Thesis submitted to the

UNIVERSITY OF MOHAMED BOUDIAF – MSILA

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

In partial fulfillment of the requirements for the degree of

Master in **Computer Science**

Specialty: Information Systems and Software Engineering **(ISSE)**

By

Djeblahi Fatima Ez Zahra

Metarfi Khawla

Title of the thesis

Development of an identification system for customer journey in a mobile company (Case study: Mobilis).

Under the supervision of

Tahar Mehenni

Composition of the jury

Rabah Mokhtari	University of Msila	President
Tahar Mehenni	University of Msila	Reporter
Nour El Houda Chalabi	University of Msila	Examiner

JUNE, 2024

DEDICATIONS

"I dedicate this memory to my parents for their love, patience, and moral support throughout my studies. Your trust in me has been my greatest source of motivation.

To my friends, for their presence and encouragement during difficult times.

To my Supervisor, ***Dr.Taher Mehenni*** *for his rigorous guidance and valuable advice.*

To my dear teachers, without exception.

To my partner ***Khawla Metrefi,*** *for her moral support, patience, and understanding throughout this project.*

Finally, I dedicate it to all those who know me."

Djeblahi Fatima Ez Zahra

أهدي عملي هذا لأمي على كل الحب على السهر و التشجيع و التصديق بأني سأكون يوما شخصا عظيما ، أهديه إلى أبي على كد يده على تعبه و كونه دائما صديقي حبيبي و أبي . أهديه لأخوتي و أخواتي على كونهم السند الذي لا يميل على صحبتهم و على كل الحب الذي غمروني به، أهديه إلى بلوطتي على كل خطوة كانت فيها معي على كل الحب الذي كانت تظهره في تشجيعي و دعمي و الاهتمام بي ، اهديه إلى كل فرد في عائلتي و كل صديقة من صديقاتي و أيضا إلى صديقتي في المشروع و المشوار من المرحلة الإبتدائية.

خولة مطرفي

ACKNOWLEDGMENTS

Above all,

I thank ALLAH, who gave us the strength, courage, and hope to do this modest work.

It was made possible only thanks to the informed guidance of our supervisor, Dr. Mehenni Tahar. We wish to express our perfect gratitude and sincere thanks for the quality of his supervision and his wise advice.

We also thank the jury members for making the pleasure to agree to review this work.

We dedicate this memory to all those who contributed to our training and supported us in our studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GENERAL INTRODUCTION

The continual evolution of mobile phone companies has resulted in a substantial reinterpretation of the relationship between customers and these companies. The customer journey is a consumer's path from discovering a service to developing loyalty, including all interactions with the mobile phone company. The interplay between these factors is essential for a company to become successful in a competitive market with ever-evolving consumer demands.

Digitization has significantly altered the consumer journey in the mobile phone companies sector. Consumers nowadays want streamlined experiences, personalized solutions, and uninterrupted connections. Mobile phone companies must adjust their strategy to suit evolving demands and provide services beyond basic connectivity.

Customers typically start their trip by finding offers and services through various sources, such as advertisements, social media, or recommendations. Mobile firms must invest in intelligent marketing methods to capture consumers' interest and persuade them to select their services.

After obtaining the customer, the activation and installation phase becomes critical. Registration, configuration, and activation steps should be straightforward, intuitive, and quick. An initial positive interaction enhances customer loyalty and builds the basis for a long-term relationship.

During the customer's lifetime, they interact with the mobile carrier several times for account management, customer care, package upgrades, and problem-solving. The quality of these experiences significantly impacts the customer's perception of the company.

Customer loyalty is now a strategic concern. Mobile phone companies spend on providing personalization, rewards programs, and value-added services to sustain client engagement. Retaining loyal customers is as important as gaining new ones since the expenses of keeping a loyal client are typically less than those of getting new consumers.

The connection between the customer's journey and mobile phone companies is essential. Companies that grasp consumer expectations, adjust to technological advancements, and provide outstanding customer experiences can gain a substantial competitive edge in these ever-evolving companies.

Thesis organization: The thesis consists of the following chapters:

- Chapter I presents the customer journey in mobile phone companies.
- Chapter II describes the different customer journey approaches.
- Chapter III presents the design of the model for predicting the customer journey.
- Chapter IV presents the implementation of the model of predicting the customer journey. and a discussion of the results.

CHAPTER I

CUSTOMER JOURNEY IN MOBILE PHONE COMPANIES

1. Introduction

Analyzing the customer's journey has recently become essential for companies to boost their sales. The loss of consumers or clients is a significant issue for mobile phone company providers as customers often unsubscribe or switch operators if their needs are met. Customers desire competitive prices and, above all, high-quality service. Attracting new customers is more costly than retaining current clients. Companies study client behavior to anticipate, adapt, and optimize each interaction with subscribers. [1]

2. What is meant by the term customer journey?

Customer journeys refer to the process, path, or sequence through which a customer accesses or uses a service, focusing on the customer's perspective. This customer-centric perspective involves various touchpoints and activities related to the service's delivery.

Many authors describe the customer journey as "the process of experiencing service through different touchpoints from the customer's point of view", while others define it as "a series of touchpoints involving all activities and events related to the delivery of the service from the customer's perspective". In [2], authors report that "case study companies often referred to a series of touchpoints as the customer journey. The customer journey involves all activities and events related to service delivery from the customer's perspective".

3. Customer journey stages

Each potential customer progresses through multiple stages before becoming a devoted client. Enhancing the client experience at every level increases the likelihood of retaining leads [3]. There are five stages in a customer's journey presented in the Figure 1.1.

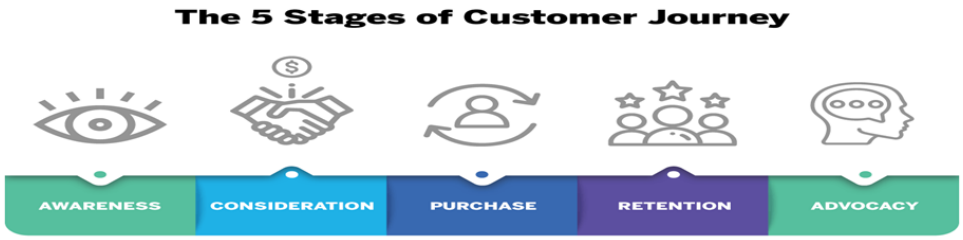


Figure 1.1: Stages of customer journey

3.1 Awareness:

The customer's journey starts before their purchase. They are individuals online searching for a solution, and it is the responsibility of the company to demonstrate that it is the most

suitable to address their issue. Initially, people must be made aware of the existence of the company, making the awareness phase crucial.

Marketers should adopt a multi-channel strategy to increase awareness. Using paid ads, press releases, news stories, social posts, and email campaigns to capture audience attention. It is crucial to understand the target audience to allocate resources effectively to the areas that people are most likely to engage with.

3.2 Consideration

Prospective customers are comparing the product with offerings from competitors during the consideration phase. The duration of this period will vary according to other factors, such as the price of the product. Significant purchases may need weeks or even months of consideration, but smaller purchases can be made in seconds or minutes. At this juncture, providing crucial information is essential to assist in their decision-making process and can be presented in several ways.

3.3 Purchase decision

The company should ensure that the work completed in the initial two stages culminates in the decision phase to prevent losing the customer at this crucial juncture. An intricate purchasing process with numerous steps could lead a user to revert to the consideration phase and choose to buy from other competitors.

Moreover, it must consider providing a discount code or free shipping to enhance the value and assist decision-making. Customer service is vital, so the team company is sure there are convenient ways to contact corporate representatives. This team should be able to view the customer journey history across all channels to avoid needing individuals to catch up.

3.4 Retention

A customer's journey might extend beyond a purchase. According to the classic marketing adage, acquiring a new customer is five times more expensive than keeping an existing one. As mentioned in [4], the previous business model was outdated, but modern predictive analytics methods may now anticipate a customer's lifetime value.

Understanding the lifetime value of both kept and acquired customers, which varies throughout companies, enables the company to make more informed decisions regarding where resources are located. Personalization, maintaining communication with existing customers, and soliciting feedback are powerful strategies for remarketing to those currently in the buyer funnel.

3.5 Advocacy

Word-of-mouth referrals from satisfied consumers are more effective than reselling to existing clients. Advocacy can take various forms, such as writing reviews online or personally recommending your company to family and friends.

Advocacy strategy can incorporate other tactics as well. Encourage referrals by providing exclusive affiliate connections that grant discounts on future purchases or inclusion into a prize drawing for both the giver and the receiver, and sharing feedback on social media and publish services for discussion with potential buyers.

4. Benefits of the customer journey

Understanding the customer journey provides insight into the expectations and requirements of the target audience. Indeed, 80% of companies primarily compete based on customer experience. Optimizing the customer experience will enhance client loyalty and increase competitiveness in attracting new business [4]. More specifically, acknowledging the customer journey can help the company in many sides including (Figure 1.2):

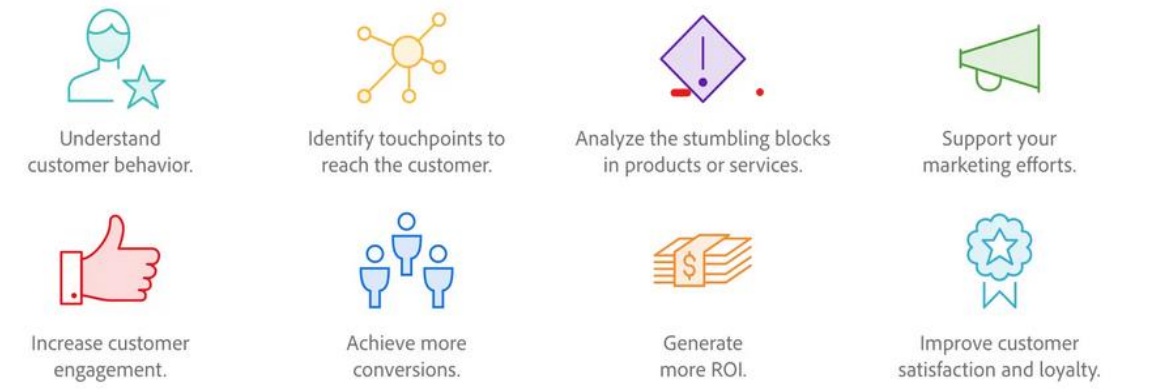


Figure 1.2: Benefits of the customer journey

- **Understand customer behavior:** Understanding each consumer's behavior. We can gain a clearer insight into their underlying motivations. Knowing the rationale behind a buyer's decisions allows us to address their requirements more.
- **Identify touchpoints to reach the customer:** Determine the touchpoints to engage with the customer. Several companies engage in multichannel marketing, although not all touchpoints are beneficial. By concentrating on the client journey, the team company will determine which of these channels is most efficient in producing sales. Businesses can save time and money by concentrating on the most efficient channels.
- **Analyze the stumbling blocks in products or services:** If potential customers often skip the purchasing process, it may indicate issues with your product or the purchase experience. Understanding the customer journey can assist in proactively addressing product or service issues to prevent them from escalating into more costly problems.
- **Support your marketing efforts:** Marketing necessitates a profound understanding of the target audience. Tracking the customer journey facilitates the marketing team in meeting buyers' expectations and addressing their issues.
- **Increase customer engagement:** Examining the customer journey enables companies to focus on the most pertinent audience for their products or services. Additionally, it enhances the customer experience. 30% of customers will reject using branded digital channels if they have a bad experience, making it crucial to improve relationships with clients.

- **Achieve more conversions:** Creating customer journey maps helps enhance conversion rates by personalizing strategies and communications to provide the audience with their desired content
- **Generate more ROI:** We need a concrete return on investment for marketing activities. Investing in the customer journey leads to increased return on investment(ROI) in all areas. Brands that provide a positive customer experience can boost revenue by 2–7%.
- **Improve customer satisfaction and loyalty:** 94% of customers indicate that a positive experience encourages them to make future purchases. Enhancing the customer journey aids in meeting shopper expectations, leading to higher satisfaction and loyalty [4].

5. Concepts relevant to the customer's journey:

5.1 Customer

General term to refer to the purchaser or user of a product or service, and therefore, often the target of a client's activities and the subject of market research. The term consumer usually refers to a product or service's purchaser or end user, mainly when obtained for personal or domestic use.

However, organizations may have customers who do not buy for personal use, such as in business-to-business research. In addition, as more organizations such as educational institutions and government departments, re-define their users as 'customers,' the scope of market research expands, and the more general term becomes more prevalent and accurate [5].

5.2 Customer loyalty

the degree to which customers will likely continue doing business with an organization. Loyalty builds up over multiple interactions and results from customer satisfaction, positive customer experiences, and the value realized from using the organization's goods or services.

Customer loyalty also drives existing customers to select one company's products or services over its competitors, yielding collateral benefits for retention, growth, and brand advocacy [6] [9].

5.3 Customer satisfaction

In broad terms, customer satisfaction is the state of a customer's being pleased with their decision to do business with you. In other words, customer satisfaction is the degree of success your business has at meeting customer expectations. Based on this, you might think customer satisfaction is easily attainable with an excellent product, right? Unsurprisingly, the rabbit hole goes way deeper.

A product can be very good at its job and still fail to satisfy the target customers. This can be due to not meeting customers' specific needs, misleading resources, lacking knowledge of practical use cases, etc [7] [9].

5.4 Customer Experience

Customer experience refers to customers' internal and subjective reactions to any direct or indirect interaction with a company. Direct interaction and firsthand involvement typically happen during buying, using, and receiving service and are usually started by the customer. Indirect contact typically occurs through chance meetings with a company's products, services, or brands through word-of-mouth referrals, advertising, reports, or reviews.

Customer Experience is defined as holistic, based on personal interactions between a customer and a brand, service provider, or product that is composed of various cognitive, affective, emotional, social, and physical responses of this customer [8].

Authors in [9] explain more concretely that customer experience encompasses every aspect of a company's offering, for example, advertising, packaging, features of products and services, ease of use, reliability, or the quality of customer care.

6. Customer Journey Analysis (CJA)

6.1 What is Customer Journey Analysis?:

Customer Journey Analysis (CJA) is analyzing how customers interact with a company at each touchpoint in the customer journey, including the different channels and touchpoints of the customer's interaction with a brand.

In other words, CJA is a data-driven approach that provides in-depth insight into how customer experiences effectively Influence a company's revenue growth. With CJA, the company can measure and optimize the effectiveness of the customer experience [11] [12].

6.2 Why is customer journey analysis critical?

Customer journey analysis should consider the channels customers use to communicate and how they use them, along with defining channels based on customer type and interaction mode.

CJA usually possesses a sophisticated analytics toolbox that includes real-time analytics, client segmentation, and predictive analytics. Incorporating the CJA into program company aims to offer practical insights that can directly influence results.

CJA involves examining and comprehending the interactions between a customer and a brand, from the first awareness to the evaluation after a purchase. The main focus is monitoring a client's route to achieve a specific goal, like completing a purchase or submitting a form [9].

6.3 Why do companies use CJA?

Companies use CJA for several benefits including [12] [13]:

- Understanding clients' objectives and determining what brings them satisfaction or dissatisfaction.
- Identifying discrepancies between a business's perceived delivery and a customer's experience.
- Enhancing consumer relations with the company.

- Obtaining insights into the series of events that result in excellent or unsatisfactory consequences for a consumer.
- Minimizing customer grievances and decrease turnover rates.
- Defining successful performance.
- Enhancing performance by prioritizing tasks that align with customers' highest values and ensuring a smooth and effortless transition between every process stage.
- Reducing expenses by production times and time required to bring a product to market.

7. Customer Journey Analysis in Mobile Phone Companies

7.1 Importance of Customer Analytics In Telecoms

Clients in the current digital age are sometimes overwhelmed by the influx of new services and packages. This is exceptionally accurate in the competitive mobile phone sectors. Customers can choose from many mobile phone operators and instantly get all the necessary information, allowing them to compare pricing and packages. Customers are increasingly price-conscious and discerning product value, leading them to explore offerings from multiple mobile phone operators.

What distinguishes one service provider from another?. The solution is for mobile phone companies to fulfill consumer requirements, offer competitive pricing, and deliver an exceptional customer experience. While this may appear standard information, only a few mobile phone companies can efficiently do it. Customer analysis is essential at that point.

Conducting customer research for mobile phone companies may provide them with a precise understanding of their client patterns by integrating data from several sources. This pertains to how clients utilize mobile, digital, and social services, and thus, conducting to apply this information to make well-informed business decisions on new services and products, as well as depending on the specific behavior of the client.

With this vital information, companies may customize services for various consumer categories and efficiently reach these customers through marketing efforts. This information can serve consumers' requirements, interacts with them, and enhances the customer experience simultaneously [14].

7.2 Using Customer Analysis for mobile phone Companies

Based on the information provided, it has been concluded that a company mobile can enhance customer satisfaction and increase interaction and overall satisfaction through customer analysis. Undoubtedly, this valuable data may greatly aid the company in generating products and services its clients desire.

Satisfied customers are less likely to switch to other service providers, leading to a decrease in turnover rates. It also restricts customers from using services offered by other companies. Utilizing customer analytics may be a significant asset for the company to optimize its operations, boost customer happiness, and stimulate business growth. Mobile phone companies may utilize customer analytics in several ways [14]. We can list the most important in the following:

- ✓ **Customer Segmentation:**
  - Analyze data to classify customers according to their usage habits, spending, revenue, goods, services, demographics, and behavior.
  - Subsequently, tailored marketing plans and promotions may be developed for each category to optimize their impact.
- ✓ **Churn Prediction and Prevention:**
  - By analyzing historical data, the company can identify patterns that precede customer churn.
  - The company can then use predictive modeling to forecast which customers are at a high risk of leaving and monitor the revenue implications of these subscribers' churning.
  - The company can implement targeted retention strategies to reduce churns, such as personalized offers and proactive customer support.
- ✓ **Customer Journey Analysis:**
  - Customer analytics enable a company to track the customer journey, including beginning, continuing usage, and possible churn points.
  - This will assist in detecting points of dissatisfaction and possibilities for improvement in the customer experience.
  - The company can optimize touchpoints to improve client satisfaction and loyalty.
- ✓ **Product and Service Recommendations:**
  - By analyzing client usage data, the company can comprehend their preferences and behavior.
  - The company can offer suggestions for extra services, enhancements, or characteristics that match each customer's requirements and use patterns.
- ✓ **Network Optimization:**
  - By analyzing network performance data, the company can pinpoint regions with high use or frequent troubles by analyzing network performance data.
  - The information may be utilized to optimize network architecture, distribute resources more efficiently, and enhance overall service quality.
- ✓ **Customer Feedback Analysis:**
  - The company can employ analytics to identify unexpected patterns of usage that may suggest fraudulent behavior.
  - They may then use real-time monitoring and service orchestration to detect and prevent fraudulent activities on the network, securing the company and its clients.
- ✓ **Dynamic Pricing Strategies:**
  - The company can employ analytics to establish flexible pricing strategies according to demand, client actions, and market dynamics.
  - The revised pricing models will enhance income generation for the company is while ensuring competitiveness in the market.

- ✓ **Cross-selling and Upselling:**
  - The company can uncover cross-selling and upselling possibilities using customer data.
  - This allows for developing focused promotions and campaigns to motivate clients to embrace extra services or enhance their current plans.

8. Conclusion

Using customer journey analysis effectively helps Mobilis clients. Customer analysis is important to Mobilis company as it facilitates educated decision-making to boost revenue, decrease customer churn, and improve customer experience.

Mobilis can enhance customer satisfaction by using consumer research to customize services to individual needs and address any customer concerns effectively. Developing a loyal client base is crucial for success in this competitive business. The next chapter will describe the different customer journey approaches in the literature.

CHAPTER II:

CUSTOMER JOURNEY ANALYSIS APPROACHES

1. Introduction

In today's information world, a vast quantity of data is readily accessible across various businesses and organizations. The utility of this extensive dataset is contingent upon its conversion into meaningful and valuable information. Alternatively, we are inundated with data yet lacking in knowledge. Data mining is a viable approach for addressing this issue, as it involves the extraction of valuable insights from vast quantities of available data [17].

2. What is data mining?

- Def 1: « Data mining is the set of methods and techniques for the exploration and analysis of computer databases (often large), automatically or semi-automatically, to detect rules, associations, unknown or hidden trends, specific structures that render most of the useful information while reducing the amount of data.» [18]
- Def 2: «Data mining is a non-trivial process that consists in identifying, in said data, new, valid, potentially useful and above all understandable and usable schemes.» [18]. (see Figure 2.1)

Figure 2.1: What is Data Mining

3. The importance of data mining

The importance of data mining lies in its ability to extract valuable knowledge and information from vast quantities of data. This data can assist companies in making well-informed decisions and enhancing their overall performance. The application of data mining techniques enables the identification of concealed trends and patterns, the anticipation of future events, the assessment of risks, and the enhancement of product and service quality.

Data mining can be employed in the battle against fraud to discover dubious behavior and ascertain possible perpetrators. In essence, data mining has the potential to enhance firms' decision-making processes, leading to enhanced consumer happiness and heightened profitability [19].

4. Historical context of data mining

4.1 Etymology

The initial terminology related to data mining emerged throughout the 1960s. Statisticians employ terminology such as "Data Fishing" to denote a data analysis method deemed unsound due to the absence of assumptions. Data mining emerged during the 1990s.

The phrase "KNOWLEDGE Discovery in Databases" was coined by Gregory Piatesky-Shapiro and has gained significant traction in community learning. The phrase "Data Mining" was introduced in 1991 and has since gained considerable usage within the business and journalism communities. Currently, the concepts of Data Mining and Knowledge Discovery are employed interchangeably [20].

4.2 Context

Data extraction has existed for centuries, with past efforts focused on developing methods for identifying data. The proliferation of new technologies and their growing capabilities have significantly expanded data gathering, management, and storage capacity. The magnitude and intricacy of the data have experienced significant growth, which can be attributed to advancements in methodologies such as neural networks and decision trees.

Data mining is the application of various technologies to uncover concealed patterns. The field of Data ming effectively facilitates the integration of applied statistics with artificial intelligence database management. The employed methodologies are of a limited origin. The Galton linear regression method was introduced in the year 1875.

Concurrently, several methodologies and approaches are being developed, including factorial analysis introduced by Guttman in 1941, neural networks developed by Mac Culloch and Pitts in 1943, and decision trees. In the year 1984, the utilization of these methodologies facilitated the exploitation and exploration of progressively more accurate models.

4.3 Today

In today's world, data mining has emerged as a crucial instrument inside marketing teams and plays a pivotal role in companies' decision-making procedures. Data mining compiles a collection of statistical methodologies that ought to be employed by descriptive or decision-making concerns.

Data mining typically involves systematically prioritizing ideas, models, and actions. Furthermore, there has been a notable improvement in storage and computational capacity efficiency over time.

5. Objectives of data mining

Data mining has many objectives. Companies can use data mining to [21]:

- **Detect hidden patterns:** Data mining facilitates the identification of concealed patterns, trends, and correlations within data that may not be readily apparent.
- **Predict future results:** Data mining enables predicting future outcomes by analyzing past data. For instance, a retail enterprise can employ data mining techniques to forecast future sales by analyzing historical sales data.
- **Improve decision-making:** Data mining enables decision-makers to enhance decision-making by offering precise and dependable data insights.
- **Optimize processes:** Data mining techniques can facilitate the identification of bottlenecks, inefficiencies, and potential areas for enhancement within corporate processes.
- **Improve the quality of products and services:** Data mining may effectively detect and address product and service quality concerns by analyzing customer satisfaction and usage data.
- **Detect fraud:** Data mining enables the identification of dubious behavior and fraudulent trends in financial transactions.

Data mining aims to get essential insights and facts from vast data, enabling firms to make well-informed decisions, enhance their performance, and maintain competitiveness in a dynamic business landscape.

6. Data mining tasks:

Various intellectual, economic, or commercial issues can be categorized into one of the following tasks based on their formalization **[22] [23]**:

- **Classification:** Classification refers to analyzing the attributes of a recently introduced element to categorize it into a specific class within a predetermined set.
- **Estimation:** In contrast to classification, an estimate yields a continuous variable. This is achieved by combining input data using one or more functions. The outcome of an estimation enables the progression of classification using a scale.
- **Prediction:** Prediction exhibits similarities to classification and estimation while operating on a distinct temporal dimension. Similar to the preceding tasks, this activity is grounded in historical and current contexts; however, its outcome is typically oriented towards a broadly defined future.
- **Segmentation:** Segmentation involves segmenting a diverse population into distinct groups that exhibit similar characteristics. Contrary to the classification, sub-populations still need to be pre-established.
- **Description:** Implementing this task is frequently one of the initial requirements for a data mining tool. The individual is requested to describe the data contained within an intricate database, which often results in further exploitation to offer explanations.
- **Optimization:** It is customary to incorporate an evaluation function into each proposed solution to address numerous problems. In optimization, the objective is to maximize or

minimize the given function. Sure, experts argue that this particular issue does not fall under the purview of Data Mining.

7. Data mining process:

Data mining is often synonymous with knowledge discovery from data (KDD), while some individuals perceive it as a crucial component of the discovery process. Figure 2.2 illustrates the iterative actions involved in the knowledge discovery process [24].

1. **Data cleaning:** To eliminate noise and inconsistent data.
2. **Data integration:** where it is possible to merge multiple data sources.
3. **Data selection:** retrieving data pertinent to the analysis task from the database.
4. **Data transformation:** The process involves transforming and consolidating data into formats suitable for mining by executing summary or aggregate processes.
5. **Data mining:** Data pattern extraction is a crucial step that involves the application of intelligent algorithms.
6. **Pattern evaluation:** Interestingness metrics are necessary to ascertain the genuinely intriguing patterns that constitute knowledge.
7. **Knowledge presentation:** Visualization and knowledge representation techniques are utilized to display acquired knowledge to consumers effectively.

Steps 1 through 4 encompass several methods of data preprocessing, which involve preparing data for mining. The data mining can include user interaction or interaction with a knowledge source. The user is presented with intriguing patterns that can potentially be stored as novel knowledge within the knowledge base.

The perspective above portrays data mining as a crucial component of the knowledge acquisition process, as it reveals concealed patterns for assessment. Nevertheless, within the realms of industry, media, and research, the phrase data mining is frequently employed to encompass the entirety of the process of uncovering knowledge compared to the term "Knowledge Discovery from Data" or KDD.

Hence, we embrace a comprehensive perspective on the usefulness of data mining. Data mining is systematically extracting valuable patterns and insights from vast quantities of data. The data sources encompass a variety of options, including databases, data warehouses, the Internet, other repositories of information, or data that is dynamically fed into the system.

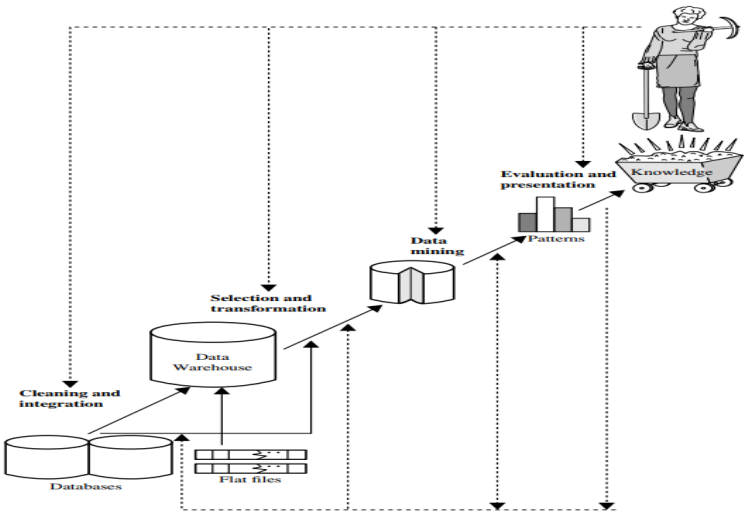


Figure 2.2: The iterative actions involved in The knowledge discovery process

8. Data mining techniques

8.1 Supervised Learning Techniques

Machine learning involves acquiring a function that enables the generation of predictions based on a collection of labeled examples and corresponding prediction values. The tags function as instructors and oversee the acquisition of the algorithm [25]. There are two distinct categories of tasks in the field of supervised learning:

8.1.1 Classification

Classification is a fundamental task in data mining, wherein data is organized into predetermined groups and classes. This approach is commonly referred to as supervised learning. It comprises two sequential stages:

- **Model construction:** involves creating a predefined collection of classes. Each tuple or sample is expected to be a member of a predetermined class. The training set is the collection of tuples used for model development. Classification rules, decision trees, or mathematical formulas are commonly used to express the model. Figure 7 displays this model.
- **Model utilization:** This model categorizes forthcoming or unfamiliar entities. The label of the test sample that is already known is compared to the categorized result obtained by the model.

The accuracy rate refers to the proportion of items in the test set that the model accurately classifies. The test set should be independent of the training set to prevent overfitting. The model depicted is presented in Figures 2.3 and 2.4.

In educational data mining, predicting a student's final grade based on their submitted work is possible. The decision tree depicts the logical principles governing students' final grades.

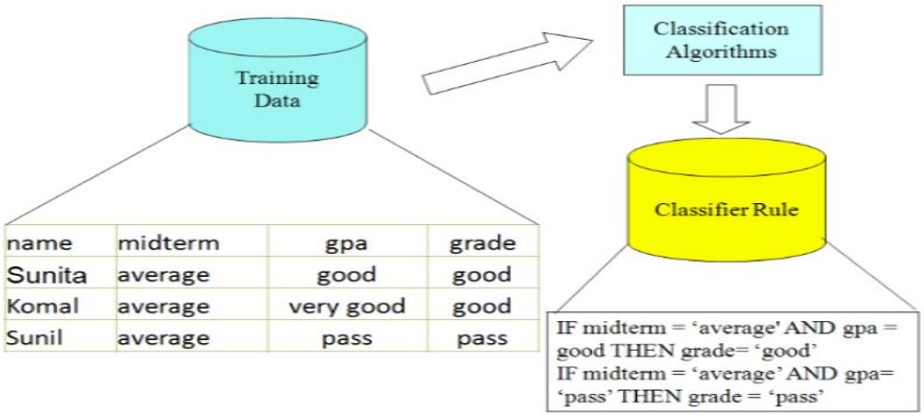


Figure 2.3: Learning step or construction model

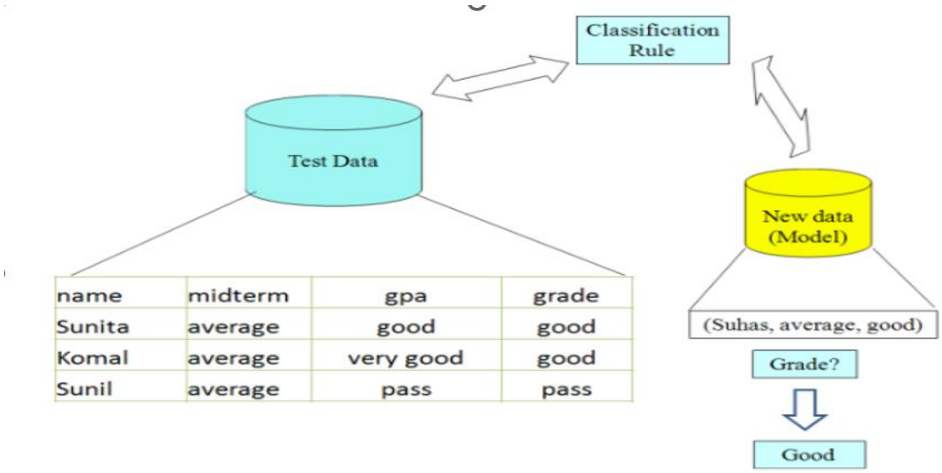


Figure 2.4: Model Usage (Classification)

8.1.2 Regression

Regression is a statistical technique that models data with low error. This statistical approach is commonly employed for numerical prediction. Regression analysis is a commonly employed statistical technique for prediction and forecasting, exhibiting significant convergence with the domain of machine learning.

Regression analysis is a statistical technique employed to ascertain the associations between independent variables and the dependent variable and investigate the nature of these associations. Regression analysis can be employed in limited situations to deduce causal connections between the independent and dependent variables. However, this can result in illusions or erroneous connections. Hence, it is prudent to exercise caution, such as recognizing that correlation does not necessarily indicate causation [26].

8.2 Unsupervised Learning Techniques

Unsupervised learning involves the absence of tags on the data. Next, analyze the observations to have a deeper comprehension of them. There is no need for any examples.

The algorithm is required to autonomously ascertain the structure based on the available data, hence independently determining the appropriate weights [25].

8.2.1 Clustering

Clustering refers to identifying groupings of things to ensure similarity among objects within one group while distinguishing them from other groups. The technique of clustering holds significant importance in the field of unsupervised learning. Figure 2.5 displays the process of clustering and its classification [27].

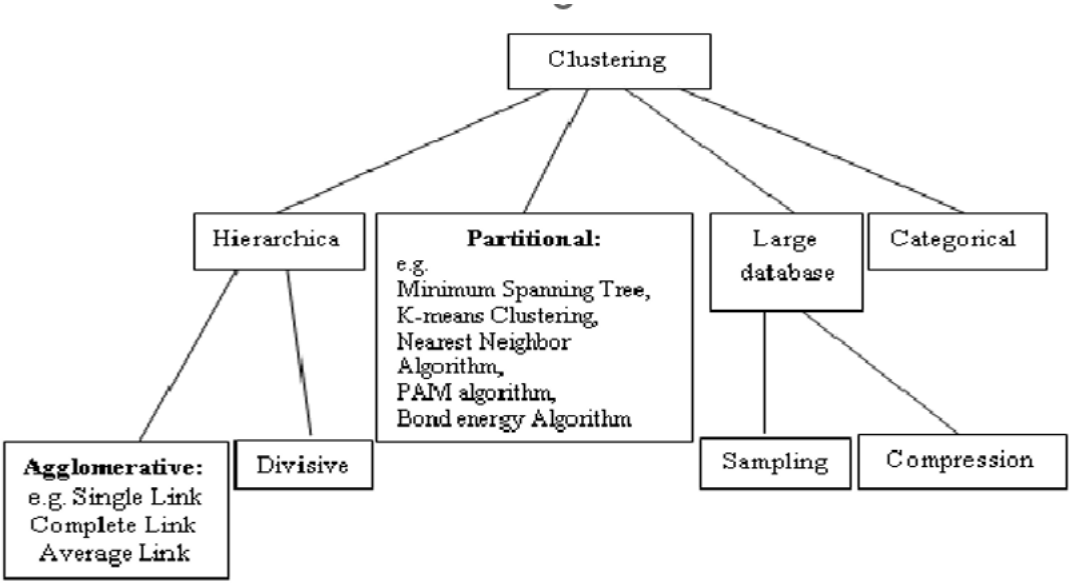


Figure 2.5 : Classification of clustering algorithm

8.2.2 Association Rules

Association rules illustrate the correlation between various data elements. Mining association rules enable the identification of regulations in the following format: If the antecedent is a set of one or more things, then the consequent is likely to be one or more items.

Generating association rules is typically divided into two distinct stages: Initially, minimum support is employed to identify all frequent item sets within a given database. Furthermore, rules are formed using frequent item sets and minimum confidence constraints [27]. Figure 2.6 depicts the creation of itemsets and frequent itemsets, with the minimum support count set to 2.

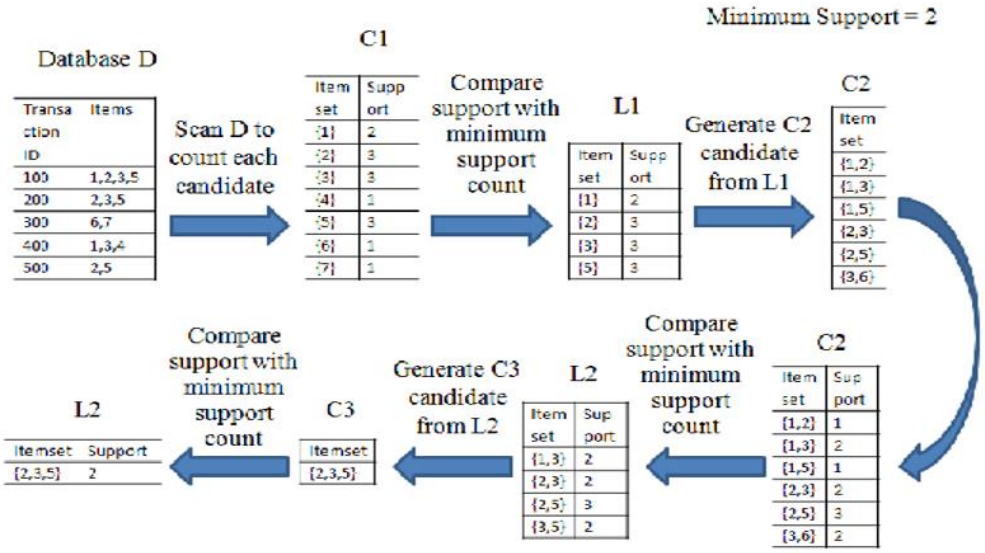


Figure 2.6 : Generation of itemsets and frequent itemsets

The conventional approach to assessing an association rule's quality is through support and confidence. Endorsement of the association rule The variable X->Y represents the proportion of transactions in the database that include XUY. The confidence level for the association rule, denoted as X->Y, is determined by calculating the ratio of transactions containing XUY to transactions containing X (Figure 2.7).

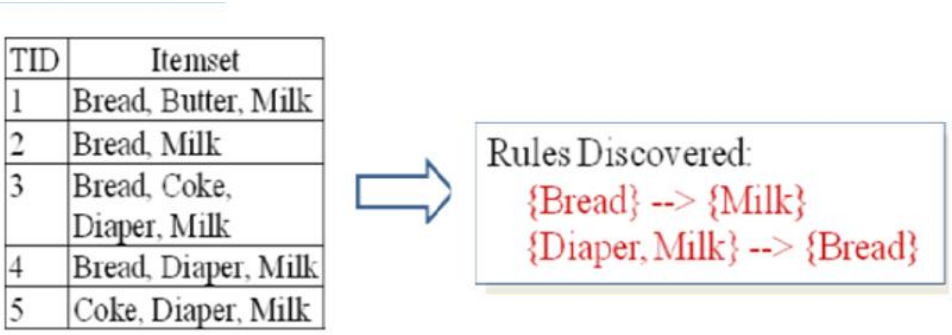


Figure 2.7 : Association Rules

9. Data mining Algorithms

9.1 Classification Algorithms

9.1.1 Decision Tree Learning

- **How It Works:**
  - ✓ **Starting Point:** Begin with the entire dataset as the root node.
  - ✓ **Feature Selection:** Choose the best feature that splits the data into subsets that are more homogenous in terms of the target variable.
  - ✓ **Splitting:** Split the data based on the chosen feature.

- ✓ **Recursive Process:** Repeat steps 2-3 for each subset until a stopping criterion is met (e.g., maximum tree depth, minimum samples per leaf).
- ✓ **Leaf Nodes:** Create leaf nodes where no further splitting is needed and assign a class or regression value to each leaf based on majority voting or mean value, respectively **[28].**

- **Strengths:**
  - ✓ **Interpretability:** Decision trees are easy to interpret and visualize, making them helpful in understanding the decision-making process.
  - ✓ **Handling Non-Linearity:** They can model complex relationships without assuming linearity in the data.
  - ✓ **Feature Importance:** Decision trees can provide insight into feature importance, helping select features.

- **Limitations:**
  - ✓ **Overfitting:** They are prone to overfitting, especially when the tree is deep and complex.
  - ✓ **Sensitive to Data Variations:** Small changes in the data can lead to significantly different trees.
  - ✓ **Bias towards Dominant Classes:** In classification, decision trees tend to favor dominant classes if they are present in the training data.

- **Pseudocode example in python :**

```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
from sklearn.tree import plot_tree
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
# Load dataset
iris = load_iris()
X = iris.data
y = iris.target
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
# Create Decision Tree classifier object
clf = DecisionTreeClassifier()
# Train Decision Tree Classifier
clf = clf.fit(X_train, y_train)
# Predict the response for the test dataset
y_pred = clf.predict(X_test)
# Model Accuracy, how often is the classifier correct?
print("Accuracy:", metrics.accuracy_score(y_test, y_pred))
# Visualize the decision tree
plt.figure(figsize=(20,10))
plot_tree(clf, filled=True, feature_names=iris.feature_names, class_names=iris.target_names)
plt.show()
```

Figure 2.8: Decsion Tree code implementation

9.1.2 K-Nearest Neighbors (KNN):

K-Nearest Neighbors (KNN) is a simple yet powerful machine learning algorithm used for both classification and regression tasks.

- **How It Works**
  - ✓ **Data Preparation:** Collect your dataset with labeled data points. Normalize or standardize the features if necessary.
  - ✓ **Choosing K:** Decide on the number of neighbors (K) to consider. Typically, K is an odd number to avoid ties in voting.
  - ✓ **Calculating Distances:** Compute the distance between the new data point and all other data points in the dataset. Common distance metrics include Euclidean distance, Manhattan distance, etc.
  - ✓ **Finding Neighbors:**
  - ✓ Select the K nearest data points based on the calculated distances.
  - ✓ **Voting:**
  - ✓ For classification, let the K neighbors vote on the class of the new data point (e.g., by majority voting).
  - ✓ For regression, take the average (or weighted average) of the K nearest neighbors' target values.
  - ✓ **Prediction:** Assign the predicted class (classification) or predicted value (regression) to the new data point based on the voting or averaging.

- **Strengths:**
  - ✓ Simplicity: Easy to understand and implement.
  - ✓ Non-Parametric: Does not assume any underlying distribution of data.
  - ✓ Adaptability: Can handle multi-class classification and regression tasks.
  - ✓ Robustness to Noise: Noise in the data does not significantly impact the algorithm's performance.
  - ✓ No Training Phase: KNN is instance-based, so there is no explicit training phase.

- **Limitations:**
  - ✓ Computationally Intensive: Calculating distances for large datasets can be slow.
  - ✓ Memory Usage: Requires storing the entire training dataset in memory.
  - ✓ Sensitive to Feature Scaling: Features with different scales can bias the distance calculation.
  - ✓ Choosing K: The performance of KNN can be sensitive to K's choice.
  - ✓ Curse of Dimensionality: Performance may degrade as the number of dimensions (features) increases due to the increased sparsity of the data.

- **Pseudocode example in python :**

```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.datasets import load_iris
# Load the Iris dataset
iris = load_iris()
X = iris.data
y = iris.target
# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
# Initialize the KNN model with k=3
knn = KNeighborsClassifier(n_neighbors=3)
# Train the model
knn.fit(X_train, y_train)
# Make predictions on the test set
y_pred = knn.predict(X_test)
# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy * 100:.2f}%')
# Generate a confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
print('Confusion Matrix:')
print(conf_matrix)
# Generate a classification report
class_report = classification_report(y_test, y_pred)
print('Classification Report:')
print(class_report)
```

Figure 2.9 :KNN code implementation

9.2 Regression Algorithms

9.2.1 Linear Regression:

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the variables, meaning the change in the dependent variable is proportional to changes in the independent variables.

- **How It Works:**
  - ✓ **Data Collection:** Gather a dataset with observations for the dependent and independent variables.
  - ✓ **Model Creation:** Choose a linear regression model based on the number of independent variables (Simple Linear Regression for one variable, Multiple Linear Regression for more than one).
  - ✓ **Model Training:** Use the data to estimate the coefficients (slope and intercept for simple linear regression) that minimize the difference between the actual and predicted values.
  - ✓ **Model Evaluation:** Evaluate the model's performance using metrics like R-squared, Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).
- **Strengths:**
  - ✓ **Interpretability:** Easy to interpret and understand the relationship between variables.

- ✓ **Computational Efficiency:** Efficient for large datasets and quick predictions once trained.
- ✓ **Versatility:** Can be used for both regression (predicting continuous values) and classification tasks (logistic regression).

- **Limitations:**
  - ✓ **Assumption of Linearity:** Assumes a linear relationship between variables, which might not always hold true.
  - ✓ **Sensitive to Outliers:** Outliers can significantly impact the model's performance.
  - ✓ **Assumption of Independence:** Assumes that the independent variables are independent of each other, which might not always be the case.

- **Pseudocode example in python:**

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
# Generating a synthetic dataset
np.random.seed(0)
X = 2 * np.random.rand(100, 1)
y = 4 + 3 * X + np.random.randn(100, 1)
# Convert to pandas DataFrame
data = pd.DataFrame(data=np.hstack((X, y)), columns=["X", "y"])
# Explore the dataset
print(data.head())
# Visualize the data
plt.scatter(data["X"], data["y"])
plt.xlabel("X")
plt.ylabel("y")
plt.title("Scatter plot of X vs y")
plt.show()
# Split the data into training and testing sets
X = data[["X"]]
y = data["y"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Figure 2.10: Linear Regression code implementation part1

```python
# Create and train the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)
# Make predictions
y_pred_train = model.predict(X_train)
y_pred_test = model.predict(X_test)
# Evaluate the model
mse_train = mean_squared_error(y_train, y_pred_train)
mse_test = mean_squared_error(y_test, y_pred_test)
r2_train = r2_score(y_train, y_pred_train)
r2_test = r2_score(y_test, y_pred_test)
print(f"Training MSE: {mse_train}")
print(f"Testing MSE: {mse_test}")
print(f"Training R²: {r2_train}")
print(f"Testing R²: {r2_test}")
# Visualize the regression line
plt.scatter(X_train, y_train, color="blue", label="Training data")
plt.scatter(X_test, y_test, color="green", label="Testing data")
plt.plot(X_train, y_pred_train, color="red", linewidth=2, label="Regression line")
plt.xlabel("X")
plt.ylabel("y")
plt.title("Linear Regression")
plt.legend()
plt.show()
```

Figure 2.11:Linear Regression code implementation part2

9.2.2 Lasso Regression

Lasso Regression, also known as L1 regularization. It's a type of linear regression that adds a penalty to the absolute values of the coefficients, encouraging simpler and more interpretable models by pushing some coefficients to zero.

- **How It Works:**

  - ✓ **Objective Function:** The Lasso regression minimizes the sum of squared errors between the predicted values and actual values, but adds a penalty term based on the absolute values of the coefficients.
  - ✓ **Penalty Term:** The penalty term is the product of a regularization parameter (λ) and the sum of absolute values of the coefficients. This penalty term is added to the ordinary least squares (OLS) objective function.
  - ✓ **Optimization:** The goal is to find coefficients that minimize the combined loss function of the sum of squared errors and the penalty term. This encourages the model to find a balance between fitting the data well and having simpler coefficients.

- **Strengths:**
  - ✓ **Feature Selection:** Lasso regression can perform feature selection by driving some coefficients to exactly zero, effectively removing those features from the model.
  - ✓ **Prevents Overfitting:** The penalty term helps prevent overfitting by penalizing large coefficients, making the model more generalizable to new data.
  - ✓ **Handles Collinearity:** Lasso can handle multicollinearity by selecting one feature among highly correlated features and setting the others' coefficients to zero.

- **Limitations:**
  - ✓ **Feature Selection Bias:** Lasso tends to arbitrarily select one feature among a group of highly correlated features, which can introduce bias.
  - ✓ **Sensitivity to Scaling:** Lasso is sensitive to the scale of features, so it is important to standardize or normalize the data before applying Lasso regression.
  - ✓ **Difficulty with Large Datasets:** For very large datasets, the computational cost of Lasso regression can be significant.

- **Pseudocode example in python :**

```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import Lasso
from sklearn.metrics import mean_squared_error, r2_score
# Sample dataset creation
np.random.seed(42)
X = np.random.rand(100, 5)
y = X @ np.array([1.5, -2., 3., 0., 0.]) + np.random.randn(100) * 0.5
# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Initialize the Lasso model with a regularization parameter alpha
lasso = Lasso(alpha=0.1)
# Fit the model on the training data
lasso.fit(X_train, y_train)
# Make predictions
y_pred_train = lasso.predict(X_train)
y_pred_test = lasso.predict(X_test)
# Calculate and print the performance metrics
train_mse = mean_squared_error(y_train, y_pred_train)
test_mse = mean_squared_error(y_test, y_pred_test)
train_r2 = r2_score(y_train, y_pred_train)
test_r2 = r2_score(y_test, y_pred_test)
print(f'Training MSE: {train_mse}')
print(f'Testing MSE: {test_mse}')
print(f'Training R^2: {train_r2}')
print(f'Testing R^2: {test_r2}')
# Print the coefficients
print("Lasso coefficients:", lasso.coef_)
```

Figure 2.12: Lasso Regression code implementation

9.3 Clustering Algorithms

9.3.1 K-means clustering

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning data into clusters based on their similarity.

- **How It Works:**
  - ✓ Initialization: Randomly select K cluster centroids.
  - ✓ Assigning data points to clusters:
    - Calculate the distance between each data point and each centroid.
    - Assign each data point to the cluster with the closest centroid.
  - ✓ Updating centroids:
    - Calculate the mean of all data points in each cluster.
    - Update the centroids to the calculated means.
  - ✓ Repeat steps 2 and 3 until convergence (centroids stop changing or a predefined number of iterations is reached).
- **Strengths:**
  - ✓ Simple and easy to implement.
  - ✓ Efficient for large datasets.
  - ✓ Works well when clusters are well-separated and have similar sizes.
- **Limitations:**
  - ✓ Requires the number of clusters (K) to be specified in advance.
  - ✓ Sensitive to initial centroid placement, which can lead to different results for different initializations.

- ✓ Not suitable for clusters with non-spherical shapes or varying sizes.
- ✓ May converge to the local optima instead of the global optimum.

- **Pseudocode example in python :**

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs
# Generate sample data
n_samples = 1500
random_state = 170
X, y = make_blobs(n_samples=n_samples, random_state=random_state)
# Plot the sample data
plt.scatter(X[:, 0], X[:, 1], s=50)
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.title('Sample Data for Clustering')
plt.show()
# Define the number of clusters
n_clusters = 3
# Create a KMeans instance with the desired number of clusters
kmeans = KMeans(n_clusters=n_clusters, random_state=random_state)
# Fit the model to the data
kmeans.fit(X)
# Predict the cluster for each data point
y_kmeans = kmeans.predict(X)
# Plot the clustered data
plt.scatter(X[:, 0], X[:, 1], c=y_kmeans, s=50, cmap='viridis')
centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], c='red', s=200, alpha=0.75, marker='X')
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.title('K-means Clustering')
plt.show()
```

Figure 2.13 :K-means code implementation

9.3.2 DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise, is a popular clustering algorithm used in data mining and machine learning.

- **How It Works:**
  - ✓ **Density-Based Clustering:**
    - DBSCAN groups together points that are closely packed based on a density criterion. It does not require the number of clusters as input.
    - It identifies "core points" (points with a minimum number of neighbors within a specified radius) and expands clusters from these core points.
  - ✓ **Algorithm Steps:**
    - **Core Point Identification:** For each point in the dataset: Determine if the point has at least minPts neighbors (including itself) within a specified radius eps. If it does, mark it as a core point.
    - **Cluster Expansion:** For each core point: Expand the cluster by recursively adding reachable points (points within eps distance) to the cluster.

- **Noise Points:** Points that are not core points and are not reachable from any core point are considered noise points.

- **Strengths:**
  - ✓ **Robust to Noise:**
    - DBSCAN can handle noise points and outliers effectively by classifying them as noise rather than assigning them to clusters.
  - ✓ **Automatic Cluster Detection:**
    - It automatically detects the number of clusters based on the data's density distribution, making it suitable for datasets with irregular cluster shapes and varying densities.
  - ✓ **Doesn't Require Predefined Number of Clusters:**
    - Unlike algorithms like K-means that require the number of clusters as input, DBSCAN determines the clusters based on the data's characteristics.
- **Limitations:**
  - ✓ **Sensitive to Parameters:**
    - DBSCAN's performance can be sensitive to the choice of parameters eps (neighborhood radius) and minPts (minimum number of points in a neighborhood to consider a core point).
    - Choosing appropriate values for these parameters can be challenging, especially for datasets with varying densities.
  - ✓ **Difficulty with Varying Density and High-Dimensional Data:**
    - DBSCAN may struggle with datasets where clusters have varying densities or in high-dimensional spaces due to the curse of dimensionality.

- **Pseudocode example in python :**

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import DBSCAN
from sklearn.datasets import make_blobs
# Generate sample data
X, _ = make_blobs(n_samples=750, centers=[[4, 4], [-2, -1], [1, 1], [10, 10]], cluster_std=0.4, random_state=0)
# Fit DBSCAN
dbscan = DBSCAN(eps=0.3, min_samples=10)
dbscan.fit(X)
# Retrieve labels and core sample indices
labels = dbscan.labels_
core_samples_mask = np.zeros_like(labels, dtype=bool)
core_samples_mask[dbscan.core_sample_indices_] = True
# Number of clusters in labels, ignoring noise if present
n_clusters = len(set(labels)) - (1 if -1 in labels else 0)
unique_labels = set(labels)
colors = [plt.cm.Spectral(each) for each in np.linspace(0, 1, len(unique_labels))]
for k, col in zip(unique_labels, colors):
    if k == -1:
        # Black used for noise
        col = [0, 0, 0, 1]
    class_member_mask = (labels == k)
    xy = X[class_member_mask & core_samples_mask]
    plt.plot(xy[:, 0], xy[:, 1], 'o', markerfacecolor=tuple(col),
             markeredgecolor='k', markersize=14)
    xy = X[class_member_mask & ~core_samples_mask]
    plt.plot(xy[:, 0], xy[:, 1], 'o', markerfacecolor=tuple(col),
             markeredgecolor='k', markersize=6)
plt.title(f'Estimated number of clusters: {n_clusters}')
plt.show()
```

Figure 2.14:DBSCAN code implementation

9.4 Association Rule Algorithms

9.4.1 Apriori Algorithm

The Apriori algorithm is a popular algorithm used for association rule mining in data mining and machine learning. It is particularly useful for finding patterns in large datasets, especially in market basket analysis, where you want to discover relationships between items that frequently co-occur in transactions.

- **How It Works:**
  - ✓ **Generate Candidate Itemsets:**
    - Start by creating a list of all possible itemsets of size 1 (single items) from the dataset.
    - Then, iterate through larger itemset sizes (2, 3, ...) by joining itemsets that share the same (k-1) items.
  - ✓ **Support Counting:**
    - Count the support (frequency of occurrence) of each itemset in the dataset. This involves scanning the dataset and checking which itemsets are present in each transaction.
  - ✓ **Pruning:**
    - Remove itemsets with support below a minimum support threshold. This reduces the search space and eliminates infrequent itemsets.
  - ✓ **Generate Association Rules:**
    - From the frequent itemsets, generate association rules based on a minimum confidence threshold. Association rules are in the form $\{A\} \rightarrow \{B\}$, where A and B are itemsets, indicating that if A occurs, B is likely to occur as well.
- **Strengths:**
  - ✓ **Scalability:** It can handle large datasets efficiently because of its candidate generation and pruning techniques**.**
  - ✓ **Interpretability:** The generated association rules are easy to interpret and understand.
  - ✓ **Versatility:** It can be applied to various domains like market basket analysis, recommendation systems, and more.
- **Limitations:**
  - ✓ **High Memory Usage:** It requires a lot of memory to store candidate itemsets and support counts, especially for large datasets with many unique items.
  - ✓ **Computationally Intensive:** Generating frequent itemsets and association rules can be computationally expensive, especially for datasets with high dimensionality.
  - ✓ **Need for Tuning:** Selecting appropriate support and confidence thresholds requires domain knowledge and tuning, which can be time-consuming.

- **Pseudocode example in python :**

```python
import pandas as pd
from mlxtend.frequent_patterns import apriori, association_rules
# Sample data: list of transactions
transactions = [
    ['milk', 'bread', 'butter'],
    ['beer', 'bread'],
    ['milk', 'bread', 'butter', 'beer'],
    ['bread', 'butter'],
    ['milk', 'bread'],
    ['milk', 'bread', 'butter', 'beer'],
    ['milk', 'bread', 'butter'],
    ['milk', 'bread', 'beer']
]
# Convert transactions to a DataFrame
df = pd.DataFrame(transactions)
# Perform one-hot encoding of the DataFrame
df_encoded = df.stack().str.get_dummies().sum(level=0)
# Apply the Apriori algorithm
frequent_itemsets = apriori(df_encoded, min_support=0.5, use_colnames=True)
# Generate the association rules
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
# Display the frequent itemsets and the rules
print(frequent_itemsets)
print(rules)

```

Figure 2.15 : Apriori Algorithm code implementation

9.4.2 FP-Growth (Frequent Pattern Growth)

FP-Growth (Frequent Pattern Growth) is a popular algorithm used for frequent itemset mining in data mining and association rule learning. It's particularly efficient for finding frequent patterns in large datasets.

- **How It Works:**
  - ✓ Building the FP-Tree:
    - Scan the database to count the frequency of each item.
    - Sort items in decreasing order of frequency.
    - Construct the FP tree by inserting transactions into the tree according to the sorted order of items. Each path from the root to a leaf node represents a transaction, and nodes with the same item are linked together.
  - ✓ Generating Frequent Itemsets:
    - Perform a depth-first traversal of the FP-tree to identify frequent itemsets.
    - Start with the least frequent item, then recursively grow the itemset by considering its conditional pattern base (sub-tree of the FP-tree containing the item's occurrences).
- **Strengths:**
  - ✓ Efficient and scalable due to compact data structure and avoidance of candidate generation.
  - ✓ Requires fewer database scans, leading to better performance.
  - ✓ Flexible and can handle large datasets effectively.

- **Limitations:**
  - ✓ Complex to implement and understand.
  - ✓ High memory consumption, especially with large or dense datasets.
  - ✓ Sensitive to data order and requires significant preprocessing.
  - ✓ Challenging to parallelize effectively.

- **Pseudocode example in python :**

```python
def load_sample_data():
    return [['milk', 'bread', 'beer'],
            ['milk', 'bread'],
            ['milk', 'beer'],
            ['bread', 'beer'],
            ['milk', 'bread', 'beer'],
            ['milk', 'bread']]
def create_init_set(data_set):
    ret_dict = {}
    for trans in data_set:
        ret_dict[frozenset(trans)] = 1
    return ret_dict
# Load the dataset
data_set = load_sample_data()
init_set = create_init_set(data_set)
# Build the FP-Tree
min_support = 2
fp_tree, header_table = create_tree(init_set, min_support)
# Mine the FP-Tree
freq_items = []
mine_tree(fp_tree, header_table, min_support, set([]), freq_items)
# Print the frequent itemsets
print(freq_items)
```

Figure 2.16 :FP-Growth code implementation

10. Data mining application domains

Numerous disciplines employ data mining technologies in order to extract valuable information from immense amounts of data and to provide quick access to it. Data mining technologies have been effectively implemented in numerous sectors, including finance, medicine, marketing, telecommunications, and fraud detection, among others. An excerpt from the application is provided below [29].

10.1 Financial Data Analysis

The banking and financial industry typically possess dependable and superior quality financial data, which enables the implementation of methodical data extraction and analysis. The following are a few typical cases: Design and construction of data repositories for data extraction and multidimensional data analysis. Prediction of loan payments and analysis of customer credit policies.

Customer clustering and classification for the purpose of targeted marketing. Money laundering and additional financial crime detection.

10.2 Retail Industry

The retail industry can benefit greatly from data mining due to the vast quantities of information it gathers regarding sales, consumer purchasing history, products, transportation,

consumption, and services. It is inevitable that the volume of data gathered will further proliferate due to the web's growing accessibility, prominence, and simplicity of use.

Data mining in the retail sector facilitates the identification of trends and purchasing patterns among customers. Enhanced customer retention and satisfaction result from this, as well as enhanced service quality. The following are instances of data mining within the retail sector:

- Construction and design of data warehouses in consideration of the advantages of data mining.
- Examining sales, consumers, products, time, and region from multiple dimensions.
- A critical evaluation of the efficacy of sales campaigns.
- Customer Retention is the objective.
- I provide product recommendations and perform item cross-referencing.

10.3 Telecommunication Industry

Currently, the telecommunication sector is one of the most growing industries, offering a wide range of services including fax, pager, mobile phone, Internet messenger, image and email transmission, web data transmission, and more. The proliferation of emerging computer and communication technologies has fueled the exponential growth of the telecommunications sector. For this reason, data mining has become an indispensable tool for assisting and comprehending businesses.

Data mining plays a pivotal role in the telecommunications industry by facilitating the identification of recurring patterns, detecting fraudulent activities, optimizing resource utilization, and enhancing service quality. The following are some instances in which data mining has enhanced telecommunication services :

- The application of multidimensional analysis to telecommunication data.
- Analysis of fraudulent patterns.
- The process of recognizing atypical patterns.
- Analysis of multidimensional associations and sequential patterns.
- Services for mobile telecommunication.
- Visualization tools are applied to the analysis of telecommunications data.

10.4 Biological Data Analysis

Presently, the domains of biology, including biomedical research, proteomics, genomics, and functional genomics, are experiencing tremendous expansion. Mining biological data is a critical component of bioinformatics.

The subsequent points outline the ways in which data mining contributes to the analysis of biological data:

- The process of semantically integrating distributed, heterogeneous genomic and proteomic databases.
- Alignment, indexing, comparative analysis, and similarity search of multiple nucleotide sequences.

- The identification of structural patterns and the examination of protein pathways and genetic networks.

10.5 Other Scientific Applications

The statistical techniques are typically applicable to relatively small, homogeneous data sets, which are the nature of the applications discussed previously. Profound quantities of data have been amassed across various scientific disciplines, including but not limited to geosciences and astronomy.

A substantial volume of data sets is being produced across diverse domains, including chemical engineering, fluid dynamics, climate and ecosystem modeling, and numerical simulations. The applications of data mining in the field of scientific applications are as follows**.**

11. Conclusion

In this chapter, we have presented an overview of data mining techniques, processes, and tasks (classification, regression, etc.) of data mining, and the different algorithms that we used in order to solve the problem of our study, which will be presented in the next chapter.

CHAPTER 3

DESIGN OF THE PROPOSED SYSTEM

1. Introduction

Two critical components in customer journey identification systems play a significant role in understanding and analyzing customer behavior - Unified Modeling Language (UML) and datasets.

UML is a standardized visual modeling language that facilitates effective communication and documentation of intricate systems for system designers and developers. However, datasets are essential to these systems as they serve as the foundation for information, supplying the required raw material for analysis and generating valuable insights. This chapter examines using UML and datasets in customer journey identification systems. It investigates how these tools work together to chart and evaluate the different points of contact a customer experience when interacting with a product or service.

By utilizing UML for visual depiction and statistics for empirical proof, firms may thoroughly comprehend their consumers' experiences and use vital insights to improve their complete customer journey.

The chapter will describe the design of the proposed system of identification of customer journey in Mobilis Company. However, we should first present the state of the art of works and systems which previously studied CJA in the mobile companies.

2. State of the art

Given the growing importance of customer behavior in the business market nowadays, telecom operators focus not only on customer profitability to increase market share but also on highly loyal customers as well as customers who are churning.

Customer journey analysis is a critical area of study in the mobile telephony industry, focusing on understanding and optimizing the entire lifecycle of customer interactions with a service provider. This analysis aims to map out the various touchpoints and experiences a customer encounter through service usage and, potentially, termination of the service. This approach allows companies to enhance customer satisfaction, reduce churn, and improve overall service delivery.

The telecommunications market is well-developed but is characterized by oversaturation and high levels of competition. Based on this, authors in [30] compared different approaches and methods for customer churn prediction and constructed different Data Science models to classify customers according to the probability of their churn from the company's client base and predict potential customers who could stop using the company's services. However, in [31], the study developed a robust customer churn prediction model in the communications industry using various machine learning and analysis methods. Forecasting customer churn

for a telecommunications company was also studied by [32], where authors provided research on churn forecasting using 11 different machine learning methods on training data containing 20 different information parameters about clients were used.

Telecom network often encounters a large number of tweets based on the user experience for a network. Based on this, [33] addressed the social media review challenges in telecom companies. The study trained a random forest classifier to predict/classify the customer's satisfaction into positive, negative, and neural.

Several studies on customer journey analytics tackled specific countries. For instance, [34] proposed a predictive model using data mining techniques to analyse customer behaviour in order to identify and classify customers with a higher risk of defection in a Peruvian telecommunications company. In [35] authors provided a methodology for telecom companies to target different-value customers by appropriate offers and services. The methodology was implemented and tested using a dataset that contains about 127 million records for training and testing supplied by Syriatel Corporation. By way of descriptive and comparative analysis, Russian and Vietnamese telecommunications markets for the period of 2015-2019 were analyzed [36]. The study's results revealed similar global customer trends in the telecommunications markets.

In Algeria, we aim to perform a customer journey analysis in a telecommunications company in order to ensure the stable operation of the company and to assist the customer by giving him more satisfaction and less intention to make the transition (churn) to another telecommunications company.

3. Presentation of the Case Study: ATM Mobilis

3.1 Identification

Mobilis or Mobilis ATM (ATM acronym of Algeria Telecom Mobile) (Figures 1.3, 1.4 and 1.5) is an Algerian mobile operator and a subsidiary of Algeria Telecom group. It is one of the three major Algerian mobile operators. It became independent in August 2003 [15].

On December 15, 2004, Mobilis launched the first experimental UMTS (Universal Mobile Telecommunication System) network in Algeria. With its GPRS «Mobi+» offer, Mobilis is a multimedia operator in Algeria.

Mobilis has launched a vast project to deploy its GSM network. Today, the network serves almost 80% of the Algerian population. Mobilis's subscriber base (GSM + 3G) stood at 19 million in December 2020. In December 2019, Mobilis obtained a global telecommunications license (2G, 3G, and 4G) to deploy in Mali (Sef Identification card in Figure 1.6).

Figure 1.3: Logo from 2003 to 2010

Figure 1.4: Current logo since 2010

Figure 1.5: Variant of logo actual

Figure1.6: Identification card of Mobilis

3.2 Geographical location

Mobilis is distributed throughout the national territory to exercise its activities [16]. It is represented by:

- A headquarters located in Algiers. It has been located in the Business District Bab Ezzouar since November 2011.
- A distribution and sales network comprising sales agencies, distributors, and points of sale.
- Eight Regional Directorates:
  - Algiers Regional Direction covers the following wilayas: Algiers, Blida, Tipaza, Tizi Ouzou, Boumerdes, and Bouira.
  - Oran Regional Management: which covers the following wilayas: Oran, Sidi Bel Abbes, Mostaganem, Tlemcen, Ain T'émouchent, Saida and Mascara.
  - Annaba Regional Direction covers the following wilayas: Annaba, Tébessa, Guelma, Skikda, El Taref, and Souk Ahras.
  - Constantine Regional Direction: which covers the following wilayas: Constantine, Batna, Oum El Bouagui, Mila, and Khenchela.
  - Chlef Regional Management covers the following wilayas: Chlef, Relizane, Tissemsilt, Tiaret, Médéa, Djelfa, and Ain Defla.
  - Sétif Regional Direction: which covers the following wilayas: Sétif, Jijel, Bejaïa, M'sila, and Bordj Bou Arreridj.
  - Béchar Regional Direction : which covers the following wilayas: Béchar, Naàma, El bayadh, Tindouf and Adrar.

- Ouargla Direction Régionale covers the following wilayas: Ouargla, Tamanrasset, Illizi, Ghardaia, Laghouat, El Oued, and Biskra.

3.3 Places of existence in the digital world

3.3.1 Official Website

The official website of Mobilis Algeria is www.mobilis.dz (Figure 1.7), where you can find information about the products and services of the first mobile operator in Algeria. The website also offers online recharge, customer service, news, and events. Mobilis is a partner of the national football team.

The website is available in Arabic, French, and English and has a simple and user-friendly design. The website features different sections for individual and business customers and a store where you can buy devices, accessories, and SIM cards. The website also allows you to access digital services such as Naghmati, Mobilis Store, and Mobilis Cloud.

Figure 1.7: The official website of Mobilis

3.3.2 Social Networks

- **Facebook:** Mobilis Funs' preferred social network has over 3 million followers. This medium is mainly utilized to disseminate Mobilis news, goods, and social activities. Official Facebook page link: https://web.facebook.com/MobilisOfficielle
- **Instagram:** Mobilis, with over 229 K followers, uses this page to promote its offers and activities around national or religious events. Customers may participate by viewing photographs, sharing comments, and expressing their thoughts through likes and shares. Official Instagram page link: https://www.instagram.com/mobilis.dz/
- **Twitter:** Mobilis publishes news about its activities, goods, and services with 567,5 K subscribers. Official Twitter page link: https://twitter.com/ATM_Mobilis
- **YouTube:** a channel showcasing the products and services of the first mobile operator in Algeria. It also features videos about the latest digital trends, the national football team, and various cultural and social events. The channel has 113K subscribers and 375 videos. Official YouTube page link: https://www.youtube.com/user/TVMobilis
- **LinkedIn:** is a platform connecting Algeria's first mobile operator with its customers, partners, and employees. It showcases the company's vision, values, achievements, and opportunities. It also updates the latest news, events, and innovations in the telecommunications sector. The page has 99 K followers and 1,011 posts. Official LinkedIn page link: https://www.linkedin.com/company/atmmobilis/

4. What is UML?

The Unified Modeling Language (UML) is a general-purpose visual modeling language that is intended to provide a standard way to visualize the design of a system.

UML provides a standard notation for many types of diagrams, which can be roughly divided into three main groups: behavior diagrams, interaction diagrams, and structure diagrams.

The creation of UML was originally motivated by the desire to standardize the disparate notational systems and approaches to software design. It was developed at Rational Software in 1994–1995, with further development led by them through 1996.

In 1997, UML was adopted as a standard by the Object Management Group (OMG) and has been managed by this organization ever since. In 2005, UML was also published by the International Organization for Standardization (ISO) and the International Electro-technical Commission (IEC) as the ISO/IEC 19501 standard. Since then the standard has been periodically revised to cover the latest revision of UML.

In software engineering, most practitioners do not use UML but instead produce informal hand-drawn diagrams; these diagrams, however, often include elements from UML" [37].

5. UML diagrams

The UML uses elements and combines them in different ways to form diagrams that represent the static or structural aspects of a system, as well as behavioral diagrams that capture the dynamic aspects of a system [38].

- **Structural diagrams;**

- Class diagram: The most frequently utilized UML diagram and the fundamental basis of any object-oriented solution. The system consists of classes, each having properties and operations. Additionally, linkages were established between the classes. Class diagrams are formed by grouping classes together to model larger systems.
- Component diagram: Represents the hierarchical connection between the various components of a software system, typically employed in complex systems consisting of multiple components. Components interact with each other through interfaces.
- Deployment Diagram: illustrates a system's infrastructure's physical and software components. They are helpful when a software solution is implemented on numerous machines with unique setups.
- Object Diagram: Illustrates the connections between items using concrete illustrations from the actual world and enables you to observe the state of a system at any specific moment. Data is contained within things, enabling it to elucidate connections between items.

- **Behavioral diagrams:**
  - Activity diagrams: Business or operational processes are illustrated visually to illustrate the actions performed by each of the system components. Activity diagrams substitute for state-transition diagrams.
  - Communication diagram: Similar to a sequence diagram, this diagram focuses on communicating messages between objects. A sequence diagram can represent the same information using various items.
  - Sequence diagram: Illustrates how things associate with one another and the sequence in which these interactions occur. They show the interplay of a specific scenario.
  - Use Case Diagram: A system diagram is a visual representation that shows the interconnections between various aspects and showcases their internal and external controllers or actors.

6. Collecting Data for Customer journey Analysis

Customer journey analysis is a powerful tool for understanding the experiences and behaviors of customers as they interact with a company's products or services. In the context of a mobile company, collecting data for customer journey analysis can provide valuable insights into how customers engage with the company's offerings, identify pain points, and uncover opportunities for improvement.

It's important to note that these datasets often contain sensitive customer information, so proper data anonymization, pseudonymization, and compliance with relevant data privacy regulations are crucial.

Additionally, some datasets may need to be generated or assembled internally by the mobile company, while others may be available from third-party providers or open data sources (with appropriate licensing and usage considerations).

6.1 Customer service dataset

The Figure 3.1 presents a screenshot of the dataset collected about the customer service, while Table 3.1 describe all the columns of the dataset.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Custmer ID	Age	Gender	Agent	Date	Time	Topic	Answered (Y/N)	Resolved	Waiting Time	AvgTalkDuration	Satisfaction rating	Recall
2	ID0001	57	1	Reda	2021-01-01	09:12:58	Contract related	Y	Y	49	00:02:23	3	1
3	ID0002	32	1	Mourad	2021-01-01	09:12:58	Technical Support	Y	N	10	00:04:02	3	1
4	ID0003	30	0	Said	2021-01-01	09:47:31	Contract related	Y	Y	10	00:02:11	3	0
5	ID0004	36	1	Toufik	2021-01-01	09:47:31	Contract related	Y	Y	53	00:00:37	2	0
6	ID0005	29	1	Mourad	2021-01-01	10:00:29	Payment related	Y	Y	35	00:01:00	3	0
7	ID0006	53	0	Said	2021-01-01	10:00:29	Technical Support	N	N	0			1
8	ID0007	28	1	Reda	2021-01-01	10:22:05	Payment related	Y	Y	24	00:03:40	2	0
9	ID0008	47	1	Reda	2021-01-01	10:22:05	Payment related	Y	Y	22	00:00:38	4	0
10	ID0009	17	0	Toufik	2021-01-01	11:13:55	Admin Support	Y	Y	15	00:06:38	4	0
11	ID0010	62	0	Djamel	2021-01-01	11:13:55	Offers related	Y	Y	18	00:01:04	3	0
12	ID0011	45	1	Leyla	2021-01-01	11:15:22	Payment related	N	N	0			1
13	ID0012	58	1	Toufik	2021-01-01	11:15:22	Payment related	Y	Y	50	00:00:32	4	0
14	ID0013	50	0	Leyla	2021-01-01	11:52:48	Payment related	Y	Y	24	00:03:34	3	0
15	ID0014	74	1	Mouna	2021-01-01	11:52:48	Contract related	Y	Y	29	00:05:44	3	0
16	ID0015	62	0	Mourad	2021-01-01	11:55:41	Admin Support	Y	Y	48	00:03:47	4	0
17	ID0016	40	1	Mourad	2021-01-01	11:55:41	Admin Support	Y	Y	3	00:05:26	2	0
18	ID0017	25	1	Toufik	2021-01-01	11:57:07	Technical Support	Y	Y	45	00:05:32	5	0
19	ID0018	48	1	Mourad	2021-01-01	11:57:07	Admin Support	N	N	0			1
20	ID0019	33	1	Djamel	2021-01-01	12:01:26	Offers related	N	N	0			0
21	ID0020	57	0	Djamel	2021-01-01	12:01:26	Contract related	Y	Y	41	00:02:27	3	0
22	ID0021	64	0	Djamel	2021-01-01	12:02:53	Technical Support	Y	Y	14	00:05:22	5	0
23	ID0022	28	1	Samia	2021-01-01	12:02:53	Admin Support	Y	Y	29	00:05:50	5	0
24	ID0023	30	1	Mouna	2021-01-01	12:02:53	Technical Support	N	N	0			1
25	ID0024	60	0	Leyla	2021-01-01	12:02:53	Technical Support	Y	Y	8	00:05:25	2	0
26	ID0025	71	1	Reda	2021-01-01	12:30:14	Offers related	Y	Y	37	00:04:09	3	0

Figure 3.1 : Screenshot of Customer Service dataset

Feature	**Description**
Customer ID	This is a unique identifier assigned to each customer
Age	The age of the customer
gender	The gender of the customer (e.g., male or female)
Agent	This refers to the customer service representative or agent who handled the interaction with the customer
Date	The date of the customer interaction
Time	The time of day of the customer interaction
Topic	Describes the type of interaction (e.g., inquiry, complaint, technical support)
Answered	The customer inquiry or issue was answered by the agent:Yes,No
Resolved	The customer inquiry or issue was fully resolved
Waiting Time	The duration of time the customer had to wait before being attended to by an agent
Avg Talk	The average duration of the conversation or interaction between the

Duration	customer and the agent
Satisfaction Rating	Represent customer satisfaction ratings (e.g., 1 for very dissatisfied, 2 for neutral, 3 for very satisfied)
Recall	Recall the customers after the interaction to gather feedback about the company services: Yes, No.

Table 3.1 : Customer Service dataset description

6.2 Customer Segmentation Study Dataset:

The Figure 3.2 presents a screenshot of the dataset collected about the customer segmentation, while Table 3.2 describe all the columns of the dataset.

	A	B	C	D	E	F	G	H	I	J	K	L
1	region	tenure	age	marital	address	income	ed	employ	retire	gender	reside	custcat
2	2	13	44	1	9	64	4	5	0	0	2	1
3	3	11	33	1	7	136	5	5	0	0	6	4
4	3	68	52	1	24	116	1	29	0	1	2	3
5	2	33	33	0	12	33	2	0	0	1	1	1
6	2	23	30	1	9	30	1	2	0	0	4	3
7	2	41	39	0	17	78	2	16	0	1	1	3
8	3	45	22	1	2	19	2	4	0	1	5	2
9	2	38	35	0	5	76	2	10	0	0	3	4
10	3	45	59	1	7	166	4	31	0	0	5	3
11	1	68	41	1	21	72	1	22	0	0	3	2
12	2	5	33	0	10	125	4	5	0	1	1	1
13	3	7	35	0	14	80	2	15	0	1	1	3
14	1	41	38	1	8	37	2	9	0	1	3	1
15	2	57	54	1	30	115	4	23	0	1	3	4
16	2	9	46	0	3	25	1	8	0	1	2	1
17	1	29	38	1	12	75	5	1	0	0	4	2
18	3	60	57	0	38	162	2	30	0	0	1	3
19	3	34	48	0	3	49	2	6	0	1	3	3
20	2	1	24	0	3	20	1	3	0	0	1	1
21	1	26	29	1	3	77	4	2	0	0	4	4
22	3	6	30	0	7	16	3	1	0	1	1	2
23	1	68	52	1	17	120	1	24	0	0	2	1
24	3	53	33	0	10	101	5	4	0	1	2	4
25	3	55	48	1	19	67	1	25	0	0	3	1
26	3	14	43	1	18	36	1	5	0	0	5	3

Figure 3.2 : Screenshot of customer segmentation dataset

Feature	**Description**
Region	The geographical region associated with each customer
Tenure	The length of time that a customer has been associated with service
Age	The age of the customer

Marital	Indicates if the customer is married: Yes, No
Address	Contains information related to the customer's address
Income	The income level of the customer
Ed	The educational background of the customer
Employ	Number of years employed
Retire	The customer is retired: Yes, No
Gender	The gender of the customer (e.g., male or female)
Reside	The number of people residing in the household
Custcat	Categorization of customers (e.g., high-value, medium-value, low-value)

Table 3.2: Customer Segmentation Dataset Description

6.3 Customer Churn Dataset

The Figure 3.3 presents a screenshot of the dataset collected about the customer churning, while Table 3.3 describe all the columns of the dataset.

	A	B	C	D	E	F	G	H	I	J	K
1	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
2	0	128	1	1	2,7	1	265,1	110	89	9,87	10
3	0	107	1	1	3,7	1	161,6	123	82	9,78	13,7
4	0	137	1	0	0	0	243,4	114	52	6,06	12,2
5	0	84	0	0	0	2	299,4	71	57	3,1	6,6
6	0	75	0	0	0	3	166,7	113	41	7,42	10,1
7	0	118	0	0	0	0	223,4	98	57	11,03	6,3
8	0	121	1	1	2,03	3	218,2	88	87,3	17,43	7,5
9	0	147	0	0	0	0	157	79	36	5,16	7,1
10	0	117	1	0	0,19	1	184,5	97	63,9	17,58	8,7
11	0	141	0	1	3,02	0	258,6	84	93,2	11,1	11,2
12	1	65	1	0	0,29	4	129,1	137	44,9	11,43	12,7
13	0	74	1	0	0,34	0	187,7	127	49,4	8,17	9,1
14	0	168	1	0	0	1	128,8	96	31	5,25	11,2
15	0	95	1	0	0,44	3	156,6	88	52,4	12,38	12,3
16	0	62	1	0	0	4	120,7	70	47	15,36	13,1
17	1	161	1	0	0	4	332,9	67	84	15,89	5,4
18	0	85	1	1	3,73	1	196,4	139	95,3	14,05	13,8
19	0	93	1	0	0	3	190,7	114	51	10,91	8,1
20	0	76	1	1	2,7	1	189,7	66	78	10,64	10

Figure 3.3 : Screenshot of Customer Churn dataset

Feature	**Description**
Churn	The customer has left the service (churned) or is still an active customer.
AccountWeeks	The number of weeks a customer has been subscribed to the service
ContractRenewal	The customer has renewed their contract:Yes,No

DataPlan	The customer has a data plan:Yes,No
DataUsage	The amount of data used by the customer (e.g., in gigabytes)
CustServCalls	The number of customer service calls made by the customer
DayMins	Total minutes of voice calls used by the customer during the day
DayCalls	The number of voice calls made by the customer during the day
MonthlyCharge	The monthly charge with the service subscribed by the customer
OverageFee	Additional fees incurred if the customer exceeds their allocated limits (e.g., data usage, minutes)
RoamMins	The number of roaming minutes used by the customer

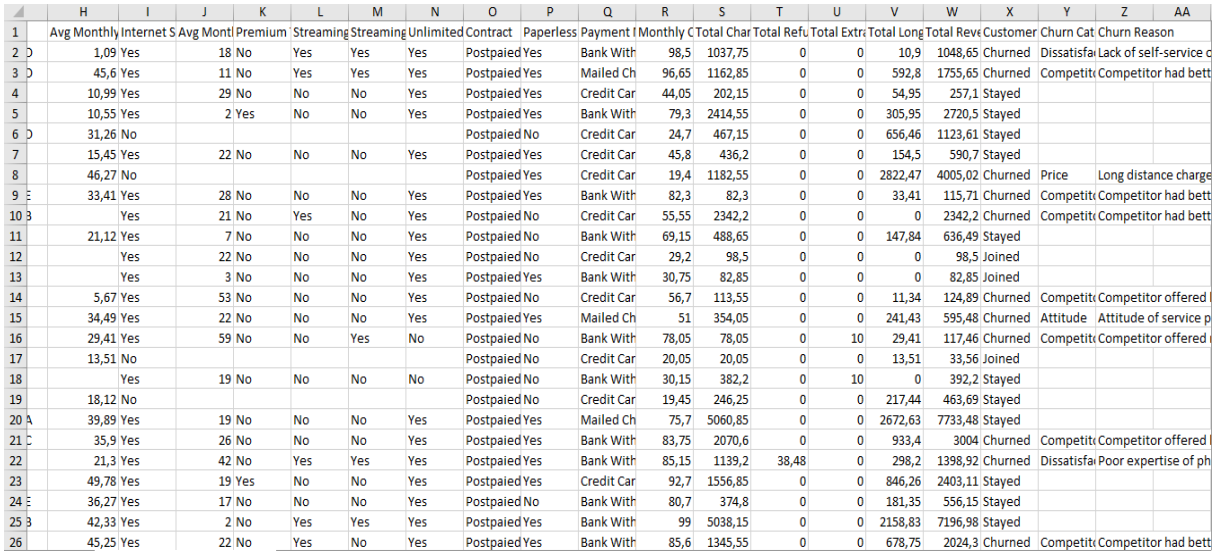
Table 3.3 Customer Churn Dataset Description

6.4 Customer Care Dataset

The Figures 3.4 and 3.5 presents a screenshot of the dataset collected about the customer care service, while Table 3.4 describe all the columns of the dataset.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Customer ID	Gender	Age	Married	Zip Code	Tenure in	Offer	Avg Monthly	Internet S	Avg Monti	Premium	Streaming	Streaming	Unlimited	Contract	Paperless	Payment I	Monthly C	Total Char	Total Refu
2	1455-UGQVH	Male	35	Yes	28220	10	Offer D	1,09	Yes	18	No	Yes	Yes	Yes	Postpaied	Yes	Bank With	98,5	1037,75	0
3	8844-TONUD	Male	54	Yes	28220	13	Offer D	45,6	Yes	11	No	Yes	Yes	Yes	Postpaied	Yes	Mailed Ch	96,65	1162,85	0
4	1635-FJFCC	Female	46	No	28510	5	None	10,99	Yes	29	No	No	No	Yes	Postpaied	Yes	Credit Car	44,05	202,15	0
5	3312-ZWLGF	Male	69	Yes	28510	29	None	10,55	Yes	2	Yes	No	No	Yes	Postpaied	Yes	Bank With	79,3	2414,55	0
6	3486-KHMLI	Male	53	No	28301	21	Offer D	31,26	No						Postpaied	No	Credit Car	24,7	467,15	0
7	6143-JQKEA	Male	34	No	28301	10	None	15,45	Yes	22	No	No	No	Yes	Postpaied	Yes	Credit Car	45,8	436,2	0
8	9063-ZGTUY	Female	22	Yes	28301	61	None	46,27	No						Postpaied	Yes	Credit Car	19,4	1182,55	0
9	0742-NXBGR	Female	64	No	28006	1	Offer E	33,41	Yes	28	No	No	No	Yes	Postpaied	Yes	Bank With	82,3	82,3	0
10	6397-JNZZG	Female	70	Yes	28006	43	Offer B		Yes	21	No	Yes	No	Yes	Postpaied	No	Credit Car	55,55	2342,2	0
11	8993-IZEUX	Male	54	No	28006	7	None	21,12	Yes	7	No	No	No	Yes	Postpaied	No	Bank With	69,15	488,65	0
12	0975-VOOVL	Female	34	No	28301	3	None		Yes	22	No	No	No	Yes	Postpaied	No	Credit Car	29,2	98,5	0
13	1078-TDCRN	Female	75	Yes	28301	3	None		Yes	3	No	No	No	Yes	Postpaied	Yes	Bank With	30,75	82,85	0
14	8450-JOVAH	Male	62	Yes	28301	2	None	5,67	Yes	53	No	No	No	Yes	Postpaied	No	Credit Car	56,7	113,55	0
15	2207-OBZNX	Male	56	No	28536	7	None	34,49	Yes	22	No	No	No	Yes	Postpaied	Yes	Mailed Ch	51	354,05	0
16	2777-PHDEI	Female	20	No	28601	1	None	29,41	Yes	59	No	No	Yes	No	Postpaied	No	Bank With	78,05	78,05	0
17	8749-CLJXC	Male	56	No	28601	1	None	13,51	No						Postpaied	No	Credit Car	20,05	20,05	0
18	0811-GSDTP	Female	50	No	28502	13	None		Yes	19	No	No	No	No	Postpaied	No	Bank With	30,15	382,2	0
19	3750-RNQKR	Female	38	No	28501	12	None	18,12	No						Postpaied	No	Credit Car	19,45	246,25	0
20	9943-VSZUV	Male	75	No	28507	67	Offer A	39,89	Yes	19	No	No	No	Yes	Postpaied	Yes	Mailed Ch	75,7	5060,85	0
21	8436-BJUMM	Male	62	Yes	28706	26	Offer C	35,9	Yes	26	No	No	No	Yes	Postpaied	Yes	Bank With	83,75	2070,6	0
22	9185-TQCVP	Male	27	Yes	28706	14	None	21,3	Yes	42	No	Yes	Yes	Yes	Postpaied	Yes	Bank With	85,15	1139,2	38,48
23	2074-GUHPQ	Female	37	No	28410	17	None	49,78	Yes	19	Yes	No	No	Yes	Postpaied	Yes	Credit Car	92,7	1556,85	0
24	6458-CYIDZ	Female	72	No	28410	5	Offer E	36,27	Yes	17	No	No	No	Yes	Postpaied	No	Bank With	80,7	374,8	0
25	5649-RXQTV	Male	34	No	28511	51	Offer B	42,33	Yes	2	No	Yes	Yes	Yes	Postpaied	Yes	Bank With	99	5038,15	0
26	8519-IMDHU	Male	74	Yes	28511	15	None	45,25	Yes	22	No	Yes	No	Yes	Postpaied	Yes	Bank With	85,6	1345,55	0

Figure 3.4 : Customer Care Dataset part1



	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	Avg Monthly	Internet S	Avg Monti	Premium	Streaming	Streaming	Unlimited	Contract	Paperless	Payment I	Monthly C	Total Char	Total Refu	Total Extr	Total Long	Total Reve	Customer	Churn Cat	Churn Reason	
2	1,09	Yes	18	No	Yes	Yes	Yes	Postpaied	Yes	Bank With	98,5	1037,75	0	0	10,9	1048,65	Churned	Dissatisfa	Lack of self-service c	
3	45,6	Yes	11	No	Yes	Yes	Yes	Postpaied	Yes	Mailed Ch	96,65	1162,85	0	0	592,8	1755,65	Churned	Competit	Competitor had bett	
4	10,99	Yes	29	No	No	No	Yes	Postpaied	Yes	Credit Car	44,05	202,15	0	0	54,95	257,1	Stayed			
5	10,55	Yes	2	Yes	No	No	Yes	Postpaied	Yes	Bank With	79,3	2414,55	0	0	305,95	2720,5	Stayed			
6	31,26	No						Postpaied	No	Credit Car	24,7	467,15	0	0	656,46	1123,61	Stayed			
7	15,45	Yes	22	No	No	No	Yes	Postpaied	Yes	Credit Car	45,8	436,2	0	0	154,5	590,7	Stayed			
8	46,27	No						Postpaied	Yes	Credit Car	19,4	1182,55	0	0	2822,47	4005,02	Churned	Price	Long distance charge	
9	33,41	Yes	28	No	No	No	Yes	Postpaied	Yes	Bank With	82,3	82,3	0	0	33,41	115,71	Churned	Competit	Competitor had bett	
10		Yes	21	No	Yes	No	Yes	Postpaied	No	Credit Car	55,55	2342,2	0	0	0	2342,2	Churned	Competit	Competitor had bett	
11	21,12	Yes	7	No	No	No	Yes	Postpaied	No	Bank With	69,15	488,65	0	0	147,84	636,49	Stayed			
12		Yes	22	No	No	No	Yes	Postpaied	No	Credit Car	29,2	98,5	0	0	0	98,5	Joined			
13		Yes	3	No	No	No	Yes	Postpaied	Yes	Bank With	30,75	82,85	0	0	0	82,85	Joined			
14	5,67	Yes	53	No	No	No	Yes	Postpaied	No	Credit Car	56,7	113,55	0	0	11,34	124,89	Churned	Competit	Competitor offered	
15	34,49	Yes	22	No	No	No	Yes	Postpaied	Yes	Mailed Ch	51	354,05	0	0	241,43	595,48	Churned	Attitude	Attitude of service p	
16	29,41	Yes	59	No	No	Yes	No	Postpaied	No	Bank With	78,05	78,05	0	10	29,41	117,46	Churned	Competit	Competitor offered	
17	13,51	No						Postpaied	No	Credit Car	20,05	20,05	0	0	13,51	33,56	Joined			
18		Yes	19	No	No	No	No	Postpaied	No	Bank With	30,15	382,2	0	10	0	392,2	Stayed			
19	18,12	No						Postpaied	No	Credit Car	19,45	246,25	0	0	217,44	463,69	Stayed			
20	39,89	Yes	19	No	No	No	Yes	Postpaied	Yes	Mailed Ch	75,7	5060,85	0	0	2672,63	7733,48	Stayed			
21	35,9	Yes	26	No	No	No	Yes	Postpaied	Yes	Bank With	83,75	2070,6	0	0	933,4	3004	Churned	Competit	Competitor offered	
22	21,3	Yes	42	No	Yes	Yes	Yes	Postpaied	Yes	Bank With	85,15	1139,2	38,48	0	298,2	1398,92	Churned	Dissatisfa	Poor expertise of ph	
23	49,78	Yes	19	Yes	No	No	Yes	Postpaied	Yes	Credit Car	92,7	1556,85	0	0	846,26	2403,11	Stayed			
24	36,27	Yes	17	No	No	No	Yes	Postpaied	No	Bank With	80,7	374,8	0	0	181,35	556,15	Stayed			
25	42,33	Yes	2	No	Yes	Yes	Yes	Postpaied	Yes	Bank With	99	5038,15	0	0	2158,83	7196,98	Stayed			
26	45,25	Yes	22	No	Yes	No	Yes	Postpaied	Yes	Bank With	85,6	1345,55	0	0	678,75	2024,3	Churned	Competit	Competitor had bett	

Figure 3.5 : Customer Care Dataset part2

Feature	**Description**	**Feature**	**Description**
CustomerID	A unique ID that identifies each customer	Offer	the last marketing offer that the customer accepted: None, Offer A, Offer B, Offer C, Offer D, Offer E
Gender	The customer's gender: Male, Female	AvgMonthlyLong DistanceCharges	The average monthly long-distance charges for the offer
Age	The customer's current age, in years	AvgMonthlyGBD ownload	the customer's average monthly download volume in gigabytes
Married	he customer is married: Yes, No	UnlimitedData	unlimited data is part of the offer: Yes, No
ZipCode	The zip code of the customer's primary residence	PaymentMethod	how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check
TenureInMonths	the total amount of months that the customer has been with the company	MonthlyCharge	the customer's current total monthly charge for all their services from the company
TotalCharges	the customer's total charges	TotalLongDistanc eCharges	the customer's total charges for long distance above those specified in their plan
CustomerStatus	The status of the customer (e.g., “Churned” or “Stayed”or”Joined”).	TotalExtraDataCh arges	the customer's total charges for extra data downloads above those specified in their plan
ChurnCategory	The category of churn (e.g., “High Churn,” “Medium Churn,” or “Low Churn”)	TotalRevenue	the company's total revenue from this customer
ChurnReason	A customer's specific reason for leaving the company		

Table 3.4: Customer Care Dataset description

7. System Design using UML

7.1 Use Case diagram

In the use case diagram (Figure 3.6), we show three actors: Manager, System and Analyst working on our system.

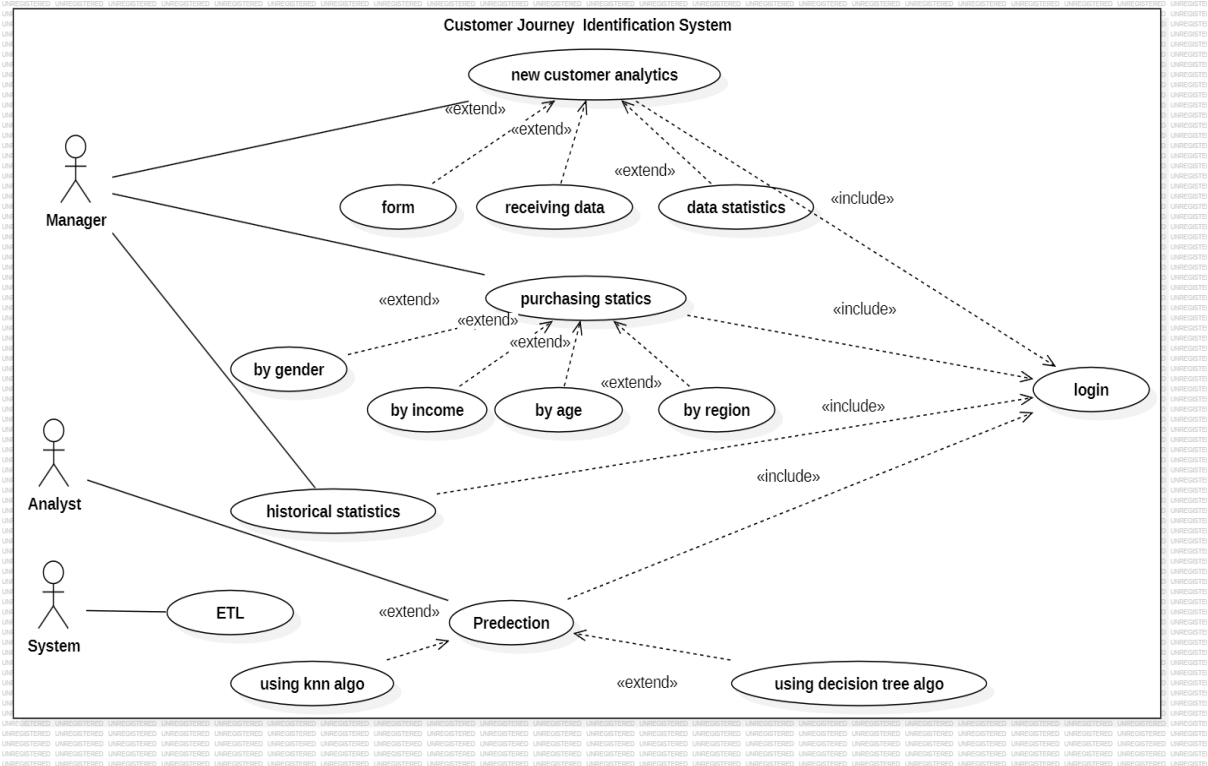


Figure 3.6: Use case diagram of Customer journey identification system

7.2 Sequence diagrams

7.2.1 Sequence diagram of Login

In the login diagram in Figure 3.7, we show the sequence of processing steps from the user's authentication request to entering it into the system, testing that information, and finally determining whether the user can enter the system**.**

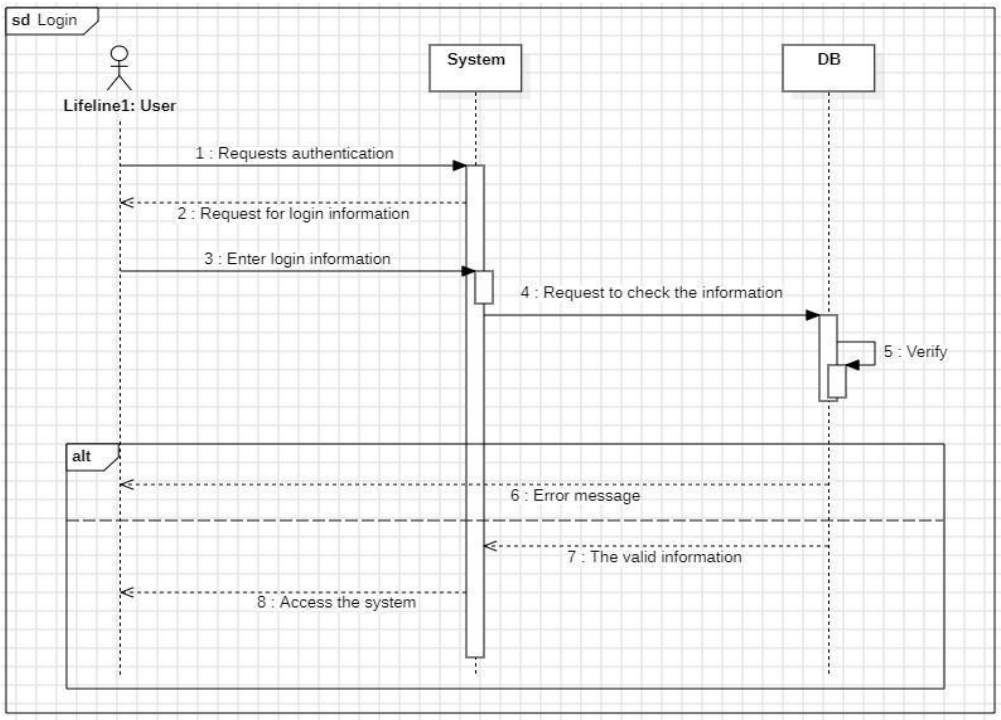


Figure 3.7 : Sequence diagram of User Login

7.2.2 Sequence diagram of New Customer Analytics

In this section, we will show the sequence diagram for received data statistics for new customer analytics in Figure 3.8. The process starts by requesting the manager for received data statistics, then entering the data. In the end, the system performs the statistics and displays them to the manager.

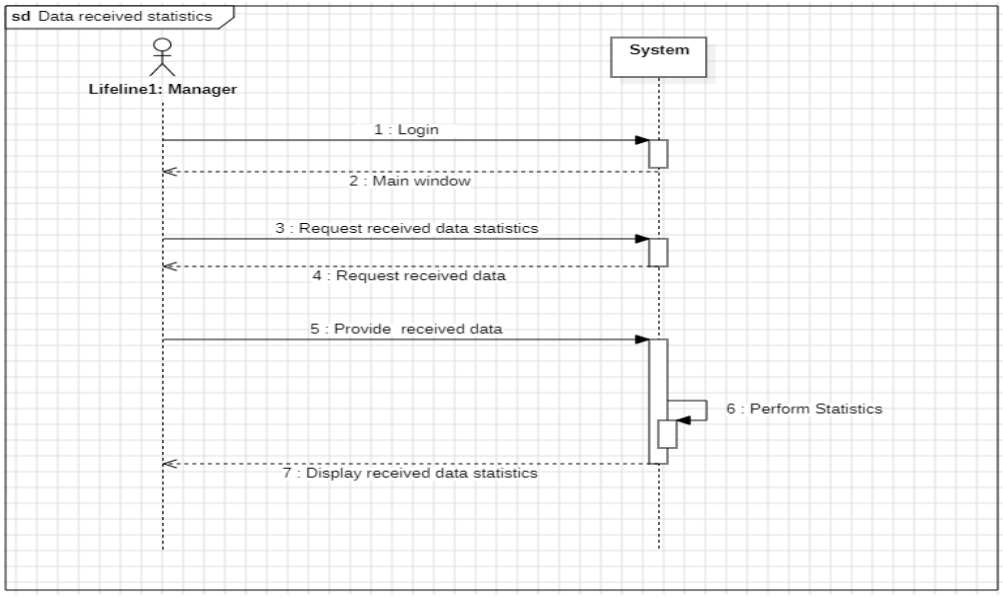


Figure 3.8: Sequence Diagram of received data statistics for new customers

7.2.3 Sequence diagram of Purchase statistics

In this section, we will show the sequence diagrams for purchase statistics by age, gender, income, and region. The process starts by requesting the manager for purchase statistics, and then entering the data. In the end, the system performs the statistics and displays them to the manager.

Purchase Statistics by age diagram is shown in the figure 3.9.

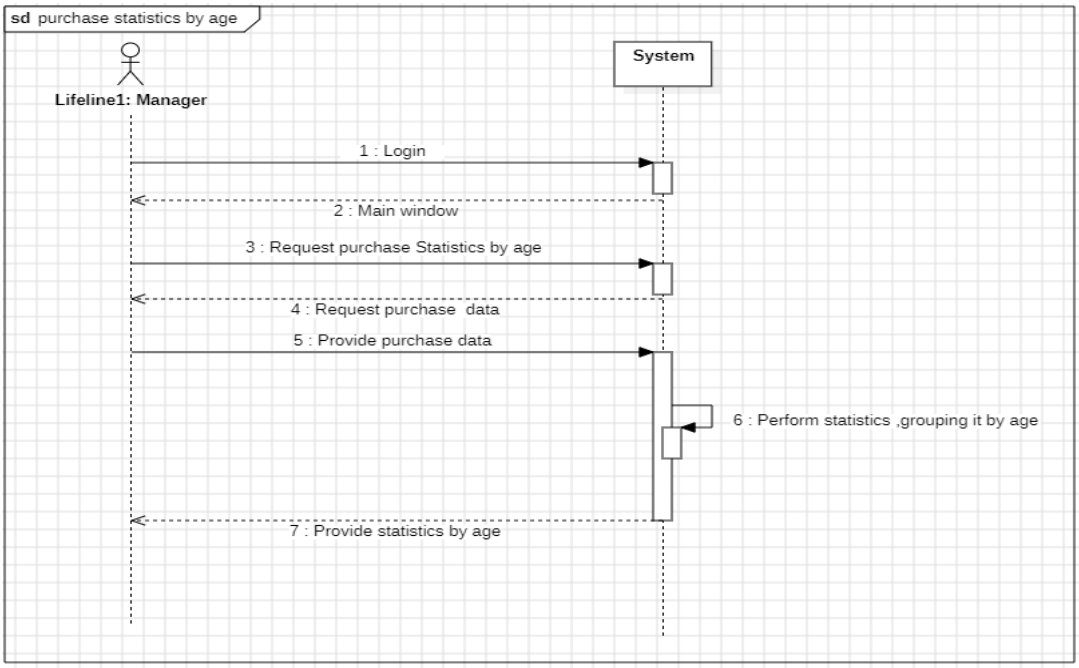


Figure 3.9: Sequence Diagram of purchase statistics by age

Purchase Statistics by region diagram is shown in figure 3.10

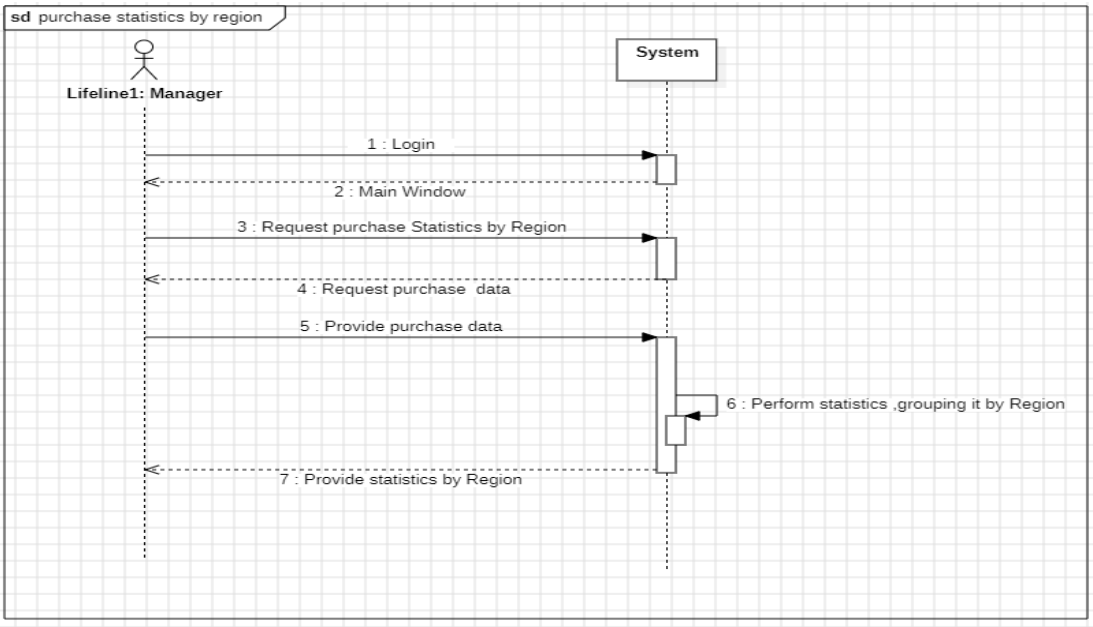


Figure 3.10 : Sequence Diagram of purchase statistics by Region

Purchase Statistics by gender diagram are shown in the figure 3.11.

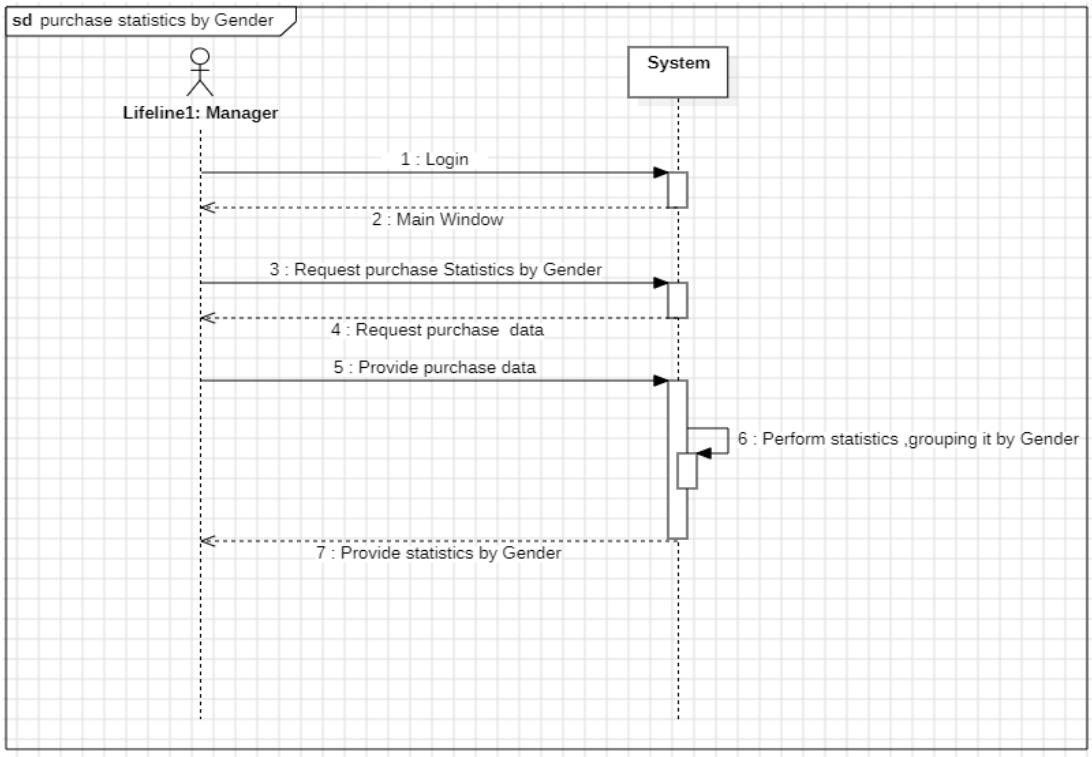


Figure 3.11 : Sequence Diagram of purchase statistics by Gender

Purchase Statistics by income diagram is shown in figure 3.12.

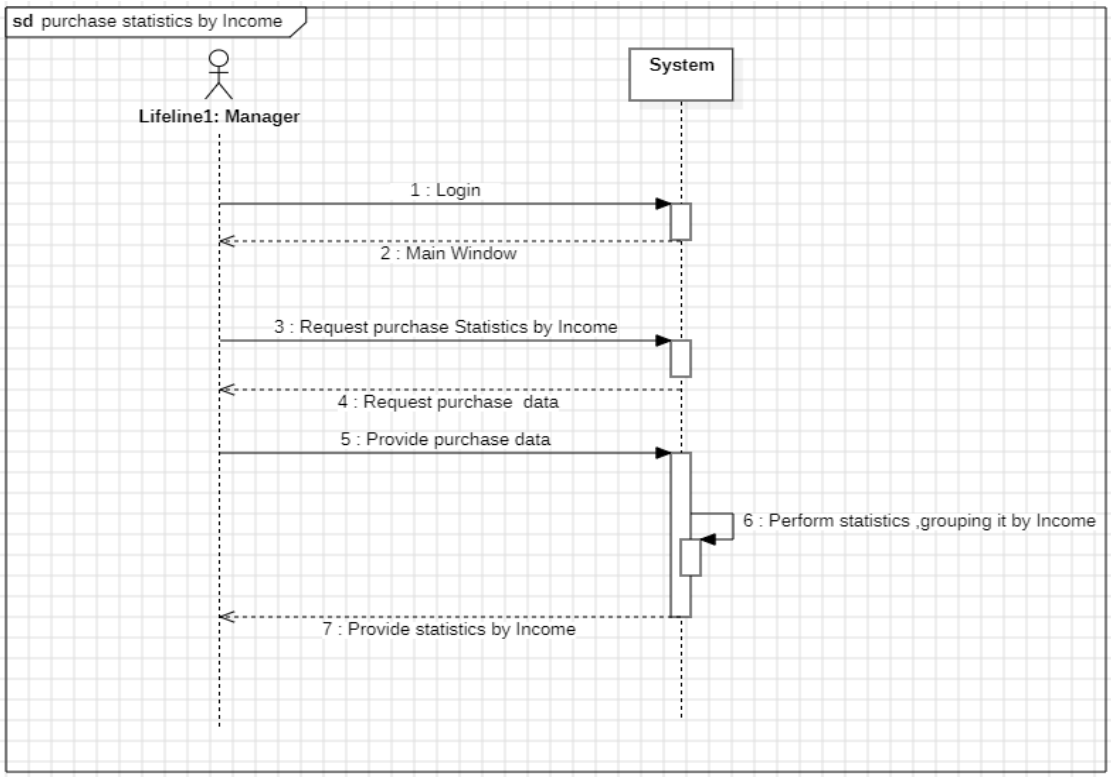


Figure 3.12: Sequence Diagram of purchase statistics by Income

7.2.4 **Sequence diagram of Historical statistics**

In this section, we will show the sequence diagram for historical statistics in Figure 3.13. The process starts by requesting the manager for historical data statistics, then entering the data. In the end, the system performs the statistics and displays them to the manager.

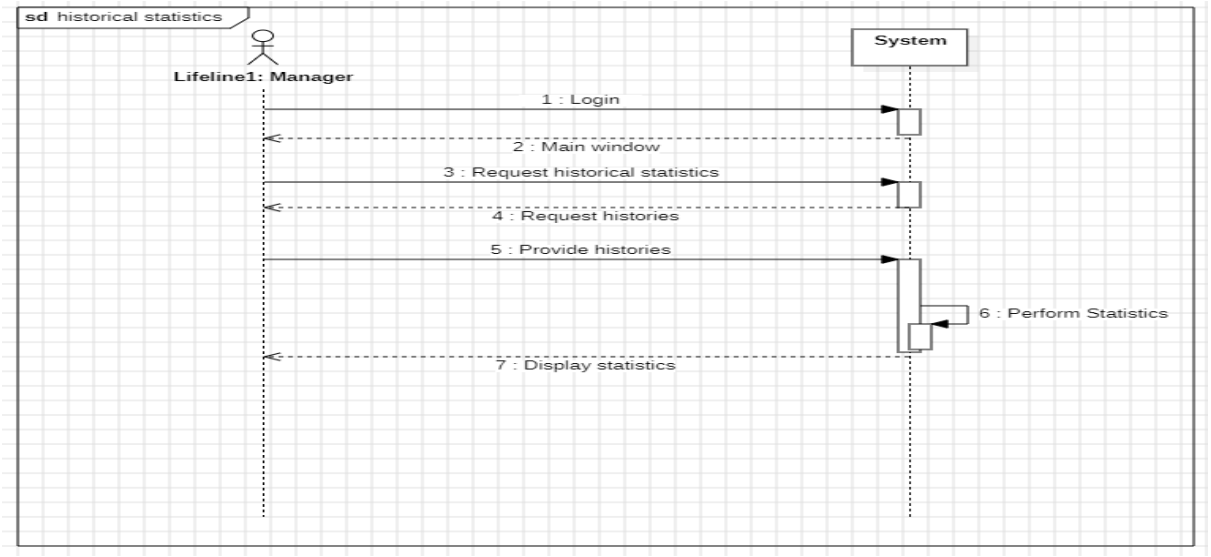


Figure 3.13 8: Sequence Diagram of Historical Statistics

7.2.5 **Sequence diagram of Prediction**

In this section, we will show the sequence diagrams for the Classification algorithms used in our system for prediction.

Apply Classification algorithm diagram is shown in the figure 3.14:

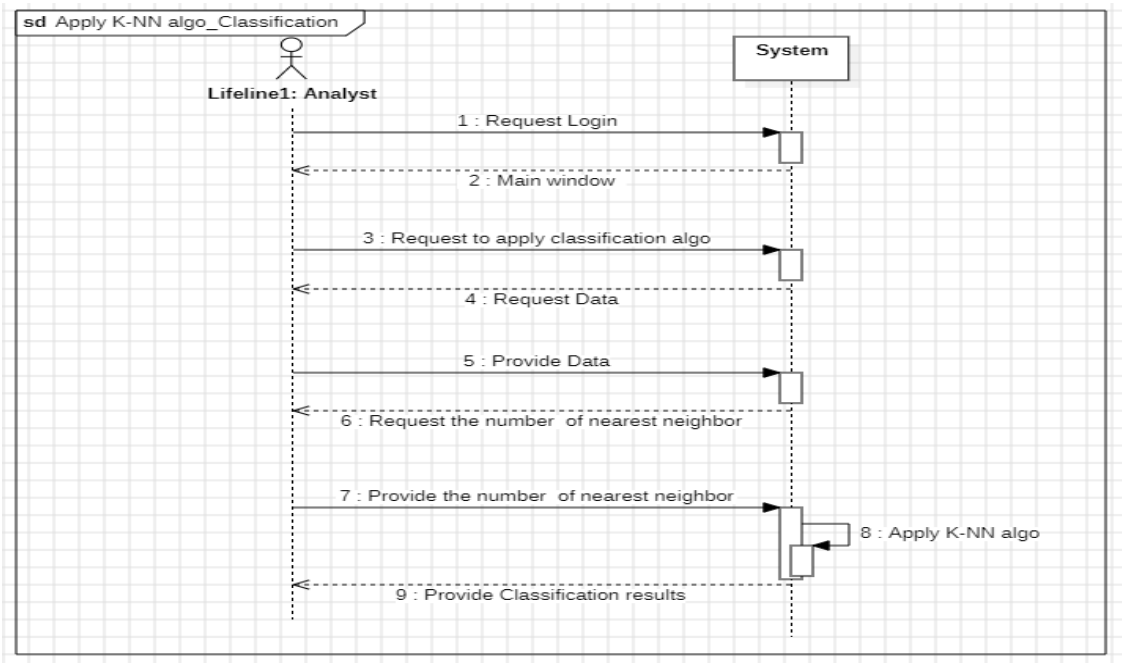


Figure 3.14: Sequence Diagram of Historical Statistics

Apply Decision tree algorithm diagram is shown in the figure 3.15:

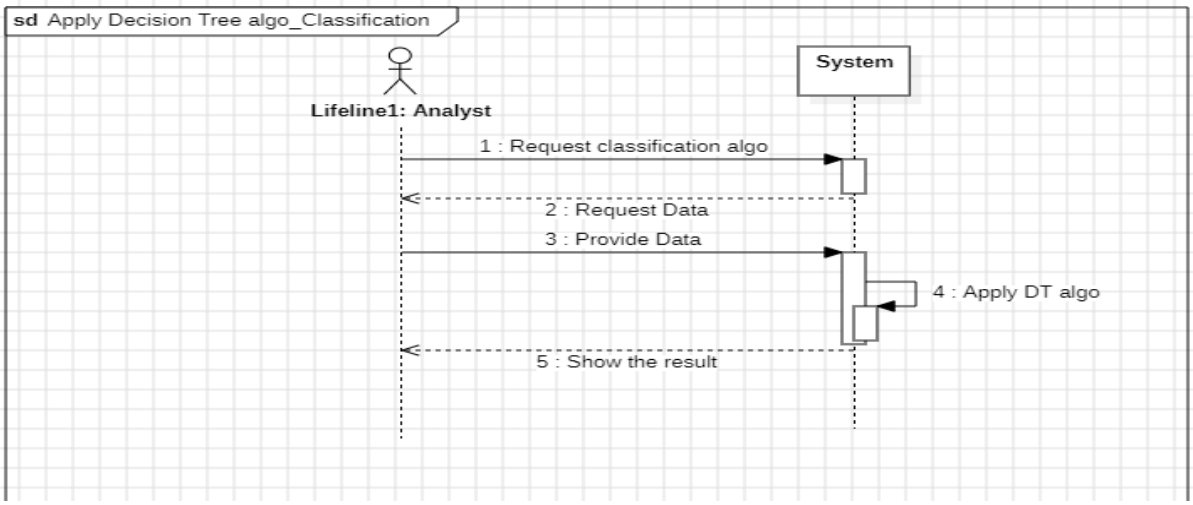


Figure 3.15: Sequence Diagram for apply *Decision tree*

7.2.6 **Sequence diagram of ETL**

ETL (Extract, Transform, Load) processes are crucial as they facilitate data flow between sources and systems, providing that datasets are appropriately organized and ready for analysis. Although not the main emphasis, ETL plays a crucial role in enabling the efficient utilization of datasets. It accomplishes this by extracting pertinent data, transforming it into a usable format, and loading it into the proper systems for subsequent processing and analysis.

In the Figure 3.16, we present the sequence diagram for the ETL process used in our system. The system starts by extracting data from the database of the customer management system of the company, then data extracted will be transformed and loaded in our system.

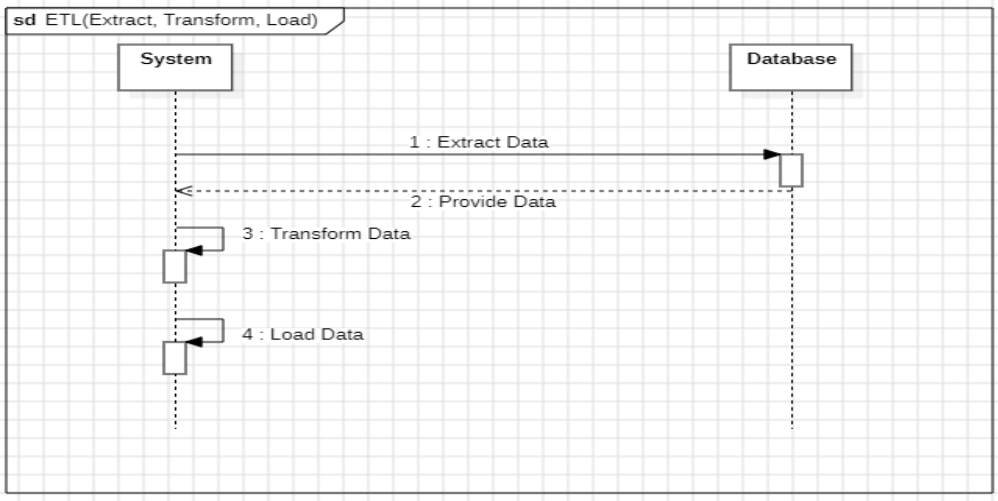


Figure 3.16 : Sequence Diagram for ETL process

7.3 Class diagram

In Figure 3.17, we present the class diagram of the customer management system use by the company, which helps us performing ETL process (see 6.2.6) in order to extract data to perform customer journey analysis in our proposed system.

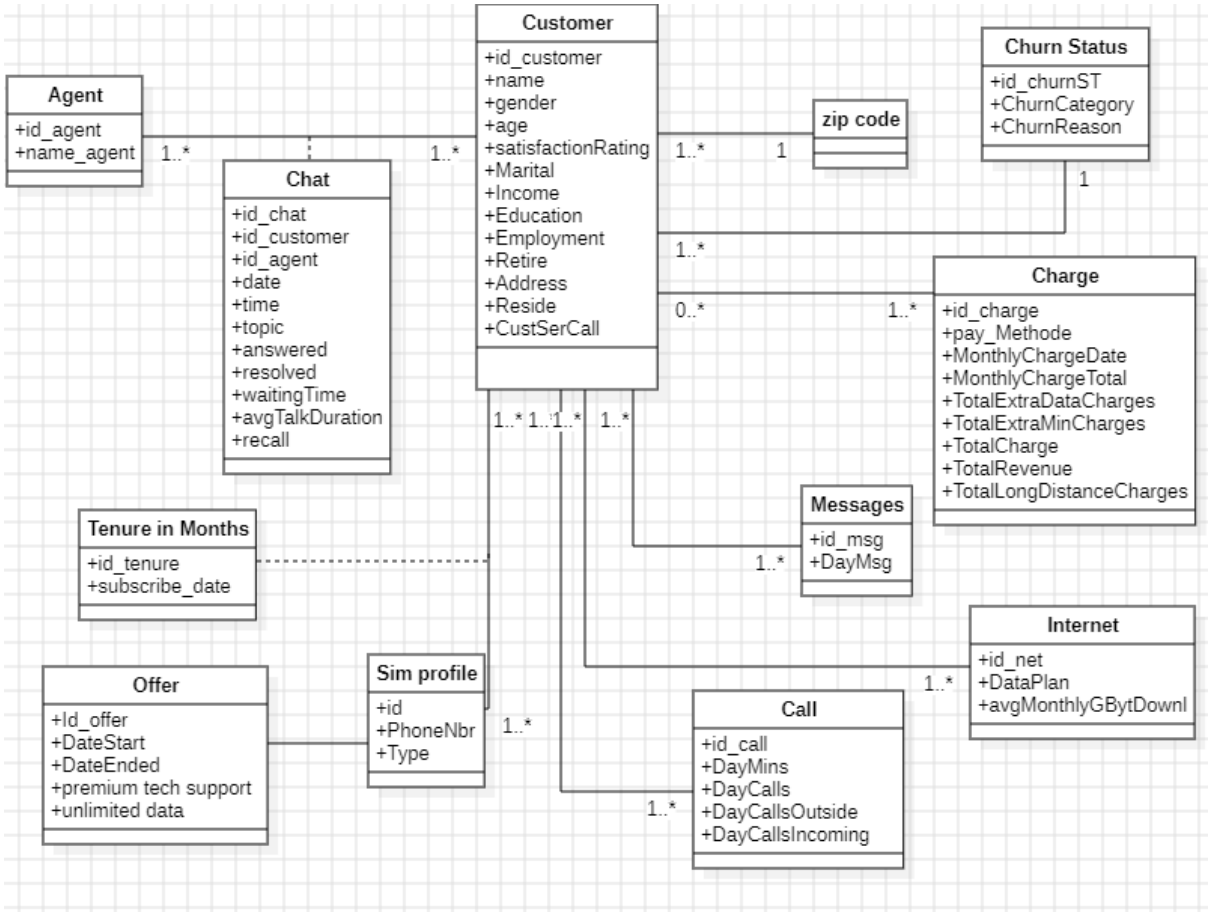


Figure 3.17: Class Diagram of customer management system for ETL process.

8. Conclusion

In this chapter, we presented the state of the art of works that proposed customer journey analysis for mobile companies in the world, then introduced UML language we have addressed before presenting in detail all the uses cases and sequence diagrams for the proposed system.

In the next chapter, we will present the implementation stage, including the environment, technical tools, and the application developed, and we will discuss the result.

CHAPTER 4

IMPLEMENTATION OF THE PROPOSED SYSTEM

1. Introduction

In this chapter, we examine the comprehensive environment that supports the development and deployment of our application. We discuss the hardware configuration and main programming tools used, emphasizing the programming language selected for its versatility and strong support community.

The chapter covers the main software tools and methodologies used, including data preprocessing and implementation of classification models. In addition, we introduce the application's various graphical interfaces and demonstrate its functionality through authentication, dashboards, and visualization of statistics and predictions.

2. Application Development Environment

2.1 Hardware

The main characteristics of the devices we utilized to test and implement our application:

- **PC 1:**
  Machine type: LENOVO THINKPAD.
  Processor: AMD Ryzen 7 PRO 5850U with Radeon Graphics 1.90 GHz.
  RAM:16,0 Go.
  Operating system: Windows 10 Professional (64-bit).
- **PC 2:**
  Type of machine: DELL.
  Processor: Intel(R) Core (TM) i5-4300U CPU @ 1.90GHz 2.50 GHz.
  RAM: 12,0 Go.
  Operating system: Windows 10 (64-bit).

2.2 Software

2.2.1 StarUML

StarUML is a software engineering tool for system modeling using the Unified Modeling Language, as well as Systems Modeling Language and classical modeling notations. It is published by MKLabs and is available on Windows, Linux, and MacOS [41].

2.2.2 Python

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented, and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library [39].

Guido van Rossum began working on Python in the late 1980s as a successor to the ABC programming language and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000. Python 3.0, released in 2008, was a major revision not completely backwards-compatible with earlier versions. Python 2.7.18, released in 2020, was the last release of Python 2.

Python consistently ranks as one of the most popular programming languages and has gained widespread use in the machine-learning community [40].

2.2.3 PyCharm

PyCharm is an integrated development environment (IDE) used for programming in Python. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems, and supports web development with Django. PyCharm is developed by the Czech company JetBrains.

It is cross-platform, working on Microsoft Windows, macOS, and Linux. PyCharm has a Professional Edition, released under a proprietary license, and a Community Edition, released under the Apache License. PyCharm Community Edition is less extensive than the Professional Edition [42].

2.2.4 Python Libraries

- **NumPy** [43]: NumPy is an abbreviation for “Numerical Python. " It is an open-source library in the Python language. It is used for scientific programming in Python, particularly for programming in Data Science, engineering, mathematics, or science.
- **Pandas** [44]: The Pandas open-source software library is specifically designed for data manipulation and analysis in Python. It is powerful, flexible, and easy to use. The name “Pandas” is actually a contraction of the term “Panel Data” for data sets that include observations over multiple periods. This library was created as a high-level tool for analysis in Python.
- **Matplotlib** [45]: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.
    - Create publication-quality plots.
    - Make interactive figures that can zoom, pan, and update.
    - Customize visual style and layout.

- Export to many file formats.
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib.

- **SciPy** [46]: SciPy is a free and open-source Python library used for scientific computing and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering.

- **scikit-learn [**47]: Scikit-learn is a key library for the Python programming language that is typically used in machine learning projects. Scikit-learn is focused on machine learning tools, including mathematical, statistical and general-purpose algorithms that form the basis for many machine learning technologies. As a free tool, Scikit-learn is tremendously important in many different types of algorithm development for machine learning and related technologies.
- **Tkinter** [48]: Tkinter is a Python binding to the Tk GUI toolkit. It is the standard Python interface to the Tk GUI toolkit, and is Python's de facto standard GUI. Tkinter is included with standard Linux, Microsoft Windows and macOS installs of Python. The name Tkinter comes from Tk interface. Tkinter was written by Steen Lumholt and Guido van Rossum, then later revised by Fredrik Lundh. Tkinter is free software released under a Python license.

3. Methodology

3.1 Data pre-processing

Data preprocessing, which involves organizing, converting, and cleaning the data before analysis, is an essential stage in data analysis. It aims to eliminate outliers and missing values, enhance data quality, and make sure the data is in an analytically-friendly format. Analysts can improve the precision and dependability of their results and extract valuable insights from the data by executing data preparation.

- **check for missing values :**

We write the following code (Figure 4.1) to show the missing data. The result is shown in Figure 4.2.

```
# Check for missing values
print(data.isna().sum())
```

Figure 4.1: Check for missing values

```
Gender                              0
Age                                 0
Married                             0
Tenure in Months                    0
Offer                               0
Internet Service                    0
Avg Monthly GB Download             0
Contract                            0
Paperless Billing                   0
Payment Method                      0
Total Charges                       0
Total Refunds                       0
Total Extra Data Charges            0
Total Long Distance Charges         0
Total Revenue                       0
Customer Status                     0
dtype: int64
```

Figure 4.2 : check for missing values

- **Convert Columns to Float by Replacing Commas with Dots using code shown in Figure 4.3.**

```
# Convert columns to float by replacing commas with dots and casting
float_columns = ['Total Charges', 'Total Refunds', 'Total Long Distance Charges', 'Total Revenue']
for column in float_columns:
    data[column] = data[column].str.replace(',', '.').astype(float)
```

Figure 4.3 : Convert Columns to Float by Replacing Commas

- **Binary Conversion for Specific Columns (Figure 4.4):**

```
# Binary conversion for specific columns
binary_columns = ['Married', 'Internet Service', 'Paperless Billing']
for column in binary_columns:
    data[column] = data[column].replace({'No': 0, 'Yes': 1}).astype(int)
# Binary conversion for Gender column
data['Gender'] = data['Gender'].replace({'Male': 0, 'Female': 1}).astype(int)
```

Figure 4.4 : Binary Conversion for Specific Columns code

- **Convert other categorical columns to numerical (Figure 4.5)**

```
# One-hot encode categorical variables
categorical_columns = ['Offer', 'Contract', 'Payment Method']
data = pd.get_dummies(data, columns=categorical_columns, drop_first=True)
```

Figure 4.5 : Convert other categorical columns to numerical code

- **Split the Dataset into Features and Target (Figure 4.6)**

```
# Split the data into features (X) and target (y)

X = data.drop( labels: "Customer Status", axis=1)
y = data["Customer Status"]
```

Figure 4.6 : Split the Dataset into Features and Target code

3.2 Selected classification Models

For predicting customer status in the customer journey analysis, we selected two classification models: Decision Trees and K-Nearest Neighbors (K-NN) (For more details see 9.1.1 and 9.1.2 in chapter 2). These models have demonstrated their effectiveness and yielded strong results in prediction tasks.

4. Presentation of our application

The chapter ends with an extensive presentation of our desktop application, including its dashboard, statistical data visualization, and user authentication graphical interfaces. The application gives users actionable insights by offering interactive and intuitive features for data analysis and prediction visualization.

We will present the different interfaces of our application and explain the utility of each of them.

4.1 Authentication interface

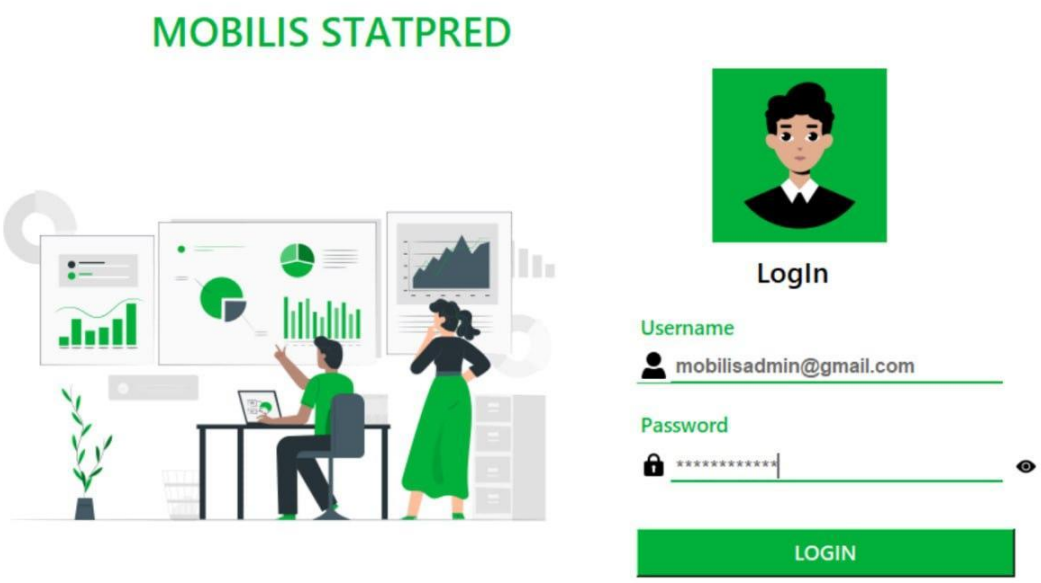


Figure 4.7 : Login Interface

4.2 Dashboard interface

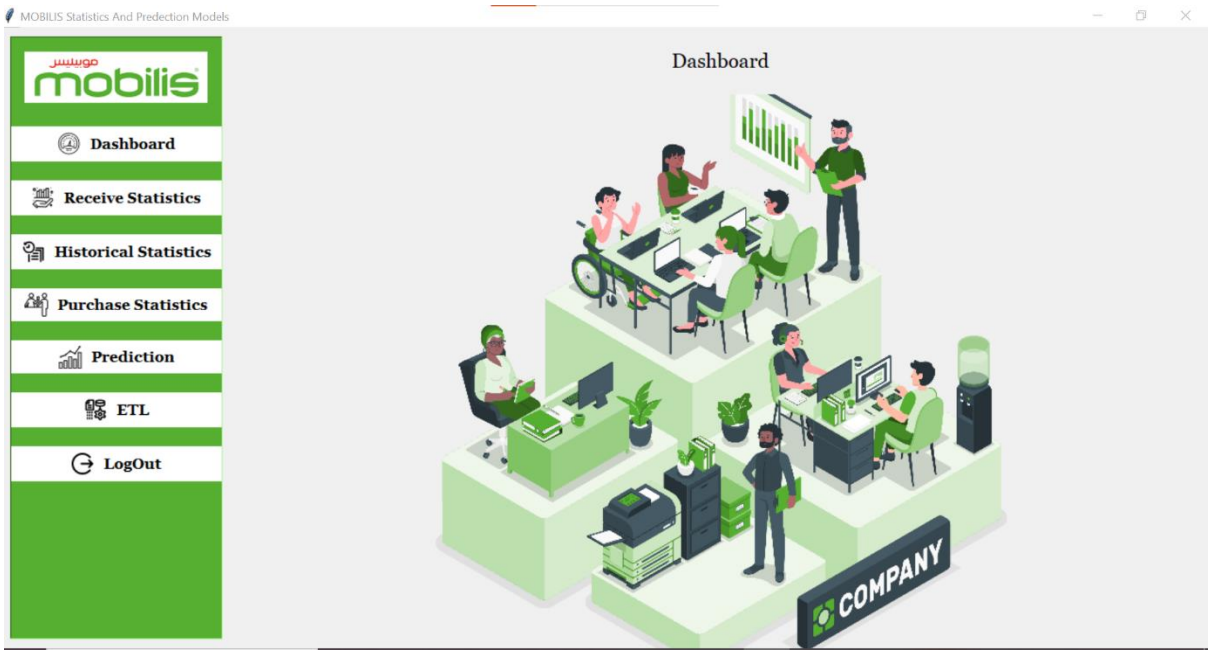


Figure 4.8 : Dashboard Interface

4.3 Statistics visualization

4.3.1 Receive data statistics

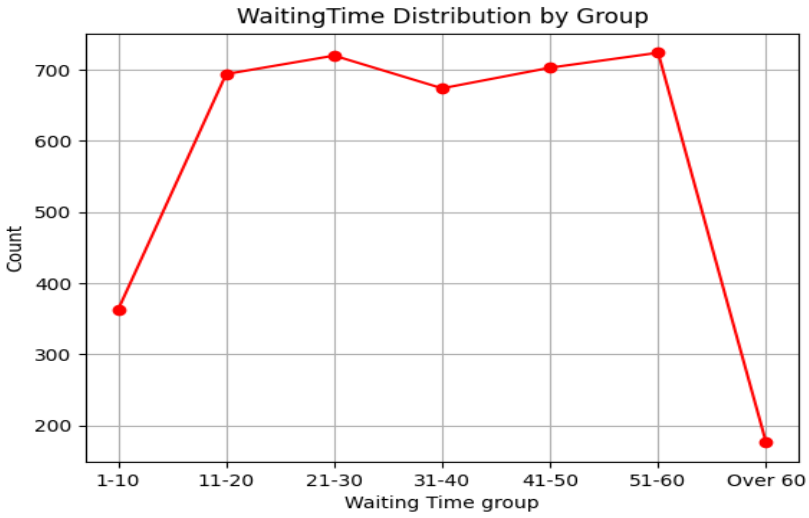


Figure 4.9 : Statistic visualization interface of waiting time

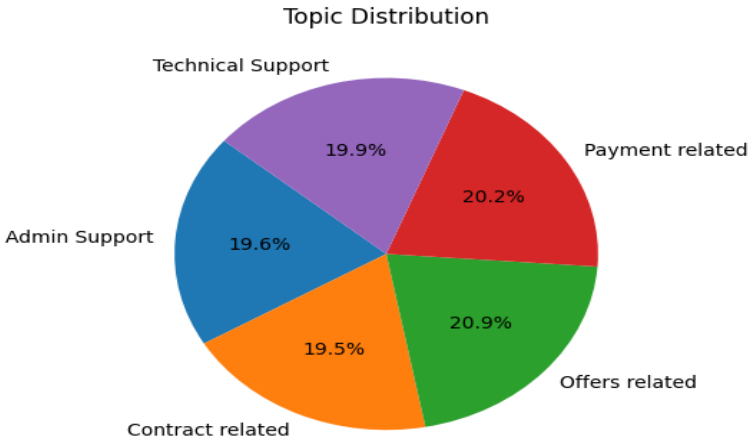


Figure 4.10 : Statistic visualization interface of Topic

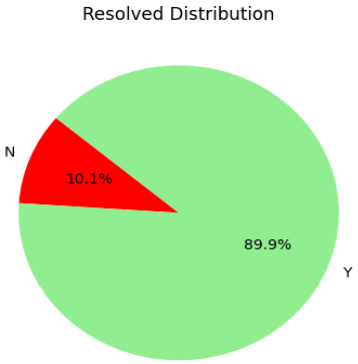


Figure 4.11 : Statistic visualization interface of Resolved

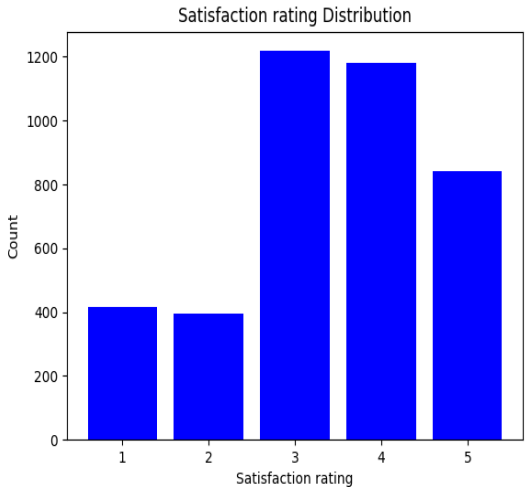


Figure 4.12 : Statistic visualization interface of Satisfaction rating

4.3.2 Purchase data statistics

➢ **By Age :**

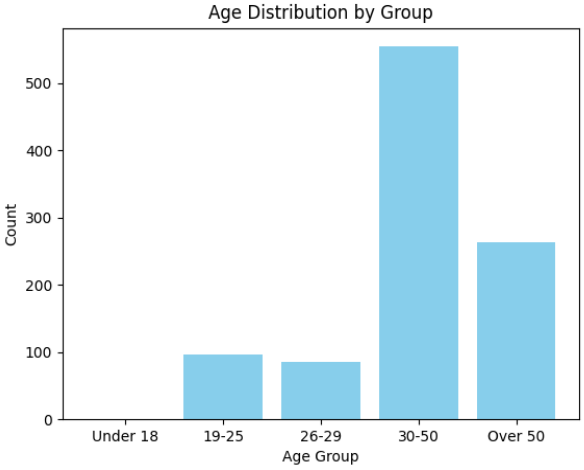


Figure 4.13 : Statistic visualization interface of Age in bar

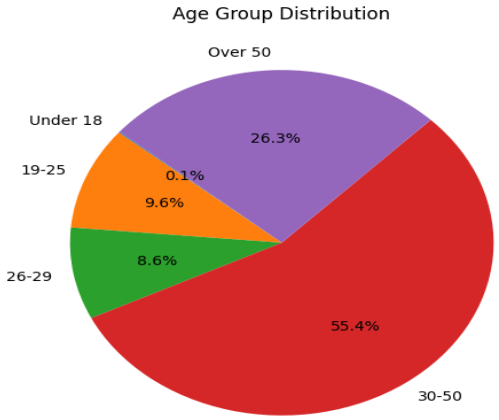


Figure 4.14 : Statistic visualization interface of Age in pie

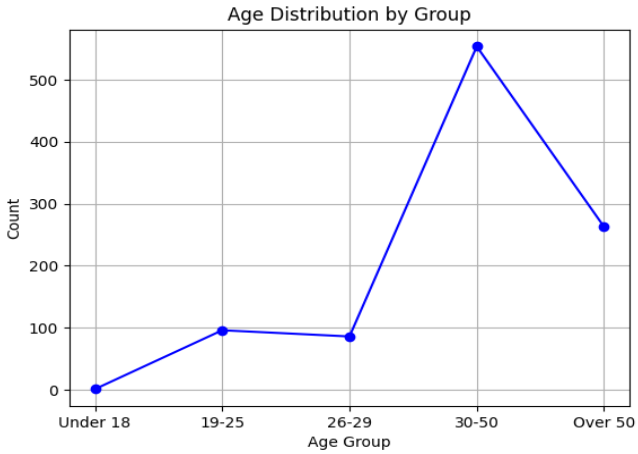


Figure 4.15 : Statistic visualization interface of Age in plot

- **By Gender :**

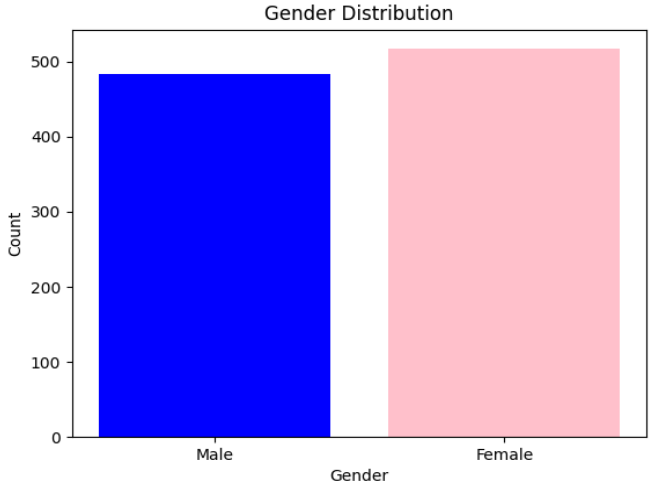


Figure 4.16: Statistic visualization interface of Gender in bar

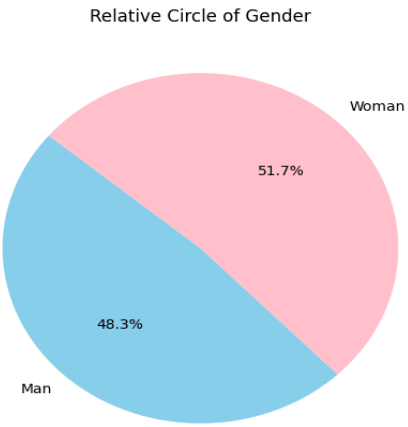


Figure 4.17: Statistic visualization interface of Gender in pie

➢ **By Income :**

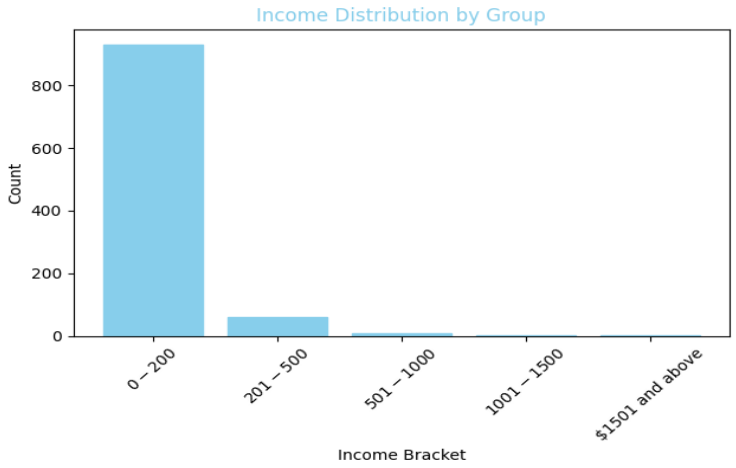


Figure 4.18 : Statistic visualization interface of Income

4.3.3 Historical data statistics

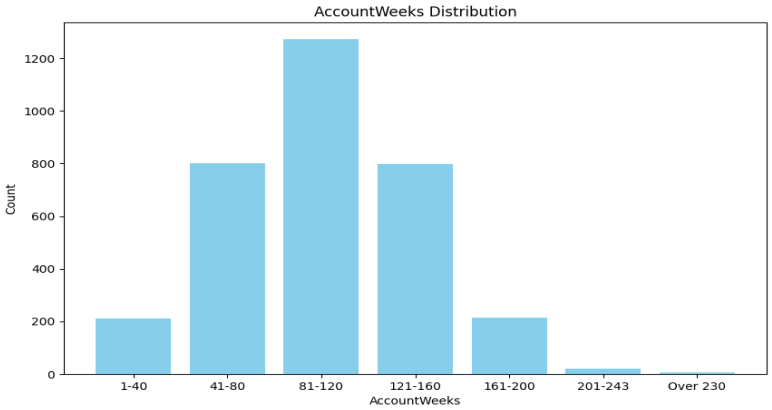


Figure 4.19: Statistic visualization interface of AccountWeeks

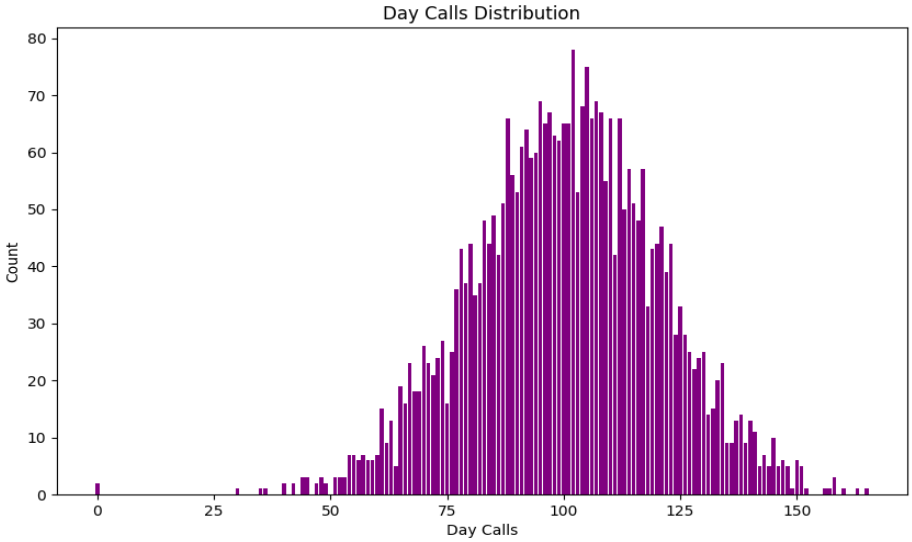


Figure 4.20:Statistic visualization interface of Day Calls

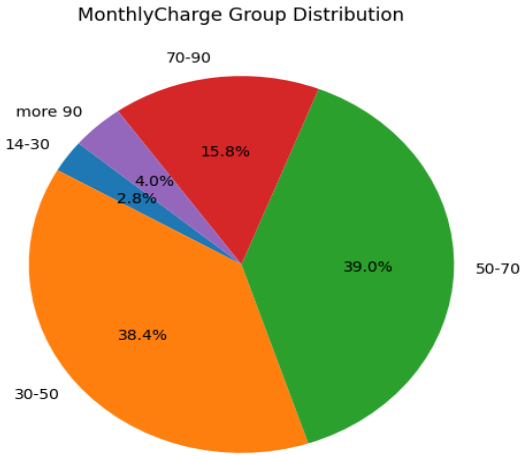


Figure 4.21:Statistic visualization interface of Monthly Charge

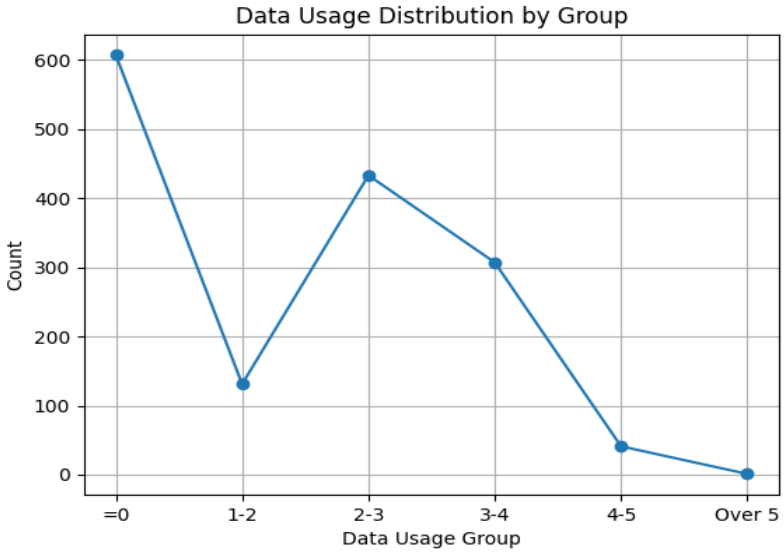


Figure 4.22: Statistic visualization interface of Data Usage

4.4 Visualization of Prediction results

4.4.1 KNN classification model

```
Accuracy: 0.75
Classification Report:
              precision    recall  f1-score   support

     Churned       0.59      0.58      0.59       380
      Joined       0.38      0.30      0.34        76
      Stayed       0.83      0.85      0.84       953

    accuracy                           0.75      1409
   macro avg       0.60      0.58      0.59      1409
weighted avg       0.74      0.75      0.74      1409

Confusion Matrix:
[[222  10 148]
 [ 37  23  16]
 [119  27 807]]
```

Figure 4.23: Accuracy and classification report using K-NN

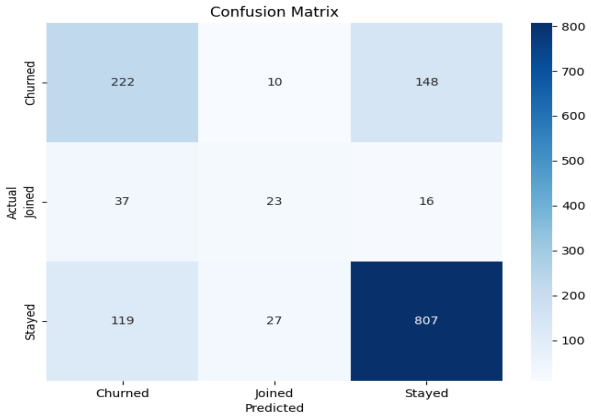


Figure 4.24: Confusion Matrix of K-NN

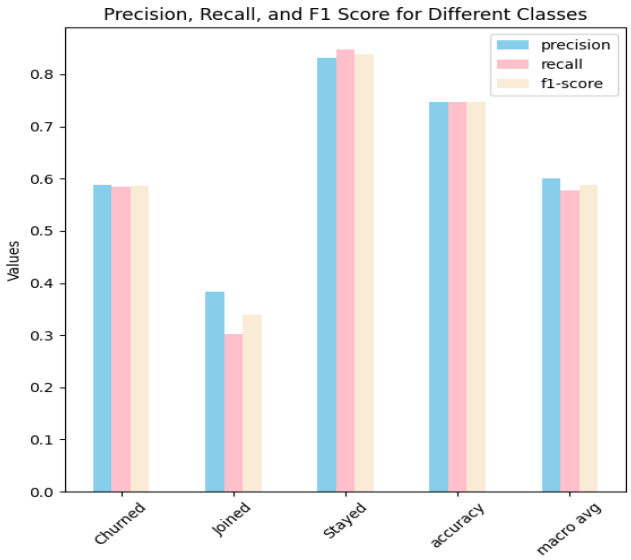


Figure 4.25:precision, recall, and F1 score plot

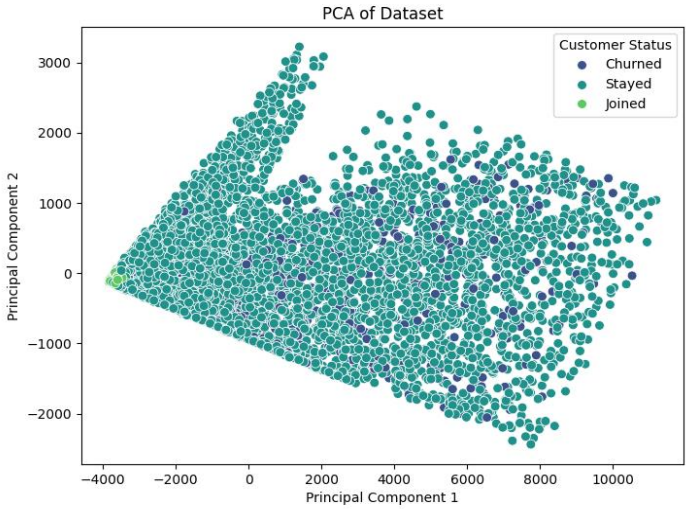


Figure 4.26: PCA for 2D visualization

4.4.2 Decision Tree classification model

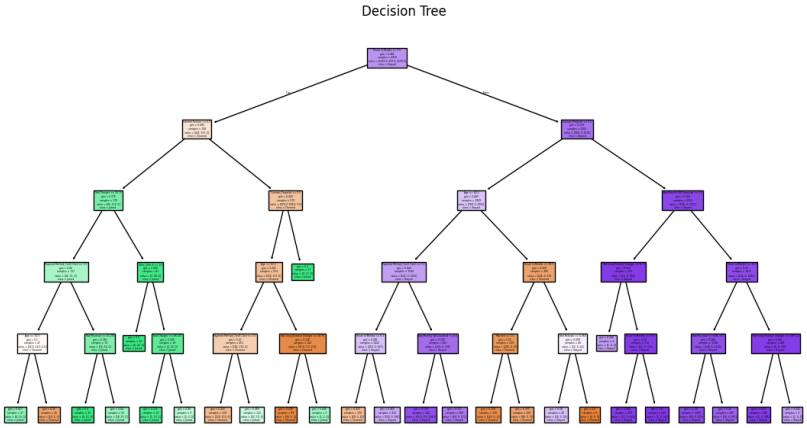


Figure 4.27: Structure of Decision Tree for customer status prediction

```
Accuracy: 0.81
Classification Report:
              precision    recall  f1-score   support

     Churned       0.69      0.45      0.55       544
      Joined       0.62      0.63      0.62       127
      Stayed       0.85      0.96      0.90      1442

    accuracy                           0.81      2113
   macro avg       0.72      0.68      0.69      2113
weighted avg       0.79      0.81      0.79      2113

Confusion Matrix:
[[ 245   49  250]
 [  47   80    0]
 [  61    0 1381]]
```

Figure 4.28: Accuracy and classification report using Decision trees

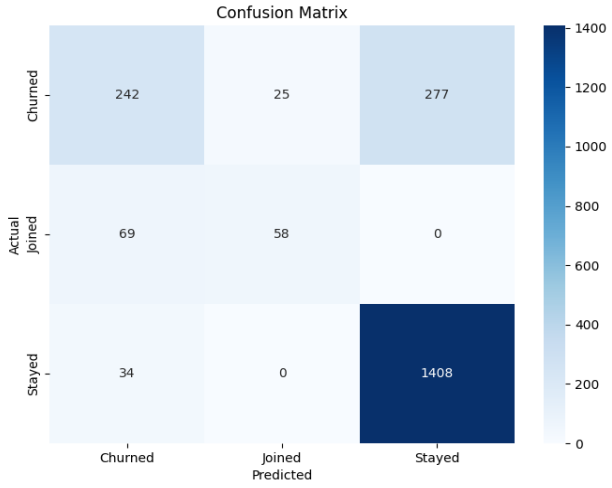


Figure 4.29 : Confusion Matrix of Decision Tree

5. Conclusion

This chapter details the development environment and tools used for our application, starting with the hardware specifications of the test machines, and the importance of Python and its libraries.

We introduced the methodology of data preprocessing, focusing on steps such as handling missing values and transforming categorical data. This chapter also described the classification models used: K-Nearest Neighbors (K-NN) and Decision Trees and their evaluation metrics.

Finally, we presented our desktop application and showed the user interface for authentication, dashboards, and statistical visualization. This chapter provides a comprehensive overview of the technical principles and implementation strategies of our application.

GENERAL CONCLUSION

In conclusion, our theme highlights the critical importance of analyzing the customer journey within a mobile phone company. A deep understanding of customer interaction and expectations allows mobile companies to develop effective strategies to attract and retain their customers.

Predictive models such as K-Nearest Neighbors (K-NN) and Decision Trees can be used in combination with data visualization tools to provide valuable insights into customer behavior and preferences. These technologies not only improve service quality, but also optimize business operations by providing actionable information.

The integration of ETL (Extract, Transform, Load) processes enhances the ability of enterprises to organize and analyze data efficiently. As a result, businesses can better meet the evolving needs of consumers and provide a personalized and engaging experience.

Ultimately, mobile phone companies that have successfully integrated these tools and approaches into their overall strategy will be put in a better position to gain significant competitiveness in the ever-changing market. Customer retention is a key strategic goal to ensure the growth and sustainability of this dynamic sector, along with the acquisition of new users.

BIBLIOGRAPHY

[1] Benmessaoud, M. & Gacem, T. (2020). Conception Et Réalisation D'un Système De Prédiction De L'attrition Client Pour Mobilis [Autre, École Supérieure En Informatique - Sidi Bel Abbès].

[2] Customer journeys: a systematic literature review Asbjørn Følstad, Knut Kvale (2018) · Journal of Service Theory and Practice.

[3] A quick guide to navigating the 5 stages of customer journey. (n.d.). https://www.insightsforprofessionals.com/marketing/customer-experience/understanding-customer-journey-stages.

[4] Team, A. E. C. (n.d.). The customer journey — definition, stages, and benefits. https://business.adobe.com/blog/basics/customer-journey.

[5] Skill Zone Ltd, https://www.skillzone.net. (n.d.). Definition: customer. Association for Qualitative Research (AQR). https://www.aqr.org.uk/glossary/customer.

[6] https://www.gartner.com/en/marketing/glossary/customer-loyalty.

[7] Siddiqe, R. (2024, February 5). Customer Satisfaction: Definition, Importance & Examples. Fluent Support. https://fluentsupport.com/customer-satisfaction/.

[8] Havir, D. (2016). Customer Experience Management Overview. ResearchGate. https://www.researchgate.net/publication/335566243_Customer_Experience_Management_Overview.

[9] Lemon, K. N., and Verhoef, P. C. (2016). Understanding Customer Experience throughout the Customer Journey. Journal of Marketing.

[10] Tueanrat, Y., Papagiannidis, S., Alamanos, E. (2021), "Going on a journey: A review of the customer journey literature", Journal of Business Research, 125 (C), 336-353.

[11] Author version of a manuscript published in Journal of Service Theory and Practice, vol. 26, Issue 6, pp. 840-867 (2016). DOI: 10.1108/JSTP-05-2015-0111 © Emerald Group Publishing Limited.

[12] https://www.questionpro.com/blog/fr/analyse-du-parcours-du-client/.

[13] du Toit, Gerard, Rob Markey, Jeff Melton, and Frédéric Debruyne. "Running the Business through Your Customer's Eyes." Bain.com, February 9, 2017.

[14] https://telecoms.adaptit.tech/blog/what-is-customer-analytics-in-telecoms/.

[15] https://fr.wikipedia.org/wiki/Mobilis

[16] Attaoua, S. & Guesmia, C. & Supervisor Mehenni, T. (2022). A Multilingual Chatbot For Supporting Mobile Companies Complaints. Case Study: Atm Mobilis Of Algeria. [Mémoire de Master, Université Mohamed Boudiaf - M'sila].

[17] Data Mining and Data Warehousing, Principles and Practical Techniques, pp. 17 - 27

DOI https://doi.org/10.1017/9781108635592.003, Publisher: Cambridge University Press, 2019

[18] L. Chaabane,« fusion et fouille de donneesguidees par les connaissances : application a l'analyse d'image»,Doctorat,Universite Mohamedkhider – biskra,2013.

[19] https://analytics.fr/definitions/data-mining/

[20] https://fr.wikiversity.org/wiki/Datamining/Historique

[21] https://analytics.fr/definitions/data-mining/

[22] MOUNA Azzeddine, Mémoire de magister en Informatique, titre : datamining distribue dans les grilles :approche règles d'association, Université des sciences et technologie d'Oran, 2012/2013

[23] Benyounes, S. (2016). Génération Des Règles D'association [Mémoire de Master, Université Mohamed Boudiaf - M'sila].

[24] The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011

:[25] Chloé-Agathe Azencott. Introduction au machine learning. Dunod, 2019. 8, 9, 10

[26] https://www.researchgate.net/profile/Santosh-Tondare/publication/275638874_Data_Mining_Task_Tools_Techniques_and_Applications/links/5540ceb00cf2718618daaf7f/Data-Mining-Task-Tools-Techniques-and-Applications.pdf

[27] https://www.researchgate.net/profile/Sunita-Dol-Aher/publication/266602921_Data_Mining_in_Educational_System_using_WEKA/links/57aee3f708ae95f9d8f15867/Data-Mining-in-Educational-System-using-WEKA.pdf

[28] OpenAI. (2024). ChatGPT (3.5) [Large language model]. https://chat.openai.com

[29] https://www.researchgate.net/profile/Santosh-Tondare/publication/275638874_Data_Mining_Task_Tools_Techniques_and_Applications/links/5540ceb00cf2718618daaf7f/Data-Mining-Task-Tools-Techniques-and-Applications.pdf

[30] Fareniuk, Yana, et al. "Customer churn prediction model: a case of the telecommunication market." Economics 10.2 (2022): 109-130.

[31] Wanikar, Padmanabh, et al. "Telco Customer Churn Prediction Using ML Models." International Journal of Intelligent Systems and Applications in Engineering 12.2 (2024): 644-653.

[32] Bazhenov, R., et al. "Applying machine learning methods to forecasting customer churn for a telecommunications company." CEUR Workshop Proceedings. 2021.

[33] Palimote, J., C. Aloy-Okwelle, and O. T. Olise. "Analyzing Customer Satisfaction Level on Telecommunication Usage Using Big Data." International Journal Of Engineering And Computer Science 12.09 (2023).

[34] López, Mirko Bruno Vela, et al. "Application of a Data Mining Model to Predict Customer Defection. Case of a Telecommunications Company in Peru." J. Wirel. Mob. Netw. Ubiquitous Comput. Dependable Appl 14 (2023): 144-158.

[35] Wassouf, Wissam Nazeer, et al. "Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study." Journal of Big Data 7.1 (2020): 29.

[36] Nguyen, Quang Dang, Khoa Van Nguyen, and Tatyana Sakulyeva. "An Analysis of Consumer Trends in the Telecommunications Markets of Russia and Vietnam." Journal of Telecommunications and the Digital Economy 9.3 (2021): 87-109.

[37] https://en.wikipedia.org/wiki/Unified_Modeling_Language

[38] https://www.lucidchart.com/pages/fr/langage-uml

[39] Airouche, K. & Alouane, K. & Mir, F. (2022). Conception Et Réalisation D'un Systéme

D'analyse Et De Prédiction De Ventes [Mémoire de Master, Université Abderrahmane Mira - Bejaia].

[40] https://fr.wikipedia.org/wiki/Python_(language)

[41] https://en.wikipedia.org/wiki/StarUML

[42] https://en.wikipedia.org/wiki/PyCharm

[43] https://datascientest.com/en/numpy-the-python-library-in-data-science

[44] https://datascientest.com/en/pandas-the-python-library

[45] https://matplotlib.org/

[46] https://en.wikipedia.org/wiki/SciPy

[47] https://www.techopedia.com/definition/33860/scikit-learn

[48] https://en.wikipedia.org/wiki/Tkinter,

الملخص:
تواجه شركة موبيليس للهواتف المحمولة تحديات كبيرة في الحفاظ على العملاء و جذب عملاء جدد بسبب التنافس الشديد ، فتعمل على التحليل الدقيق لرحلة العميل التي تبدأ من اكتشاف الخدمة وصولا إلى تطوير الولاء، فتحليلها يساعد الشركة في فهم سلوك العميل و تخصيص خدمات بشكل فعال لكل فئة من العملاء ، و من خلال هذا التحليل أيضا نستطيع التنبؤ بحالات الفقدان المحتملة للعملاء . يقوم بحثنا على جمع بيانات التفاعل مع العملاء و تحليل البيانات باستخدام تقنيات التحليل الاحصائي و التنقيب في البيانات لفهم سلوك العميل ، و قمنا بإنشاء نموذج يعتمد على التعلم الآلي و الذكاء الاصطناعي للتنبؤ برحلة العميل و الذي بدوره أثبت كفاءته في التنبؤ بسلوك العميل و تقديم توصيات لتحسين تفاعلهم مع الشركة ، بهذا يمكن لشركة موبيليس تحسين إستراتيجياتها التسويقية و خدمة العملاء ، مما يساهم في زيادة ولاء و تعزيز النمو المستدام.
الكلمات المفتاحية : رحلة العميل ، تحليل البيانات ، التنبؤ بسلوك العميل ، رضا العملاء ، شركة الاتصالات المحمولة ، ذكاء اصطناعي ، تحليل احصائي ، الاحتفاظ بالعملاء ، python , pycharm, K-NN , Decision Tree , data mining

Abstract:
Mobilis, the mobile phone company, faces significant challenges in retaining customers and attracting new ones due to intense competition. The company conducts a precise analysis of the customer journey, which starts from discovering the service and extends to developing loyalty. This analysis helps the company understand customer behavior and effectively customize services for each customer segment. Through this analysis, we can also predict potential customer churn. Our research involves collecting customer interaction data and analyzing it using statistical analysis and data mining techniques to understand customer behavior. We have developed a model based on machine learning and artificial intelligence to predict the customer journey, which has proven effective in forecasting customer behavior and providing recommendations to improve their interaction with the company. This enables Mobilis to enhance its marketing strategies and customer service, contributing to increased loyalty and sustainable growth.

Keywords: customer journey, data analysis, customer behavior prediction, customer satisfaction, mobile telecommunications company, artificial intelligence, statistical analysis, customer retention, Python, PyCharm, K-NN, Decision Tree, data mining.

Résumé :
Mobilis, la société de téléphonie mobile, est confrontée à des défis importants pour fidéliser ses clients et en attirer de nouveaux en raison de la concurrence intense. L'entreprise effectue une analyse précise du parcours client, qui part de la découverte du service et s'étend au développement de la fidélité. Cette analyse aide l'entreprise à comprendre le comportement des clients et à personnaliser efficacement les services pour chaque segment de clientèle. Grâce à cette analyse, nous pouvons également prédire le taux de désabonnement potentiel des clients. Notre recherche consiste à collecter des données d'interaction client et à les analyser à l'aide d'analyses statistiques et de techniques d'exploration de données pour comprendre le comportement des clients. Nous avons développé un modèle basé sur l'apprentissage automatique et l'intelligence artificielle pour prédire le parcours client, qui s'est avéré efficace pour prévoir le comportement des clients et fournir des recommandations pour améliorer leur interaction avec l'entreprise. Cela permet à Mobilis d'améliorer ses stratégies marketing et son service client, contribuant ainsi à une fidélisation accrue et à une croissance durable.

Mots-clés : parcours client, analyse de données, prédiction du comportement client, satisfaction client, société de télécommunications mobiles, intelligence artificielle, analyse statistique, rétention client, Python, PyCharm, K-NN, arbre de décision, data mining.