was split into a test and Training dataset, using 20% for testing. Eight machine learning steps were implemented, they are:

* Text Cleaning
* Tokenization
* Stop- Word Removal
* Stemming and Lemmatization
* Handling Contractions
* Removing Numbers and Extra Whitespace
* Text Vectorization
* Optional- Spelling Correction

**1. Text Cleaning**

* **Purpose:**  
  Remove unwanted elements such as HTML tags, punctuation, and special characters to minimize noise. Lowercase conversion ensures uniformity for word matching.
* **Techniques and Tools:**
  + **Regular Expressions (regex):** Remove non-alphabetic characters.
  + **BeautifulSoup (optional):** Strip HTML tags if processing web data.

**2. Tokenization**

* **Purpose:**  
  Break the cleaned text into individual words (tokens) for analysis. Tokenization is essential for applying subsequent steps like removing stopwords and lemmatization.
* **Techniques and Tools:**
  + **NLTK's word\_tokenize:** A common choice for splitting text into words.
  + **spaCy:** Another powerful tool for tokenization, especially if you need more advanced processing.

**3. Stop-Word Removal**

* **Purpose:**  
  Eliminate common words (like "the", "is", and "and") that don’t provide significant meaning for sentiment classification.
* **Techniques and Tools:**
  + **NLTK's stopwords:** Provides a list of common English stopwords.

**4. Stemming & Lemmatization**

* **Purpose:**  
  Reduce words to their base or root form, which helps in grouping similar words (e.g., "running", "ran" → "run") and reduces the feature space.
* **Techniques and Tools:**
  + **Stemming:**
    - **PorterStemmer:** Quicker but may produce non-dictionary roots.
  + **Lemmatization:**
    - **WordNetLemmatizer:** More sophisticated as it considers the word's context.

If you choose stemming, you could use nltk.PorterStemmer instead.

**5. Handling Contractions**

* **Purpose:**  
  Expand contractions (e.g., "can't" to "cannot") to standardize words, which helps improve tokenization and further processing.
* **Techniques and Tools:**
  + **Contractions Library:** This library can automatically handle most contractions.

**6. Removing Numbers and Extra Whitespace**

* **Purpose:**  
  Remove numerals if they are not contextually important and condense multiple spaces to a single space for cleaner data.

**7. Text Vectorization**

* **Purpose:**  
  Convert text (tokens) into numerical features that machine learning models can process.
* **Techniques and Tools:**
  + **Bag-of-Words (BoW):** Counts of words.
  + **TF-IDF (Term Frequency-Inverse Document Frequency):** Weights terms based on their frequency across documents.
  + **Word Embeddings:** Methods like Word2Vec or GloVe for capturing contextual similarities.

**8. Optional — Spelling Correction**

* **Purpose:**  
  Correct misspelled words to prevent feature dilution. Use this step cautiously, as over-correction might alter intended sentiment.
* **Techniques and Tools:**
  + **TextBlob Spell Correction:** A simple way to correct spelling.

***Next Steps***

Once your text is pre-processed:

* **Vectorization:** Transform your corpus into numerical features (using TF-IDF, count vectors, or embeddings) for training your machine learning model.
* **Model Training:** Use the transformed features to fit models such as SVM, Naive Bayes, or Logistic Regression for sentiment classification.

This comprehensive pre-processing pipeline helps ensure that your text data is clean, normalized, and ready for robust sentiment analysis.

Chapter 4: Evaluation and Result

A variety of results were obtained from the models trained with 80% of the data and tested with 20% of the data. In this section, we evaluate our models developed using various metrics. The metrics used for evaluation include:

* Accuracy
* Precision
* Recall
* F1-score

**Accuracy:**

This is the total proportion of observations that have been correctly predicted mathematically, Accuracy is defined as:

TP + TN

TP + FP + TN + FN

where TP= True positive, TN= True Negative, FP= False positive and FN = false negative.

**Precision:**

This is the percentage of the positive instances predicted that were correct, mathematically, it is defined as

TP

TP + FP

where TP= True positive, FP= False Positive

**Recall:**

This is the percentage of the positive instances out of the total actual positive, mathematically it is defined as:

TP

TP + FN

where TP= True positive, FN= False Negative

**F1-score:**

This is the harmonic mean of precision and recall metrics, it is the overall correctness the model has achieved, mathematically it is defined as

2 2 \* Precision \* recall



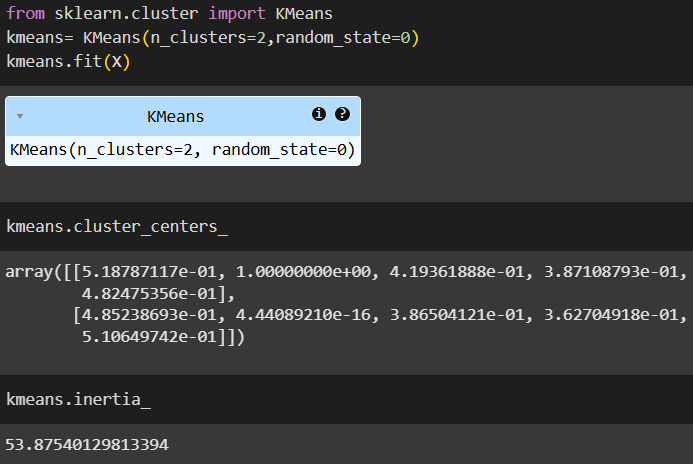
1 1 recall + Precision

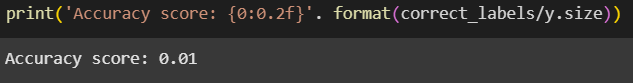


Precision recall

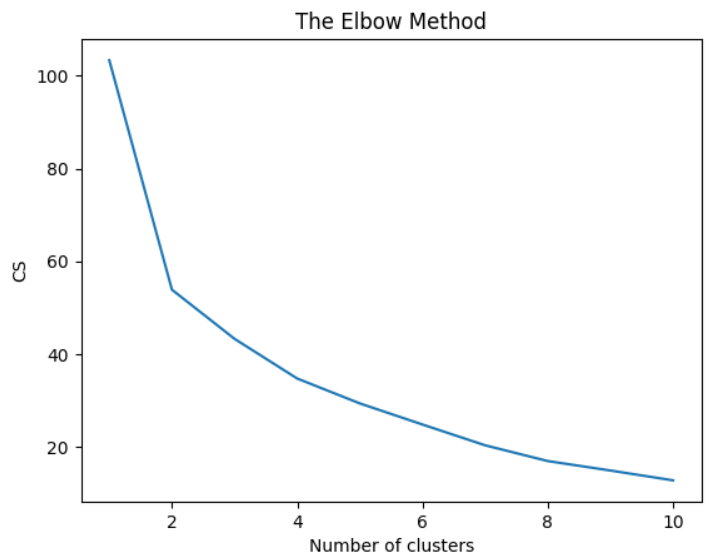
* **K-Means Clustering**

After implementing the K- Means Clustering algorithm, we find the inertia, represent the cluster centres with an Array and then, find the Accuracy Score, as shown in Fig 4.1 below.

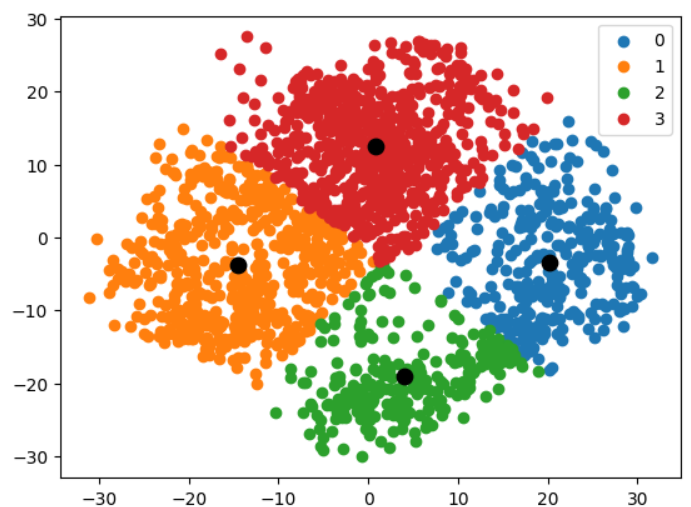




Then, we implement the Elbow Method to find the optimal number of clusters, as shown in Figure 4.2 below.

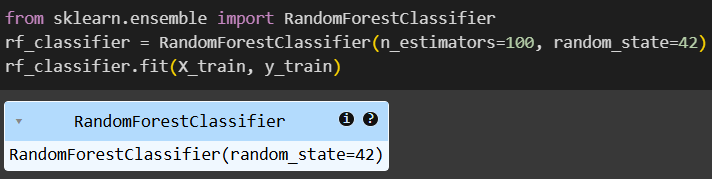
****

Finally, we label the Centroids along with applying different colours to the clusters for ease of differentiating between them, as shown in Figure 4.3 below.

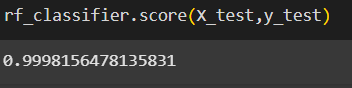


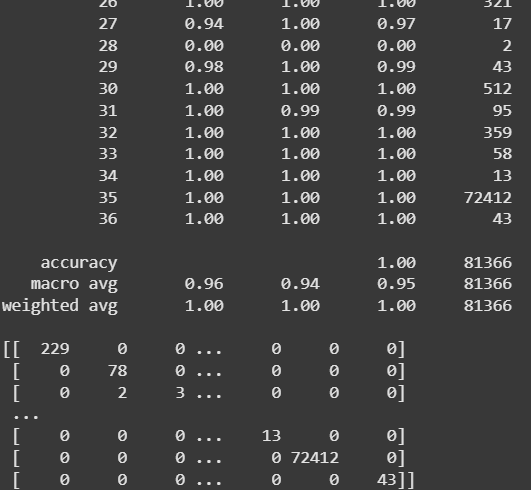
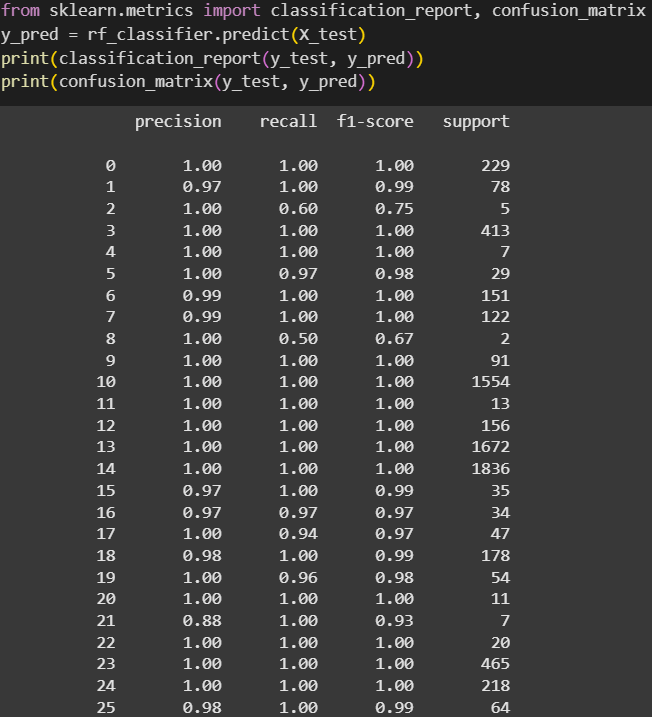
* **Random Forest Classifier**

For the random forest classifier, accuracy, precision, recall, and f1- score were generated, as it is shown in figure 4.4 below.

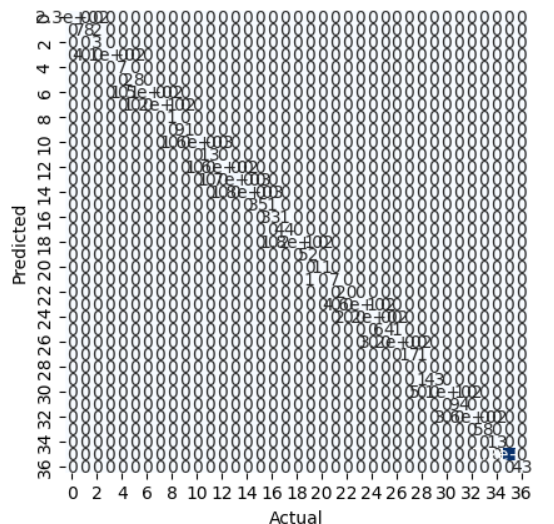


Then, we calculated its accuracy and built the classification report, shown in Fig. 4.5 below.



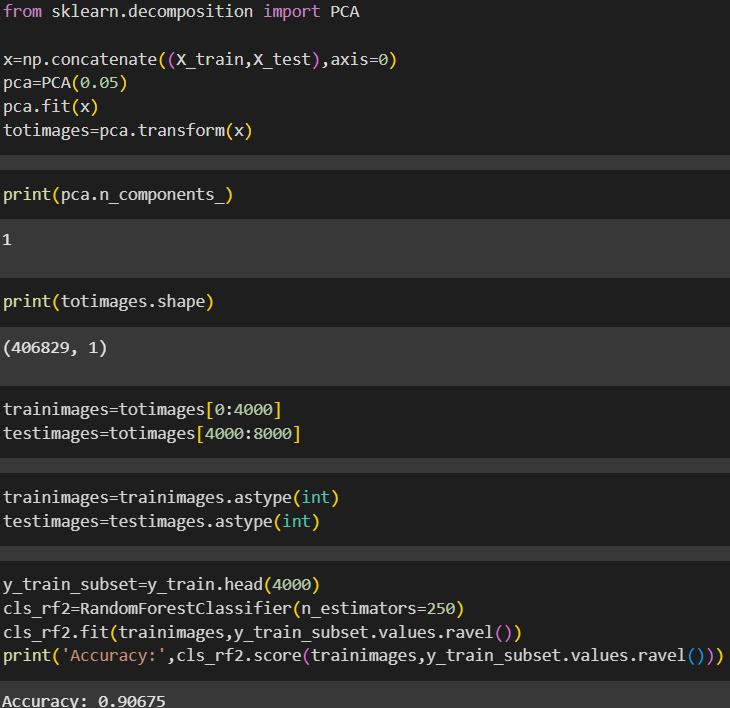


Then, we generated the Confusion Matrix based on the Random Forest Classifier, as shown in Figure 4.6 below.



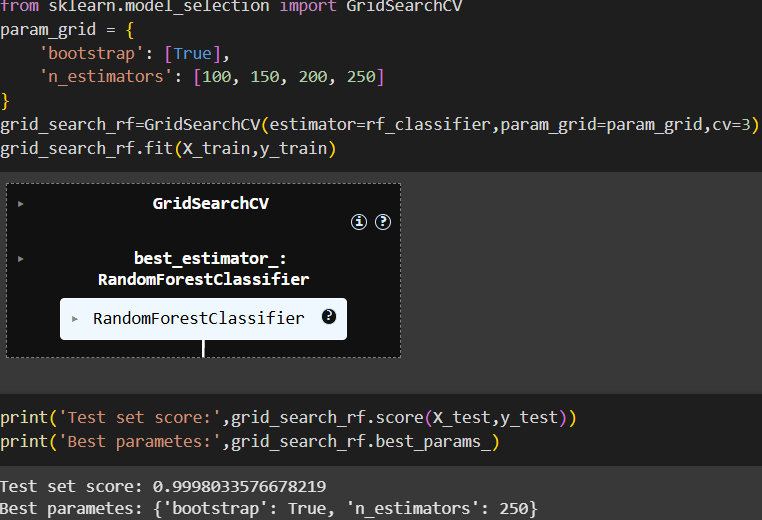
* **Principal Component Analysis (PCA)**

After implementing the Principal Component Analysis algorithm, we find the number of components to be 1, as shown in Figure 4.7 below. Then, we calculate the accuracy to be 90.6%.

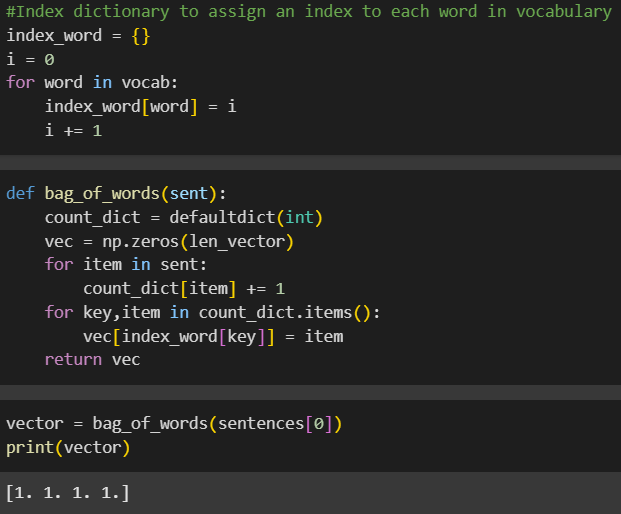
****

* **Grid Search CV**

After implementing the Grid Search Cross-Validation algorithm on the Random Forest Classifier, we find the best parameters of the Test data, as shown in Figure 4.8 below. Then, we calculate the accuracy to be 99.9 %.



In case of Sentiment Analysis of Customer reviews, after all the pre-processing steps are followed, we form a bag\_of\_words, as shown in Figure 4.9 below.



Chapter 5: Result Analysis

* + 1. ***For Customer Segmentation***:
* **Mall Customers**- To segment the customers based on their features, we will use k-means clustering. This algorithm starts by selecting k objects from the dataset randomly that will serve as the initial centres for our clusters, known as centroids. Then at each step, the algorithm seeks to minimize the intra distance, which is the distance between individuals within each cluster and maximize the inter-distance, which is the distance between clusters.

 With the help of clustering, we can understand the variables much better, prompting us to take careful decisions. With the identification of customers, companies can release products and services that target customers based on several parameters like income, age, spending patterns, etc. Furthermore, more complex patterns like product reviews are taken into consideration for better segmentation.

* **Online Retail**- How customer segmentation allows you to make data-driven decisions that influence business growth and customer satisfaction by allowing for:

1. **Personalization**: Segmentation allows businesses to tailor their marketing messages, product recommendations, and promotions to each customer group's specific needs and interests.
2. **Improved Targeting**: By identifying *high-value* and *at-risk* customers, businesses can allocate resources more efficiently, focusing efforts where they are most likely to yield results.
3. **Customer Retention**: Segmentation helps businesses create retention strategies by understanding what keeps customers engaged and satisfied.
   * 1. ***For Sentiment Analysis***

The main thing learnt while creating this project was the amount of information we can extract from simple text reviews, what seems to be chaos of jumbled words at first can give us a lot of insights about the consumers perspective towards anything.

Chapter 6: Conclusions and Future Work

* 1. ***For Customer Segmentation***

1. **Enhanced Understanding of Customer Behaviour:**  
   The project demonstrates that leveraging unsupervised learning methods—such as K-Means, hierarchical clustering, or DBSCAN—can uncover latent patterns within customer data that traditional rule-based methods often miss. By rigorously preprocessing data and engineering robust features, the segmentation model successfully

distinguishes distinct customer personas based on behavioural and demographic attributes.

1. **Actionable Insights for Personalized Marketing:**  
   Through effective segmentation, businesses gain the ability to tailor marketing strategies for specific customer groups. This leads to more targeted campaigns, optimized resource allocation, and improvements in customer retention strategies. The dynamic nature of the clustering approach allows for periodic updates as customer behaviours and market trends evolve, ensuring continued relevance.
2. **Scalability and Adaptability:**  
   The project underscores the need for continuous model training and validation. By integrating batch updates or online learning mechanisms, the segmentation model can evolve in real time, thereby supporting agile decision-making in fast-changing market environments.

**FURTHER IMPROVEMENT**

* Product Category could be incorporated into the segmentation. Further product description can be used to derive the product categories with the help of NLP.
* Conducting deeper segmentation on customers based on their geographical location, demographic and psychographic factors.
* Taking other factors such as geographical location, demographic, psychographic factors and purchase history into consideration we can build a predictive model to predict their next purchase for them and for the customers with similar attributes.

***2. For Sentiment Analysis***

As we stand on the cusp of technological evolution, sentiment analysis stands as a testament to our ability to imbue machines with the power of understanding emotions. From aiding businesses in crafting effective marketing strategies to enabling governments to gauge public sentiment, the potential applications are boundless.

In the grand tapestry of technology, sentiment analysis threads a story of insight and understanding, weaving together the worlds of data, language, and emotion. As we continue to refine this art, we unveil new layers of understanding that bring us closer to the essence of human communication. The future holds the promise of deeper insights, richer experiences, and a world where technology not only understands us but truly empathizes with our sentiments.

**FURTHER IMPROVEMENT**

We can use this function in future to make predictions on any reviews to analyse the sentiment of the user and learn their opinion quantitatively.

* It can also be used to deploy a simple web app that works for a single review at a time or a batch of reviews together.

**Integrated Conclusions and Future Perspectives**

* **Holistic Customer Insights:**  
  When combined, these projects provide a comprehensive 360-degree view of the customer landscape. While segmentation identifies “who” the customers are, sentiment analysis delivers insights on “what” they feel and speak. This dual approach equips businesses with a finely tuned understanding of both customer identity and sentiment.
* **Data-Driven Decision Making:**  
  Together, the models empower organizations to make informed decisions. The

The segmentation aspect ensures that marketing interventions are precisely targeted, while sentiment analysis facilitates agile responses to customer feedback, thus enabling continuous enhancements to products and services.

* **Scope for Future Enhancements:**  
  Both projects highlight the importance of scalability and regular adaptability. Future work could integrate these models further by developing advanced dashboards, real-time alert systems, and even predictive analytics frameworks that forecast customer transitions and sentiment shifts. Additionally, combining other data modalities (like images or interaction metadata) could enrich insights and offer a more nuanced understanding of customer behaviour.

In summary, the research and application of these projects provide robust frameworks for leveraging machine learning and NLP in the marketing domain. They contribute a strategic advantage by transforming raw data into actionable insights that drive personalized customer experiences and sustained business growth.

**REFERENCES**

* Mall Customers-

Dataset: <https://www.kaggle.com/datasets/ramisashararnidhi/mall-customer-csv>

* Online Retail-

Dataset: <https://www.kaggle.com/datasets/tunguz/online-retail>

* Sentiment Analysis-

Dataset: <https://www.kaggle.com/datasets/bensonmoses/amazon-unlocked-mobile>

1. **Customer Segmentation Using Machine Learning – IJCRT**  
   [Link to Paper](https://ijcrt.org/papers/IJCRT_196519.pdf)  
   *This paper explores the use of clustering algorithms—such as K-Means—for segmenting customers based on behavioural and demographic data.*
2. **Customer Segmentation Using Machine Learning: A Literature Review – AIP Conference Proceedings**  
   [Link to Paper](https://pubs.aip.org/aip/acp/article/2481/1/020036/2826844/Customer-segmentation-using-machine-learning-A)  
   An extensive review that discusses the evolution from traditional methods to ML-based segmentation techniques, emphasizing data preprocessing and dynamic segmentation.
3. **Customer Segmentation Using Machine Learning: A Comprehensive Research Study – IJNRD**  
   [Link to Paper](https://www.ijnrd.org/papers/IJNRD2306572.pdf)  
   A study that focuses on segmenting customers using ML models, highlighting the benefits of advanced clustering approaches for targeted marketing.
4. **Aggarwal, C. C. (2014). *Data Clustering: Algorithms and Applications*.**  
   *This book provides an in-depth look into various clustering algorithms and serves as an excellent theoretical foundation for customer segmentation studies.*
5. **Literature Review on Sentiment Analysis – IJSTR**  
   [Link to Paper](https://www.ijstr.org/final-print/apr2020/Literature-Review-On-Sentiment-Analysis.pdf)  
   A comprehensive overview of sentiment analysis techniques, from lexicon-based methods to modern machine learning and deep learning approaches applied to customer reviews.
6. **Sentiment Analysis Literature Survey – IIT Bombay**  
   [Link to Survey](https://www.cfilt.iitb.ac.in/resources/surveys/SentimentAnalysis-Vinita.pdf)  
   A detailed survey that reviews methodologies and challenges related to sentiment analysis, providing insights into the state-of-the-art techniques in the field.
7. **Natural Language Processing for Analysing Online Customer Reviews – PeerJ**  
   [Link to Article](https://peerj.com/articles/cs-2203/)  
   This study examines the application of various NLP models—including transformer-based approaches—for analysing and classifying sentiments in online customer reviews.
8. **Pang, B., & Lee, L. (2008). "Opinion Mining and Sentiment Analysis."**  
   *A seminal paper that lays down the foundation for modern sentiment analysis techniques and has been widely cited in subsequent research.*
9. **Hutto, C. J., & Gilbert, E. (2014). "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text."**  
   Introduces the VADER model, which is well-suited for sentiment analysis in social media and customer review contexts due to its effectiveness in handling informal language.

These references provide a solid foundation for exploring the methodologies and applications related to customer segmentation and sentiment analysis. They cover theoretical underpinnings, practical implementations, and comprehensive reviews that should prove useful for advancing future AI-ML projects in the Marketing domain.