



**KasdiMerbah Ouargla University**



Faculty of New Information and Communication Technologies

Department of Computing and Information Technology

Master's

Domain: Computer Science

Specialty: Fundamental IT

Presented by: Azizi Rayane , DjouhriAhlam

Subject

---

Predicting customer satisfaction using machine learning

---

Evaluation Date: 19 \06\2023

**In front of the jury composed of:**

President:MahdjoubMed Bachire

Rapporteur:MezatiMessaoud

Examiner:Ameurkhadidja



# THANKS

First of all, we thank our Creator for giving us the strength to accomplish this work.

We would like to thank our supervisor Mezati Messaoud who raised this topic memory, as well as for his advice and guidance.

We also have the honor to thank the management of the department Department of Information to facilitate the selection process for theses and

Supervisor as well as good organization and coordination.

A big thank you to the candidates for agreeing to read them and

Evaluate our memory.

Finally, we would like to thank everyone who contributed to this letter.

# Dédicace

*I dedicate this work to my dear parents*

*To my brothers and my sisters*

*To my dear friends*

*With the help of Almighty God, I was able to complete this*

*modest work that I dedicate*

*To my parents, brothers and sisters, and*

*To all those who are dear to me.*

## ملخص :

يلعب رضا العملاء دورًا مهمًا في نجاح منصات التجارة الإلكترونية. وتوفر إدارة علاقات العملاء إطارًا وأدوات للشركات لإدارة علاقات العملاء بفعالية ، وتعزيز رضا العملاء ، وتحسين الأداء العام على منصات التجارة الإلكترونية . يركز عملنا على تحليل رضا العملاء باستخدام تقنيات التعلم الآلي المطبقة على مجموعة بيانات تم الحصول عليها من منصة "Olist" ، وهي سوق عبر الإنترنت في البرازيل. تتكون مجموعة البيانات من 100,000 طلب تم تقديمها في الفترة من 2016 إلى 2018 عبر أسواق متعددة في البرازيل. يتضمن عوامل مختلفة ، بما في ذلك تفاصيل الطلب وخصائص المنتج وطرق الدفع والتركيب السكانية للعملاء. من خلال تحليل البيانات الاستكشافية ، نكشف عن أنماط واتجاهات رضا العملاء ، باستخدام اختيار الميزات وتقنيات المعالجة المسبقة لتحديد العوامل المؤثرة ، ومقارنة النتائج بالدراسات الأخرى التي استخدمت نفس قاعدة البيانات مع اختلاف في تحديد المدخلات . تُستخدم خوارزميات التعلم الآلي مثل الانحدار اللوجستي وأشجار القرار والغابات العشوائية لتطوير نموذج تنبؤي. تسلط النتائج الرئيسية لدراستنا الضوء على تأثير طرق الدفع وأوقات الشحن ومواقع العملاء على رضا العملاء. نحن نقدم تداعيات وتوصيات عملية لتعزيز رضا العملاء ، بما في ذلك تحسين عمليات الخروج ، وتقليل أوقات الشحن ، وإضفاء الطابع الشخصي على تجارب العملاء.

**الكلمات المفتاحية:** رضا العملاء , إدارة علاقات العملاء, التعلم الآلي , تنبؤ.

## Résumé:

La satisfaction client joue un rôle important dans le succès des plateformes de commerce électronique. Le CRM fournit un cadre et des outils permettant aux entreprises de gérer efficacement les relations clients, d'améliorer la satisfaction client et d'améliorer les performances globales sur les plateformes de commerce électronique. Notre travail porte sur l'analyse de la satisfaction client à l'aide de techniques d'apprentissage automatique appliquées à un ensemble de données obtenu à partir de la plateforme "Olist", une place de marché en ligne au Brésil. L'ensemble de données se compose de 100 000 applications placées entre 2016 et 2018 sur plusieurs marchés au Brésil. Il comprend divers facteurs, notamment les détails de la commande, les caractéristiques du produit, les méthodes de paiement et les données démographiques des clients. Grâce à l'analyse exploratoire des données, nous révélons les modèles et les tendances de la satisfaction client, en utilisant des techniques de sélection et de prétraitement des caractéristiques pour identifier les facteurs d'influence, et comparons les résultats à d'autres études qui ont utilisé la même base de données avec une différence dans l'identification des entrées. Des algorithmes d'apprentissage automatique tels que la régression logistique, les arbres de décision et les forêts aléatoires sont utilisés pour développer un modèle prédictif. Les principaux résultats de notre étude mettent en évidence l'impact des modes de paiement, des délais d'expédition et de la localisation des clients sur la

satisfaction des clients. Nous fournissons des implications pratiques et des recommandations pour améliorer la satisfaction client, notamment en améliorant les processus de paiement, en réduisant les délais d'expédition et en personnalisant l'expérience client.

**Mots clés :** satisfaction client, gestion de la relation client, apprentissage automatique, prédiction.

### **Abstract:**

Customer satisfaction plays an important role in the success of e-commerce platforms. CRM provides a framework and tools for companies to effectively manage customer relationships, enhance customer satisfaction, and improve overall performance on e-commerce platforms. Our work focuses on analyzing customer satisfaction using machine learning techniques applied to a dataset obtained from the platform "Olist", an online marketplace in Brazil. The dataset consists of 100,000 apps placed between 2016 and 2018 in multiple markets in Brazil. It includes various factors, including order details, product characteristics, payment methods, and customer demographics. Through exploratory data analysis, we reveal patterns and trends in customer satisfaction, using feature selection and pre-processing techniques to identify influencing factors, and compare results to other studies that used the same database with a difference in identifying inputs. Machine learning algorithms such as logistic regression, decision trees, and random forests are used to develop a predictive model. The main findings of our study highlight the impact of payment methods, shipping times, and customer locations on customer satisfaction. We provide practical implications and recommendations for enhancing customer satisfaction, including improving checkout processes, reducing shipping times, and personalizing customer experiences.

**Key words:** customer satisfaction, customer relationship management(CRM), machine learning, predict.



**Contents**

- General Introduction ..... 1
- Chapter One: Customer Relationship Management and Customer Satisfaction: ..... 3
  - 1.1 introduction:..... 4
  - 1.2 customer relationship management (CRM):..... 4
    - 1.2.1 Definitions of CRM:..... 4
    - 1.2.2 Evolution and Growth of CRM..... 5
    - 1.2.3 the key components of CRM:..... 7
    - 1.2.4 Types of CRM..... 10
      - 1.2.4.1 Operational CRM: ..... 10
      - 1.2.4.2 Analytical CRM: ..... 11
      - 1.2.4.3 Collaborative CRM: ..... 11
    - 1.2.5 The Benefits of CRM ..... 11
  - 1.3 Customer satisfaction: ..... 13
    - 1.3.1 Definitions of customer satisfaction..... 13
    - 1.3.2 Factors that Influence Customer Satisfaction..... 13
    - 1.3.3 How does CRM affect customer satisfaction? ..... 14
  - 1.4 Conclusion ..... 15
- 2 Chapter two: Machine Learning ..... 17
  - 2.1 Introduction: ..... 18
  - 2.2 Definition of Artificial Intelligence: ..... 18
    - 2.2.1 Definition of machine learning: ..... 19
    - 2.2.2 Principle of operation of machine learning: ..... 19
    - 2.2.3 Types of machine learning: ..... 20
    - 2.2.4 Machine learning algorithms:..... 24
    - 2.2.5 Application of machine learning: ..... 28
  - 2.3 The impact of machine learning on CRM: ..... 30
  - 2.4 Similar works:..... 31
    - 2.4.1 Predicting negative customer reviews Using Data Science to Predict Negative Customer Reviews: ..... 31



2.4.2	EDA + Satisfaction Prediction + NLP + RFM + Deployment: .....	31
2.4.3	Customer Satisfaction Prediction Using Machine Learning: .....	32
	Conclusion: .....	33
3	Chapter Three: construction and implementation .....	35
3.1	Introduction: .....	36
3.2	Project objective: .....	36
3.3	The development environment and tools: .....	36
3.3.1	Google Colab: .....	36
3.3.2	Python: .....	37
3.3.3	Python Librarys: .....	37
3.4	Design: .....	38
3.4.1	System structure: .....	38
3.4.2	The dataset : .....	38
3.5	Implementation: .....	40
3.5.1	Preprocess the data: .....	40
3.5.2	Feature Selection: .....	42
3.5.3	Model Selection: .....	50
3.6	Results and discussion: .....	51
3.7	Comparison with other works: .....	52
	Conclusion: .....	53
	General conclusion: .....	55

List of figures:

Figure 1:the Timeline of the CRM evolution. <sup>6</sup> .....	5
Figure 2:Characteristics on which customers can be classified <sup>7</sup> .....	8
Figure 3:Technologies that can be used in marketing automation <sup>11</sup> .....	9
Figure 4:Types of CRM <sup>14</sup> .....	10
Figure 5:Benefits of CRM <sup>18</sup> .....	12
Figure 6:Fields of work with artificial intelligence <sup>29</sup> .....	18
Figure 7:operation of machine learning. <sup>32</sup> .....	19
Figure 8:Applications of machine learning. <sup>34</sup> .....	30
Figure 9:types of machine learning . <sup>36</sup> .....	20
Figure 10:supervised learning workflow. <sup>36</sup> .....	22
Figure 11:The two main subcategories of supervised learning. <sup>34</sup> .....	22
Figure 12:unsupervised learning workflow. <sup>38</sup> .....	23
Figure 13:Reinforcement learning workflow. <sup>38</sup> .....	24
Figure 14:Machine Learning Algorithms. <sup>39</sup> .....	25
Figure 15:logistic regression exemple <sup>40</sup> .....	26
Figure 16:Diagram showing how decision tree algorithm works <sup>43</sup> .....	28
Figure 17:A graphic showing the importance of attributes using the method "SelctKBest" <sup>49</sup> .....	32
Figure 18:System structure.....	38
Figure 19:Brazilian E-Commerce Public Dataset by Olist. <sup>44</sup> .....	39
Figure 20:Graph showing positive reviews for each payment type .....	43
Figure 21:Preferred payment methods of online shoppers worldwide as of March 2017. ....	44
Figure 22:Graph showing positive reviews for each state. ....	45
Figure 23:the distributions of "freight_value" for positive and negative classes.....	46
Figure 24:the distribution of the "payment_value" for both the positive and negative classes. .....	46
Figure 25:Distribution of the number of orders per percentage of customers. ....	47

List de Table:

Table 1:the results we obtained.....	51
--------------------------------------	----

## General Introduction

---

### General Introduction:

The e-commerce sector has witnessed significant growth due to the widespread accessibility of the internet. This transformation has reshaped commercial activities and become an integral part of people's lives. The e-commerce has gained popularity across a wide range of product categories, and this trend is expected to continue with the potential for further expansion.

E-commerce refers to the use of the internet as a platform for customers to access online stores, browse products, and make purchases using various devices. The convenience of e-commerce lies in its availability 24/7, offering customers the flexibility to shop at any time and access a diverse selection of products from multiple sellers. While e-commerce brings advantages such as accessibility and product variety, it also faces challenges. Limited consumer services in the online space make it challenging to adequately demonstrate products, and delivery times can be a potential drawback. In the field of e-commerce, machine learning has emerged as a valuable tool, offering a range of applications and benefits. It enables accurate sales prediction, predicts the next consumer order, analyzes reviews for sentiment analysis, provides personalized recommendations, and enhances the shopping experience with features like voice and image search and interactive chatbots. Additionally, Machine Learning can augment the in-store experience through technologies like augmented reality. For better understanding of customer satisfaction in e-commerce platforms and provide practical recommendations to improve customer experiences and drive business growth in this ever-evolving industry. Our solution for this problem WE USED Machine Learning techniques to predict customer satisfaction in e-commerce. By analyzing various factors, including order details, product characteristics, payment methods, and customer demographics, the aim is to gain insights into the factors influencing customer satisfaction. The collected data will be processed and used to develop a predictive model that effectively forecasts customer satisfaction levels.

Our project includes three chapters, Chapter One provides an introduction to customer relationship management (CRM) and its impact on customer satisfaction. Chapter two discusses various aspects of machine learning (ML) and its impact on customer relationship management (CRM). Chapter Three provides an overview of the project's construction and implementation. It begins with an introduction, presenting the problem statement and the

software environment used in the implementation. The objective of the project is to predict customer satisfaction scores based on historical data. The problem can be framed as a classification problem, where the goal is to predict whether a customer will be satisfied or dissatisfied with their purchase.

# **Chapter One: Customer Relationship Management and Customer Satisfaction:**

## 1.1 introduction:

In today's highly competitive business landscape, companies must seek out advantages that can help them stand out. This is where customer relationship management (CRM) comes into play. Knowing one's customers and their preferences is a critical competitive advantage, which has led many organizations to adopt CRM as a strategy for building and maintaining long-lasting relationships with their customers. By understanding customers' needs, companies can work to improve their satisfaction and loyalty, which ultimately helps the company achieve its goals and improve its image. CRM also facilitates communication with customers, enabling businesses to better serve their needs. Overall, CRM is a powerful tool that can significantly impact customer satisfaction, thereby enhancing a company's position in the marketplace as a distinct competitor.

## 1.2 customer relationship management (CRM):

In this section, we will delve into the concept of Customer Relationship Management (CRM) and its evolutionary stages, explore the key components that comprise CRM systems, examine the various types of CRM that exist, and ultimately conclude by highlighting the benefits that CRM brings to businesses.

### 1.2.1 Definitions of CRM:

Over the past decade and a half, Customer Relationship Management (CRM) has evolved into an area of great interest. However, there is a wide difference in academic and administrative terminology about what is meant by CRM .

here are some academic definitions of CRM:

"CRM is a cross-functional process for managing customer relationships in order to create value for the customer and profit for the company".[1]

"CRM is an organizational strategy designed to build long-term customer relationships and enhance organizational performance by integrating customer-focused processes and technology across the organization." [2]

"CRM is a business approach that seeks to understand, anticipate and manage the needs of an organization's current and potential customers through an ongoing, mutually beneficial relationship in order to create long-term shareholder value." [3]

So in general, Customer Relations Management (CRM) refers to a work strategy aimed at managing interactions and analyzing them with customers throughout their entire life, with the aim of improving customer satisfaction, retaining and profitability. It includes the use of technology to organize, automate, synchronize sales, marketing, customer service and technical support.

1.2.2 Evolution and Growth of CRM

Over the past few decades, marketing tools have expanded significantly, with the growth of the internet driving an increased reliance on customer relationship management (CRM) across many industries. To illustrate this trend, Figure 1 provides a timeline of the evolution of CRM over time, highlighting the various milestones and developments that have shaped the field.

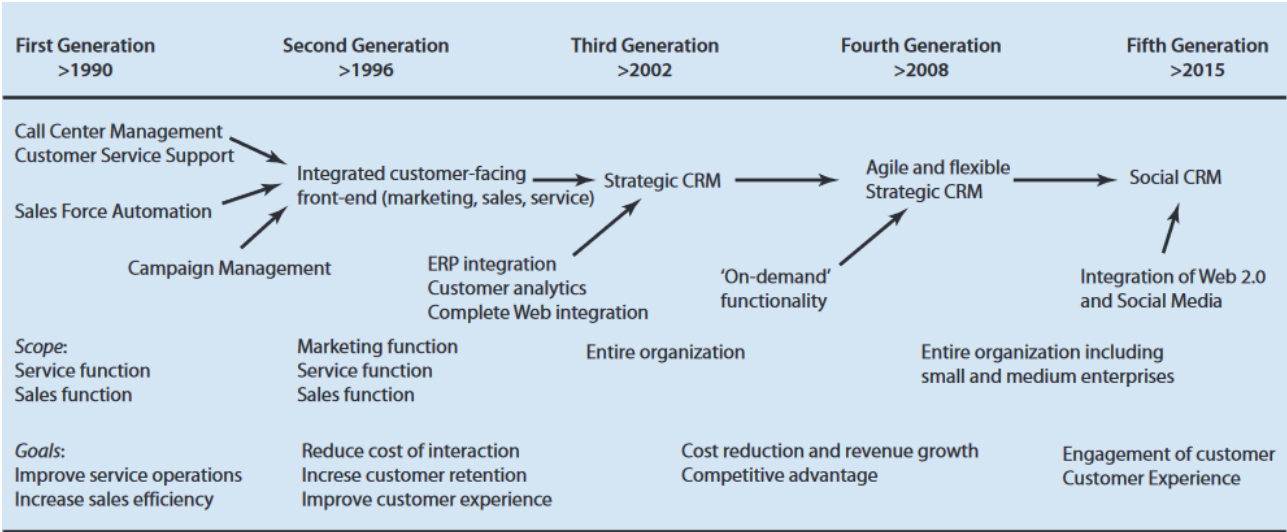


Figure 1: the Timeline of the CRM evolution.[4]

### 1.2.2.1 First Generation (Functional CRM)

In the past, CRM was developed as two separate product offerings: Sales force automation (SFA) and customer service and support (CSS). SFA handled pre-sales functions, such as maintaining customer data, telemarketing, generating leads, creating sales quotes, and placing sales orders. CSS, on the other hand, focused on post-sales activities, such as help desks, contact and call centers, and field service support. These early applications promised improvements in sales and service, but their market niche remained small due to fragmentation and poor integration with back-office systems. Meanwhile, the market for enterprise resource planning (ERP) tools was growing, designed to integrate all company departments and functions within a single computer system[4].

### 1.2.2.2 Second Generation (Customer Facing Front-End Approach)

The aim of CRM technology was to address the business requirements of the customer-facing front-end and fill in the gaps left by ERP functionality. The goal was to create a unified view of all customer interactions, regardless of the purpose of the contact (pre-sales, sales transactions, or after-sales service) or the mode of communication. However, this objective was not achieved during the 1990s. Customer expectations in this period surpassed the benefits offered by CRM technology. Industry experts began discussing the potential downfall of CRM. It became evident that implementing, realizing, and measuring revenue growth through technology was difficult without a more strategic comprehension of the process[4].

### 1.2.2.3 Third Generation (Strategic Approach)

Towards the end of 2002, as the CRM market began to rebound, leading organizations started concentrating on integrating customer-facing front-end systems with back-end systems and systems utilized by suppliers and partners. The integration of internet technology proved to be a catalyst in boosting CRM. Many companies have come to realize that a strategic approach to CRM can be advantageous. They have realized that the ultimate objective of CRM is not just to control costs but also to increase revenue[4].

### 1.2.2.4 Fourth Generation (Agile and Flexible Strategic CRM)

Fourth Generation (Agile and Flexible Strategic CRM)  
As we approach the end of the 2000s, the fourth generation of CRM is emerging. Strategic



CRM has been broadly accepted and established as a critical component of marketing strategy. Companies of all sizes are gradually adopting this management tool to boost their business. The increasing prevalence of web-based services and the emergence of social media and self-service indicate that customer empowerment is becoming a significant focus. Pay-per-use CRM technology, in particular, can provide on-demand functionality[4].

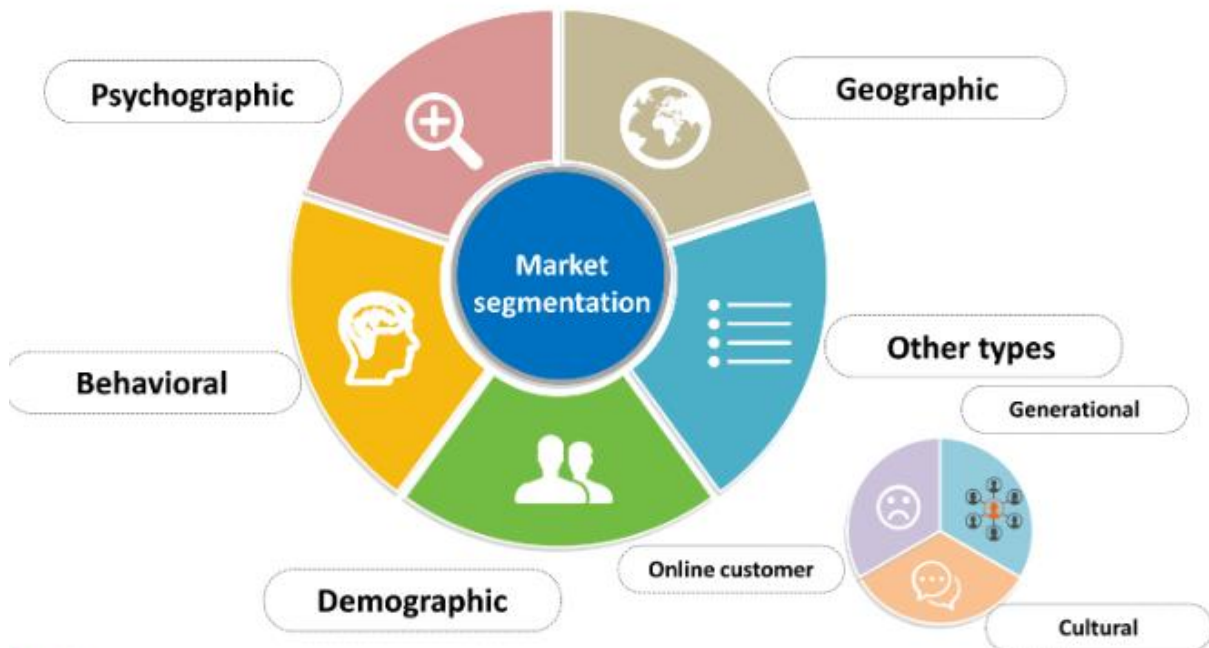
#### 1.2.2.5 Fifth Generation (Social CRM):

The fifth generation of CRM has emerged, fueled by new technological advancements and unprecedented access to social media. Social CRM revolves around customer engagement through web and social media integration and leverages data-driven insights to enhance the overall customer experience. Companies are encouraging active customer engagement online and using software applications to monitor social data in real-time. With this information, companies can deliver tailored content and personalized messages to specific customers and improve the customer experience at every touchpoint in the customer journey. Moreover, combining data across various social media platforms enables companies to assess customer value based not just on profitability but also on online behavior such as referrals, knowledge dispersion, and influence on other members of the social media community[4].

#### 1.2.3 the key components of CRM:

Customer Relationship Management involves various components that work together to support effective customer interactions and enhance customer satisfaction. The key components of CRM include:

### 1.2.3.1 Customer Segmentation:



*Figure 1 Characteristics on which customers can be classified[5]*

Customer segmentation refers to the process of categorizing and classifying customers based on common attributes. This approach enables businesses to customize and personalize their marketing, sales, and service efforts to meet the specific requirements of each group. Consequently, it can lead to higher customer retention rates and increased conversion rates[6].

### 1.2.3.2 Data Management:

CRM data management involves organizing large volumes of critical data based on a company's current and prospective clients. By analyzing and utilizing this data for the company's benefit, customer relationships are established. This data is highly significant for both marketing and sales teams. The CRM system provides information on almost everything related to a particular industry. The CRM data management simplifies the task of finding details for team members without having to dig through piles of files, as all the details are well-organized[7].

### 1.2.3.3 Sales Force Automation:

Sales force automation (SFA) is a technology solution designed to automate and streamline repetitive, administrative tasks in the sales process, such as lead management, sales forecasting, contact management, and sales order processing. SFA software can also provide

sales representatives with real-time access to customer data, historical sales information, and sales performance metrics to help them make informed decisions and improve their overall productivity. By automating these tasks, SFA can help businesses reduce costs, increase efficiency, and improve their overall sales performance[8].

#### 1.2.3.4 Marketing Automation:

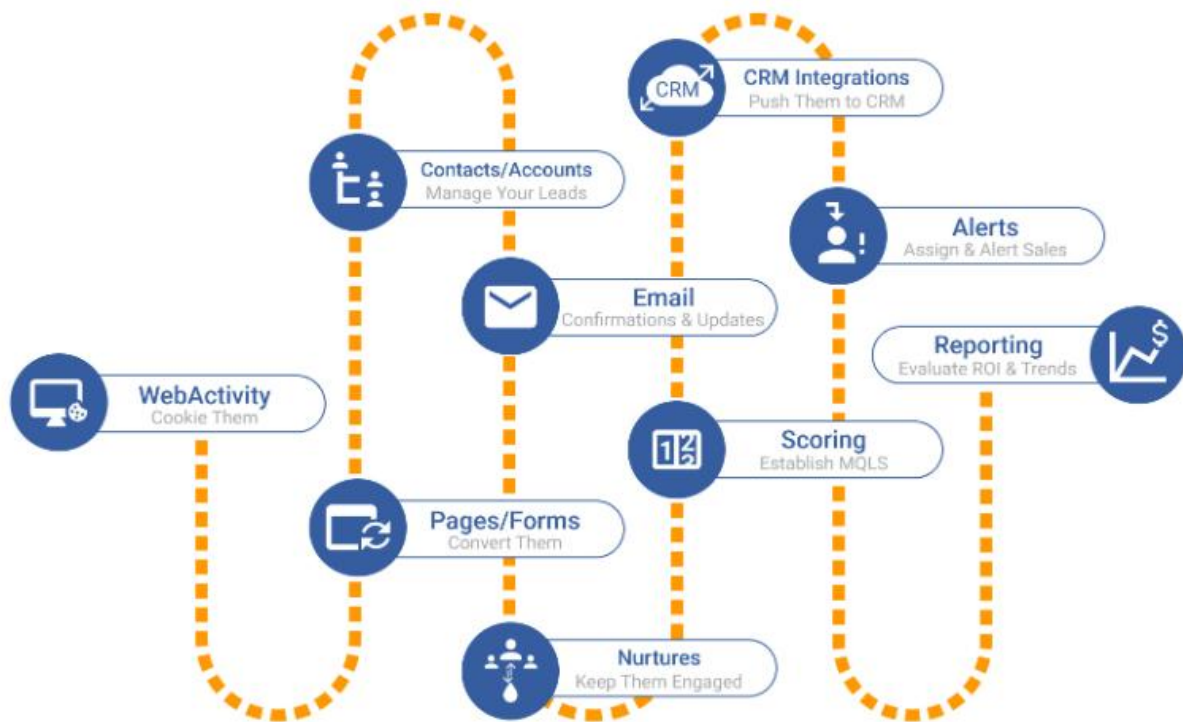


Figure 2: Technologies that can be used in marketing automation [9]

Marketing automation refers to the use of technology and software to automate marketing processes and campaigns across multiple channels such as email, web, social media, and text messaging. It involves creating and sending targeted messages to customers

It is used by marketing and sales teams to increase efficiency, reduce human error, and free up time to focus on higher-order problems. Effective marketing automation systems are designed to scale alongside the growth and complexity of an organization[10].

#### 1.2.3.5 Customer Service and Support:

Customer service and support can be defined as the process of assisting customers in meeting their needs and concerns related to a product or service. Academically, it is an

essential component of the customer experience and a critical factor in customer satisfaction and loyalty.

The significance of customer service and support in enhancing customer relationships and loyalty is highlighted in a research paper published in the journal *Service Research*. The paper suggests that efficient customer service and support can result in higher levels of customer satisfaction, trust, loyalty, and favorable word-of-mouth referrals, ultimately leading to repeated business. Overall, customer service and support plays a crucial role in shaping the customer experience, and businesses should prioritize providing top-notch customer service and support to cater to their customers' requirements and establish enduring customer relationships[11]

#### 1.2.4 Types of CRM

There are three main types of CRM:

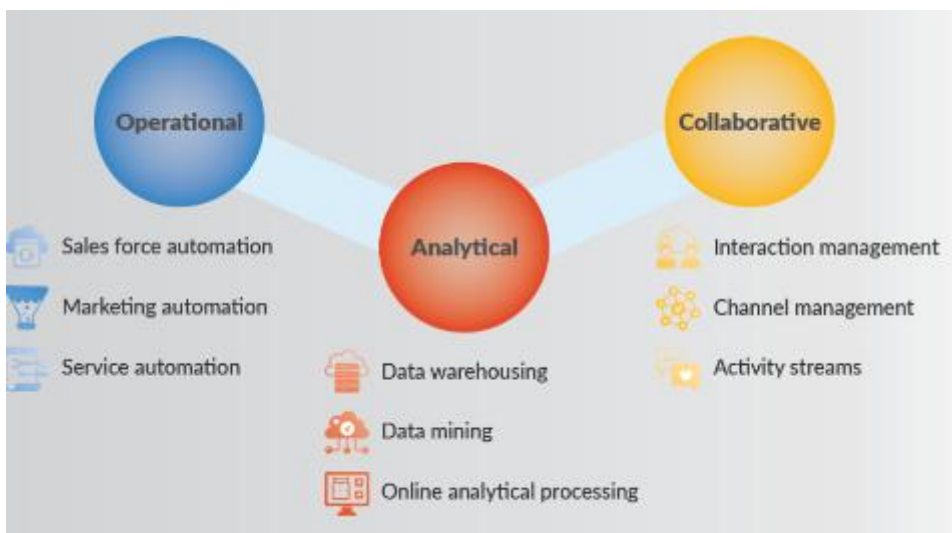


Figure 3: Types of CRM[12].

##### 1.2.4.1 Operational CRM:

This type of CRM is focused on improving the efficiency and effectiveness of business processes that involve customer interactions, such as sales, marketing, and customer service. Operational CRM typically involves the use of software and automation tools to streamline processes and improve productivity[13].

#### **1.2.4.2 Analytical CRM:**

This type of CRM focuses on collecting, analyzing, and interpreting customer data to gain insights into customer behavior, preferences, and needs. Analytical CRM can help businesses identify trends, predict customer behavior, and make data-driven decisions to improve customer engagement and retention.

Analytical CRM “provides analysis of customer data and behavioral patterns to improve business decisions.” This includes the underlying data warehouse architecture, customer profiling/segmentation systems, reporting, and analysis[14].

#### **1.2.4.3 Collaborative CRM:**

This type of CRM focuses on improving collaboration and communication between different departments and teams involved in customer interactions, such as sales, marketing, and customer service. Collaborative CRM typically involves the use of shared databases, communication tools, and other technologies to facilitate collaboration and ensure a consistent, coordinated customer experience[15].

#### **1.2.5 The Benefits of CRM**

Customer Relationship Management (CRM) is a popular business strategy that has gained significant attention from academics in recent years. Here are some of the key reasons why CRM is important:



Figure 4: Benefits of CRM [16].

a) Facilitates customer-focused management:

CRM helps companies to focus on the needs and preferences of their customers, rather than just their products or services. This approach is supported by the marketing concept, which emphasizes the importance of understanding and satisfying customer needs[17].

b) Improves customer retention and loyalty:

CRM enables companies to build stronger relationships with their customers by providing personalized experiences and proactive support. This can lead to increased customer loyalty and retention, which are critical for long-term business success[18].

c) Enhances marketing effectiveness:

CRM provides companies with valuable customer data and insights that can be used to develop more effective marketing campaigns. By targeting the right customers with the right messages at the right time, companies can improve their marketing ROI and generate more revenue.

d) Supports cross-functional collaboration:

CRM requires collaboration and communication across different departments within a company, such as sales, marketing, and customer service. This cross-functional approach can lead to better customer outcomes and improved organizational performance[19].

### 1.3 Customer satisfaction:

In this section, we will explore the realm of customer satisfaction management, examining the factors that influence it and delving into the impact of CRM on enhancing customer satisfaction.

#### 1.3.1 Definitions of customer satisfaction

Customer satisfaction is "the customer's fulfillment response. It is a judgment that a product or service feature, or the product or service itself, provided (or is providing) a pleasurable level of consumption-related fulfillment, including levels of under- or over-fulfillment"[20].

"Customer satisfaction is the extent to which a product's perceived performance matches a buyer's expectations" [19].

In general, customer satisfaction is an evaluation of how well a product or service lives up to a customer's expectations. It represents the customer's overall sentiment and disposition towards a company or its offerings, which is informed by their interactions and experiences with them. The greater the degree of customer satisfaction, the higher the likelihood that customers will persist in their business with the company, advocate for it to others, and sustain their loyalty over time. To gauge customer satisfaction, companies typically rely on various feedback mechanisms such as surveys, reviews, and other channels that elicit input from customers.

#### 1.3.2 Factors that Influence Customer Satisfaction

There are some factors that can influence customer satisfaction. These include:

- a) **Product or service quality:** The quality of the product or service is a critical factor in determining customer satisfaction. Customers expect products and services that meet their needs and perform as advertised[21].

- b) **Customer service:** Customer service plays a critical role in determining customer satisfaction. A good customer service experience can make a customer feel valued and appreciated, leading to a positive impression of the company and increased customer loyalty. On the other hand, a poor customer service experience can lead to dissatisfaction, negative word-of-mouth, and lost business[22].
- c) **Price:** The price of the product or service can impact customer satisfaction. Customers expect fair and reasonable prices for products and services[23].
- d) **Brand reputation:** the reputation of a brand can have a significant impact on customer satisfaction. A brand's reputation is essentially the sum of the perceptions and experiences that people have had with the brand over time, and it can affect how customers perceive the quality, reliability, and value of its products or services[24].
- e) **Convenience:** The convenience of purchasing and using the product or service can impact customer satisfaction. Customers expect products and services to be easy to find, purchase, and use.

If a product or service is convenient to purchase, customers are more likely to choose it over competing products or services. For example, if a customer can easily find a product online and purchase it with just a few clicks, they are more likely to be satisfied with the purchasing experience than if they had to go to a physical store and navigate through a complex checkout process.

- f) **Personalization:** The level of personalization of the product or service can impact customer satisfaction. Customers expect products and services to be tailored to their needs and preferences[21].

### 1.3.3 How does CRM affect customer satisfaction?

CRM can have a significant impact on customer satisfaction. By providing a more personalized and engaging experience, businesses can increase customer loyalty and improve their overall satisfaction levels.

CRM can have a significant impact on customer satisfaction in several ways:

- a) **Personalization:** By using CRM tools and strategies, businesses can better understand their customers' needs and preferences and provide more personalized experiences. This can lead to higher levels of satisfaction and loyalty[25].



- b) **Responsiveness:** CRM can help businesses respond more quickly and effectively to customer inquiries, complaints, and feedback, leading to improved customer satisfaction.
- c) **Consistency:** CRM can help businesses provide consistent experiences across all channels, including email, phone, social media, and in-person interactions. This can lead to greater trust and satisfaction among customers[26].
- d) **Empathy:** CRM can help businesses develop a deeper understanding of their customers' challenges and needs, allowing them to empathize with their customers and provide more effective solutions.
- e) **Proactivity:** CRM can help businesses anticipate their customers' needs and preferences and provide proactive support and solutions. This can lead to higher levels of satisfaction and loyalty[19].

#### 1.4 Conclusion

A CRM system helps companies customize their interactions with customers, quickly resolve issues, and deliver exceptional service to enhance customer satisfaction and loyalty. To achieve this, companies must invest in the right technologies, processes, and people, including implementing CRM software, establishing clear communication channels, training employees on customer service, and continuously analyzing customer feedback and behavior. Through the use of machine learning in customer relationship management systems, customer interactions are enhanced. customers and overall business performance. By analyzing customer data, machine learning algorithms identify patterns and segment customers based on behavior, preferences, demographics, or purchase history. This enables personalized marketing, personalized product recommendations, and targeted promotions to a specific customer segment.



# **Chapter two: Machine Learning**

2.1 Introduction:

Indeed, Turing's article entitled "Computing Machinery and Intelligence," published in 1950, is often regarded as the starting point for research in artificial intelligence. Chapter two introduces the concept of machine learning (ML) within the broader field of artificial intelligence (AI).

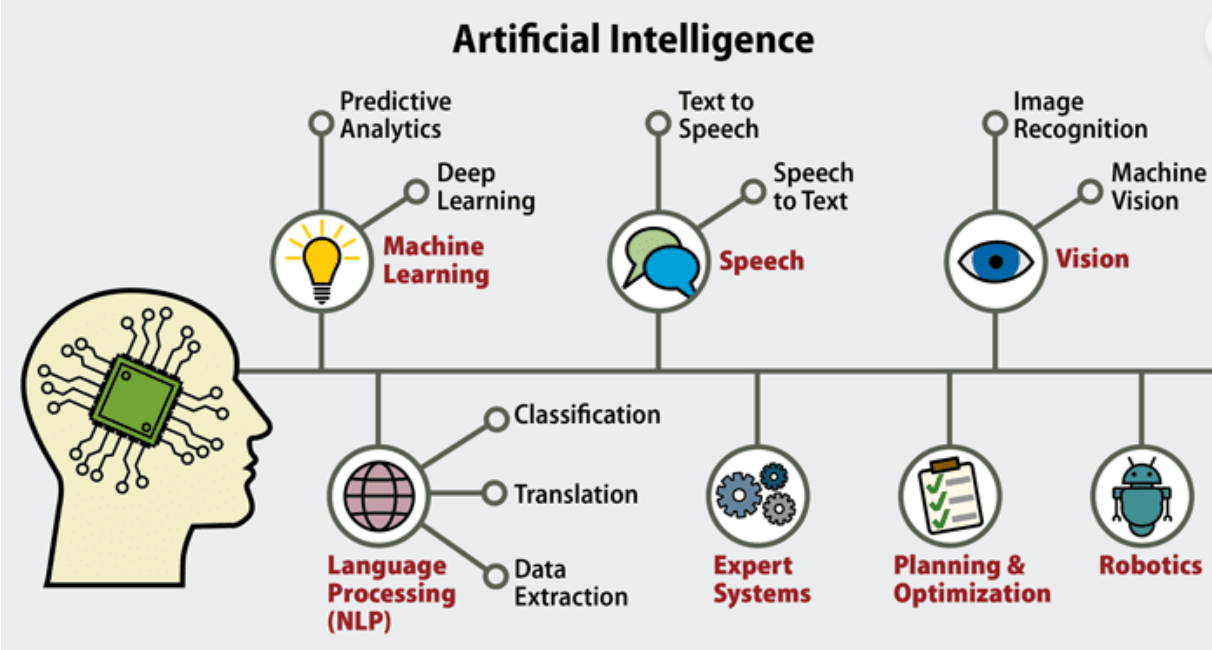


Figure 5:Fields of work with artificial intelligence [27].

2.2 Definition of Artificial Intelligence:

Artificial Intelligence (AI) is a field of computer science that focuses on creating intelligent machines that can perform tasks that normally require human intelligence. AI involves the simulation of human intelligence processes, including perception, cognition, decision-making, and learning, using formal methods such as logical reasoning, planning, search, machine learning, and natural language processing. Intelligent machines can be designed to perform a wide variety of tasks, from speech and image recognition to complex decision-making and autonomous vehicle driving. The ultimate goal of AI is to create machines capable of autonomously solving complex problems using reasoning and learning methods, without direct human intervention.[28]

2.2.1 Definition of machine learning:

Machine learning is a subfield of artificial intelligence (AI) that focuses on creating computer programs that can learn and improve from experience, without being explicitly programmed. This involves developing algorithms and statistical models that enable machines to recognize patterns in data and make predictions or decisions based on that analysis.

Essentially, machine learning involves creating systems that can learn from data and improve their performance over time. This is typically done by introducing large amounts of data into the system, which are then used to train a model capable of identifying patterns and making predictions or decisions based on new data. Machine learning is used in a wide range of applications, including image and speech recognition, natural language processing, fraud detection, recommendation systems, and many others.[29]

2.2.2 Principle of operation of machine learning:

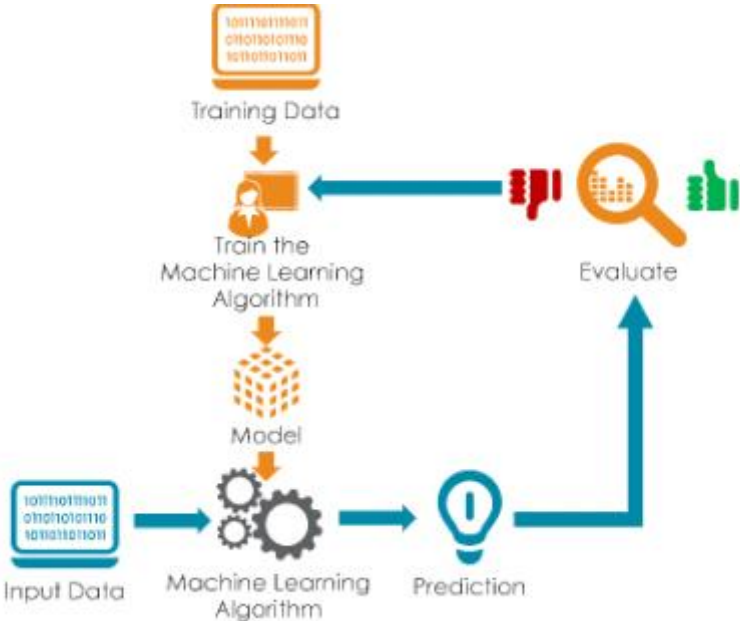


Figure 6: operation of machine learning. [30]

The principle of machine learning is to enable a machine or computer to learn from data using algorithms. In other words, instead of explicitly programming a computer to perform a given task, a set of data is provided to the computer and it is allowed to draw conclusions from that data.

The process of machine learning begins with the acquisition of training data, which is used to train a mathematical model or algorithm that can be used to predict outcomes on new data. The mathematical model is adjusted and optimized using optimization algorithms such as gradient descent, in order to minimize prediction errors.

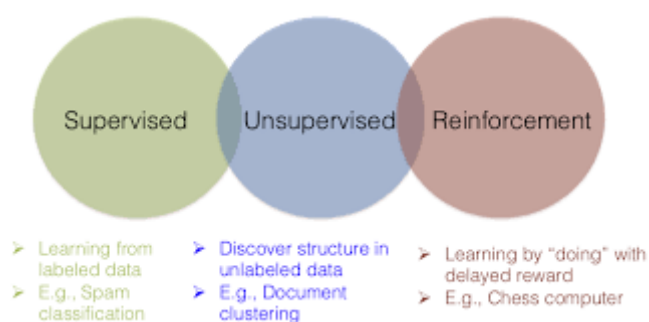
Once the model is trained, it can be used to predict outcomes on new data. Input data is fed into the model, which provides an output prediction. The accuracy of the prediction depends on the quality of the training data and the model's ability to generalize to new data.

Machine learning also includes an evaluation step, where the model's performance is measured on test data that was not used during training. If the model performs well on the test data, it can be deployed to make predictions on new data.

In summary, machine learning is an iterative process that involves training a mathematical model from training data, optimizing the model to minimize prediction errors, evaluating the model's performance on test data, and finally deploying the model to make predictions on new data.

### 2.2.3 Types of machine learning:

There are three types of machine learning: supervised learning, unsupervised learning, and reinforcement learning.[31]



*Figure 7:types of machine learning .[32]*

#### 2.2.3.1 Supervised Learning:

In supervised learning, the model is trained from a set of labeled data. This means that input examples are associated with a known output. The goal is to learn to predict the output

for new inputs. For example, a supervised learning model can be trained to predict whether an image represents a cat or a dog from a set of labeled images.

Specifically, supervised learning can be used to solve classification or regression problems. In the case of classification, the goal is to predict the class of a new example, for example, to determine if an image represents a cat or a dog. In the case of regression, the goal is to predict a continuous numerical value, for example, to predict the price of a house based on its features.

To train a supervised learning model, it is necessary to split the dataset into two sets: a training set and a validation set. The training set is used to learn the model, while the validation set is used to evaluate the performance of the model on data it has never seen before. The goal is to find a model that generalizes well to new data, meaning it can accurately predict outputs for unknown examples.

The success of learning is evaluated by its ability to generalize correctly on novel examples that are not part of the training set. When the output is a finite set of values, the problem is called classification, and if there are only two possible values, it is called "Boolean" or "binary" classification. When the output is a real number, the problem is called regression, which involves finding a mean value or a conditional expectation of the output, as the probability of finding exactly the right real value is very low.[33]

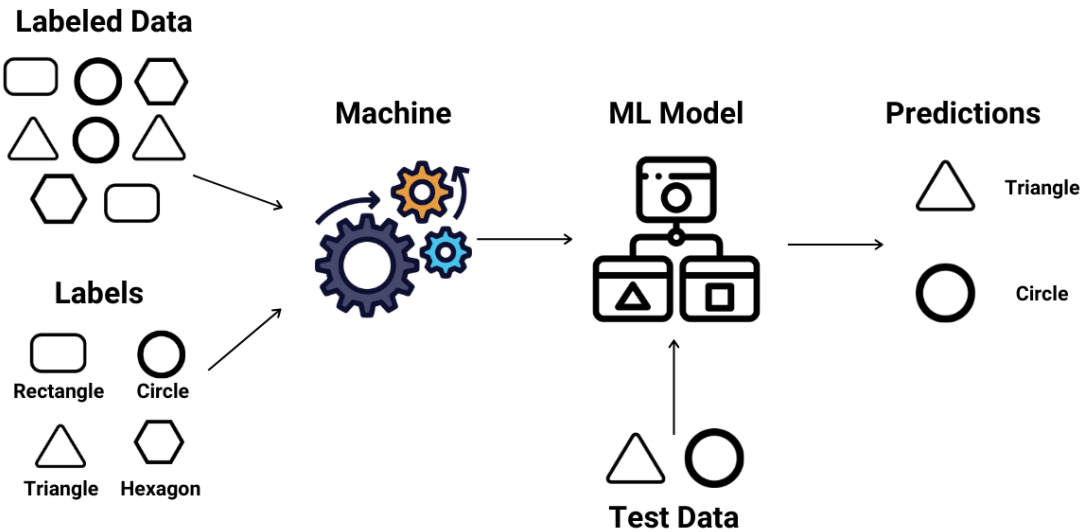


Figure 8: supervised learning workflow.<sup>36</sup>

Regression and classification are the two main subcategories of supervised learning, used to predict a continuous and discrete variable, respectively.

Regression:

Regression is used to predict a continuous numerical value based on a set of input variables. For example, predicting the price of a house based on its size, location, age, etc.

Classification:

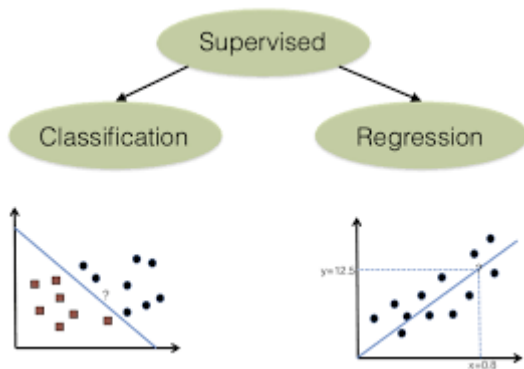


Figure 9: The two main subcategories of supervised learning.<sup>34</sup>

Classification, on the other hand, is used to predict a discrete label (category) based on input variables. For example, classifying emails as spam or not, detecting fraud on a credit card, or classifying the nature of a tumor as malignant or benign.

### 2.2.3.2 Unsupervised Learning:

Unsupervised learning is another machine learning method, where the goal is not to predict a labeled output, but rather to discover patterns and hidden structures in unlabeled data.

Unlike supervised learning, unsupervised learning does not require labels to train a model. Instead, it uses algorithms to uncover patterns and structures in data based on measures such as similarity, density, or probability.



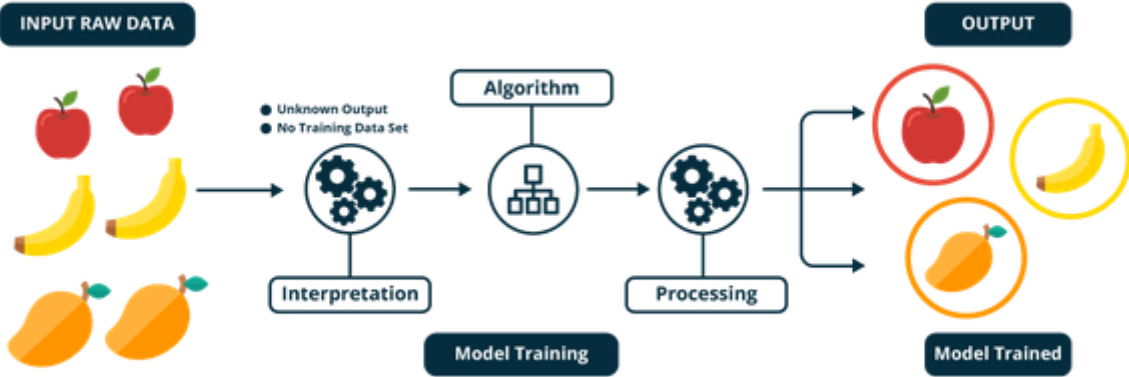


Figure 10:unsupervised learning workflow.[34]

Common examples of unsupervised learning include clustering, dimensionality reduction, data mining, and association.

**Clustering:** involves grouping similar data points into distinct clusters without any labeling information provided to the algorithm.

**Dimensionality reduction:** aims to reduce the complexity of data by reducing the number of input variables while retaining essential information.

**Data mining:** involves extracting useful information from large datasets, such as frequency patterns, correlations, trend models, etc.

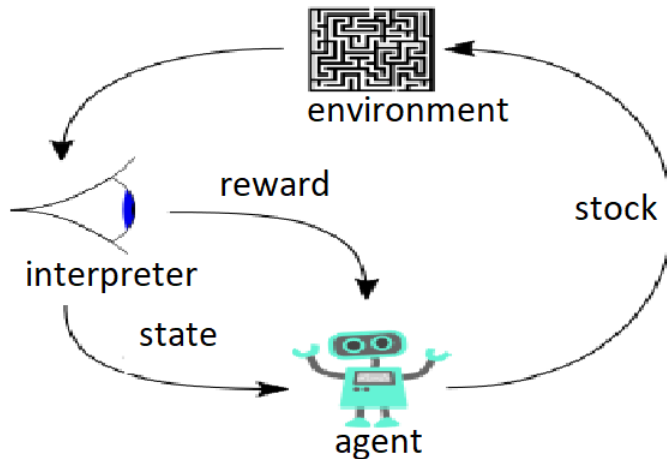
**Association:** is a problem where one desires to discover rules that describe large portions of their data. For example, in the context of a study of purchasing behavior of a group of customers, people who buy a particular product also tend to buy another specific product.[34]

2.2.3.3 Reinforcement learning:

Reinforcement learning is a method of learning where an agent learns to interact with its environment to maximize a numerical reward. The agent makes decisions based on observations it makes of its environment, and receives positive or negative rewards depending on the quality of its decisions. The goal is to maximize the total reward over a given period of

time. This method is often used in applications such as video games, robotics, and recommendation systems.

reinforcement learning is a method of learning where the agent learns to maximize a reward by interacting with the environment.



*Figure 11: Reinforcement learning workflow.[34]*

#### 2.2.4 Machine learning algorithms:

Machine learning algorithms are computer methods that allow computers to learn from data and make decisions or take actions without being explicitly programmed to do so.

There are several types of machine learning algorithms, each with its own characteristics and applications. Here are some examples:

**Artificial neural networks:** these algorithms are inspired by the functioning of the human brain and are used for tasks such as image recognition, automatic translation, or prediction of time series.

**Decision trees:** these algorithms allow decisions to be made by following a series of "yes/no" questions to arrive at a conclusion. They are often used for classification tasks.

**Clustering algorithms:** these algorithms allow similar data to be grouped into clusters. They are often used for market segmentation tasks or analysis of unstructured data.

**Regression algorithms:** these algorithms allow a numerical value to be predicted from input data. They are often used for prediction tasks, such as price or sales prediction.

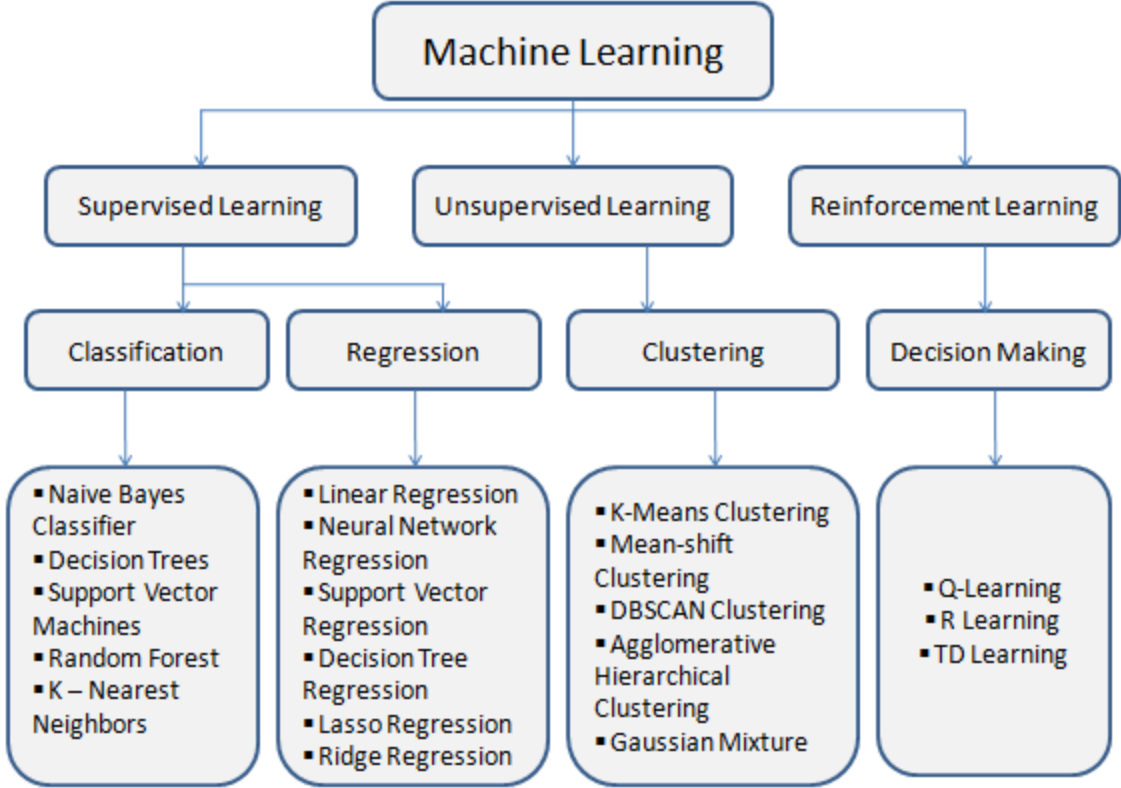


Figure 12:Machine Learning Algorithms.[35]

2.2.4.1 Logistic regression:

Is a machine learning algorithm that predicts the probability of an event occurring based on multiple input variables. This method is often used for binary classification problems, where the output variable can only take two possible values, such as "true" or "false", "positive" or "negative", etc.

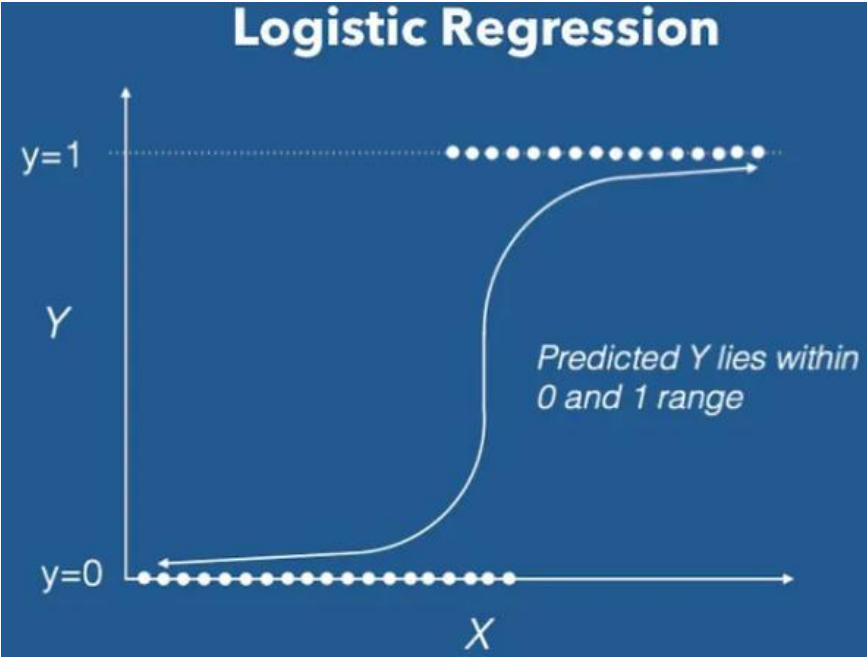


Figure 13:logistic regression exemple[36].

Logistic regression uses a logistic function to calculate the probability that the output variable takes the value "1" (as opposed to "0"). This function takes the linear combination of the input variables, weighted by regression coefficients, and returns a value between 0 and 1.

To train a logistic regression model, a training dataset is used containing examples labeled with their expected output value. The algorithm adjusts the regression coefficients to minimize the prediction error on this dataset.

Once the model is trained, it can be used to predict the probability that new examples belong to the positive class. If this probability exceeds a certain threshold (e.g. 0.5), the example is classified as belonging to the positive class; otherwise, it is classified as belonging to the negative class.[37]

2.2.4.2 Decision Tree:

The decision tree is a machine learning algorithm used for classification and regression. It is a decision model that takes the form of a tree where each node represents an input variable, each branch represents a value of that variable, and each leaf represents a prediction.

The goal of the decision tree is to create a predictive model that takes into account multiple input variables and predicts the output variable. To build the tree, the algorithm

recursively divides the data set based on the value of each input variable, in order to maximize the purity of the resulting subsets.

The purity of a subset is determined by the homogeneity of the output variable labels. For example, if a subset contains only examples of the "positive" class, then it is considered pure and will be associated with a leaf of the tree with a "positive" prediction.

The decision tree is a model that is easy to understand and interpret, as it allows for the visualization of the underlying decision logic. However, it can be sensitive to noisy or imbalanced data, and can easily overfit the training data.

There are several techniques for improving the performance of the decision tree, such as input variable selection, regularization, or tree ensemble (bagging or boosting). The use of these techniques allows for more robust and accurate models for predicting the output variable.

Decision trees are a popular and widely used method in data mining for classification and regression. Classification trees are used to predict membership in a target variable class, while regression trees are used to predict a numerical value for the target variable. Both types of trees function by recursively partitioning the target variable space into homogeneous subsets in terms of the target variable, and using decision rules to determine class membership or numerical value of the target variable. Decision trees are easy to interpret and visualize, making them useful in many areas including data analysis, predictive modeling, and machine learning.[38]

2.2.4.3 Random forest:

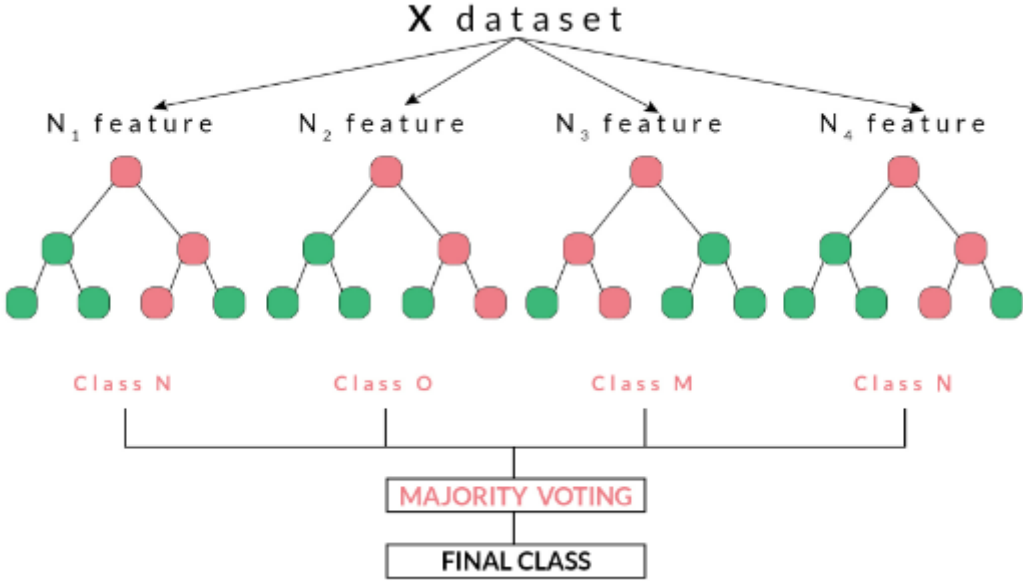


Figure 14:Diagram showing how decision tree algorithm works[39].

Random Forest is a powerful machine learning algorithm used for regression and classification. It consists of multiple decision trees, called estimators, that make individual predictions. These predictions are combined to produce a more accurate overall prediction.[40]

Random Forest overcomes the over fitting issues of standard decision trees by using an ensemble approach. It generalizes well to new and unseen data, even with missing values. It can handle large datasets with high dimensionality and features of different types (categorical and numerical).[40]

While Random Forest is excellent for classification tasks, its performance in regression problems is slightly less optimal. Unlike linear regression, a Random Forest regressor cannot predict values beyond the range of its training data.[40]

2.2.5 Application of machine learning:

Machine learning is used in a wide range of applications to give computers or machines the ability to analyze input data such as images, text, sound signals, videos, and many other types of data. The goal of machine learning is to enable machines to recognize patterns and make decisions based on the data they receive.[41]

Here are some examples of machine learning applications:

**Image recognition:** Machine learning algorithms can be trained to recognize objects or patterns in images, which is used in applications such as face detection, image classification, character recognition, traffic sign detection, and many others.

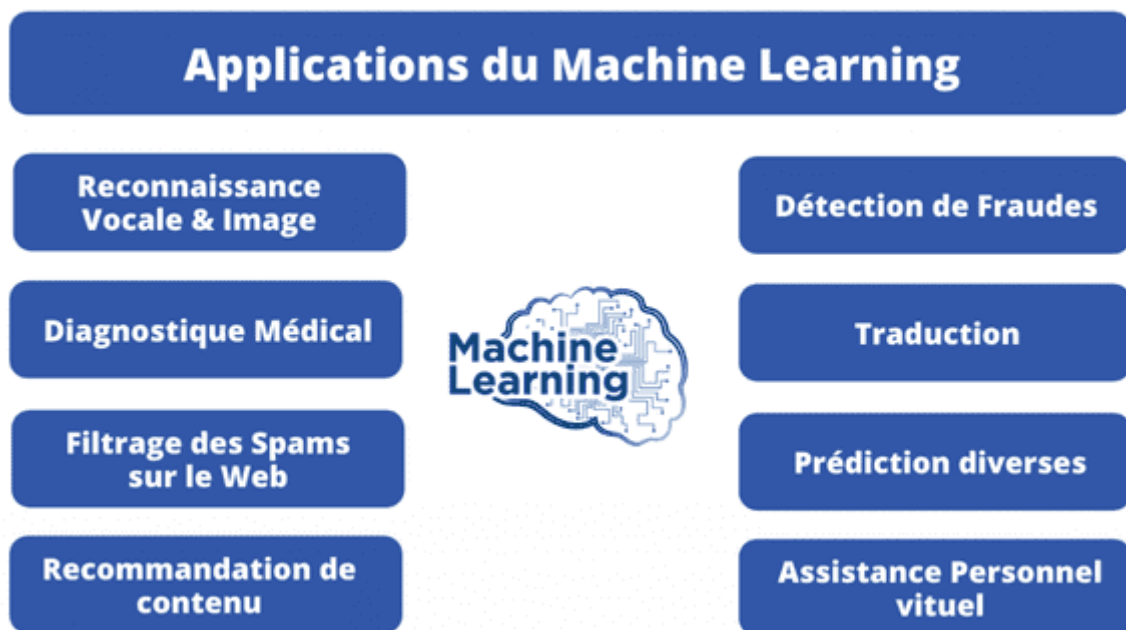
**Speech recognition:** Speech recognition systems use machine learning algorithms to understand and interpret users' voice commands. These systems are used in personal assistants such as Siri, Alexa, or Google Assistant.

**Fraud prediction:** Machine learning algorithms can be used to detect fraudulent behavior patterns in financial transactions, which can help prevent fraud before it happens.

**Product recommendation:** Machine learning algorithms are used in product recommendation applications to suggest products or services tailored to users' preferences.

**Automatic translation:** Machine translation systems use machine learning algorithms to understand linguistic patterns in one language and translate them into another language.

These examples are just a small part of the many applications of machine learning. The possibilities are endless, and more and more applications are being developed every day thanks to the constant advances in this technology.



*Figure 15: Applications of machine learning.[42]*

### 2.3 The impact of machine learning on CRM:

The impact of machine learning (ML) in customer relationship management (CRM) is significant and transformative. ML has revolutionized the way businesses manage and engage with their customers, leading to improved customer satisfaction, increased operational efficiency, and enhanced business outcomes. Here are some key impacts of ML in CRM:

**Personalization:** ML enables businesses to personalize customer experiences by analyzing vast amounts of customer data and generating tailored recommendations, offers, and communications. By understanding customer preferences, behaviors, and needs, businesses can deliver targeted and relevant interactions, resulting in higher customer satisfaction and engagement.

**Predictive Analytics:** ML algorithms can analyze historical customer data and identify patterns, trends, and correlations that may not be apparent to humans. This enables businesses to make accurate predictions about customer behavior, such as churn prediction, cross-sell and upsell opportunities, and customer lifetime value. By leveraging predictive analytics, businesses can proactively address customer needs, improve retention rates, and optimize marketing and sales strategies[43].

**Intelligent Automation:** ML-powered automation capabilities streamline CRM processes and workflows, saving time and effort for businesses. Tasks such as lead scoring, customer segmentation, and response automation can be automated using ML algorithms, allowing businesses to focus on high-value activities and provide faster and more efficient customer service.

**Sentiment Analysis:** ML techniques, such as natural language processing (NLP), can analyze customer feedback, social media posts, and customer interactions to extract sentiment and emotions. This provides businesses with valuable insights into customer satisfaction levels, brand perception, and customer sentiment towards specific products or services. By understanding customer sentiment, businesses can take proactive measures to address issues and improve customer satisfaction[44].

**Enhanced Customer Service:** ML-powered chatbots and virtual assistants can provide instant and personalized responses to customer inquiries, 24/7. These intelligent chatbots can understand natural language, interpret customer queries, and provide relevant solutions or



escalate issues to human agents when necessary. ML enables businesses to deliver fast, accurate, and consistent customer service, leading to improved customer satisfaction and loyalty[45].

**Data-driven Decision Making:** ML algorithms analyze large volumes of customer data to generate actionable insights and recommendations. This empowers businesses to make data-driven decisions, optimize marketing campaigns, improve sales forecasting, and tailor product offerings to customer needs. By leveraging ML-driven insights, businesses can align their strategies with customer preferences and enhance overall customer satisfaction[45].

## 2.4 Similar works:

### 2.4.1 Predicting negative customer reviews Using Data Science to Predict NegativeCustomer Reviews:

The study by Chuck Utterback highlights the importance of predicting and addressing negative customer reviews to improve customer satisfaction and mitigate potential business risks.

**Exploratory Data Analysis:** Various visualizations are used to gain insights into the data. Key findings include the impact of late deliveries on bad reviews, the relationship between seller quality and reviews, variation in bad review rates by product category, and geographic differences in review patterns.[46]

**Correlation Analysis:** Positive correlations are observed for delivery intervals, product and seller bad review rates, and certain order-related metrics.[46]

**Predictive Modeling:** Seven supervised learning classification models are trained and optimized on the precision metric, with a focus on achieving reasonable recall (the proportion of bad reviews correctly predicted). The Gradient Boost model is recommended as the best-performing model, with a precision of 84%, recall of 29%, accuracy of 88%, and AUC of 64%. The most influential features in the model include delivered days late, days variance, product and seller bad review rates, and order-related metrics.[46]

### 2.4.2 EDA + Satisfaction Prediction + NLP + RFM + Deployment:

the project don by "MOHAB ALLAM" aims to leverage the data from Olist to understand customer satisfaction in the e-commerce sector and provide recommendations for improving the customer experience and driving business growth[47].

By analyzing the data set, extract some characteristics that can be useful in forecasting, and choose a set of features related to the product, such as price, category, product description such as payment type and installments[47].

In order to choose the appropriate attributes for the prediction, use the "SelectKBest" method in Python to select the most important attributes to include in the prediction model, and select 13 important attributes for the prediction[47].

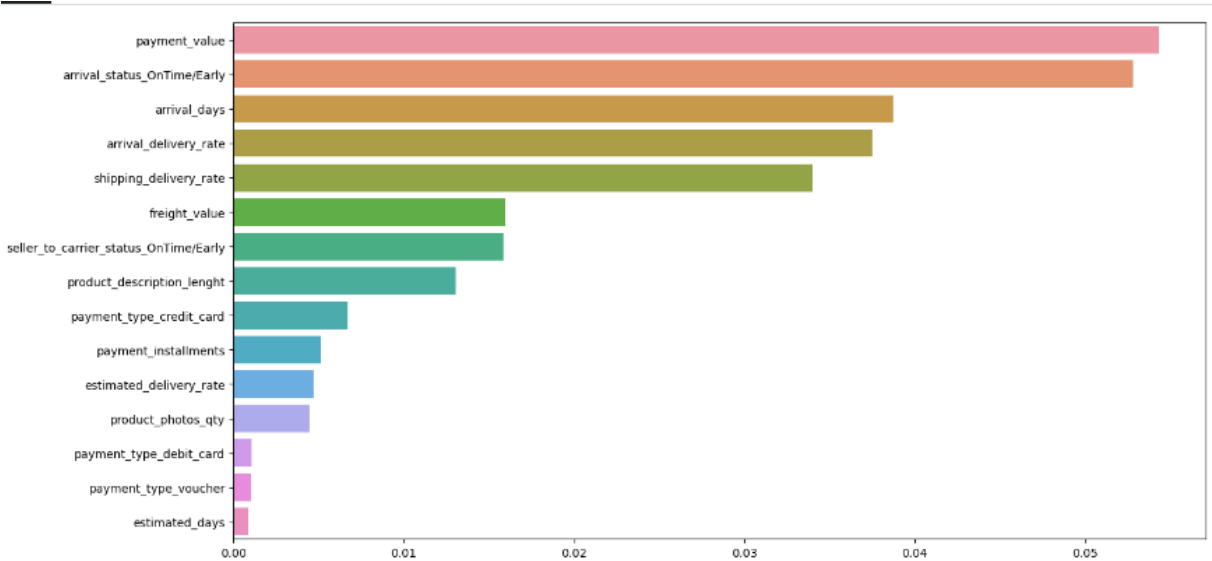


Figure 16:A graphic showing the importance of attributes using the method "SelctKBest"[47].

For prediction, he chose a set of binary classification algorithms such as random forest, Ada Boost,KNN Classifier,Decision Tree....,And the best result he getit is 88% acuraccy by using "MultinomialNB" algoithem[47].

### 2.4.3 Customer Satisfaction Prediction Using Machine Learning:

The aim of this work is to predict customer satisfaction through the customer information in the data set " olist", and this work was applied by "paritoshmahto"[48].

And choose a set of attributes that can be included in the forecast model, product-related attributes such as price, location-related attributes such as customer status and customer postcode prefix, shipping-related attributes such as shipping values, and customer-related attributes such as comment messages...[48]

And he extracted a set of new features that could be useful in his work, for example, he made a calculation RFM score (recency, frequency, and monetary value) for each customer...[48]

For the prediction process, a set of algorithms such as: Linear SVM Model, Random Forest, Logistic Regression..., and the best result obtained using VotingClassifier (vot\_hard)algorithm , with an accuracy of 80%,and the best result obtained using deep learning, with an accuracy of 83%[48].

### Conclusion:

In this chapter, we presented what artificial intelligence and machine learning are, we also talked about the principle of learning and mentioned some use cases. We mentioned the different types of learning that exist with their respective algorithms. It highlights the significance of machine learning in advancing AI and its impact on CRM.

The next chapter represents the creation of a machine learning model for predicting customer satisfaction using machine learning algorithms.



## **Chapter Three: construction and implementation**

### 3.1 Introduction:

In this chapter, we will provide an overview of our project, including the problem statement and the software environment used in the implementation. We will outline the specific steps involved in implementing our solution and discuss the dataset we used, highlighting its key characteristics. Finally, we will present the results obtained from the different algorithms used in our project and compare them with similar works.

### 3.2 Project objective:

Our project problem is to predict the review score of customer satisfaction, based on historical data of the customer. The given task involves working with data from the Brazilian e-commerce site Olist and involves using various features such as price, item description, on-time delivery and delivery status, etc., to predict the satisfaction score of a customer. The problem can be framed as a classification problem, where the objective is to predict whether a customer will be satisfied or dissatisfied with their purchase. The prediction can be based on various factors such as the customer's past buying behavior, feedback, reviews. The goal is to build a predictive model that can accurately predict customer satisfaction and help businesses improve their services and offerings to ensure better customer satisfaction.

Our work can be approached as a multi-layered rating problem where the goal is to predict a customer satisfaction score as [1,2,3,4,5], or as a binary rating problem where scores greater than 3 are positive and scores less than or equal to 3 are negative.

### 3.3 The development environment and tools:

#### 3.3.1 Google Colab:

Google Colab, short for Google Collaboratory, is a cloud-based development environment for writing, running, and sharing code in Python. It is designed to provide easy access to powerful computing resources and tools for data analysis, machine learning, and other related tasks.[49]

Colab allows users to create and run Jupyter notebooks, which are interactive documents that can contain live code, text, equations, and visualizations. It also provides access to a wide range of data science libraries, including TensorFlow, Keras, PyTorch, Scikit-learn, Pandas, and NumPy.[49]

One of the key advantages of Colab is that it provides free access to a GPU and TPU for running computationally intensive workloads, which can be very expensive to run on a personal computer. Additionally, Colab allows users to easily share their notebooks with others, making it a useful tool for collaboration and reproducible research.[49]

Colab is hosted by Google, which means that users do not need to install any software on their local machines, and can access their work from anywhere with an internet connection. However, this also means that Colab notebooks are subject to Google's terms of service, which may limit certain types of usage or access to certain data sources.[49]

- We can use “Google Colab” through this link: <https://colab.research.google.com>

### 3.3.2 Python:

Python is a high-level, interpreted programming language that is widely used in a variety of domains, including web development, data analysis, machine learning, artificial intelligence, scientific computing, and more. It was first released in 1991 by Guido van Rossum and has since become one of the most popular programming languages in the world, thanks to its simplicity, readability, and versatility.

- We can download the python from their original website via this link: <https://www.python.org>

### 3.3.3 Python Librarys:

Python has a vast ecosystem of libraries, which are collections of pre-written code that can be imported into your Python program to extend its functionality. Here are some commonly used Python libraries:

- **Pandas:** A library for data analysis and manipulation(`read_csv()`,`marge()`,`get_dummies()`,`to_datetime()`).
- **Matplotlib:** A library for data visualization and plotting(`figure()`).
- **Scikit-learn:** A library for machine learning, including classification, regression, and clustering algorithms(`model_selection()`, `feature_selection()`).

## 3.4 Design:

### 3.4.1 System structure:

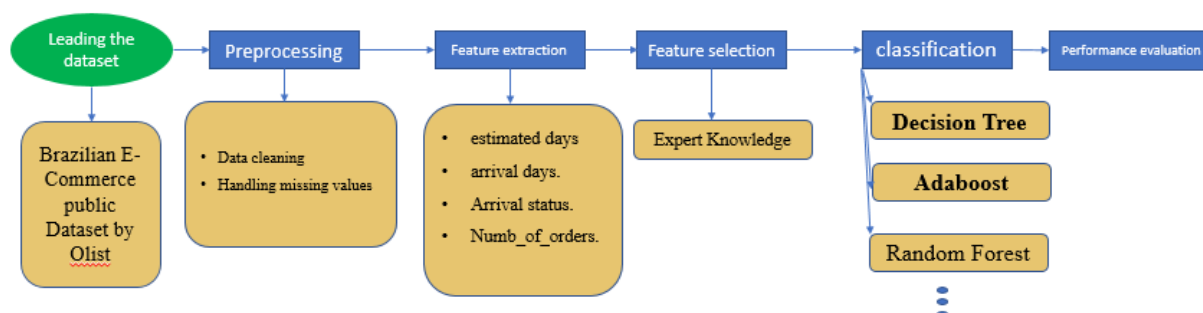


Figure 17: System structure.

We obtained a dataset from the Olist e-commerce platform containing customer information. The dataset was carefully chosen to ensure representation of diverse customer demographics. For data preprocessing, we handled missing values. Additionally, we performed feature engineering to create new relevant features, normalized numerical features, and encoded categorical variables. Through exploratory data analysis, we gained insights into the dataset, identifying distribution, correlation, and patterns using visualizations such as histograms and scatter plots. Based on domain knowledge and expert input, we selected the final set of features. We then split the dataset into training and testing sets and developed predictive models using various machine learning algorithms, including logistic regression and random forests. The models were evaluated and compared using accuracy.

### 3.4.2 The dataset

a general description of the Olist dataset

#### 3.4.2.1 Description of the data set:

The Olist dataset is a publicly available e-commerce dataset consists of 100,000 orders placed in the period from 2016 to 2018 across multiple markets in Brazil. that contains information about orders, customers, products, and sellers, which connects sellers to customers through its platform. The dataset provides valuable insights into customer behavior, product sales, and seller performance[50].

**Order Data:** Each order is represented by a unique order ID and contains information such as the order purchase date, order status, and the customer's ID. Order-related features include the



order value, freight value, payment type, and estimated delivery date. The dataset also includes information about the customer's location, including the state and city[50].

**Customer Data:** Each customer is identified by a unique customer ID and some customer-related features Customer location, such as state and city.[50]

**Product Data:** Each product is represented by a unique product ID and includes information such as the product category, product name, and product description. The dataset also provides details about the seller associated with each product, including the seller ID, seller name, and seller location[50].

**Review Data:** Customers can leave reviews and ratings for products they have purchased. The review data includes information such as the review score, review comment, and the date of the review[50].

**Seller Data:** Each seller is identified by a unique seller ID and includes information such as the seller's name, location, and business segment[50].

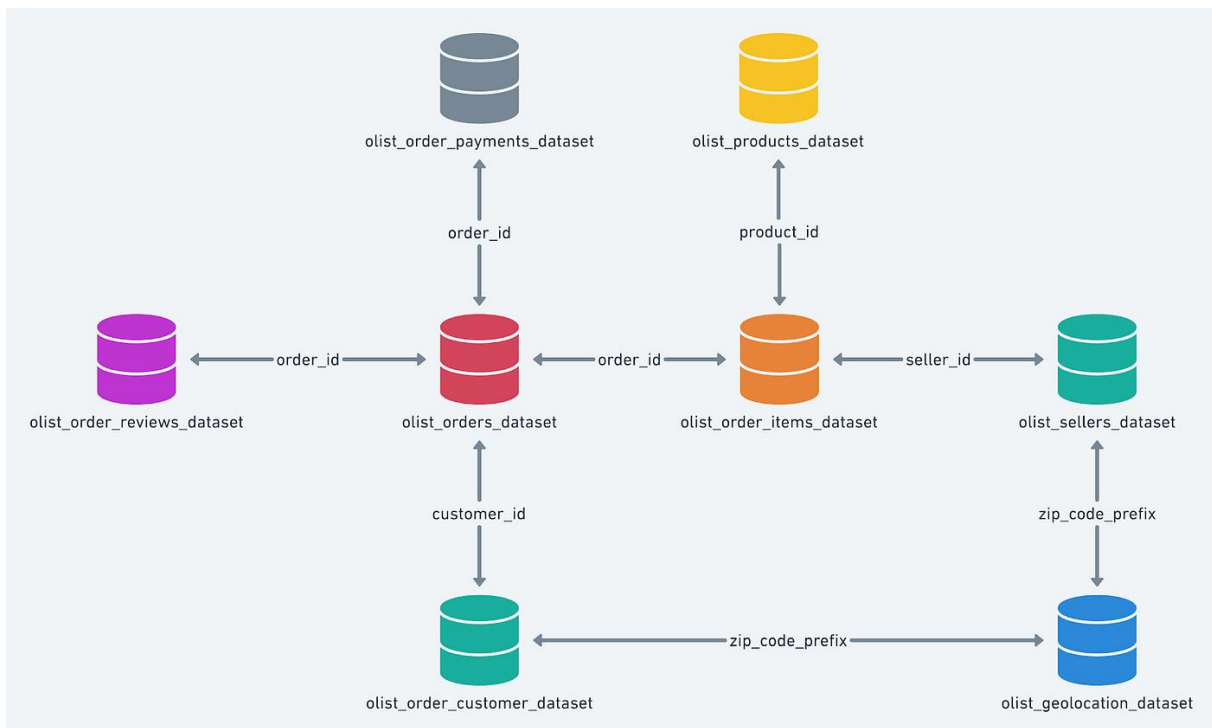


Figure 18: Brazilian E-Commerce Public Dataset by Olist.<sup>44</sup>

The data is divided into 9 CSV files, and it is necessary to combine them to obtain the final data needed for analysis, feature selection, and model training.

- We can download and use this dataset from the following link:<https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

### 3.4.2.2 Justification for our selection of this dataset:

- The dataset contains various features related to the customer's shopping experience, such as product ratings, seller ratings, review scores, and order status. These variables can be used as inputs to predict customer satisfaction accurately.
- The dataset is curated by Olist, a well-known Brazilian e-commerce platform, which indicates that it is likely to be of high quality and completeness.
- The dataset consists of thousands of orders, covering a wide range of products, customers, and sellers. This diversity helps in capturing different scenarios and variations that can affect customer satisfaction, making it more representative and suitable for building robust machine learning models.
- The dataset is publicly available on the Olist GitHub repository, making it easily accessible for researchers and practitioners. Its availability fosters transparency, reproducibility, and comparability of results across different studies.
- The "Brazilian e-commerce generic dataset by Olist" has been used in various studies, showcasing its suitability for customer satisfaction prediction and related research tasks.

## 3.5 Implementation:

### 3.5.1 Preprocess the data:

we used merge method in library pandas to merge each csv dataset according to the specified data schema.

```
#df: data Frame
#pd: pandas library
#merge: merge is a function in the pandas library to marge datasets together
df= pd.merge(customers_df, orders_df, on="customer_id", how='inner')
df= df.merge(reviews_df, on="order_id", how='inner')
df= df.merge(items_df, on="order_id", how='inner')
df= df.merge(products_df, on="product_id", how='inner')
df= df.merge(payments_df, on="order_id", how='inner')
df= df.merge(sellers_df, on='seller_id', how='inner')
```

#### 3.5.1.1 Handling missing values:

- To address the issue of missing data, we discovered that a more effective approach is to fill in the empty values rather than simply excluding them. This is important as excluding missing data can adversely impact the efficacy of machine learning.
- To address the missing values, we employ a specific method for filling in the empty values.
- We replaced the missing values in the *'order\_approved\_at'* column with the corresponding values from the *'order\_purchase\_timestamp'* column.

To clarify:

“*order\_approved\_at*”: Timestamp of the order approval.

“*order\_purchase\_timestamp*”:Timestamp of the purchase.

- We substituted the missing values in the *'order\_delivered\_customer\_date'* column with the corresponding values extracted from the *'order\_estimated\_delivery\_date'* column.

To clarify:

"*order\_delivered\_customer\_date*": date at which customer got the product.

"*order\_estimated\_delivery\_date*": estimated delivery date of the products.

#### 3.5.1.2 Handlingthe duplicated values:

Handling duplicated values in a dataset is an important preprocessing step when working with machine learning models. Duplicated values can introduce biases, affect the performance of models, and lead to incorrect results.

We did not encounter any instances of repeated values, thus there was no need for us to address this situation.

#### 3.5.1.3 Datetime features:

converting datetime features to the datetime format facilitates proper processing and analysis of time-related data, allowing you to take advantage of the full range of capabilities offered by Python libraries and perform meaningful feature engineering to improve model performance.

```
#Convert Datetime features from Object to Datetime
df['order_purchase_timestamp'] = pd.to_datetime(df['order_purchase_timestamp'])
df['order_delivered_customer_date'] =
pd.to_datetime(df['order_delivered_customer_date'])
df['order_estimated_delivery_date'] =
pd.to_datetime(df['order_estimated_delivery_date'])
df['shipping_limit_date'] = pd.to_datetime(df['shipping_limit_date'])
df['order_delivered_carrier_date'] =pd.to_datetime(df['order_delivered_carrier_date'])
```

### 3.5.1.4 Feature Engineering:

- a) To derive the estimated shipping time, we computed the difference between the 'order\_estimated\_delivery\_date' and the 'order\_purchase\_timestamp', resulting in a new feature called '*estimated\_days*'.
- b) To calculate the shipping time from the purchase date to the customer's delivery date, we obtained the difference between the "order\_delivered\_customer\_date" and the "order\_purchase\_timestamp". This calculation yielded the feature "*arrival\_days*".
- c) And to find out if the delivery process was on time or late, we subtracted the day the customer received the product "order\_delivered\_customer\_date" from the estimated delivery day "order\_estimated\_delivery\_date", and if the production is greater or equal to 0, then the delivery process is on time or early, otherwise it is The order arrived late, and these results were shown in the new feature "*arrival\_status*".
- d) To find out how often each customer orders from the merchant's website, we counted the frequency of each customer's '*customer\_unique\_id*'; And the output we showed in the new feature "*numb\_of\_orders*".

### 3.5.2 Feature Selection:

- a) Paymenttype:

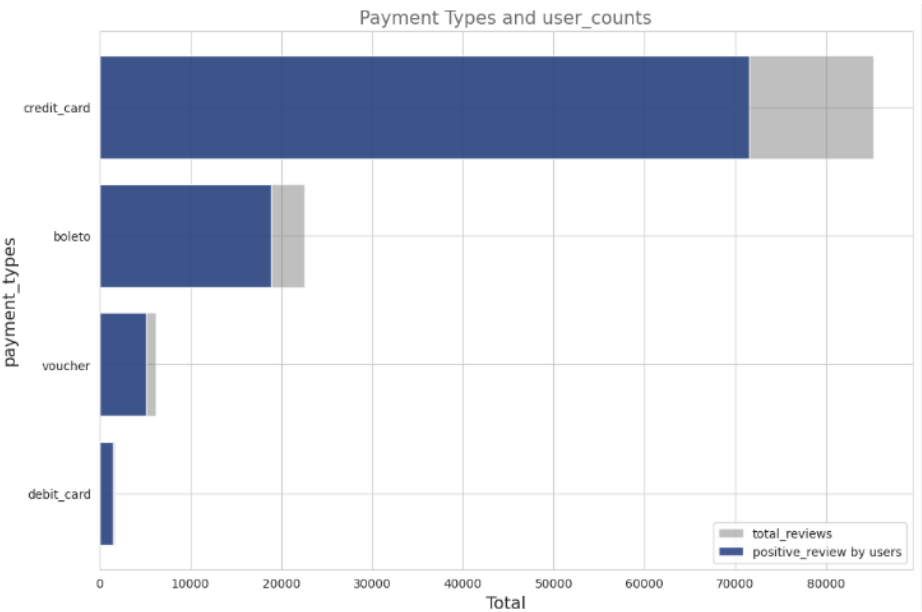


Figure 19: Graph showing positive reviews for each payment type

This eCommerce dataset contains four primary payment methods that are commonly used: credit card, boleto (a type of payment slip used in Brazil), voucher, and debit card.

Based on the plots provided above, we can observe that the majority of orders in this eCommerce dataset are paid using a credit card. The second most commonly used payment method is boleto, which is used for a smaller proportion of orders compared to credit card payments.

Based on the stacked plot shown above, it is evident that the majority of customers who used credit cards to make payments have given positive reviews. The same is also observed for customers who used boleto, voucher, and debit cards.

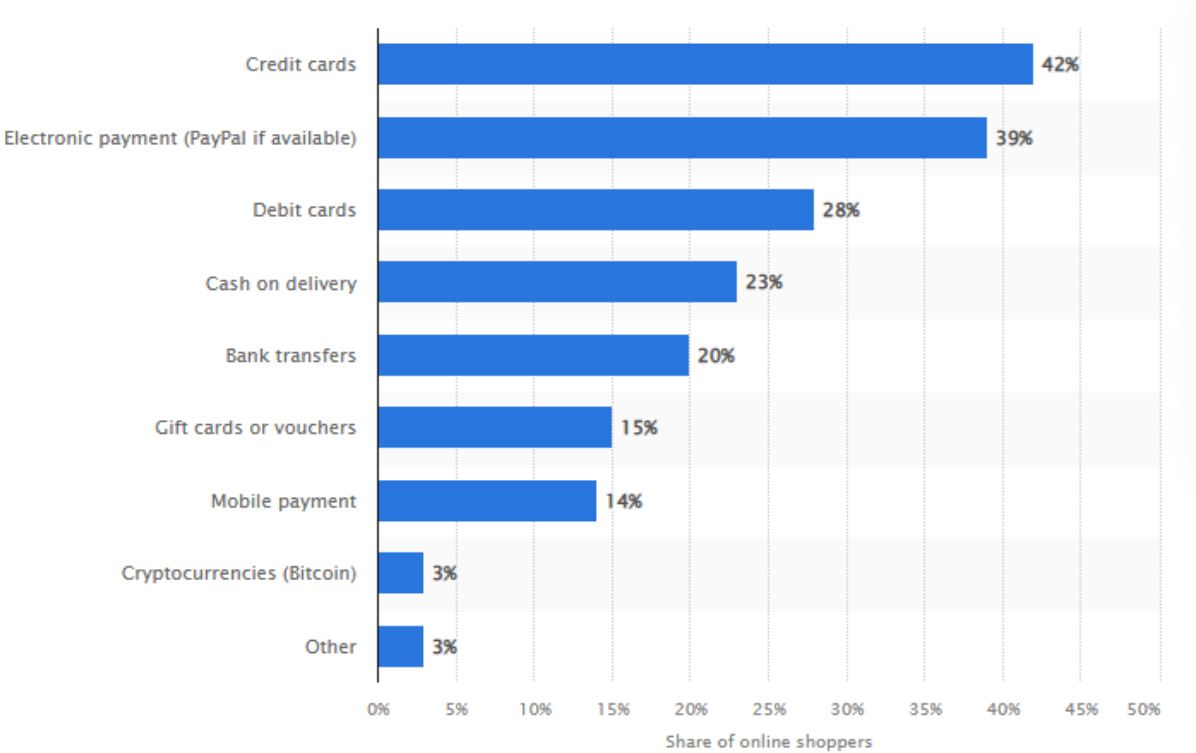


Figure 20: Preferred payment methods of online shoppers worldwide as of March 2017.

According to a survey conducted in March 2017, it was found that among global online shoppers, 42 percent of respondents preferred to pay via credit card. This indicates that credit cards were the most favored payment method among the options available, surpassing electronic options like PayPal[51], it is worth noting that payment preferences may have evolved since the survey was conducted in 2017. As the online payment landscape continues to evolve, new payment methods and platforms have emerged, providing consumers with more options and convenience. To gain a comprehensive understanding of current payment preferences, it would be beneficial to refer to more recent surveys and studies[51].

The results of the "Usage of E-Payment and Customer Satisfaction" research aims to understand the levels of customer satisfaction with e-payment usage, their readiness for adopting a changed lifestyle, and the factors that encourage or discourage their usage of e-payments. Based on these findings, the study proposes necessary strategies and approaches to enhance the e-payment environment[52].

This suggests that the payment method used by the customers could be an important categorical feature for the problem being analyzed.

b) Consumersstate:

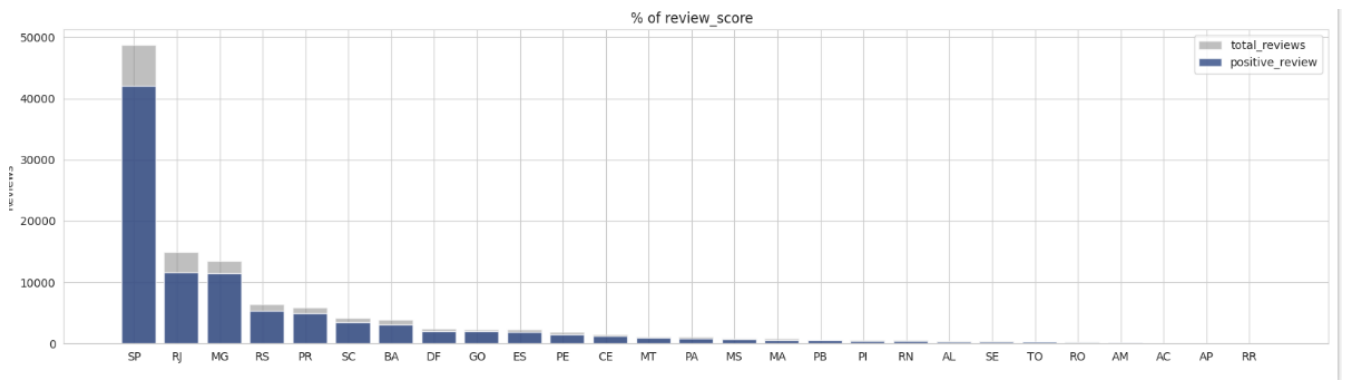
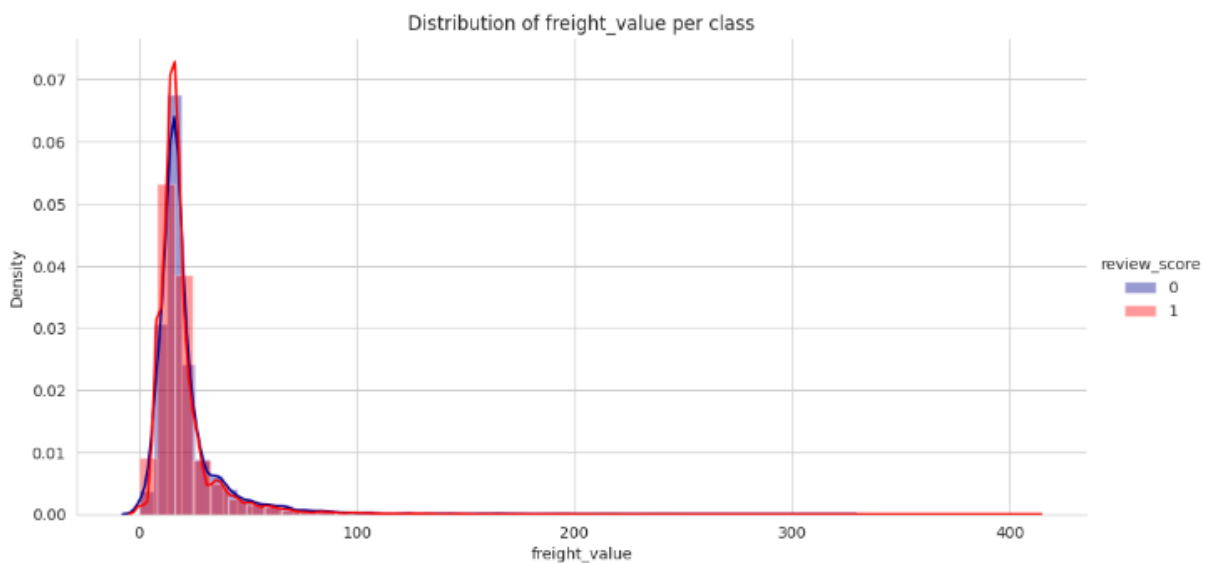


Figure 21: Graph showing positive reviews for each state.

The stacked plot of reviews per state shown below indicates that the majority of consumers from each state have given positive reviews. For example, in the SP state, out of a total of 40800 reviews, 35791 reviews are positive. Similarly, in the RJ state, 9968 reviews are positive out of a total of 12569 reviews. Based on these observations.

we can conclude that the consumer\_state feature could be important for the problem being analyzed.

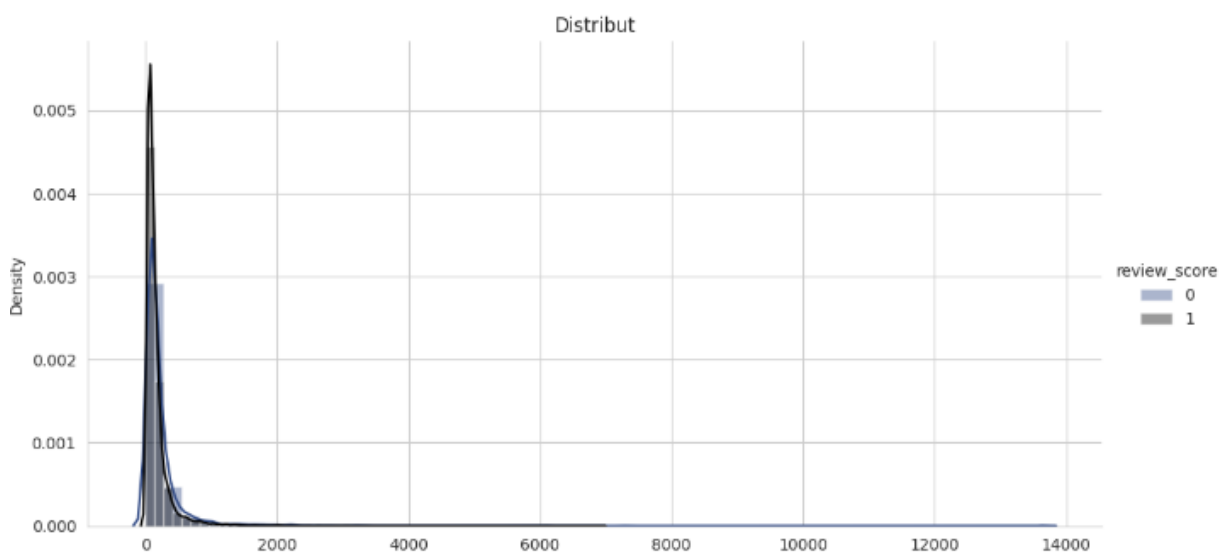
c) Freightvalue:



*Figure 22: the distributions of "freight\_value" for positive and negative classes.*

The distribution plot indicates the distributions of the "freight\_value" feature for both the positive and negative classes. Based on the plot, it appears that there is significant overlap between the two distributions, indicating that the "freight\_value" feature alone may not be sufficient to accurately classify or differentiate the positive and negative classes.

d) Paymentvalue:



*Figure 23: the distribution of the "payment\_value" for both the positive and negative classes.*

Based on the description of the distribution plot, it appears that the distribution of the "payment\_value" feature for both the positive and negative classes exhibits significant overlap. This indicates that using the "payment\_value" feature alone may not be sufficient to accurately classify or differentiate between the positive and negative classes.

e) Numboforders:



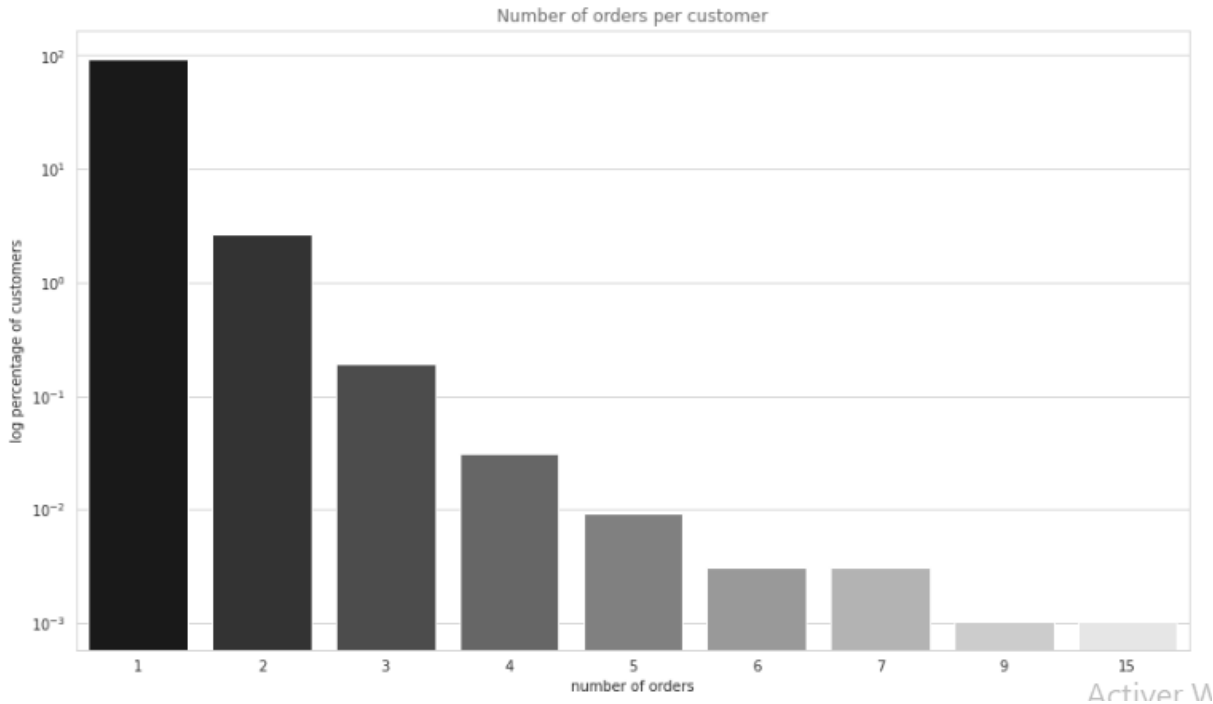
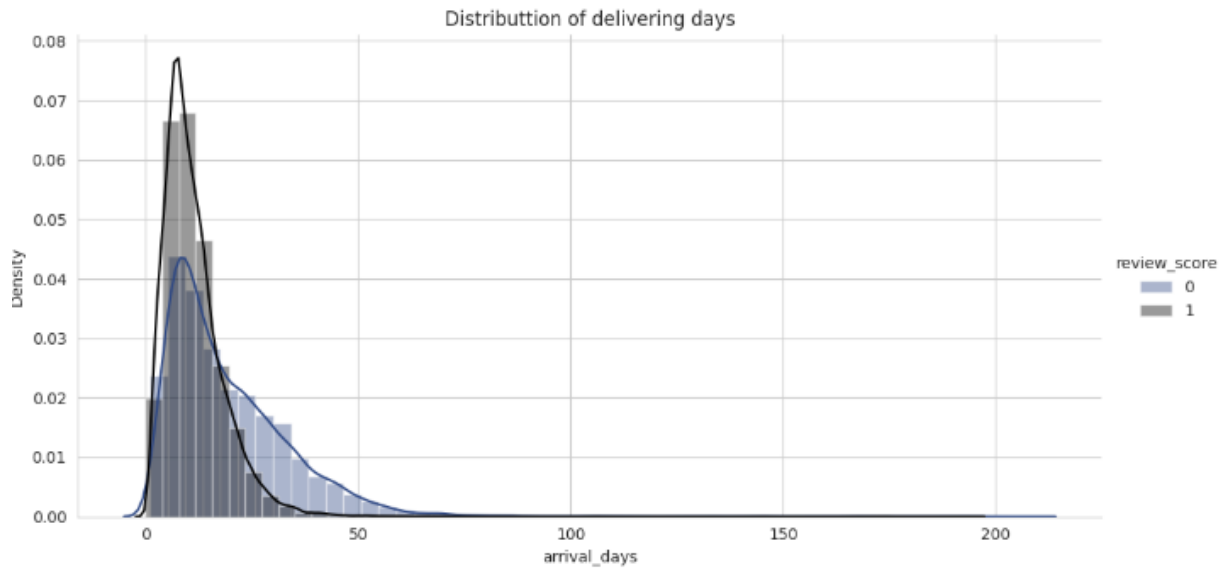


Figure 24: Distribution of the number of orders per percentage of customers.

the number of orders can indeed be an important feature in certain problems. where most consumers only place an order once and a few consumers place orders more than 35 times, the order frequency can provide valuable information for classification or analysis.

f) Arrival days:

Arrival days are the number of days from the date of order to the date of receipt of the product ('order\_purchase\_timestamp') by the customer (order\_delivered\_customer\_date).



*Figure 27: Distribution of delivering days for positive and negative classes.*

Based on the description of the distribution plot, the distribution of the “arrival\_days” feature for the positive and negative categories seems to show that the fewer the arrival days, the more positive the review.

In the article "The impact of online shopping attributes on customer satisfaction and loyalty: Moderating effects of e-commerce experience" indicates that the entire delivery process, including the logistics of returned items, can affect customer patience and contribute to overall online shopping dissatisfaction. Therefore, online retailers need to pay attention to the entire delivery procedure, including return logistics, to ensure a positive customer experience. By prioritizing fast and accurate delivery, retailers can meet customer expectations and boost satisfaction and loyalty in the competitive online retail landscape[53].

we can conclude that the “arrival\_days” feature could be important for the problem being analyzed.

g) Estimateddays:

The article "The impact of online shopping attributes on customer satisfaction and loyalty: Moderating effects of e-commerce experience" indicates that Perceived delivery time is a critical factor that positively affects customer satisfaction. Timely and reliable delivery enhances satisfaction levels and encourages repeat purchases, highlighting the impact of timely delivery on customer satisfaction. Conversely, delays in delivery can lead to customer

anxiety and discontentment. The longer customers have to wait between ordering and receiving a product, the higher the chances of dissatisfaction[53].

we can conclude that the “estimated\_days” feature could be important for the problem being analyzed

h) Arrivalstatusontime/early:

In the article "The Effect Of On-Time Delivery On Customer Satisfaction And Loyalty In Channel Integration" Logistics and distribution activities are emphasized as being of increasing importance in this context. It is suggested that companies review their distribution processes and conclude comprehensive agreements with logistics service providers to ensure that products are delivered to customers on time. This means that timely delivery plays an important role in meeting customer expectations and enhancing satisfaction[54].

we can conclude that the “arrival\_status\_ontime/early” feature could be important for the problem being analyzed

i) Productdescriptionlength:

information quality is a critical factor for e-retail success. E-retailers must ensure the accuracy, relevance, timeliness, completeness, consistency, and format of the information presented on their websites. By doing so, they can positively influence customer perceptions, satisfaction, trust, and loyalty[53].

Customers rely heavily on information provided on websites to make informed purchasing decisions, especially when they are unable to actually evaluate the product. Therefore, it is essential for electronic retailers to ensure that the information provided on their websites is accurate, relevant, timely, complete, consistent, and presented in an easy-to-use format[53].

Hence, we conclude that “producte\_description\_length” is an important feature for predicting customer satisfaction and can be used.

j) Productphotosqty:

The article of "The Importance of Product Photos in ecommerce" highlights the importance of high-quality and sufficient quantity of product images in e-commerce to

convert users into customers. The image of a product serves as an introduction to the customer, and it is crucial for it to be visually appealing, accurate, and representative of the actual product[55].

Regarding the quantity of photos, it is stated that the number of images may vary depending on the product itself. However, it is recommended to include at least one front picture of the full product. Additionally, including multiple perspectives and detailed shots can be beneficial. For example, when showcasing a shoe, it is suggested to include four photographs showing the front view, side view, rear view, and close-ups of details like laces, instep, and toe[55].

The statistics mentioned from recent studies by Shopify provide insights into online consumer preferences. Only a small percentage (0.52%) of consumers prefer to see a single product photo, while a significant portion (33.16%) prefer multiple photos. The majority (58.03%) expressed a preference for photographs that offer a 360° view of the product, enabling them to have a comprehensive understanding of its appearance[55].

These findings indicate the significance of visual content in influencing consumer behavior in e-commerce[55].

we can conclude that the “product\_photos\_qty” feature could be important for the problem being analyzed

### 3.5.3 Model Selection:

**Random Forest:** Random Forest is a versatile ensemble algorithm known for its robustness and high performance in various classification tasks. It combines multiple decision trees and employs bagging and feature randomness to improve accuracy and reduce overfitting. Random Forest is effective in handling complex datasets with a large number of features and can capture non-linear relationships[56].

**Decision Tree:** Decision Tree algorithms are intuitive and easy to interpret. They provide a clear decision-making process based on splitting criteria, making them useful for gaining insights into the data and identifying important features. Decision Trees can handle both categorical and numerical data and are capable of handling non-linear relationships. However, they are prone to overfitting on complex datasets[57].

**Logistic regression:** It is important to note that logistic regression has its limitations. It assumes that the relationship between input and output variables is linear, which may not always hold in complex datasets. In such cases, more advanced techniques such as decision trees, random forests, or neural networks may be more appropriate. Nevertheless, logistic regression remains a powerful and widely used algorithm for classification tasks, particularly when the assumptions of linearity and simplicity are reasonable[58].

**Adaboost:** Adaboost, or Adaptive Boosting, is a widely used algorithm for classification problems. It improves accuracy by combining multiple weak classifiers in an iterative manner, focusing on misclassified instances. Adaboost is flexible and can work with various weak classifiers, allowing it to adapt to different data types. It handles complex relationships by capturing nonlinear patterns through the ensemble of weak classifiers. Moreover, Adaboost reduces overfitting and provides insight into feature importance. However, it can be sensitive to noisy data and outliers, and its computational cost may increase with complex models or large datasets. Overall, Adaboost is a powerful algorithm that enhances classification accuracy and handles intricate data relationships[59].

### 3.6 Results and discussion:

Now we will discuss the results we obtained through our work:

Algorithms \ Evaluation	Random forest	Decision Tree	logistic regression	AdaBoost	KNN Classifier	xgboost
Training recall	1.00	1.00	0.67	0.76	0.91	0.92
Testing recall	0.77	0.73	0.67	0.68	0.72	0.69
Training precision	1.00	1.00	0.68	0.76	0.92	0.92
Testing precision	0.84	0.71	0.63	0.64	0.64	0.78
Training F1Score	1.00	1.00	0.67	0.76	0.91	0.92
Testing F1Score	<b>0.90</b>	0.84	0.77	0.77	0.75	0.87

*Table 1: the results we obtained.*

The table presents the evaluation results of various machine learning algorithms, including Random Forest, Decision Tree, Logistic Regression, AdaBoost, KNN Classifier,

and XGBoost. The results are measured in terms of recall, precision, and F1 score, both for training and testing datasets.

In terms of recall, which measures the ability of the models to correctly identify positive instances, the Random Forest algorithm achieved the highest recall on the training set (1.00), followed closely by the Decision Tree (1.00), KNN Classifier (0.92), and XGBoost (0.92). On the testing set, Random Forest still performed well with a recall of 0.77, followed by Decision Tree (0.73) and KNN Classifier (0.72).

Regarding precision, which measures the ability of the models to correctly classify positive instances, Random Forest had the highest precision on the training set (1.00), followed by Decision Tree (1.00), KNN Classifier (0.92), and XGBoost (0.92). However, on the testing set, Random Forest (0.84) and XGBoost (0.78) outperformed the other algorithms in terms of precision.

The F1 score, which considers both precision and recall, indicates the overall performance of the models. Random Forest achieved a perfect F1 score of 1.00 on the training set, followed by Decision Tree (1.00), KNN Classifier (0.92), and XGBoost (0.92). On the testing set, Random Forest (0.90), XGBoost (0.87), and Decision Tree (0.84) had the highest F1 scores.

Based on these results, the Random Forest algorithm demonstrated the highest testing accuracy (0.90), indicating its effectiveness in predicting the target variable. XGBoost also performed well with an accuracy of 0.87. Decision Tree showed competitive performance with an accuracy of 0.84. Logistic Regression and AdaBoost algorithms achieved lower accuracy on the test data, with logistic regression performing slightly better.

These results suggest that the Random Forest algorithm may be the most suitable choice for the given problem, followed by XGBoost and Decision Tree. However, further analysis and consideration of other factors, such as computational efficiency and interpretability, may be necessary to make a final decision on the algorithm selection.

### 3.7 Comparison with other works:

The best result we got using Random Forest algorithm with 90% accuracy, and for similar works, the first (EDA + Satisfaction Prediction + NLP + RFM + Deployment) using Random Forest algorithm may get 75% accuracy, and the second (Predicting Customer

Satisfaction using Machine Learning) using Random Forest Algorithm may get you an accuracy of 80%.

The result we got using logistic regression algorithm with 77% accuracy, and for similar works, the first (EDA + Satisfaction Prediction + NLP + RFM + Deployment) using logistic regression algorithm may get 72% accuracy, and the second (Predicting Customer Satisfaction using Machine Learning) using logistic regression algorithm may get you an accuracy of 80%.

The result we got using Decision Tree algorithm with 84% accuracy, and for similar works, the first (EDA + Satisfaction Prediction + NLP + RFM + Deployment) using Decision Tree algorithm may get 73% accuracy, and the second (Predicting Customer Satisfaction using Machine Learning) using Decision Tree algorithm may get you an accuracy of 80%.

The result we got using AdaBoost algorithm with 77% accuracy, and for similar works, the first (EDA + Satisfaction Prediction + NLP + RFM + Deployment) using AdaBoost algorithm may get 71% accuracy, and the second (Predicting Customer Satisfaction using Machine Learning) using AdaBoost algorithm may get you an accuracy of 79%.

### Conclusion:

this chapter sets the foundation for the project, presenting the problem, software environment, dataset, and implementation details. It provides an overview of the project objectives, the tools and libraries used, the dataset characteristics, and the design and implementation steps taken, The results are presented for four different algorithms: logistic regression, decision tree, random forest, and AdaBoost. Each algorithm is evaluated based on its accuracy on both the training and testing data sets. the given results provide an assessment of the performance of different machine learning algorithms, the random forest algorithm achieved the highest testing accuracy (0.90), followed by the decision tree algorithm (0.84). The logistic regression and AdaBoost algorithms achieved lower accuracies, with the logistic regression algorithm performing slightly better on the testing data.





### **General conclusion:**

Our work aimed to predict customer satisfaction in the e-commerce sector using machine learning techniques. By analyzing factors such as order details, product characteristics, payment methods, and customer demographics, insights were gained into the factors influencing customer satisfaction. Exploratory data analysis revealed that payment value alone may not be sufficient for accurate classification, as there was significant overlap between positive and negative customer reviews. However, payment methods, shipping duration, and customer locations were found to have a notable impact on customer satisfaction.

A predictive model was developed using machine learning algorithms such as logistic regression, decision trees, and random forests, enabling the forecasting of customer satisfaction levels. The model's performance was evaluated using appropriate metrics, and practical implications were drawn from the findings. The results emphasize the importance of optimizing payment processes, reducing shipping durations, and personalizing customer experiences to enhance customer satisfaction in the e-commerce sector. The study highlights the significance of considering diverse factors and leveraging machine learning techniques for accurate customer satisfaction prediction. Overall, this research enhances the understanding of customer satisfaction in e-commerce platforms and provides valuable insights for businesses to enhance customer experiences, foster growth, and maintain competitiveness. Future research directions may involve exploring additional factors, refining models, and employing advanced techniques to further improve the accuracy and effectiveness of customer satisfaction prediction in the e-commerce domain.

## List of references:

- [1] F. Buttle, *Customer Relationship Management: Concepts and Tools*. Elsevier Butterworth-Heinemann, 2004.
- [2] K. MUKERJEE, *CUSTOMER RELATIONSHIP MANAGEMENT: A STRATEGIC APPROACH TO MARKETING*. PHI Learning Pvt. Ltd., 2007.
- [3] « Developing superior value propositions: a strategic marketing imperative | Emerald Insight ». <https://www.emerald.com/insight/content/doi/10.1108/JOSM-01-2014-0036/full/html> (consulté le 4 mars 2023).
- [4] V. Kumar et W. Reinartz, *Customer Relationship Management: Concept, Strategy, and Tools*. in Springer Texts in Business and Economics. Berlin, Heidelberg: Springer, 2012. doi: 10.1007/978-3-642-20110-3.
- [5] A. Bhadoria, « Succeed in SaaS by understanding the types of customers and customer segments! », *SmartKarrot / Comprehensive Customer Success*, 23 juin 2020. <https://www.smartkarrot.com/resources/blog/succeed-in-saas-with-customers-and-customer-segments/> (consulté le 20 mai 2023).
- [6] « Customer Segmentation: How to Effectively Segment Users & Clients ». <https://blog.hubspot.com/service/customer-segmentation> (consulté le 5 mars 2023).
- [7] infocleanse, « CRM Data Management - A Full Fledged Infographic », *InfoCleanse*, 23 septembre 2020. <https://www.infocleanse.com/crm-data-management/> (consulté le 5 mars 2023).
- [8] « Sales force automation (SFA) defined ». <https://www.oracle.com/cx/sales/sales-force-automation/what-is-sfa/> (consulté le 5 mars 2023).
- [9] « How Marketing Automation can create value in the customer journey ». <https://www.constructdigital.com/insights/how-marketing-automation-can-impact-create-value-in-the-customer-journey> (consulté le 20 mai 2023).
- [10] « What Is Marketing Automation? How Does It Work? », *Salesforce*. <https://www.salesforce.com/products/marketing-cloud/what-is-marketing-automation/> (consulté le 5 mars 2023).
- [11] N. Puccinelli, R. Goodstein, D. Grewal, R. Price, P. Raghubir, et D. Stewart, « Customer Experience Management in Retailing: Understanding the Buying Process », *Journal of Retailing*, vol. 85, p. 15- 30, mars 2009, doi: 10.1016/j.jretai.2008.11.003.
- [12] « Pin on CRM », *Pinterest*. <https://www.pinterest.com/pin/814729388824418257/> (consulté le 20 mai 2023).
- [13] « OperationalCRM:Definition, Importance, Examples », *Zendesk*. <https://www.zendesk.com/sell/crm/operational-crm/> (consulté le 31 mars 2023).
- [14] R. Iriana et F. Buttle, « Strategic, Operational, and Analytical Customer Relationship Management: Attributes and Measures », *Journal of Relationship Marketing*, vol. 5, n° 4, p. 23- 42, avr. 2007, doi: 10.1300/J366v05n04\_03.
- [15] « What Is Collaborative CRM? Definition, Uses & Tools », *CRM.org*, 21 septembre 2018. <https://crm.org/crmland/collaborative-crm> (consulté le 31 mars 2023).
- [16] « Important CRM information on boosting sales - yourinfobucket.com », 23 octobre 2020. <https://yourinfobucket.com/important-crm-information-on-boosting-sales/> (consulté le 21 mai 2023).

- [17] S. Rabbi, « Principles of Marketing [Philip Kotler, Gary Armstrong, John Saunders, (@Shahriar Anum, Academia.edu ) », Consulté le: 8 avril 2023. [En ligne]. Disponible sur:  
[https://www.academia.edu/43372724/Principles\\_of\\_Marketing\\_Philip\\_Kotler\\_Gary\\_Armstrong\\_John\\_Saunders\\_at\\_Shahriar\\_Anun\\_Academia\\_edu\\_](https://www.academia.edu/43372724/Principles_of_Marketing_Philip_Kotler_Gary_Armstrong_John_Saunders_at_Shahriar_Anun_Academia_edu_)
- [18] W. J. Reinartz et V. Kumar, « The Mismanagement of Customer Loyalty », *Harvard Business Review*, 1 juillet 2002. Consulté le: 8 avril 2023. [En ligne]. Disponible sur:  
<https://hbr.org/2002/07/the-mismanagement-of-customer-loyalty>
- [19] A. Payne et P. Frow, « Strategic Customer Management: Integrating Relationship Marketing and CRM », *Cambridge Core*, mars 2013.  
<https://www.cambridge.org/core/books/strategic-customer-management/7C4D36EC74B457D8A8B43DB31229B9F0> (consulté le 8 avril 2023).
- [20] R. L. Oliver, *Satisfaction: A Behavioral Perspective on the Consumer: A Behavioral Perspective on the Consumer*, 2<sup>éd</sup>. New York: Routledge, 2014. doi: 10.4324/9781315700892.
- [21] S. Shafiq, « 6 Factors that Influence Customer Experience », *High Stuff*, 27 mai 2019.  
<https://www.highstuff.com/influence-customer-experience/> (consulté le 13 avril 2023).
- [22] A. Gustafsson, M. D. Johnson, et I. Roos, « The Effects of Customer Satisfaction, Relationship Commitment Dimensions, and Triggers on Customer Retention », *Journal of Marketing*, vol. 69, n<sup>o</sup> 4, p. 210- 218, oct. 2005, doi: 10.1509/jmkg.2005.69.4.210.
- [23] Melanie, « How Your Pricing Strategy Affects Customer Satisfaction », *Unleashed Software*, 13 octobre 2017. <https://www.unleashedsoftware.com/blog/pricing-strategy-affects-customer-satisfaction> (consulté le 13 avril 2023).
- [24] M. H. N. Thi, « Marketing 4.0\_ Moving from Trad - Philip Kotler.pdf », Consulté le: 14 avril 2023. [En ligne]. Disponible sur:  
[https://www.academia.edu/37219827/Marketing\\_4\\_0\\_Moving\\_from\\_Trad\\_Philip\\_Kotler\\_pdf](https://www.academia.edu/37219827/Marketing_4_0_Moving_from_Trad_Philip_Kotler_pdf)
- [25] E. Peelen et R. Beltman, *Customer Relationship Management*. Pearson Education, 2013.
- [26] B. Hernández, J. Jiménez, et M. J. Martín, « Key website factors in e-business strategy », *International Journal of Information Management*, vol. 29, n<sup>o</sup> 5, p. 362- 371, oct. 2009, doi: 10.1016/j.ijinfomgt.2008.12.006.
- [27] S. Greengard, « What is Artificial Intelligence & How Does It Work? », *Datamation*, 24 mai 2019. <https://www.datamation.com/artificial-intelligence/what-is-artificial-intelligence/> (consulté le 21 mai 2023).
- [28] Cigref, « Livre Blanc CIGREF « Gouvernance de l'Intelligence Artificielle dans les entreprises » », *Cigref*, 28 septembre 2016. <https://www.cigref.fr/livre-blanc-cigref-gouvernance-de-l-intelligence-artificielle-dans-les-entreprises> (consulté le 19 mai 2023).
- [29] D. H. Adenane, « Etude et comparaison de modèles de prédiction basés sur l'apprentissage automatique ».
- [30] « Building the Machine Learning Infrastructure | 7wData ». <https://7wdata.be/big-data/building-the-machine-learning-infrastructure/> (consulté le 21 mai 2023).
- [31] « Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). MIT Press. Chapter 1. »
- [32] E. Viennet, « Apprentissage et Fouille de Données Visuelles », *machine learning*.

- [33] « Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* (1st ed.). Cambridge, MA: MIT Press. »
- [34] « L'apprentissage supervisé : définition et exemples - Blent.ai ». <https://blent.ai/apprentissage-supervise-definition/> (consulté le 26 février 2023).
- [35] « Biochar and Application of Machine Learning: A Review | IntechOpen ». <https://www.intechopen.com/chapters/84407> (consulté le 26 février 2023).
- [36] « Machine Learning: C++ Logistic Regression Example | by Russsun | Medium ». <https://medium.com/@dr.sunhongyu/machine-learning-c-logistic-regression-example-1bec1168db97> (consulté le 21 mai 2023).
- [37] « Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc. »
- [38] « Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann. »
- [39] « Random Forest Algorithm In Trading Using Python », *Quantitative Finance & Algo Trading Blog by QuantInsti*, 12 mars 2019. <https://blog.quantinsti.com/random-forest-algorithm-in-python/> (consulté le 21 mai 2023).
- [40] « RandomForests », *DeepAI*, 10 septembre 2020. <https://deepai.org/machine-learning-glossary-and-terms/random-forest> (consulté le 15 mai 2023).
- [41] « <https://doi.org/10.1016/j.neuron.2016.06.019><https://doi.org/10.1016/j.neuron.2016.06.019> ».
- [42] « Qu'est ce que le machine learning ? Définition, Exemples ». <https://superdatacamp.co/le-machine-learning/>. (consulté le 24 février 2023).
- [43] D. Sarvaiya, « How Machine Learning and AI Empower CRM », *Datafloq*, 22 août 2022. <https://datafloq.com/read/machine-learning-ai-empower-crm/> (consulté le 5 juin 2023).
- [44] « AI in CRM: How to Enhance Customer Relationships », *Keenethics*. <https://keenethics.com/blog/enhancing-customer-relationships-with-ai-and-crm> (consulté le 5 juin 2023).
- [45] « (PDF) Artificial intelligence in customer relationship management: literature review and future research directions ». [https://www.researchgate.net/publication/359813862\\_Artificial\\_intelligence\\_in\\_customer\\_relationship\\_management\\_literature\\_review\\_and\\_future\\_research\\_directions](https://www.researchgate.net/publication/359813862_Artificial_intelligence_in_customer_relationship_management_literature_review_and_future_research_directions) (consulté le 5 juin 2023).
- [46] C. Utterback, « Using Data Science to Predict Negative Customer Reviews », *Medium*, 23 décembre 2022. <https://towardsdatascience.com/using-data-science-to-predict-negative-customer-reviews-2abdbfb3d82> (consulté le 30 mai 2023).
- [47] « EDA+SatisfactionPrediction+NLP+RFM+Deployment ». <https://kaggle.com/code/mohabovic/eda-satisfaction-prediction-nlp-rfm-deployment> (consulté le 4 juin 2023).
- [48] « Customer Satisfaction Prediction Using Machine Learning | by ParitoshMahto | Towards Data Science ». <https://towardsdatascience.com/customer-satisfaction-prediction-using-machine-learning-66240e032962> (consulté le 4 juin 2023).

- [49] « Google Colaboratory ». <https://colab.research.google.com/notebooks/intro.ipynb#scrollTo=P-H6Lw1vyNNd> (consulté le 14 mai 2023).
- [50] « Brazilian E-Commerce Public Dataset by Olist ». <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce> (consulté le 16 mai 2023).
- [51] « Global preferred online payment methods 2017 », *Statista*. <https://www.statista.com/statistics/508988/preferred-payment-methods-of-online-shoppers-worldwide/> (consulté le 2 juin 2023).
- [52] K. V. Krishna et V. Shanmugam, « Usage of E-Payment and Customer Satisfaction », juin 2020.
- [53] T. E. Mofokeng, « The impact of online shopping attributes on customer satisfaction and loyalty: Moderating effects of e-commerce experience », *Cogent Business & Management*, vol. 8, n<sup>o</sup> 1, p. 1968206, janv. 2021, doi: 10.1080/23311975.2021.1968206.
- [54] A. O. Dünder et R. Öztürk, « The Effect Of On-Time Delivery On Customer Satisfaction And Loyalty In Channel Integration », *Business And Management Studies An International Journal*, vol. 8, p. 2675- 2693, sept. 2020, doi: 10.15295/bmij.v8i3.1520.
- [55] J. Romero, « The Importance of Product Photos in ecommerce | Sales Layer ». <https://blog.saleslayer.com/product-photos-in-ecommerce> (consulté le 2 juin 2023).
- [56] G. Louppe, « Understanding Random Forests: From Theory to Practice ». arXiv, 3 juin 2015. Consulté le: 5 juin 2023. [En ligne]. Disponible sur: <http://arxiv.org/abs/1407.7502>
- [57] Pumalin, « DecisionTrees for Classification — Complete Example », *Medium*, 5 février 2023. <https://towardsdatascience.com/decision-trees-for-classification-complete-example-d0bc17fcf1c2> (consulté le 5 juin 2023).
- [58] « ML | Why Logistic Regression in Classification ? », *GeeksforGeeks*, 3 mai 2019. <https://www.geeksforgeeks.org/ml-why-logistic-regression-in-classification/> (consulté le 5 juin 2023).
- [59] S. Patel, « Chapter6: Adaboost Classifier », *Machine Learning 101*, 3 juin 2017. <https://medium.com/machine-learning-101/https-medium-com-savanpatel-chapter-6-adaboost-classifier-b945f330af06> (consulté le 5 juin 2023).