DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

Abderrehmane Mira University of Bejaia
Faculty of exact sciences
Department of computer science

THESIS

For the Degree of
Master of Computer science

Major: Advanced information systems (AIS)

THEME :

A Hybrid recommender System for Enhanced E-commerce Recommendations

Presented on: **04/07/2024**

Presented by:

Souha SACI and Salas MERZOUK

Before the jury composed of:

TAHAKOURT ZINEB	***MCB***	***President***
KESSIRA DALILA	***MAA***	***Examiner***
AKILAL KARIM	***MCB***	***Superviser***

2023 - 2024

Contents

Abbreviations

AI	Artificial Intelligence
CF	Collaborative Filtering
RS	Recommendation System
SVD	Singular Value Decomposition
TF-IDF	Term Frequency-Inverse Document Frequency
RMSE	Root Mean Square Error
MSE	Mean Squared Error
ML	Machine Learning
NN	Neural Network
SVM	Support Vector Machines
CB	Content-Based
KNN	k-Nearest-Neighbours
MAE	Mean Absolute Error
RL	Reinforcement Learning
LR	Linear Regression
CNN	convolutional neural network
DT	Decision Trees
RNN	Recurrent Neural Network
IT	Information Technology

List of Figures

List of Tables

Acknowledgments

We would like to express our deepest gratitude to God, whose blessings, and wisdom have been instrumental in our journey towards completing this Project . His grace has provided us with strength, perseverance, and clarity of mind throughout this endeavor.

We are also immensely thankful to our supervisor, **Mr. Akilal Karim**, for his invaluable support, mentorship, and encouragement during the development of this project. His expertise, constructive feedback, and dedication have been pivotal in shaping this work and guiding us towards academic excellence.

We would like also to extend our heartfelt gratitude to **Ms. Tahakourt Zineb** and **Ms. Kessira Dalila** for graciously accepting to evaluate our work. Your willingness to dedicate your time and expertise to review our research is deeply appreciated. Your insightful feedback and constructive criticism will undoubtedly contribute significantly to the enhancement of our work. Thank you for your valuable contributions and for being a part of this important academic journey.

Additionally, we extend our appreciations to our families, friends, and colleagues for their unwavering support, understanding, and encouragement during this challenging yet rewarding journey.

Finally, we would like to acknowledge the contributions of all those who have directly or indirectly assisted us in this endeavor. Your support and encouragement have been invaluable, and we are deeply grateful for your generosity and kindness.

Dedications

To my loving family,

*Your unwavering support, boundless love, and constant encouragement have been the bedrock upon which I have built my dreams. To my parents **DJAMEL** and **MALIKA**, who instilled in me the values of hard work and perseverance, and to my siblings **KARIM, HALIM, and SARAH**, who have always believed in me and stood by my side through every challenge, thank you for your endless patience and faith. This work is as much yours as it is mine.*

*To my beloved aunty **LOUIZA**, who raised me and left us too soon,*

Your love, care, and guidance have been the foundation of my life. You nurtured me, supported me, and believed in me when I needed it most. Your spirit continues to inspire me every day, and this work is dedicated to your memory. Though you are no longer with us, your legacy lives on in my heart and in everything I do.

*To my colleague **SALAS***

Your dedication, hard work, and creativity have been instrumental in bringing this project to fruition. Thank you for the countless hours of collaboration, for sharing your insights, and for your unwavering commitment to excellence. This accomplishment is a testament to our collective efforts.

*To my dearest **friends** and **cousins**,*

Your laughter, companionship, and encouragement have been my solace and strength. Your belief in me has been a guiding light, and for that, I am eternally grateful.

*To my **mentors** and **teachers**,*

Your wisdom, guidance, and invaluable lessons have shaped me into the person I am today. Thank you for challenging me to think critically, for inspiring me to push beyond my limits, and for nurturing my passion for learning. Your influence is evident in every page of this work.

*To all the **dreamers** ,*

May this work serve as a source of inspiration and motivation. Pursue your passions with relentless determination, embrace every challenge as an opportunity to grow, and never lose sight of your dreams. Remember that with perseverance and hard work, anything is possible.

To anyone who was there for me with an effort or encouragement and whose name I did not mention, you are in my heart .

With deepest gratitude and affection

Souha

Dedications

This thesis is dedicated to the most important people in my life who have provided unwavering support and love throughout my academic journey.

First and foremost, I would like to thank my parents, ***Aicha*** *and* ***Hamid****. Your endless love, encouragement, and sacrifices have been the driving force behind all my achievements. Thank you for believing in me and nurturing my passion for learning.*

To my brother, ***Aris****, thank you for your companionship, wisdom, and for always being my greatest cheerleader. Your support has been invaluable.*

I am also deeply grateful to my beloved grandmother, ***Sadia****. Your love and prayers have been a source of strength and inspiration for me.*

A heartfelt thank you to my colleague, ***Souha****, for your support, collaboration, and for making the journey more enjoyable. Your camaraderie has been a great source of motivation.*

To my dear ***friends****, thank you for the laughter, the late-night study sessions, and for being there through thick and thin. Your friendship has been a vital pillar of support.*

Finally, I want to extend my sincere gratitude to all my ***teachers*** *who have imparted their knowledge, guided me, and shaped my academic path. Your dedication and commitment to education have played a crucial role in my success.*

This thesis is a testament to the collective support, love, and encouragement from all of you. Thank you from the bottom of my heart.

To anyone who was there for me with an effort or encouragement and whose name I did not mention, you are in my heart.

With deepest gratitude and affection

Salas

General Introduction

The rapid expansion of online platforms and the multitude of options available to consumers in the field of electronic commerce have highlighted the importance of recommendation systems for customizing user experiences and maximizing sales. Their primary objective is to satisfy the customer by offering products, services, or content that best match their tastes and needs. By analyzing data such as purchase histories, ratings, and browsing behaviors, these systems can anticipate users' expectations and enhance their overall experience. Additionally, recommendation systems aim to provide a diverse range of choices, allowing users to discover new options they might not have considered otherwise. This ability to surprise and satisfy customers helps to increase their engagement and loyalty to the platform or service using these systems. Conventional recommendation systems, like collaborative filtering or content-based filtering, encounter issues like sparse data, cold start problem, scalability challenges, restricted diversity, and absence of context awareness. That's why a new method known as hybrid recommender systems has emerged to address these issues.

Hybrid recommendation systems merge different recommendation approaches like collaborative filtering and content-based filtering to maximize their strengths and minimize their weaknesses, resulting in more precise and diverse recommendations.

Recommendation systems in e-commerce provide customized recommendations, enhancing user shopping experience and boosting sales [20](Burke, 2002). These systems use different algorithms and methods to create suggestions relying on user preferences, actions, and historical data. Content-based recommendation suggests items based on their attributes and descriptions, matching them with user interactions. To clarify, imagine a customer who buys organic food products often and gives them high ratings. A content-based recommendation system could recommend additional organic items or products within the same category to improve the shopping experience.

Collaborative filtering is also used as an recommendation method which focuses on comparing the similarities or differences between users or items. First, User-based collaborative filtering recommends items based on users' behaviors compared to other users, while item-based collaborative filtering recommends items by comparing their behaviors with other items [118](Zhang et al., 2019).

To understand, think about an internet bookshop where customers who bought fantasy novels have also displayed a liking for science fiction novels. Item-based collaborative filtering algorithms recommend science fiction books to users who purchased fantasy novels, by analyzing the correlation between these genres and the shopping habits of similar users. Incorporating a hybrid recommendation system that combines collaborative filtering to identify similar users and content-based recommendations to improve suggestions through item analysis. Using like these combinations hybrid recommendations can provide better suggestions, combining different methods' advantages and overcoming their drawbacks.

In this work, we introduce a hybrid approach which will make it possible to resolve some difficulties faced in recommendation systems following a plan of 3 chapters: Before delving into the main topic of our research , we start in chapter 1 , by introducing some general concepts like artificial intelligence, machine learning and E-commerce sector.

Next, in chapter 2, we provide a study of the state of the art.Indeed after studying articles and research papers and try to understand the strength and short-comings of the available approaches.

In chapter 3, we describe the proposed approach by representing the methods and algorithms used , the implementation,a comparison with some other techniques and an evaluation.

We conclude with a conclusion and future perspectives

Chapter 1

Introduction to e-commerce and artificial intelligence

1.1 Introduction

In the first part of this chapter, in the first part we will try to give important insights about the E-commerce sector (definition, history, evolution, forms and last statistics).Afterwards, in the second part, we will discuss the artificial intelligence domain and Machine learning including its different types, and algorithms.

1.2 E-commerce

Nowadays, leaving home has become not necessary to do any-type of shopping , consumers can make all their purchases from theirs homes thanks to the Internet and online shopping (also called E-commerce) that represents an effective way to save money and time.

1.2.1 Definition of E-commerce

E-commerce or electronic commerce a term representing and referring to buying and selling products and exchanging services using Internet and which can be done through computers, smartphones, tablets, etc.

We can also define E-commerce as the sale or purchase using methods specifically designed for order placement or receipt. Even though goods or services are ordered electronically, payment and delivery do not necessarily have to occur online. An electronic business transaction can occur between businesses, households, individuals, governments, and other public or private organizations [41].

1.2.2 History of E-commerce

The development of e-commerce has been driven by significant technological advancements and shifts in consumer behavior. In the early 1990s, the emergence of the World Wide Web paved the way for online communication and commerce. Prior to this, the minitel system introduced limited online services, including early forms of e-commerce.

A pivotal moment came in 1994 with the first remote credit card transaction when Phil Brandenberger bought a Sting album for $12.48 in the USA, demonstrating the potential of secure online transactions and laying the foundation for e-commerce growth.

During the 2000s, e-commerce grew quickly because many people started using broadband internet where companies like Amazon in the United States were important in making online shopping popular.

Then, the introduction of smartphones in the 2010s marked a shift towards mobile commerce (also called m-commerce). This allowed consumers to shop, pay, and use online stores directly from their phones, making online shopping more convenient.

Today, e-commerce has expanded beyond geographical limitations due to advancements like data centers and international rule standardization. This has made global purchasing and trade smooth and accessible to consumers everywhere. Modern e-commerce platforms offer a wide range of products, secure payment options, personalized shopping experiences, and fast delivery services, making online shopping convenient and efficient.

1.2.3 Evolution of E-commerce

The statistics reveal that global e-commerce revenue surpassed $2.382 billion in 2017. In 2018, the sector continued to grow, with global revenue estimated at over $2.928 billion. This upward trend persisted in 2019, with revenues exceeding $3.535 billion. In 2020, there was a significant increase compared to previous years, reaching $4.206 billion. The year 2021 saw a revenue of $4.927 billion, marking a 15% increase. In 2022, revenues reached $5.695 billion, and finally, in 2023, e-commerce revenue hit a staggering $6.542 billion. These numbers illustrate how online shopping has revolutionized the way we purchase goods and highlight the growing importance of this sector in our daily lives.

This rapid growth is attributed to various factors such as :

- The easy access to the Internet.
- The improved secure payment technologies.
- The diversification of offers and services available on online commerce platforms.

These developments ensure a positive future for the industry and an optimistic continuation.

1.2.4 Forms of E-commerce

E-commerce platforms use six distinct forms to connect with customers online and streamline transactions. These strategies allow businesses to reach their target audience while ensuring smooth and convenient shopping experiences.

1. **Business to Business (B2B)** Refers to buying and selling products between businesses without intermediaries, sharing information in digital form.(Alibaba, Amazon Business)

    - *Example:* A wholesale distributor of electronic components sells large quantities of products to a computer manufacturing company through an online portal.

2. **Business to Consumer (B2C)** Involves online sales to end buyers, widely used in countries like Norway, Denmark, Sweden, the UK, and the US, primarily for IT products, clothing, and digital items. (Amazon, eBay)

    - *Example:* An individual purchases a new smartphone directly from an electronics retailer's website like Amazon.

3. **Business to Administration (B2A)** Involves electronic transactions between a business and an administration, such as government procurement and licensing procedures.(e-Achats)

    - *Example:* A software company provides an online portal for a government agency to purchase software licenses and manage service contracts.

4. **Consumer to Consumer (C2C)** Platforms facilitating direct transactions between consumers, such as eBay, where transactions occur between individuals.(eBay, Leboncoin)

    - *Example:* One individual sells a used bicycle to another individual through an online marketplace like eBay or Craigslist.

5. **Consumer to Business (C2B)** Involves individual consumers offering products or services to businesses, such as freelance work or personalized products.(Upwork)

- *Example:* A graphic designer offers logo design services to a small business through a freelancing platform like Upwork.

6. **Consumer to Administration (C2A)** Encompasses electronic transactions between individuals and public administrations, such as paying taxes online or renewing government-issued licenses.(IRS.gov)

    - *Example:* A citizen renews their driver's license online through a government portal instead of visiting a physical office.

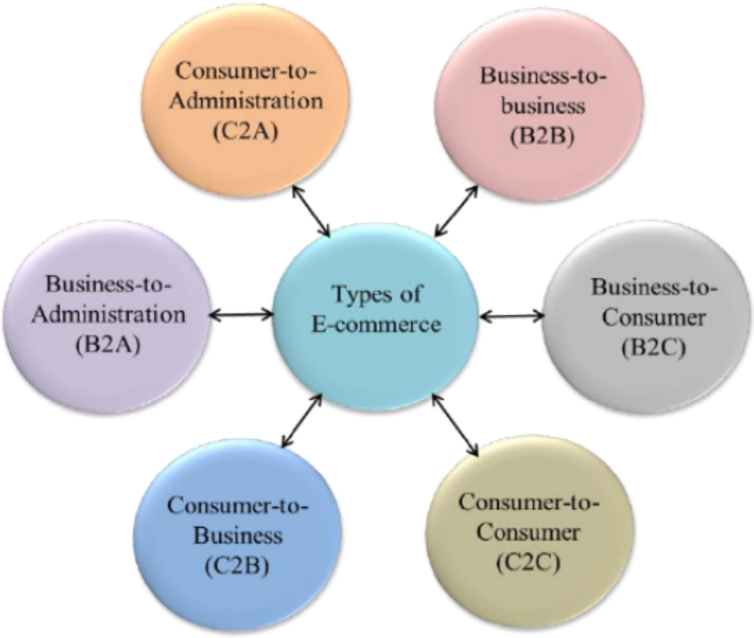

Consumer-to-Administration (C2A)
Business-to-business (B2B)
Business-to-Administration (B2A)
Types of E-commerce
Business-to-Consumer (B2C)
Consumer-to-Business (C2B)
Consumer-to-Consumer (C2C)

Figure 1.1: Types Of E-commerce [43].

1.2.5 Advantages of E-commerce

E-commerce brings several benefits that make shopping and selling easier and more effective.

- Cost-effective and efficient: Online stores don't have to pay for things like store rent and utilities, which can be expensive for physical stores. This means they can offer products at lower prices. As a result, buying and selling online is quicker and smoother because there are fewer overhead costs involved.
- Global reach without geographical limitations: With e-commerce, you can shop from anywhere in the world, and businesses can sell their products to customers

worldwide. This expands your choices and creates more opportunities for businesses to reach a larger audience.

- Convenience for consumers and businesses: Online shopping is simple and adaptable, allowing you to shop whenever it suits you. For businesses, this translates to increased sales and satisfied customers.

- Diverse product offerings and competitive pricing: Online stores offer a wide variety of products from various sellers, giving you a greater selection. The competition among sellers often results in better prices for shoppers.

- Enhanced communication and customer service: You can talk to businesses instantly through chat or email, getting help whenever you need it. This makes your shopping experience smoother and more enjoyable.

1.2.6 Recommendation engines

There are many recommendation motors or platforms , we cite some of them.

1.2.6.1 Amazon

Amazon is an American e-commerce company based in Seattle, founded by Jeff Bezos in July 1994.

Initially focused on selling books online, Amazon later expanded its scope to include various other domains such as computer science, cinematography, health clinics, and more.

Today, Amazon is a dominant force in internet commerce, representing approximately 40% of online sales.

Its revenue continues to grow steadily, reaching $574.8 million dollars for the entire year of 2023.

1.2.6.2 Jumia

Jumia is an African e-commerce company based in Lagos, Nigeria, founded by Tunde Kehinde and Raphael Afaedor in 2012.

Initially focused on online retail in Nigeria,Jumia later expanded its operations to serve customers in several African countries, offering a wide range of products including electronics, fashion, home goods, and more.

Even if Jumia's revenue for 2023 ($186.6 million dollars) also decreased by 8.3% compared to that of 2022 ($203.3 million) Jumia is a leading player in the African e-commerce market, representing a significant share of online sales across the continent.

1.2.6.3 Alibaba

Alibaba is a Chinese multinational e-commerce company founded by Jack Ma in April 1999(Hangzhou, China).

Initially as a platform for business-to-business (B2B) transactions, before diversifying its services to include business-to-consumer (B2C) and consumer-to-consumer (C2C) sales, cloud computing, digital media, and entertainment.

Alibaba is one of the world's largest e-commerce companies, with a significant presence over the world.

Its platforms(Taobao, Tmall, and AliExpress) are facilitating a wide range of transactions across various industries.

Alibaba's revenue has seen consistent growth over the years, with its financial reports reflecting substantial earnings from e-commerce activities ,In 2024 alone, Alibaba reported a revenue of 448 million (US$3,802 million), playing an important role in the global e-commerce landscape.

1.2.6.4 eBay

eBay is an American multinational e-commerce corporation founded by Pierre Omidyar in September 1995 (San Jose, California, USA).

Initially started as an online auction platform, eBay has since diversified to include fixed-price sales, making it a marketplace for both consumer-to-consumer (C2C) and business-to-consumer (B2C) transactions.

eBay is one of the world's leading e-commerce companies, facilitating millions of transactions daily across a wide array of product categories.

Its platform enables users to buy and sell a variety of goods and services worldwide, leveraging its auction-style sales, buy-it-now options, and classified advertisements.

EBay's revenue for the twelve months ending March 31, 2024 was $10.158B, underscoring its significant role in the e-commerce industry.

1.2.6.5 Shopify

Shopify is a Canadian multinational e-commerce company founded by Tobias Lütke, Daniel Weinand, and Scott Lake in 2006 (Ottawa, Canada).

Initially conceived as an online store for snowboarding equipment, Shopify evolved into a comprehensive platform that enables businesses to create and manage their own online stores.

Shopify is one of the world's leading e-commerce platforms, providing tools for businesses of all sizes to sell products online, in-store, and through various online channels.

Its platform includes customizable storefronts, payment processing, and a wide array of integrations and applications to enhance the e-commerce experience.

Shopify's revenue has experienced robust growth, reflecting its expanding merchant base and the increasing adoption of e-commerce solutions globally. Shopify's revenue for the twelve months ending March 31, 2024 was $7.413B cementing its position as a key player in the e-commerce ecosystem.

These huge numbers call for automation in general, and computers to improve them. In the next section, we discuss a computer science field that can help , namely artificial intelligence

1.3 Artificial intelligence and ML basic concepts

In this section we will introduce the basic concepts that are important to delve in artificial intelligence and machine learning domains.

1.3.1 Definition of Artificial intelligence

We can introduce Artificial intelligence as the capacity of machines and computer programs to accomplish complex tasks by generating from the intelligence of the human being mathematical models or algorithms to be compiled to help the machine to:

- Process data.
- Learn.
- Take decision.

It is based on several aspects (natural language processing, pattern recognition, machine learning, etc.)

The term "artificial intelligence", ("AI" in English: artificial intelligence),was created by John McCarthy who defines it as: "The science and engineering of making intelligent machines, especially intelligent computer programs." It is related to the similar task of using computers to understand human intelligence, but AI should not be limited to methods that are biologically observable. » [38]

Marvin Lee Minsky , who is one of its creators, defines it as: "The construction of computer programs which undertake tasks which are, for the moment, accomplished more satisfactorily by human beings because they require high-level mental processes such as [55]:

- Perceptual learning
- Memory organization
- Critical reasoning

We focus on Machine Learning which represents our domain of research :

1.3.2 Definition of Machine Learning

Arthur Samuel defines Machine Learning as :" The ability of a computer to learn without having been explicitly programmed" [93].

Also We can define Machine Learning as a field of artificial intelligence that represents the scientific study and development of algorithms and statistical models that computer systems use to perform specific tasks without explicit instructions, relying on patterns and inference instead [36] and that allow computers to :

- Do things on their own by looking at examples.
- Learning from data or patterns.
- Make decisions based on this experience,

Machine learning no longer focuses on how to find abstract objects like a probability law for example, but focuses above all on the operational side, that is to say making decisions based on data by making as few errors as possible." [7].

1.3.3 foundational concepts in ML

1.3.3.1 Classification

Consists of predicting a class or categorizing a new element (e.g: predicting whether an email is spam or not) [79].

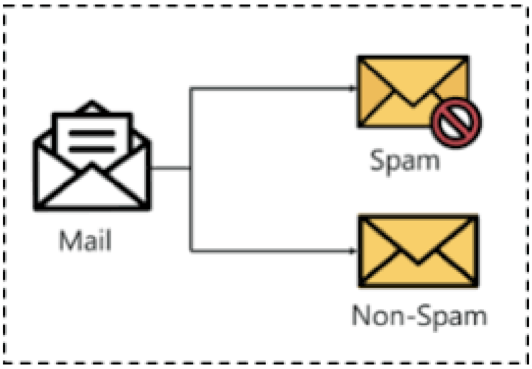


Figure 1.2: classification example [105]

Classification in machine learning involves choosing an appropriate model, training on labeled data, and evaluating the model's performance on unseen data to ensure that it can generalize correctly [10].

1.3.3.2 Regression

Regression is based on predicting a continuous numerical value (e.g: predicting the price of a house based on its characteristics) [63].

Regression in machine learning involves modeling the relationship between a dependent variable and independent variables to predict continuous values. There are several regression algorithms and techniques, and the choice of method depends on the characteristics of the data and the specific needs of the problem [92].Figure 1.4 below summarizes the most common of these methods

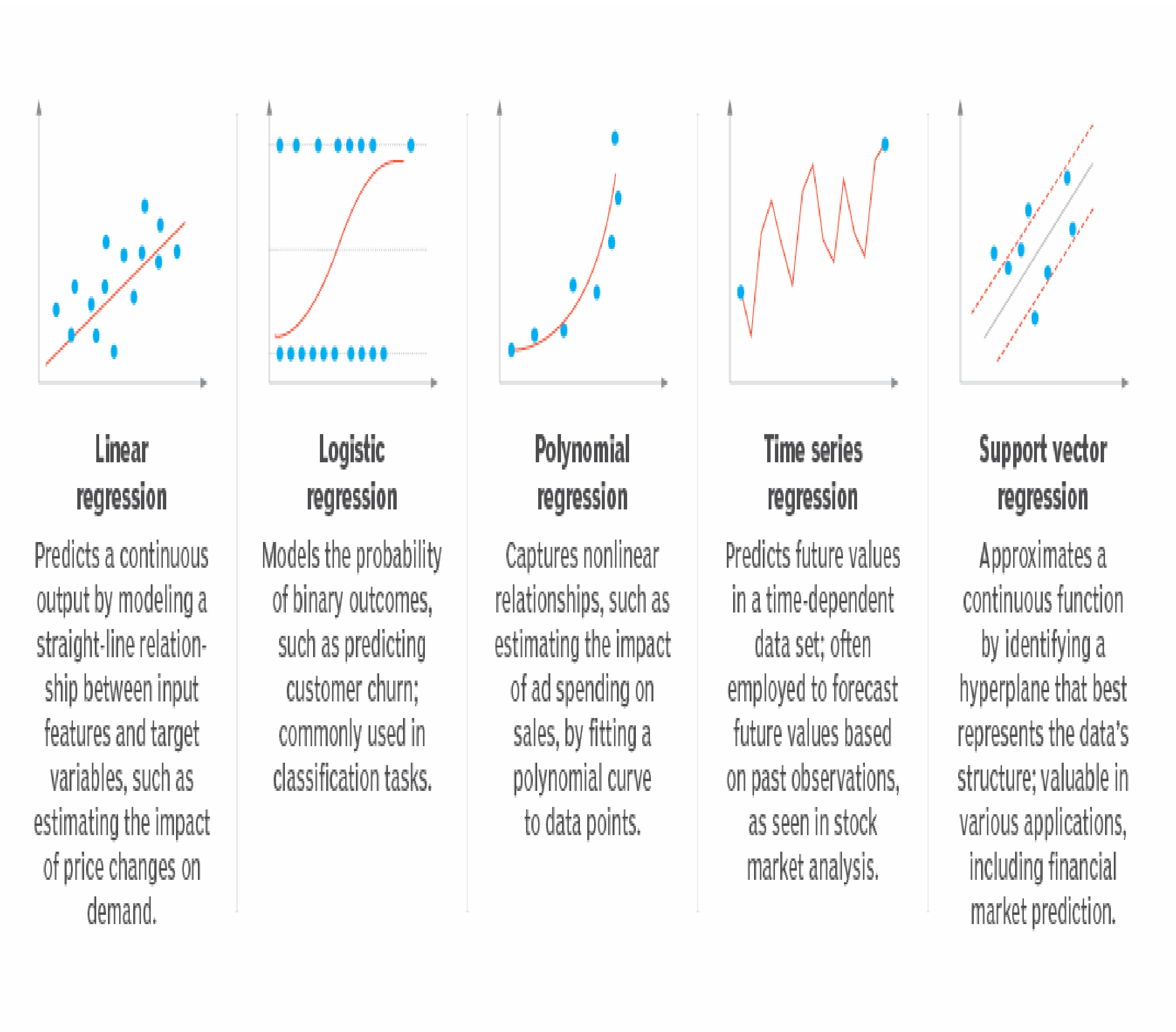

Linear regression
Predicts a continuous output by modeling a straight-line relationship between input features and target variables, such as estimating the impact of price changes on demand.
Logistic regression
Models the probability of binary outcomes, such as predicting customer churn; commonly used in classification tasks.
Polynomial regression
Captures nonlinear relationships, such as estimating the impact of ad spending on sales, by fitting a polynomial curve to data points.
Time series regression
Predicts future values in a time-dependent data set; often employed to forecast future values based on past observations, as seen in stock market analysis.
Support vector regression
Approximates a continuous function by identifying a hyperplane that best represents the data's structure; valuable in various applications, including financial market prediction.

Figure 1.3: Regression types [56]

The difference between the two concepts is shown in the figure 1.5 below :

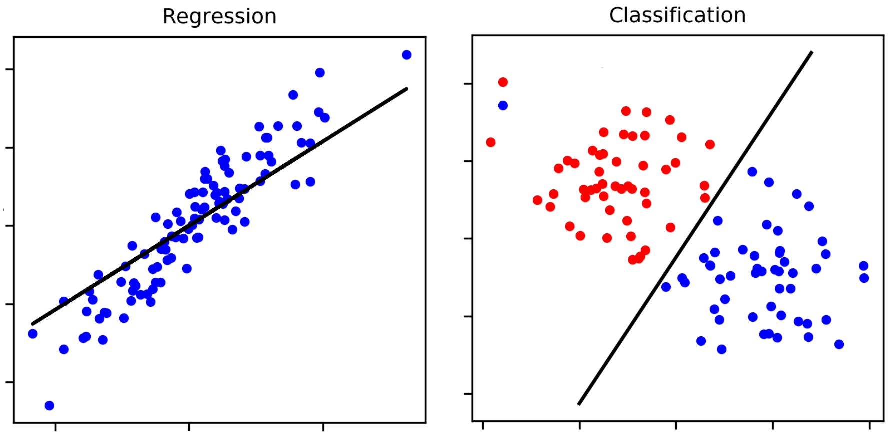


Figure 1.4: Classification vs regression [2].

1.3.4 Common ML problems

These ML methods often suffer from two common problems : over-fitting and under-fitting.In order to comprehend them the reader must be aware of some basic statistics concepts such as :

Variance : Variability of the predictions for different datasets [31].

Training error : measures how well the model has learned the training data.

Test error : (Also called validation error) measures the model's ability to generalize to new or unseen data (applying the patterns and relationships it has learned during training to new instances) [16].

Bias : refers to how a model consistently makes predictions that are either too low or too high compared to the actual values.It can happen when the model is too simple or when the training data doesn't accurately represent the real-world scenarios (lack or representation) [100].

1.3.4.1 Over-fitting

Occurs when the model's fit to the training data is too tight, and the model loses its ability to generalize correctly to new data [9].This occurs when we have:

- High variance
- Low training error.
- High Test error

1.3.4.2 Under-fitting

Occurs when the model is too simple to capture the complexity of the data and fails to generalize well [9].It happens when we have :

- High bias.
- High Test error.
- High Training error.

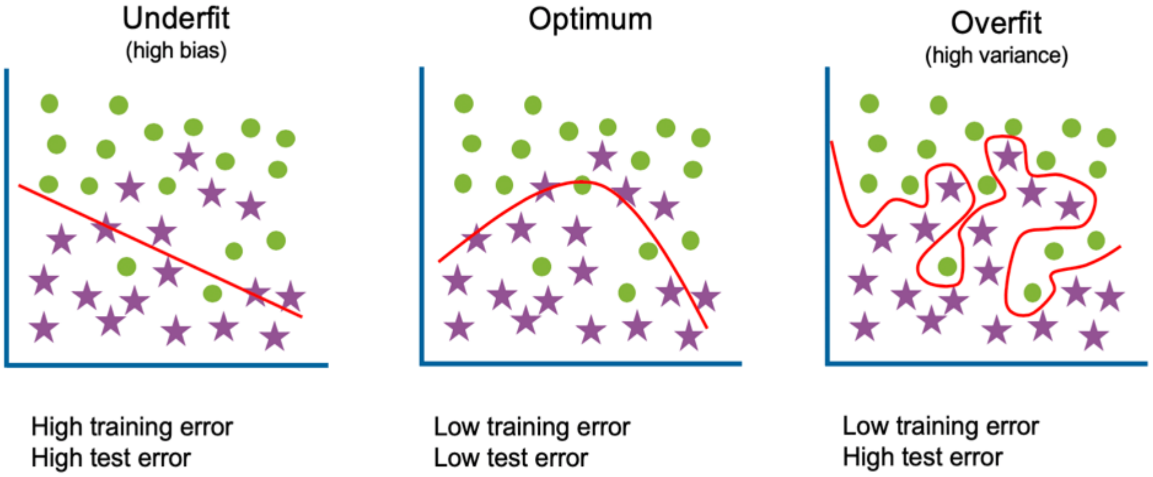


Figure 1.5: Over-fitting VS Under-fitting [66]

1.3.4.3 Solutions for under-fitting and over-fitting

There are many different techniques used to solve under-fitting and over-fitting issues.

- Solutions for Over-fitting

Regularization, which is a technique that adds a penalty term to the model's coefficients during training. This penalty term discourages large values for the coefficients, effectively reducing their magnitude and helping to generalize the model and prevents it from fitting the training data too closely, which can lead to poor performance on new, unseen data.

Model Simplification by reducing the complexity of the model (e.g., in decision tree models, simplification can involve limiting the depth of the tree or reducing the number of branches, nodes, or features used in the model) to prevent it from over-fitting the training data.

Cross-Validation, a method that checks how well a model will work on new data. It splits the dataset into parts, trains the model on some, and tests it on the rest, repeating this process to ensure reliable results. This helps prevent over-fitting, where the model fits the training data too closely and performs poorly on new data.

- Solutions for under-fitting

Increasing model complexity by using more sophisticated models or adding additional terms, such as higher-degree polynomials, to capture intricate patterns and nuances in the data. This approach aims to improve the model's ability to accurately represent the underlying relationships and trends present in the dataset.

Feature Engineering by analyzing the existing features in the dataset and using domain knowledge or statistical methods to create new features that can help the model better understand the underlying patterns and relationships in the data.The goal is to provide the model with more relevant and informative data, improving its performance and predictive accuracy.

Reducing regularization that decreases the level of penalty imposed on the model's coefficients, enabling a more flexible fitting to the data. This adjustment can lead to a model that is less constrained and more likely to capturing complex patterns and nuances in the dataset."

1.3.5 ML types

In this part , We describe different known types of Machine learning as follows :

1.3.5.1 Supervised learning

Supervised learning is one of machine learning's types where algorithms and models are trained on a set of labeled data, i.e, each example is associated with a known output (e.g. category, numerical value) to assure predictions or classifications.

Common algorithms in supervised learning include linear regression, decision trees, SVM, and neural networks. The workflow typically includes data collection, preprocessing, model selection, training, evaluation, and deployment as shown below in figure 1.7

1.3.5.2 Unsupervised learning

Unsupervised learning, contrary to supervised learning is another machine learning type where algorithms are trained on unlabeled data without specific guidance on desired outcomes. Algorithms used in this type seek to identify patterns, structures, and relationships on their own without prior labels. Common tasks in unsupervised learning include clustering similar data points together and dimensionality reduction.

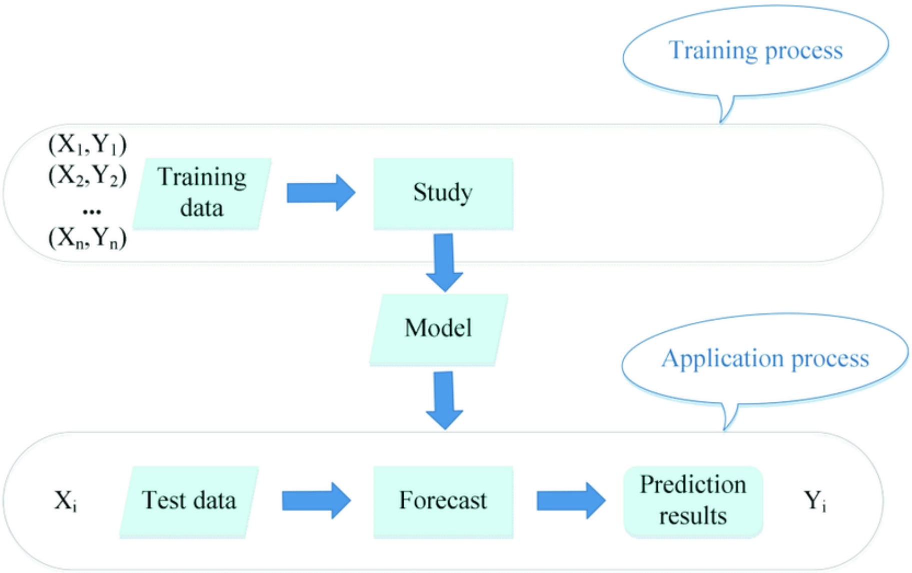


Figure 1.6: Supervised learning algorithms workflow [61].

Among those algorithms we can cite K-means and hierarchical clustering as the most that are widely used. Unsupervised learning is valuable for exploratory data analysis, anomaly detection, and discovering hidden insights in complex datasets.

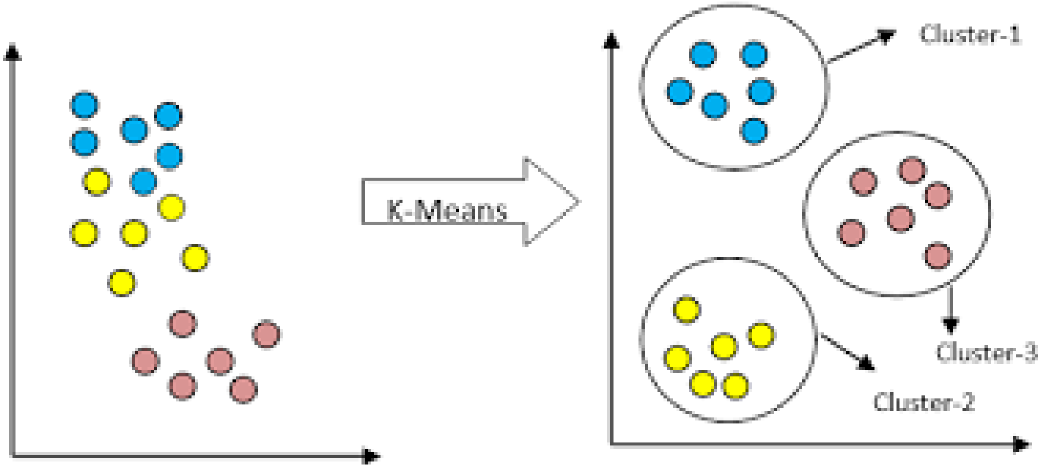


Figure 1.7: Example of Unsupervised learning : K-means [102]

1.3.5.3 Semi-supervised learning

Semi-supervised learning is another ML type that is a combination of supervised and unsupervised ML types . In which, the algorithms are trained on a dataset that contains both labeled data (known outputs) and unlabeled data (unknown outputs).

As shown below in figure 1.9 we can see the comparison of ML types concepts

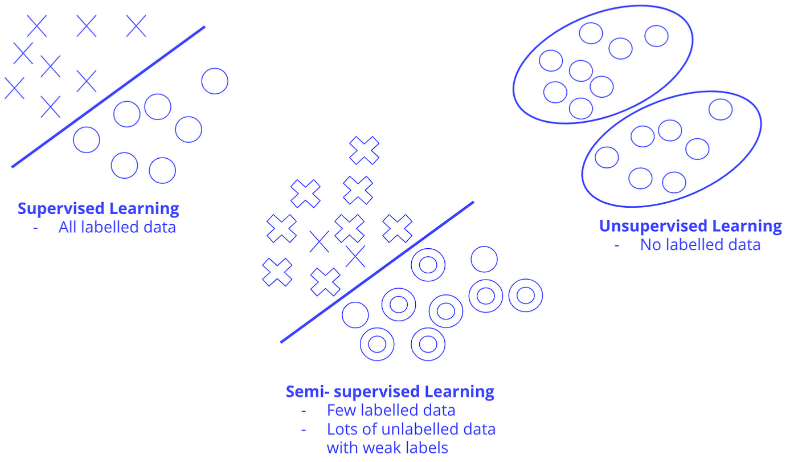


Figure 1.8: Comparison of ML types [53]

1.3.5.4 Reinforcement learning

Reinforcement learning (RL) is another ML known type where an agent learns to make decisions by interacting with an environment by taking actions based on its current state and receiveing feedbacks in the form of rewards or penalties from the environment in order to learn a policy that maximizes cumulative rewards over time . The process of this learning type involves a trade-off between exploration (trying new actions) and exploitation (leveraging known actions with high rewards) to discover optimal strategies in dynamic environments.

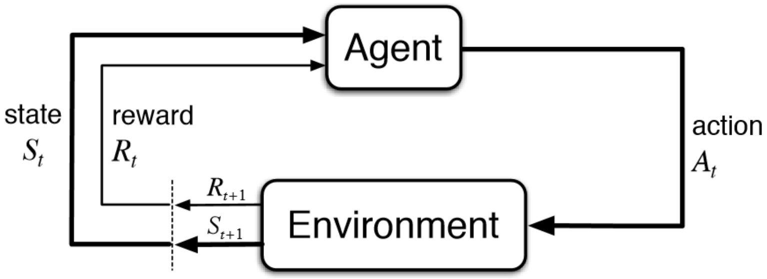


Figure 1.9: Reinforcement ML process [58].

1.3.6 ML algorithms

The classification of ML algorithms depends on several factors such as :

- The nature of the learning task.
- The availability of labeled data.

- The algorithm's approach to learning.
- The desired output or prediction.

So depending on these factors we can represent this classification as follows :

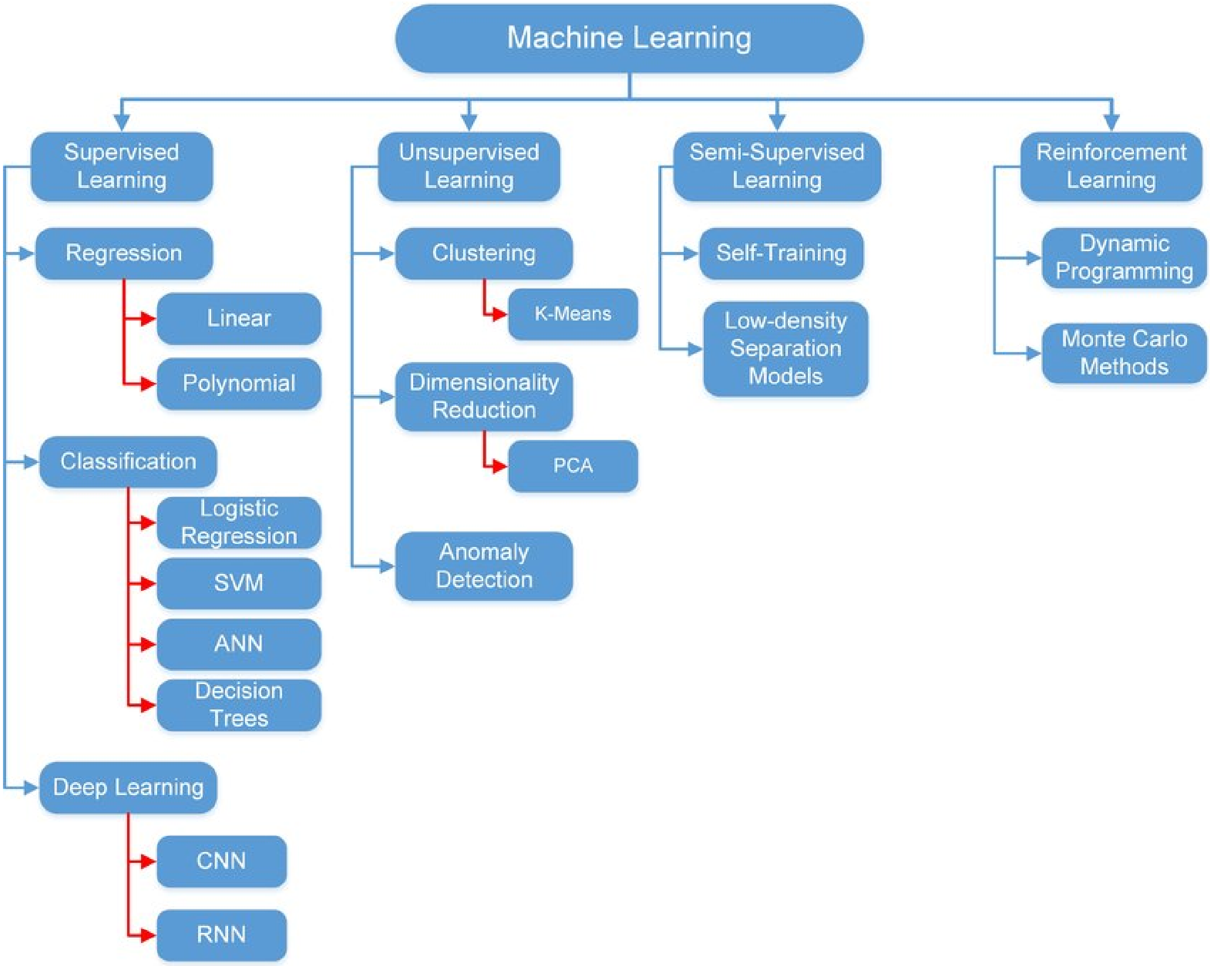


Figure 1.10: ML Algorithms [72]

1.4 AI for Optimal Performance in E-commerce

Using computer science and AI in e-commerce is all about making things work better for both customers and businesses.

For customers, it means getting personalized recommendations for products they will love, quick help from chatbots, avoiding items being out of stock, and feeling safe from fraud.

For businesses, it means selling more by suggesting the right products, keeping customers happy with great service, saving money by managing stock well, and keeping transactions secure. These improvements help make e-commerce smoother and more successful for everyone involved. We can delve into more details about how computer science and AI improve e-commerce performance for customers and businesses with the mentioned points below

Personalized Recommendations , by analyzing customer behavior, preferences, and purchase history to provide personalized product recommendations,by examining browsing habits, past purchases, and even items left in shopping carts, and using algorithms like Alternating Least Squares (ALS) [30] that decomposes the user-item interaction matrix into user and item matrices by alternating between updating these matrices to minimize prediction errors. AI can suggest products that match the customer's interests and needs. This not only enhancing the shopping experience by making it more tailored but also boosting sales by increasing the likelihood of repeat purchases and higher total spend.

Chat-bots and Virtual Assistants , that can handle a wide range of customer service tasks. They can answer frequently asked questions, provide detailed product information, and guide customers through the purchasing process. Available 24/7, these tools rely on natural language processing (NLP) algorithms [1] that extract intent and entities from user input to understand the meaning of the message then generate appropriate responses through predefined rules or machine learning models trained on extensive data to ensure that customers receive immediate assistance, reducing wait times and improving overall satisfaction and handling multiple inquiries simultaneously, making them more efficient than human representatives.

Predictive Analytics algorithms such as decision trees [74], and neural networks (NN) [84] ,forecast future trends based on historical data. This helps in managing inventory by predicting which products will be in demand. It also assists in pricing strategies by determining the optimal price points to maximize sales and profit margins. In marketing, predictive analytics can identify which campaigns are likely to be the most effective, allowing businesses to allocate resources more efficiently and achieve better results.

Fraud Detection and Prevention of fraudulent activities such as payment fraud, account hacking, and other suspicious behaviors by analyzing patterns and anomalies in transaction data using algorithms such as Isolation Forest [110] that efficiently spots unusual or suspicious activities in data by isolating them from normal patterns and other clustering techniques. It can quickly identify unusual activities that deviate from a customer's normal behavior and flag them for further investigation. This ensures high security by preventing fraudulent transactions before they occur, protecting both the business and its customers, building such a secure environment to both of them .

Image and Voice Search using technologies like image recognition and voice search enhance the shopping experience by making it easier for customers to find products. Image search allows customers to upload photos of items they are looking for, and the AI will find similar products available for purchase like convolutional neural networks (CNNs) complex algorithm [67] that breaks down images into smaller parts, then analyzing and combining these parts to understand the overall picture. This process helps them learn and recognize patterns in images .Also, voice search lets customers use natural language to search for products, making the process faster and more intuitive, especially on mobile devices.

Sentiment Analysis of customer feedback, reviews, and social media interactions to gauge customer sentiment. By classifying text data into positive, negative, or neutral sentiments, businesses can gauge customer satisfaction and sentiment towards products and services.Many algorithms used to understand how customers feel about products and services such as SVM (Support Vector Machines) [50] that separates data points into different classes using a hyperplane, classifying text into positive, negative, or neutral sentiments based on features extracted from the text. By continually refining their offerings based on customer sentiment, businesses can improve customer satisfaction and loyalty.

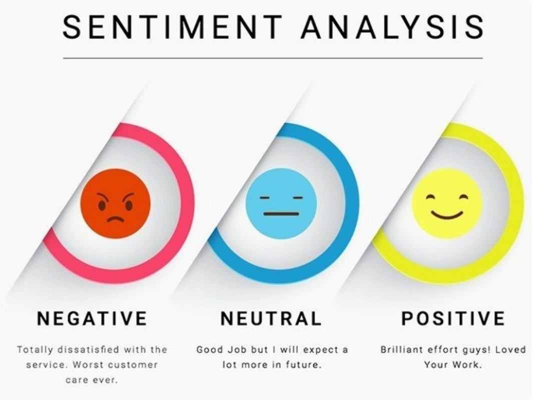

SENTIMENT ANALYSIS
NEGATIVE
Totally dissatisfied with the service. Worst customer care ever.
NEUTRAL
Good Job but I will expect a lot more in future.
POSITIVE
Brilliant effort guys! Loved Your Work.

Figure 1.11: sentiment analysis [25].

1.5 Conclusion

Nowadays, with the evolution of E-commerce sector and diversifying of its domains,integrating artificial Intelligence become necessary after studying its important impact and high performance and costumer experiences. AI-powered recommendation systems have enabled businesses to deliver personalized product suggestions, improving customer engagement and increasing sales. Additionally, AI-driven analytics and insights have empowered businesses to make data-driven decisions, optimize pricing strategies, and forecast demand accurately, leading to improved operational efficiencies and cost savings.

Moreover, AI has played a crucial role in improving customer experiences through enhanced personalization, efficient customer service, and streamlined processes. By automating repetitive tasks and providing intelligent insights, AI has freed up human resources to focus on strategic initiatives and value-added activities. This has resulted in higher customer satisfaction, loyalty, and retention rates.

Furthermore, AI has contributed to the growth and scalability of e-commerce businesses by enabling them to adapt quickly to market trends, customer preferences, and competitive dynamics. AI-driven automation, predictive modeling, and advanced analytics have empowered businesses to stay agile, responsive, and innovative in a rapidly evolving digital landscape.

For that we can say that the performance of AI in e-commerce has been transformative . Its ability to deliver personalized recommendations, optimize operations, and foster innovation has positioned AI as a strategic asset for e-commerce businesses seeking sustainable success in the digital era.

Chapter 2

State of the art

2.1 Introduction

In this second chapter we give an overview of the state of the art about recommendation systems and give more information and basic concepts on which our approach is based to be more able to have clearer ideas on what has been done and what can be improved.

2.2 Definitions and fundamental concepts

We will introduce and give the definitions of the basic concepts needed to make the ideas clear about recommendation systems.

System A system is a set of resources and input elements interconnected in an organized manner and which work together to achieve a specific objective (output elements) [11].

Recommendation The act of recommending means suggesting or offering to someone, something which can be a product, service, action according to their preferences or specific needs [108].

Item An item represents system's recommendation that can be a product or anything made available for sale or purchase on an e-commerce platform [104]. It can be described by a name, a description, a price, etc.

-Browsing lists of products.

-Making purchases.

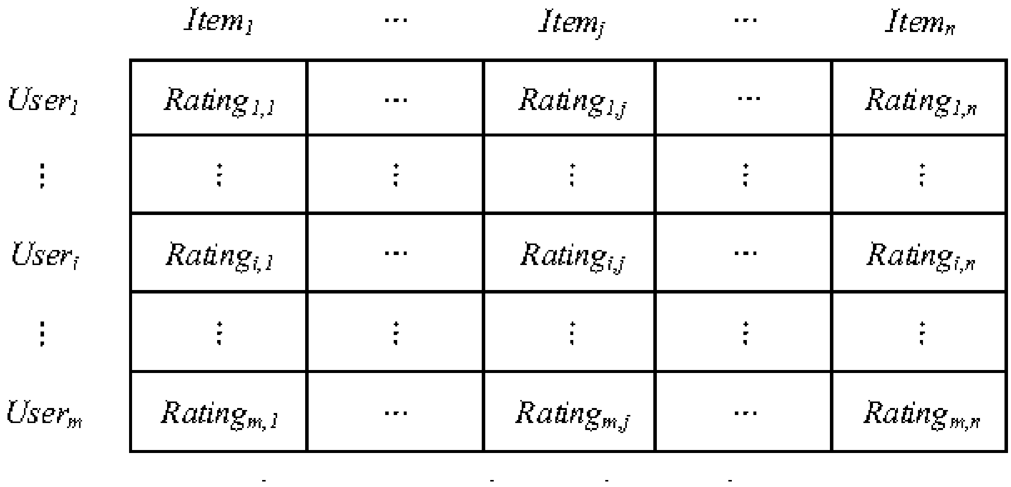
-Leaving reviews and ratings about the different proposed items.

Notes or Rating The rating reflects the user's preference for the items recommended by the platform [57].

Rating Matrix / User-Item Matrix The user-item matrix represents the ratings given by users for items recommended by the platform [101], where:

u Refers User , i Refers Item

- Elements of the system's set $\langle u, i \rangle$ are recorded.
- Each row corresponds to ratings provided by a single user. $i \in [1, n]$
- Each column corresponds to ratings received by a single item from all users.$u \in [1, m]$



	$Item_1$	$\cdots$	$Item_j$	$\cdots$	$Item_n$
$User_1$	$Rating_{1,1}$	$\cdots$	$Rating_{1,j}$	$\cdots$	$Rating_{1,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$User_i$	$Rating_{i,1}$	$\cdots$	$Rating_{i,j}$	$\cdots$	$Rating_{i,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$User_m$	$Rating_{m,1}$	$\cdots$	$Rating_{m,j}$	$\cdots$	$Rating_{m,n}$

Figure 2.1: Rating/ User-Item Matrix. [29]

2.3 Recommendation systems

2.3.1 Definition :

We find in the literature several definitions of recommendation systems.

In a general way and according to Robin Burke , recommendation systems are : "Systems which are capable of providing personalized recommendations making it possible to guide the user to interesting and useful resources within an important data space". [21]

Also, we can say that a recommendation system is defined as a system allowing to suggest or offer in a relevant and personalized way to a user products, services, or specific content based on his interests [47], his behavior during the purchase,and other contextual factors by analyzing his data ,especially the history of his experience on the platform.

2.3.2 Characteristics of Recommendation Systems

To more understand the functioning and benefits of recommendation systems, it is essential to review their key characteristics.

First **Decision Support** that involves leveraging user behavior data and item characteristics to estimate how likely a user is to rate an item positively thus aiding in personalized recommendations [77].

Comparison Assistance, by analyzing user preferences and historical interactions, the system arranges items in a ranked order, facilitating easier decision-making and improved user experience [12].

Discovery Aid through data analysis and pattern recognition, the system introduces users to new and relevant items that match their interests but may not be familiar to them yet, fostering exploration and discovery [88].

2.3.3 Recommendation Systems importance

Income boost

The main goal of a recommendation system (RS) is to make users' experiences better and help businesses make more money by suggesting personalized products or content.The RS as a digital shopping helper that looks at what the user bought before, what he likes, and what he had looked at online to suggest things he might want to buy next. For instance, an online store's RS that checks past purchases, favorite brands or styles, and size to recommend clothes that suit him. This personalized touch not only makes the shopping journey smoother but also makes the user more likely to buy something.

From a business point of view, recommending products or content that match customers' interests leads to more sales and better conversion rates. When businesses show items that customers are likely to love, they can improve their marketing strategies and earn more money. Overall, recommendation systems are essential for keeping customers happy, engaging them better, and helping businesses grow in today's competitive digital world.

Better User Experience

The recommendation system enhances the user experience by leveraging sophisticated algorithms to analyze user behavior, preferences, and interactions with the platform.

By understanding the user's unique tastes and interests, the system can recommend items that are highly likely to resonate with them. This personalized approach reduces the time and effort users spend searching for relevant content or products, leading to a smoother and more enjoyable browsing experience. Additionally, the system continuously learns from user feedback and interactions, refining its recommendations over time to further enhance user satisfaction, this saves user's time and makes his experience more enjoyable .

More Choices, Better Ads

Recommendation systems play a crucial role in introducing users to a wide range of products or services, ensuring they are not limited to seeing the same items repeatedly. This diversity is beneficial for users as it exposes them to new options and helps them discover products or services they may not have been aware of otherwise, by presenting a variety of choices, recommendation systems enrich the user's experience.

Also, from a business perspective, the diversity in recommendations enables more effective advertising strategies. Businesses can leverage recommendation systems to showcase their products or services to users who are likely to be interested based on their browsing history, preferences, and behavior ,increasing the relevance of advertisements, leading to better engagement and conversion rates then maximizing the impact of their marketing efforts.

2.3.4 Fields of application

The broad use of recommendation systems has grown in various service areas.

This study investigates how recommendation models are used in different sectors considering each sector's unique features and goals.

Through a comprehensive examination of collected research papers, the service fields using recommendation systems were categorized into seven main sectors:

2.3.4.1 Streaming Services

Netflix, Spotify, and YouTube . . . etc rely heavily on recommendation systems to enhance the user's experience by suggesting content that matches his preferences. These systems analyze viewing history, user ratings, and behavior patterns to propose movies, TV shows, music, and videos. . . .A hybrid recommender system for music streaming services has been proposed by Yu (2020) [112], in which contextual filtering is combined with collaborative filtering to enhance recommendation accuracy and diversity by considering user listening context and item-item collaborative filtering.

Example: Netflix's recommendation system uses collaborative filtering and deep learning techniques to personalize content recommendations. To learn features from movie data and provide accurate, relevant movie suggestions, Anjum(2019) [13] introduced a hybrid recommender system for movie recommendations, combining collaborative filtering for personalized recommendations, content-based filtering, and deep learning (using stacked auto encoders).

Similarly, Hassan A. Khalil's (2024) [48] proposed an approach that combines similarity-based and matrix factorization-based models in a hybrid recommendation system tailored specifically for movie recommendations.The focus were on overcoming key challenges in movie recommendation systems to provide more effectiveness.

2.3.4.2 Social Network Service

Social networks such as Facebook, Instagram.., use recommendation systems to suggest friends, pages, groups, and content,by analyzing user interactions, likes, shares, and follows to provide relevant suggestions.

Example : Recommending potential friends based on mutual connections and interests by Facebook's "People You May Know" feature.

2.3.4.3 Tourism Services

Recommendation systems in this sector are used by platforms like TripAdvisor, Airbnb, and Booking.com to suggest travel destinations, accommodations, activities, and restaurants. These systems consider user reviews, preferences, past bookings, and seasonal trends.

Also, Kang(2019) in [46] focused on a context-aware hybrid recommender system for tourist attractions, incorporating collaborative filtering, content-based filtering, and context-aware filtering techniques based on user location, weather, and time, to offer tailored recommendations for tourist spots.

Example : we found TripAdvisor which uses CB filtering to recommend attractions and dining options based on user reviews and ratings.

2.3.5 E-Commerce Services

Many E-commerce giants like Amazon, Alibaba, and eBay use recommendation systems to enhance the shopping experience by suggesting products that align with user's interests and purchasing behavior. They mostly a combination of CF, CB filtering, and also hybrid methods [8].

Example : we take Amazon's recommendation engine example which analyzes browsing history, purchase history, and items in the shopping cart to make personalized product suggestions.

2.3.5.1 Healthcare Services

In which recommendation systems assist in personalized treatment plans, medication suggestions, and health monitoring. Platforms like IBM Watson Health and personalized healthcare apps use these systems to recommend diets, exercise routines, and medical treatments based on individual health data, genetic information, and lifestyle.

Example: A health app might be used to recommend specific exercises and dietary changes based on a user's activity level and medical history.

2.3.5.2 Education Services

Educational platforms like Coursera, Duolingo leverage recommendation systems to personalize learning experiences by suggesting courses, exercises, and resources based on user progress, interests, and learning goals. For instance,Safa (2021) [91] presented an enhanced hybrid recommender system for e-learning, using a combination of collaborative filtering, content-based filtering, and deep learning (convolutional neural networks (CNNs) and recurrent neural networks (RNNs)) to offer personalized course recommendations, improving accuracy and relevance in e-learning environments.

2.3.5.3 Academic Information Services

In these services , recommendation systems help researchers and students to discover relevant papers, journals, and conferences. Platforms like Google Scholar, ResearchGate, and Mendeley use these systems to suggest academic content based on citations, co-authorship networks, and research interests.

Example : Google Scholar's recommendation system provides personalized article suggestions based on the user's publication and citation records.

These categories were established based on the increasing user base and growing business significance of services utilizing recommendation systems, also considering popular searches on Google Scholar related to 'Recommendation System' [90].

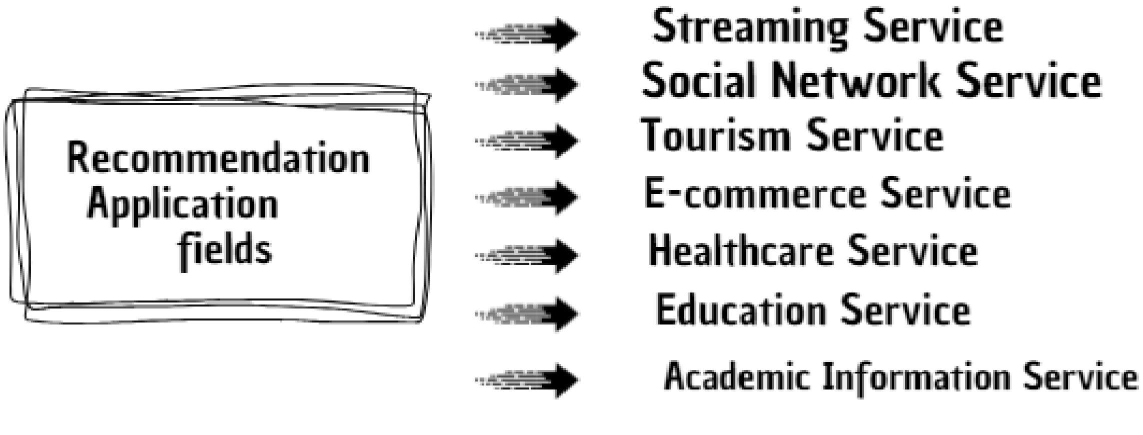


Figure 2.2: Application fields [52].

2.3.6 RS: How do they work ?

(RS) Recommendation systems try to recommend different items to the users , after an deep comprehension of user's interactions and preferences and combining different relations between them and the items recommended that we enumerate below :

2.3.6.1 Relation User-Item

The User-Item relationship is formed when certain users show a preference or interest in specific products or items that they find useful or attractive.

For example, a gardening enthusiast may have a preference for gardening-related items. Online platforms, such as social media or e-commerce sites, establish a user-item relationship by tracking user interest in products or items such as gardening tools, plants, and garden accessories.

2.3.6.2 Relation User-User

User-user relationships arise when certain customers share similar tastes for a particular product or service.

For example, they may have mutual friends, similar product or service preferences, or come from similar geographic regions .

2.3.6.3 Relation Item-Item

The Item-Item relationships form when items share similar characteristics, whether in terms of appearance or description. *For example,* books of the same genre, music with similar styles, cuisine from the same region, or news articles covering a specific event.

2.4 Recommendation Systems classification

Recommendation systems use various methods and approaches to suggest items, content, or products to users based on their preferences and behaviors.

The most commonly used recommendation system methods could be classified into : Collaborative Filtering methods, Content-Based Filtering methods, and Hybrid Methods.

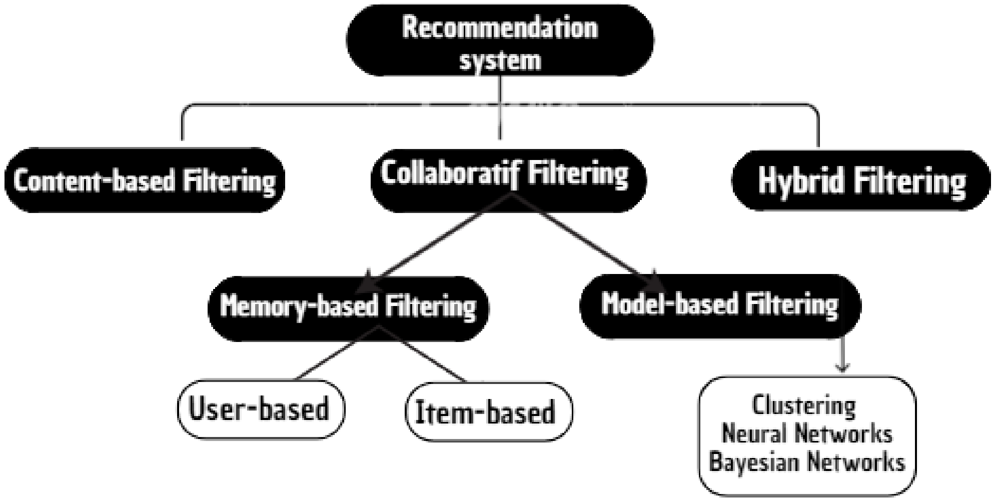


Figure 2.3: Recommendation system different techniques. [39]

2.4.1 Collaborative Filtering

Collaborative Filtering is one of the most recommendation system techniques that leverages the preferences and behaviors of users to generate recommendations [95].

Goldberg first used the term Collaborative Filtering (CF) in the recommender system Tapestry [4]. It operates under the assumption that users who have agreed on preferences in the past are likely to agree in the future.

The CF system typically goes through three basic stages: beginning by gathering user ratings for things in order to create a user-item rating matrix. Next, by calculating the similarity between users and objects, use this matrix to find the neighbors. Lastly, use an aggregation techniques for the prediction stage.

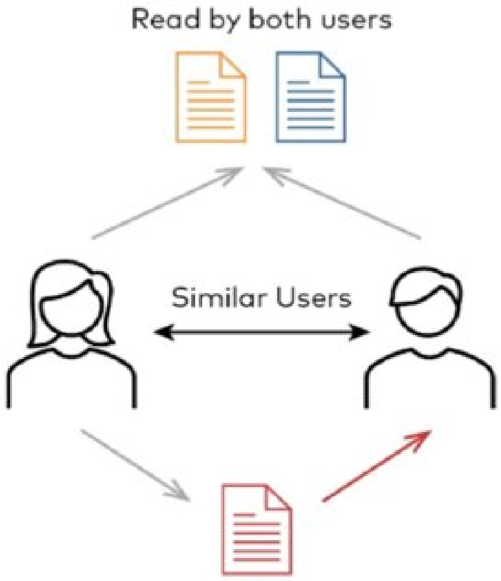


Figure 2.4: Collaboratif Filtering concept [119].

At this point, unrated things' ratings scores are predicted, their ranking is determined by these prediction scores, and the N-top items are chosen as suggestions for a certain user.

This strategy is based on the logical assumption that consumers with comparable prior preferences would likely have similar future interests [5].

For instance, in a system for suggesting movies, algorithms used in CF look for other users who share similar interests before suggesting the movies that they enjoy the most. Generally speaking, CF may be divided into two primary categories: memory-based, which uses user correlations to directly generate suggestions for the user, and model-based, which builds a model beforehand that is then used to forecast what the user would like

Memory-based techniques are further divided into two categories : item-based in which items are similar to those that have been liked by the target user, using the similarity of item preferences across users and user-based, in which items are recommended to a user based on the preferences of users with similar tastes, identifying patterns in historical interactions.

These two methods offer advantages such as : User independence and serendipity, but also face challenges like : Sparsity and scalability issues. Model-based techniques in recommender systems use mathematical models to predict user preferences by identifying patterns in historical data. They include methods like matrix factorization and deep learning, which capture latent relationships between users and items. These models are known for their accuracy and scalability, but require substantial data and computational resources.

2.4.2 Content-Based Filtering

CB Filtering is a frequently used technique in e-commerce ,it is used to suggest goods or items that are similar to those that are being liked or clicked [54]. The item description and the user's interest profile serve as the foundation for user suggestions. E-commerce platforms frequently employ recommender algorithms that are based on content.

CB Recommender Systems are a category of recommendation systems that generate personalized suggestions based on the intrinsic features and characteristics of items and the preferences of users. Content-based methods focus on the attributes of the items themselves rather than relying on user-user or item-item interactions [82]. This model has been mainly used in services that recommend items or text data items that are easy

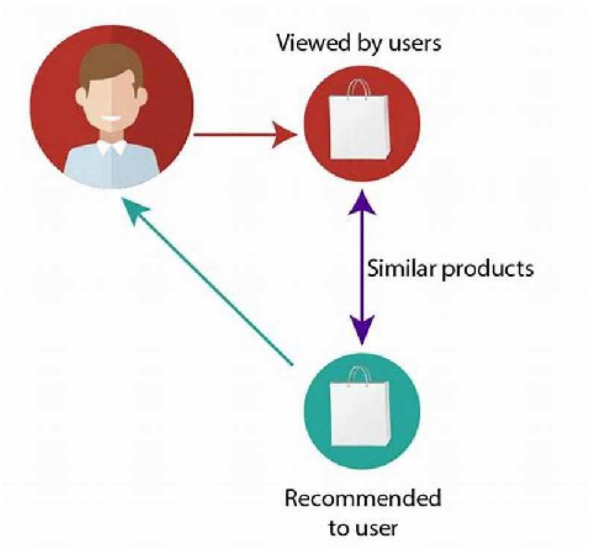


Figure 2.5: Content-based Filtering concept [49].

to recommend based on item information and user profile information.

The Content-Based Filtering Model uses many technologies to identify user preferences.The fundamental principle underlying content-based methods is the detailed analysis of item features, which can encompass a wide array of characteristics depending on the nature of the items in question.

This meticulous examination allows the recommendation system to discern patterns, themes, and intrinsic attributes to the items, creating a profile that reflects the nuanced preferences of users.

Jain et al. in [42] introduces a Journal Recommendation System (JRS) designed to assist researchers in selecting appropriate journals for publishing their work. The system addresses the challenge faced by novice authors in identifying suitable journals, which often leads to wasted time and effort for both authors and editors.

JRS uses a content-based filtering method, using a dataset prepared by the authors themselves. A distance algorithm is applied to recommend journals based on the content and characteristics of the research article. This system aims to streamline the journal selection the process and improve the efficiency of academic publishing for authors.

In essence, content-based recommendation systems stand as versatile solutions, particularly adept in domains where the inherent attributes of items significantly influence user preferences. Their reliance on advanced technologies ensures a deep understanding of content, fostering accurate and tailored suggestions that cater to the individualized tastes and preferences of users across diverse application areas.

2.4.3 Hybrid Filtering

Recommender systems have evolved significantly since their first appearance, going through several important phases, each bringing significant improvements in accuracy and personalization. Initially, collaborative recommender systems (or collaborative filtering) were

widely used. These systems work by analyzing user preferences and behaviors to identify similar users and recommend items that they liked.

- For example, if user A has similar tastes to user B, items liked by B will be recommended to A. While this approach is effective at exploiting the wisdom of crowds,also can encounter scalability and performance problems when it has to process a very large number of users and articles.

To overcome these limitations, content-based recommender systems have been introduced. These focus on item characteristics (such as movie genres, book authors, recipe ingredients, etc.) and explicit user preferences to recommend items similar to those a user already has appreciated.

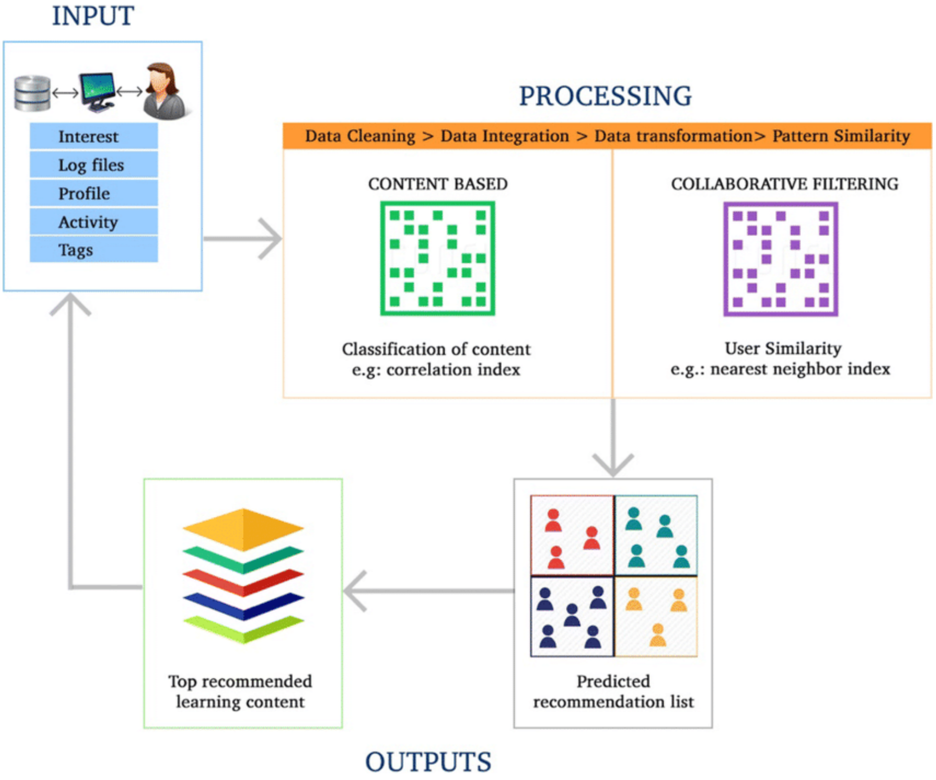


Figure 2.6: Hybrid RS example [51].

This method allows you to offer personalized recommendations without requiring a large amount of data about other users.

However, it may lack diversity in recommendations, tending to suggest articles very similar to those already consumed by the user, which may limit the discovery of new content.

In order to combine the advantages of the two previous approaches and to overcome their respective limitations, hybrid recommendation systems have been developed. These systems incorporate collaborative and content-based techniques, providing more accurate and diverse recommendations.

- For example, they can use item characteristics to improve collaborative filtering by reducing cold start effects or combine multiple recommendation models into one to increase accuracy.

Hybrid systems can also use content-enhanced collaborative filtering approaches or incorporate explicit user preferences to refine recommendations.

This fusion of methods makes it possible to benefit from the strengths of each approach while minimizing their weaknesses,they acknowledge the complexity of user preferences and behaviors,unlike singular methods [22].

Hybrid systems are particularly effective at handling common issues in recommender systems, such as cold starts for new users and items, and provide better personalization in dynamic, large-scale environments, such as e-commerce platforms, modern streaming services like Netflix or Spotify, and social networks. They allow for greater diversity in recommendations and are more robust to variations in user preferences.

In summary, the evolution of recommendation systems, from collaborative methods to content-based approaches, and ultimately to hybrid systems, has significantly improved the accuracy, relevance, and diversity of recommendations. These advancements meet the varied and growing needs of users, thereby contributing to a more satisfying and personalized user experience in various application areas.

2.5 Recommendation Techniques

This section explores a variety of recommendation techniques, including hybrid approaches, designed to enhance user experience and decision-making across diverse applications.

2.5.1 Recommendation Methods

Recommendation systems use different techniques such as:

Text mining [17] is a technique used for extracting valuable information from data by analyzing text-related information. Recent advancements in natural language processing technologies have enabled the extraction of semantically important information from text. While text analysis often relies on word frequency, limiting semantic understanding, the use of ontology has emerged to construct a conceptual schema and define a common vocabulary for accurate text interpretation.

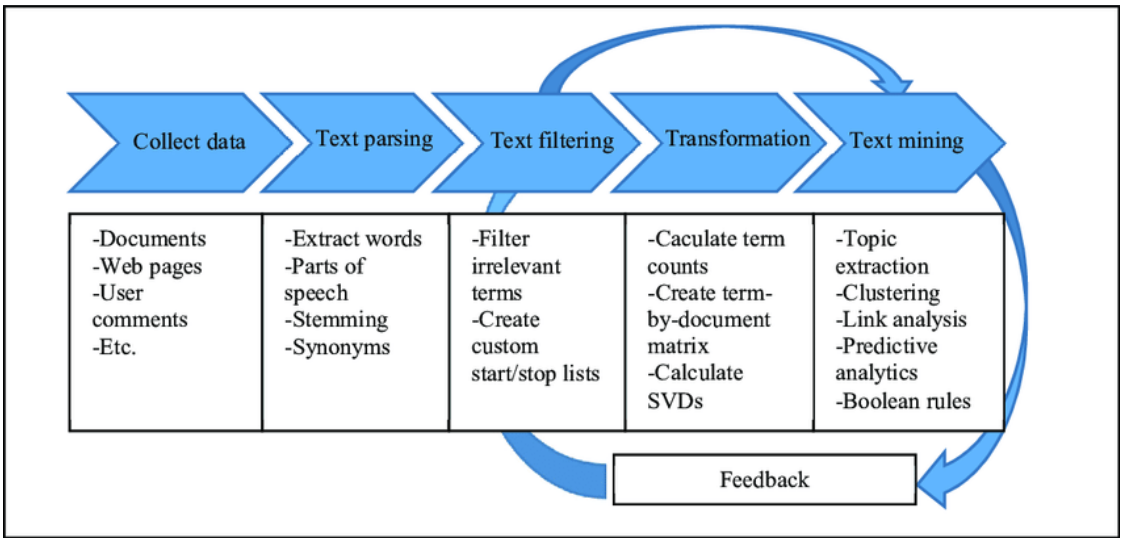


Figure 2.7: Text Mining process [45].

In the context of recommendation systems, text mining is used for semantic analysis in Content-Based Filtering models. It involves performing semantic analysis of item information to recommend similar items. Collaborative Filtering models also benefit from text mining by evaluating semantic knowledge between users, facilitating item recommendations based on similarity.

Focusing on the Content-Based Filtering recommendation model, Term Frequency–Inverse Document Frequency **(TF-IDF)** is a commonly used text mining technique. TF-IDF assigns weights to words based on their frequency, expressing document components as vectors and identifying term importance.

Also, text mining techniques, including fuzzy linguistic modeling (FLM), enhance context awareness, especially in situations where user preferences are unclear or insufficient [17].

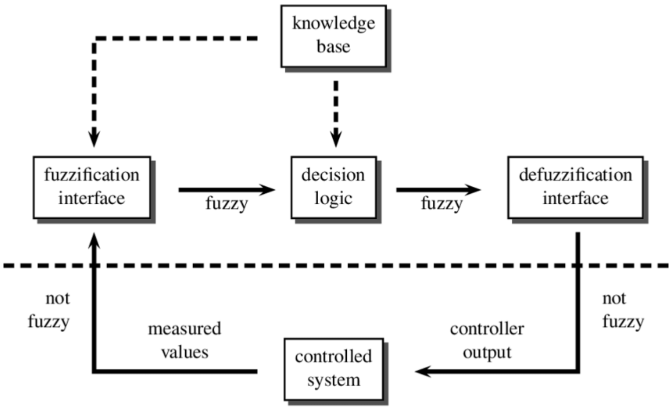


Figure 2.8: Fuzzy linguistic modeling demonstration [68].

In contrast, K-Nearest Neighbor (KNN) [14] that is an algorithm used for classifying datasets by comparing the similarity between data items, classifies items similar to users' tastes based on patterns in the user's behavior data. However, studies indicate challenges with the KNN algorithm, including the need for selecting an appropriate value for K and performance degradation with large input sizes [116].

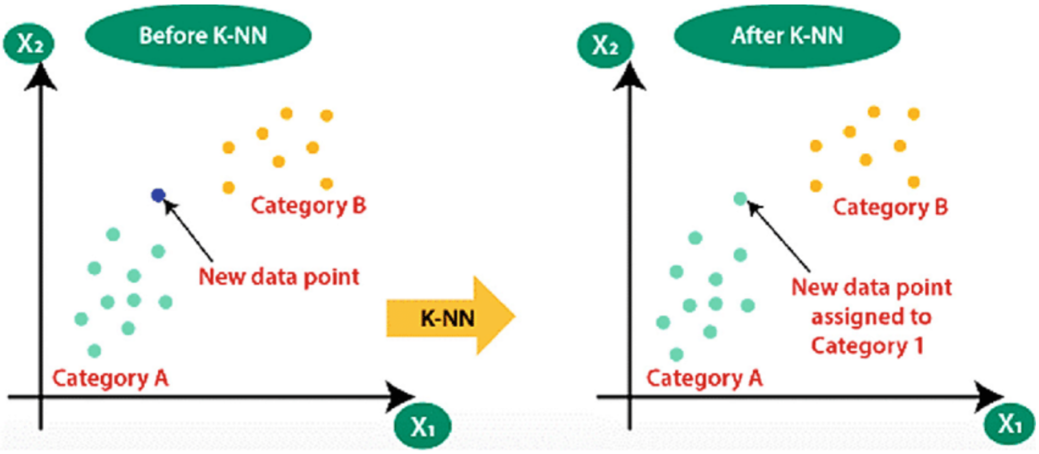


Figure 2.9: KNN mechanism [70].

Clustering,or identifying categories or clusters in data based on their features or characteristics, is widely used in recommendation systems,by identifying patterns or structures within data points that allow for the creation of meaningful clusters as K-means clustering that follows a simple principle,starting by randomly initializing cluster centroids and then iteratively assigning data points to the nearest centroid based on their distance, after that calculating the mean of all data points assigned to each cluster to update the centroids . This process continues until the centroids no longer change significantly, indicating convergence [115].

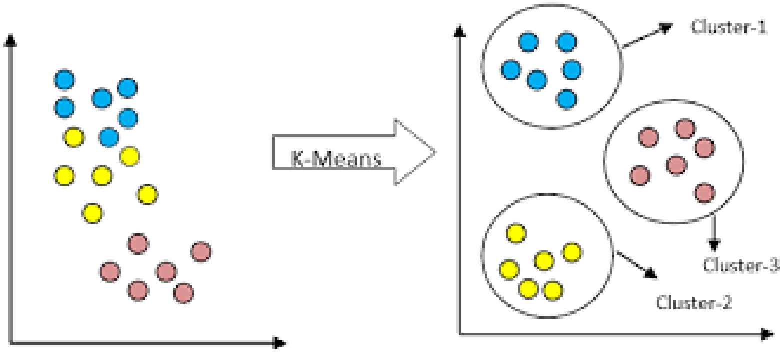


Figure 2.10: K-means clustering demonstration [103].

Matrix Factorization recommendation systems as discussed in [64], address the sparsity problem in Collaborative Filtering by characterizing items and user data through latent factors. The principle behind it is to decompose the user-item interaction matrix into low-rank matrices, representing latent features that capture underlying patterns in the data. This decomposition allows for a more efficient representation of user preferences and item characteristics. As example the Alternating Least Squares (ALS) [30] method that involves iteratively updating user and item factors, by alternating between fixing one set of factors while optimizing the other to minimize the reconstruction error of the user-item matrix. This process continues until convergence, resulting in optimized latent factors that capture user preferences and item characteristics effectively.

Finally, neural networks [80], that are widely used in various fields, are gaining attention in recommendation systems. Based on layers of interconnected nodes (neurons) that can learn complex patterns and relationships in data. In recommendation systems, neural networks can effectively capture user preferences and item characteristics, making them capable of providing highly personalized recommendations. They can handle vast amounts of data and adapt to changing user behavior, offering scalability and flexibility. Despite their complexity, neural networks are becoming a powerful tool for enhancing the accuracy, cold start, and relevance of recommendations [6] .

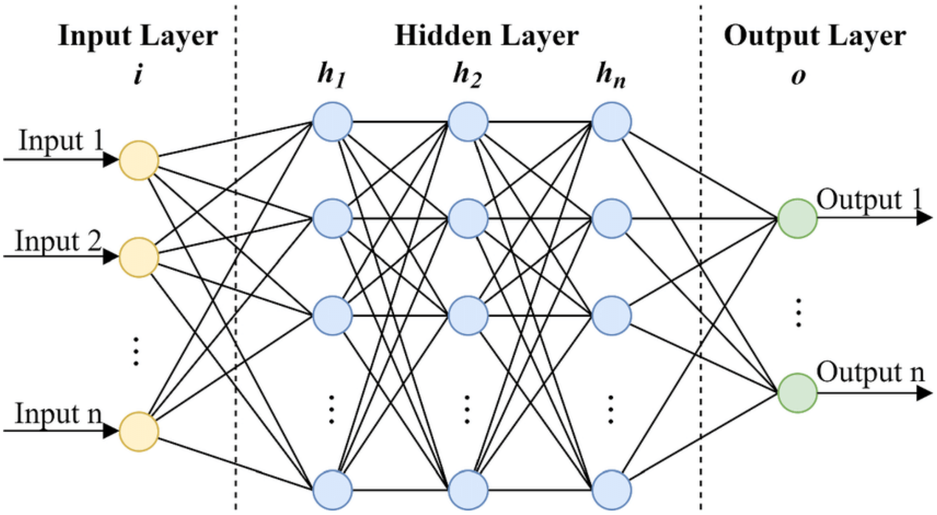


Figure 2.11: Neural Network's architecture [15]

All of these techniques are summarized in the figure 2.12 below :

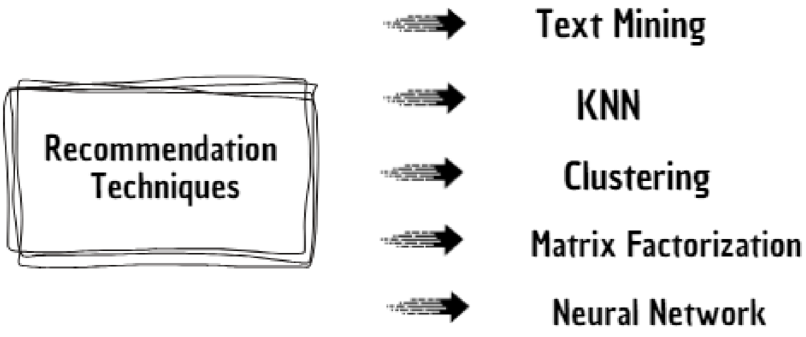


Figure 2.12: Recommendation techniques [52].

2.5.2 Hybrid Filtering techniques

Hybrid recommendation systems often use three primary techniques : weighted hybrid, feature combination, and cascade approaches.

2.5.2.1 Weighted Hybrid

The weighted hybrid method combines various recommendation techniques by assigning weights based on their performance or relevance. These weighted recommendations are then aggregated to produce the final recommendation list.

For instance, [24] proposed an improved weighted hybrid system that considers user preferences and similarities, leading to enhanced recommendation accuracy.

Similarly *Sonule et al.* [98] proposed a weighted hybrid recommendation method to tackle the challenges of cold start and data sparsity in traditional recommender systems within a big data environment by using keywords to indicate user preferences, it combines content-based and knowledge-based filtering algorithms to generate personalized service recommendations. Extensive experiments on real-world datasets show that the proposed method significantly enhances the accuracy and scalability of service recommender systems compared to existing approaches.

2.5.2.2 Feature Combination

In this approach, different recommendation methods merge their features or representations into a unified model.

By leveraging the strengths of each method, such as combining user-item collaborative filtering with feature integration, these systems offer improved recommendations, especially in scenarios with limited user interactions [59].

For instance*Andreu et al.* in [106] introduced two hybrid recommender systems that enhance collaborative filtering by incorporating song feature vectors to improve music playlist continuation.This addresses the issue of long-tailed distributions in music collections, where many songs appear in few playlists and are thus poorly represented. The proposed systems are evaluated through offline experiments, showing that they predict more accurately playlist continuations compared to traditional collaborative filtering methods.

2.5.2.3 Cascade

The cascade technique integrates recommendation methods sequentially, where one method's output influences the recommendations of the next method.

For example ; a cascade hybrid collaborative filtering system initially uses collaborative filtering for recommendations and then refines these suggestions using social influence data in a step-by-step process in [117].

This technique was used by *Rebelo et al.* in [89] in which they propose an innovative method to enhance recommender systems by combining model-based, memory-based, and content-based approaches in a cascade-hybrid system, where each approach sequentially refines the recommendations. A straightforward way to incorporate time-awareness into rating matrices is also proposed. The focus is on being intuitive, flexible, robust, and auditable. The system's performance is evaluated using metrics such as Novelty Score, Catalog Coverage, and mean recommendation price.

The success of hybrid recommendation systems depends on factors like the specific recommendation task, available data, method characteristics, and evaluation metrics.

Evaluating and comparing these systems based on specific application requirements is crucial for achieving optimal performance and relevance in recommendation tasks.

2.6 Evaluation metrics

Table 2.1: Evaluation Metrics for Recommendation Systems [97], [34]

Metric	Description and Calculation
Quantitative Evaluation Indicators	
RMSE	An index to evaluate prediction accuracy. Calculated by taking the square root of the mean squared error.
Qualitative Evaluation Indicators	
Precision	Proportion of recommended items that match the user's taste. Precision $= \frac{\text{True Positives}}{\text{True Positives+False Positives}}$
Recall (TPR)	Ratio of items recommended by the model to the total number of items that should be recommended. Recall $= \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$
Accuracy	Ratio of successful recommendations to all recommended items. Accuracy $= \frac{\text{True Positives+True Negatives}}{\text{Total Recommendations}}$
F-Measure	Harmonic average of Precision and Recall. F-Measure $= \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$
ROC Curve	Graph showing the relationship between FPR and TPR.
AUC	Area under the ROC curve, measuring the accuracy of the recommendation model.

2.7 Recommender System Challenges

Recommender systems face several significant challenges that impact their ability to effectively deliver personalized and relevant suggestions to users.

Diversity ensures that the recommended items are varied, reducing redundancy and increasing the chances of discovering different items [111]. A diverse set of recommendations can prevent user boredom and encourage exploration by presenting users with a wide range of options that span different categories or genres. This variety is particularly important for keeping users engaged and for promoting lesser-known items that might otherwise be overlooked.

Novelty measures how new or unexpected the recommended items are to the user [83]. High novelty can enhance user experience by introducing them to items they have not previously encountered or considered. Novel recommendations help users discover new interests and expand their horizons, thereby enriching their experience with the system. However, balancing novelty with relevance is crucial to ensure that recommendations are still aligned with the user's preferences [44].

Coverage measures the proportion of items in the dataset that can be recommended. High coverage ensures that a wide range of items are available for recommendation, which is essential for catering to niche markets and long-tail items [99]. By increasing coverage, a recommender system can serve a diverse user base more effectively [86], including those with unique or specific interests. High item coverage also prevents the system from becoming overly focused on a small subset of popular items, which can lead to a lack of variety and potential user dissatisfaction.

Serendipity refers to the pleasant surprise of receiving recommendations that the user did not expect but still finds relevant and enjoyable [35]. High serendipity can enhance user engagement by providing delightful discoveries, which can make the user experience more enjoyable and memorable. This aspect is crucial for creating a sense of discovery and satisfaction, as users are more likely to appreciate recommendations that introduce them to new and intriguing items that align with their interests.

Cold Start Problem One of the most significant challenges in recommender systems is the cold start problem, which occurs when there is insufficient data available to generate reliable recommendations [113]. This issue can be broken down into two main types:

- **User Cold Start**: When a new user joins the system, there is little or no historical data on their preferences, making it difficult to provide personalized recommendations [60]. Recommender systems need effective strategies to gather initial preference data, such as asking new users to rate a few items or leveraging demographic information to make initial suggestions. [107].

- **Item Cold Start**: When a new item is added to the catalog, there is limited information on how it might be received by users, which poses a challenge for the system to recommend it effectively. Addressing the item cold start problem involves integrating new items into the recommendation process, possibly by using content-based methods that rely on item attributes or metadata, or by leveraging initial user interactions with the new item [120].

In summary, successfully addressing these challenges requires a balanced approach that ensures recommendations are varied, fresh, broadly applicable, pleasantly surprising, and effectively tailored to both new and existing users and items.

2.8 Recommendation Systems Comparison

Many aspects are used to compare different filtering methods:

Aspect	**Collaborative Filtering**	**Content-Based Filtering**	**Hybrid Filtering**
Data Requirements	Relies on user-item interaction data [3].	Utilizes item features and user preferences [69].	Combines data requirements of both collaborative and content-based methods [71].
Cold Start Problem	Susceptible to the cold start problem for new users or items [81].	More robust in handling cold starts by relying on item features.	Aims to mitigate cold start issues by leveraging both collaborative and content-based methods [87].
Personalization	Offers strong personalization based on similar users' preferences.	Provides personalized recommendations by focusing on item features [37].	Combines collaborative and content-based methods to enhance personalization.
Serendipity and Diversity	May lack serendipity as it relies on user's past behaviors [27].	Can introduce serendipity by recommending items with similar features.	Aims to balance serendipity and diversity by combining collaborative and content-based methods.
Scalability	May face scalability challenges with a growing number of users and items [40].	Generally more scalable as it relies on item features [65].	Scalability depends on the integrated methods; careful design is crucial for large-scale systems.
Explaining Recommendations	Recommendations might lack transparency.	Provides transparent recommendations grounded in explicit item features.	Transparency depends on the integration method; it may offer a balance between transparency and accuracy.
Novelty	May lack novelty based on historical user interactions.	Can introduce novelty by recommending items with similar features.	Strives to balance between providing novel recommendations and aligning with user preferences.
Adaptability	Less adaptable to changes in user preferences over time.	More adaptable as it relies on explicit item features and captures evolving user preferences [73].	Offers adaptability by leveraging collaborative and content-based methods, adjusting to dynamic user behaviors [62].

Table 2.2: Comparison of Recommendation System Aspects

2.9 Conclusion

In conclusion, studying the latest trends in recommendation systems has given us valuable insights into how personalized services are evolving in online shopping. By looking at the basic ideas behind these systems, like how they work and who they're for, we have built a strong understanding of how modern recommendation systems function.

By digging into different research and methods, we have seen the problems that old recommendation systems face, such as cold start, sparsity, etc.

As technology keeps improving, artificial intelligence and machine learning are changing how people use online stores. By using advanced technologies like deep learning and combining different methods, recommendation systems are becoming more efficient, personal, and focused on what users want.

Looking forward, it's important to bring together knowledge from different areas like data science, user feedback, and new technologies to keep improving online shopping recommendations. By building on what we have learned so far, we are on a path to creating smarter, more flexible, and more interesting recommendation systems that meet the diverse needs and tastes of today's shoppers.

Chapter 3

Proposed approach

3.1 Introduction

In this chapter we will discuss our work using a description of our approach,the results obtained,an evaluation using different metrics and its added values.

3.2 Our proposed approach

Our proposed weighted approach combines collaboratif and content-based filtering algorithms : Singular Value Decomposition (SVD) and Term Frequency-Inverse Document Frequency (TF-IDF) .

We will discuss each one to clarify our hybrid approach :

3.2.1 Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) is a widely collaborative method that is based on matrix factorization ,and that reduces the dimensionality of the user-item matrix [33].

By using this decomposition, we can reduce the dimensionality of the rating matrix M while preserving important information about user preferences and item relevance. This allows us to improve the quality of recommendations in recommendation systems.

Identification of Latent Features

SVD decomposes the rating matrix M into three matrices: U, Σ (Sigma), and V^T (transpose of V). These matrices help identify important latent features that describe user preferences and product attributes [18].

Its general formula is:

$$M = U\Sigma V^T$$

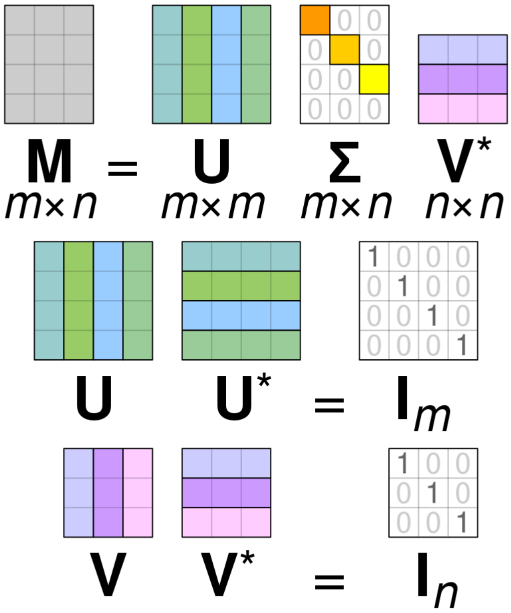


Figure 3.1: Matrix decomposition in SVD algorithm [76]

where:

- M is the original matrix we want to decompose(the preferences matrix(u:user prefers i:item),it is a hollow matrix (i.e ,contains zeros because the user rates some items and not all of them)

- U is the left singular matrix (columns are left singular vectors that represent the users). The columns of U contain eigenvectors of the matrix MM^T.

- Σ is a diagonal matrix containing singular (eigen) values.

- V is the right singular matrix (columns are right singular vectors that represent the items). The columns of V contain eigenvectors of the matrix M^TM.

After this factorisation, SVD predicts ratings, by using the matrices U, Σ, and V^T, our approach can accurately predict the ratings that users will give to products they have not yet evaluated. This allows for personalized recommendations based on each user's individual preferences.

Also, Handling Sparse Data, where most values are missing. It reliably fills in missing values, improving the overall quality of recommendations [28]. The rows of matrix U ($|U| \times k$) express users' interest in each of the k latent factors, while the rows of matrix V ($|I| \times k$) represent the relevance of each item to each latent factor.

SVD algorithm keeps the large singular values of rank K to obtain an approximation with reduced dimension but keeping the more important information [19], with a general formula :

$$M_k = U_k \Sigma_k V_k^T$$

Where:

- U_k is the matrix of truncated left singular vectors at rank k,
- Σ_k is the diagonal matrix of truncated singular values at rank k,
- V_k^T is the matrix of truncated right singular vectors at rank k.

SVD offers the recommendation using the factor matrices to fill the missing entries of the matrix M with predicted values [109] (i.e. give new values based on the previous ones) ,sorting the items (not evaluated): once all the items are evaluated and have a score (i.e. all the cases of the matrix M are filled) . SVD evaluates the items according to the predicted scores then suggests to users those with high scores.

3.2.2 Term Frequency-Inverse Document Frequency (TF-IDF)

To enhance our recommendation system's quality, we've embraced the content-based approach [23], which we'll outline as follows:

TF-IDF (Term Frequency-Inverse Document Frequency), in a broader sense, is a statistical technique gauging a term's (word or phrase) significance within a document concerning a document set. It primarily relies on two metrics:

- **Term Frequency in the Document (TF)**: The term frequency measures how often a specific term appears in a document relative to the total number of terms in that document [26]. It is used to evaluate the relative importance of a term in a specific document.

  $$\text{TF}(t,d) = \frac{\text{Number of times } t \text{ appears in } d}{\text{Total number of terms in } d}$$

  For example, if a document contains 100 words and the term "shirt" appears 5 times, the TF of the term "shirt" in that document would be $\frac{5}{100} = 0.05$.

- **Inverse Document Frequency (IDF)**: The inverse document frequency evaluates the importance of a term in the entire collection of documents by considering how often that term appears in all documents in the collection [85]. It is calculated by taking the logarithm of the ratio between the total number of documents and the number of documents containing the specific term.

  $$\text{IDF}(t) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing term } t}\right)$$

For example, if a collection of documents contains a total of 1000 documents and the term "juice" appears in 100 documents, the IDF of the term "juice" would be $\log\left(\frac{1000}{100}\right) = \log(10) = 1$.

In recommendation systems, specifically, TF-IDF calculates item similarities (like products, articles, videos, etc.) based on their textual descriptions [114], using the following formula:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

Its principle hinges on:

The document representation : by converting The product category into a TF-IDF vector, where each dimension corresponds to a term, and the value represents the term's TF-IDF weight [78], and using linear regression similarity to assess the similarity between TF-IDF vectors [32], by focusing solely on the product category document, it predicts the product rating, capturing subtle relationships between terms and enhancing the recommendation process. Also, facing Sparse data and cold start problem : by highlighting the most vital terms for document representation, TF-IDF with linear regression efficiently manages data sparsity [96]. This facilitates recommending relevant items despite data gaps (new users or items: new data).

This ultimately ensures measuring the frequency and importance of a term in a document or text (such as a product description).

At the end we can say that combining the SVD algorithm with the content-based TF-IDF algorithm in a hybrid approach for recommendation systems assures many advantages:

- Improving data sparsity management.
- Enhancing recommendation accuracy .
- Addressing cold start problems.
- Promoting recommendation diversification for a richer user experience.

3.3 Adopted hybridization technique

Our approach use both SVD and TF-IDF, each with a given importance or contribution weighth to the final recommendation .Doing so, the collaborative filtering component (SVD) provides personalized recommendations based on the patterns in user-item interactions, ensuring that the recommendations are tailored to individual user preferences.

Meanwhile, the content-based component (TF-IDF) ensures that the recommendations are relevant by focusing on the specific features. The weighting mechanism in our approach combines the predictive power of SVD with the descriptive granularity of TF-IDF. This synergy allows us to balance the influence of user behavior patterns with the distinct characteristics of the items, resulting in a comprehensive recommendation system that is both accurate and contextually relevant.

By carefully weighting the contributions from both SVD and TF-IDF, our system effectively captures the multifaceted relationships between users and items, improving the overall recommendation quality.

3.4 Hybrid Model's Performance Variation with α

The Hybrid model's performance depends on the weighting parameter α , By tuning α, the model adjusts the contribution of the SVD and TF-IDF components, allowing for a dynamic balance between collaborative filtering and content-based filtering.

This flexibility ensures that the model can optimize its performance according to the specific characteristics of the dataset and the recommendation context.

For instance, a higher α would focus more on the collaborative filtering aspect, which is effective in identifying latent user-item interactions, such as recommending movies based on viewing history.

On the other hand, a lower α might emphasize the content-based TF-IDF component, which is beneficial in scenarios where item content plays a crucial role in user preferences, such as recommending books based on their descriptions.

This capacity to fine-tune the model's parameters and balance different filtering techniques is essential for achieving superior recommendation accuracy.

It allows the Hybrid model to leverage the strengths of both approaches, ensuring robust performance across diverse datasets and use cases in an e-commerce environment. By effectively handling various recommendation challenges, such as the cold start problem and data sparsity, the Hybrid model stands out as a versatile and powerful tool for delivering personalized user experiences, with a simple and fluid architecture that is represented in the figure 3.2 below:

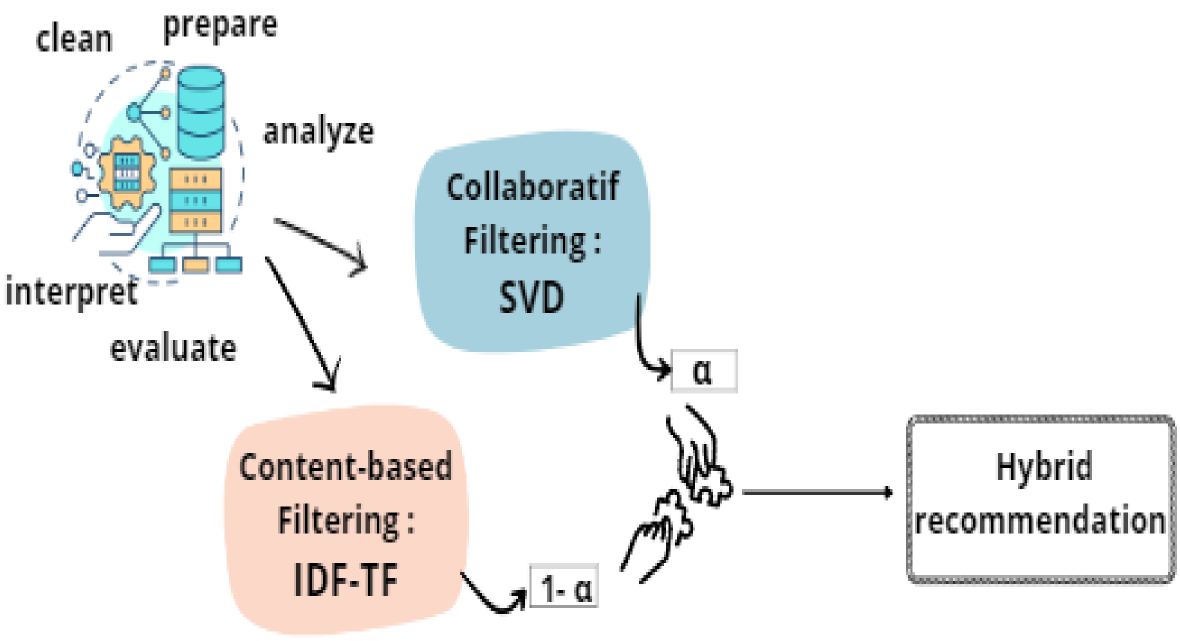


Figure 3.2: Approach's architecture

3.5 Experimental Evaluation

This section presents and analyzes the results obtained by our hybrid recommender system when tested with real data from Amazon Beauty products. We compare these results with those of other established recommender systems, namely Singular Value Decomposition (SVD), k-Nearest Neighbors (k-NN), and TF-IDF with Linear Regression. The evaluation metrics used for this analysis are, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Additionally, we discuss the effectiveness of these models in addressing common challenges in recommendation systems, such as the cold start problem, enhancing serendipity, and handling data sparsity.

3.5.1 Dataset Description

The dataset used for our experiments is an extensive collection of Amazon Beauty product reviews. It was obtained from a publicly available repository on Kaggle, which specifically focuses on reviews of Amazon beauty products (source: Kaggle, 2021) [94]. This dataset has been anonymized to protect user privacy and is provided for academic and research purposes, in compliance with applicable data usage policies and ethical standards.

The dataset on Kaggle was curated by an individual who adapted it from the study *"Justifying Recommendations Using Distantly-Labeled Reviews and Fine-Grained Aspects"* by Jianmo Ni, Jiacheng Li, and Julian McAuley, published in the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) [75]. The original research provided a framework for extracting and utilizing detailed information from product reviews, which the Kaggle contributor leveraged to compile a relevant and useful dataset for further analysis.

In summary, the Amazon Beauty product reviews dataset is an invaluable resource for conducting comprehensive analyses and developing advanced recommendation systems. Its extensive coverage and granular details make it an excellent choice for achieving the research objectives outlined in this study.

3.5.1.1 Dataset Overview

The dataset comprises 20,000 entries, each representing an individual review of a beauty product available on Amazon. The dataset captures various dimensions of user interactions and product features, facilitating an in-depth analysis of user preferences and product characteristics. The core features included in the dataset are:

- **UserId**: A unique identifier assigned to each user, facilitating the monitoring of individual behavior and preferences across various products. The dataset contains 19,586 unique user IDs, enabling a comprehensive analysis of diverse user interactions.

- **ProductId**: A unique identifier assigned to each product, aiding in the aggregation of reviews and the analysis of specific products' performance and reception. The dataset includes 10,526 unique product IDs, providing a robust foundation for detailed product-level insights.

- **ProductType**: The dataset includes 22 unique categories of beauty products, such as Nail Polish (Represents 28.32% of all products), Shower Gel (15.26%) and Sunscreen (10.11%). This feature categorizes each product by its type, enabling effective segmentation of the data and in-depth analysis of trends within specific product categories.

- **Rating**: A numerical rating given by the user, typically on a scale from 1 to 5. Ratings are pivotal for understanding user satisfaction and are a primary target variable for recommendation algorithms.

  Figure 3.3 displays the distribution of different ratings, showing how frequently each rating value (1 through 5) occurs within the dataset.

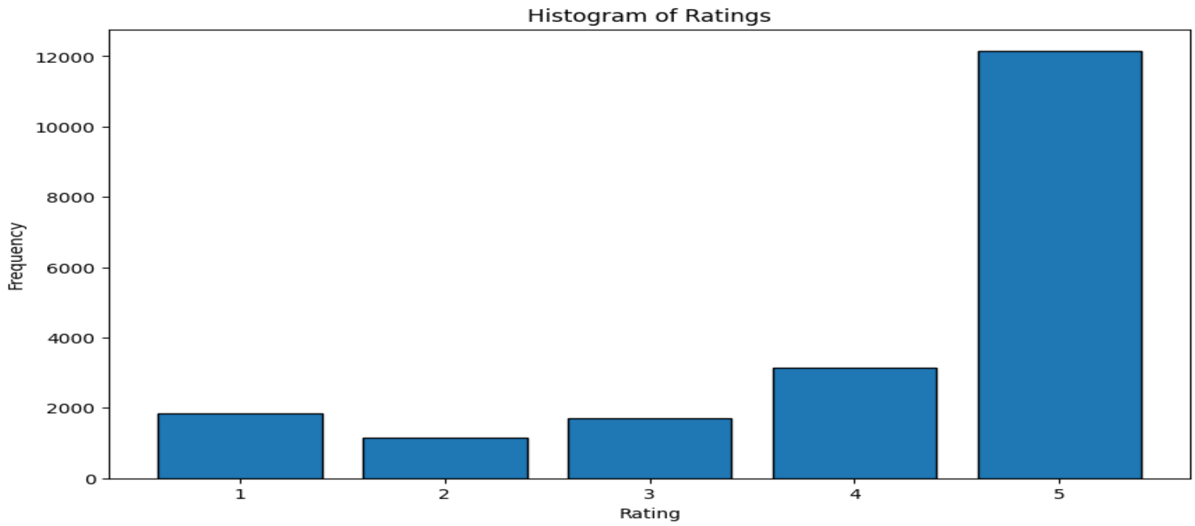
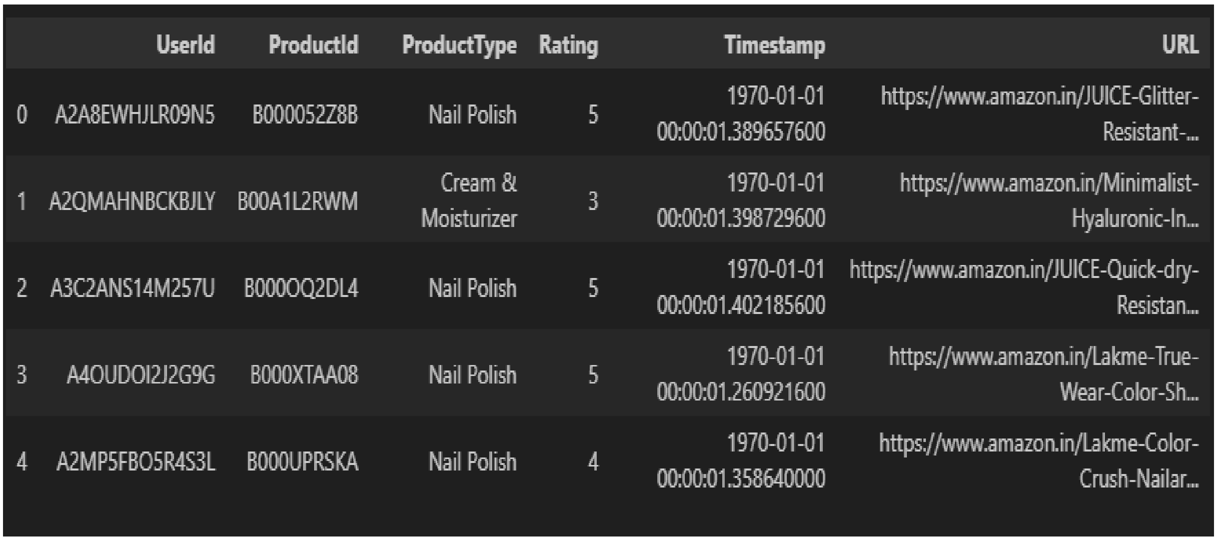


Figure 3.3: Distribution of Product Ratings

- **Timestamp**: The date and time when the review was submitted. This feature enables temporal analysis, such as tracking changes in user preferences or product popularity over time.
- **URL**: The web link to the product's page on Amazon. This feature provides direct access to the product for further exploration and validation of data.

Figure 3.4 represents a snapshot of the first few rows of the dataset



	UserId	ProductId	ProductType	Rating	Timestamp	URL
0	A2A8EWHJLR09N5	B000052Z8B	Nail Polish	5	1970-01-01 00:00:01.389657600	https://www.amazon.in/JUICE-Glitter-Resistant-...
1	A2QMAHNBCKBJLY	B00A1L2RWM	Cream & Moisturizer	3	1970-01-01 00:00:01.398729600	https://www.amazon.in/Minimalist-Hyaluronic-In...
2	A3C2ANS14M257U	B000OQ2DL4	Nail Polish	5	1970-01-01 00:00:01.402185600	https://www.amazon.in/JUICE-Quick-dry-Resistan...
3	A4OUDOI2J2G9G	B000XTAA08	Nail Polish	5	1970-01-01 00:00:01.260921600	https://www.amazon.in/Lakme-True-Wear-Color-Sh...
4	A2MP5FBO5R4S3L	B000UPRSKA	Nail Polish	4	1970-01-01 00:00:01.358640000	https://www.amazon.in/Lakme-Color-Crush-Nailar...

Figure 3.4: Snapshot of the First Few Rows of the Dataset

3.5.1.2 Dataset Composition

- **Entries and Scope**: The dataset encompasses a broad range of beauty products reviewed over several years. This diversity ensures that the dataset reflects varied user experiences and product evaluations, making it a valuable resource for comprehensive analysis.

3.5.1.3 Data Quality and Integrity

- **Data Completeness**: The authors of the dataset have taken great care to insure that it is entirely complete, with no missing values in any of the crucial fields such as UserId, ProductId, and Rating. This ensures the dataset is fully intact and ready for analysis. [94, 75]

- **Data Accuracy**: The authors have diligently ensured the integrity and reliability of the data. This process included extensive checks for duplicate entries and confirmation that all ratings are whole numbers, ranging from 1 to 5. [94, 75]

3.5.2 Implementation of Our Recommendation System

In our implementation of a hybrid recommendation system, we utilized Python and its powerful libraries such as `scikit-learn` for TF-IDF, `Surprise` for Singular Value Decomposition (SVD), and `pandas` and `numpy` for data manipulation and analysis. The process involved several key steps, including data preparation, model training, and evaluation.

3.5.2.1 Data Preparation and Splitting

We began by preparing our dataset, using `pandas` and `numpy`, we loaded and preprocessed the data to ensure it was suitable for training and testing our models. Specifically, we created a user-item ratings matrix, representing the interactions between users and items in terms of ratings. We also extracted the product categories, which were used as textual attributes for content-based filtering. The data was then split into a training set and a test set to facilitate the evaluation, ensuring that the training set was used to build the models while the test set was reserved for performance assessment.

3.5.2.2 Model Training

To create a robust hybrid recommendation system, we trained several classic recommendation models using different aspects of our dataset:

TF-IDF (Term Frequency-Inverse Document Frequency): We used `scikit-learn` to implement the TF-IDF model, which transformed the product categories into numerical vectors. This model was particularly effective for content-based filtering, allowing us to recommend items with similar attributes to those that the user had previously interacted with.

SVD (Singular Value Decomposition): We employed the `Surprise` library to implement the SVD model, which is a popular approach for collaborative filtering. SVD was trained using the user-item ratings matrix, capturing the interactions between users and items. By decomposing this matrix into latent factors, SVD identifies patterns in user preferences and item characteristics, enabling effective recommendations.

k-Nearest Neighbors (KNN): For our regression-based approach, we implemented k-Nearest Neighbors (KNN) using `scikit-learn`. The KNN model was also trained using the user-item ratings matrix. This model recommends items based on the similarity between users or items, determined by their ratings patterns. KNN identifies users or items that are close in the feature space, allowing it to suggest items that similar users have interacted with or rated highly.

3.5.2.3 Model Evaluation

Once our models were trained, we evaluated their performance using the test set. This involved predicting user ratings or interactions for the items in the test set and comparing these predictions against the actual data. We measured the accuracy of each model using metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), providing insight into their effectiveness in making recommendations.

By integrating these various recommendation techniques, we were able to leverage the strengths of each approach. The TF-IDF model excelled at identifying content similarities based on product categories, SVD effectively captured latent user-item interactions from the ratings matrix, and KNN provided robust neighbor-based recommendations using user-item rating patterns.

The combination of these models enabled our hybrid system to deliver more precise and diverse recommendations, significantly enhancing the user experience in the e-commerce context.

3.5.3 Evaluation Metrics

The following metrics were used to evaluate the performance of the recommender systems:

3.5.3.1 Root Mean Squared Error (RMSE)

The RMSE is the square root of MSE, offering a more interpretable measure of error by being in the same unit as the ratings.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where:

- The same notations as for MSE apply. - The square root applied to the average of the squared errors.

3.5.3.2 Mean Absolute Error (MAE)

The MAE is the average of the absolute errors (absolute differences between predicted and actual ratings), providing a straightforward interpretation of the magnitude of errors in the same unit as the ratings.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where:

- y_i is the true rating. - $\hat{y}_i$ is the predicted rating. - n is the total number of ratings. - The absolute value of the errors is averaged.

3.6 Performance Results

After evaluating our approach using different metrics described above ,we could summarise the results obtained in the table below :

Table 3.1: Performance Results

Recommender System	RMSE	MAE
Hybrid (α)	**1.2945**-1.3328	**1.0405**-1.0408
SVD	1.2954	1.0416
k-NN	1.3191	1.0516
TF-IDF with Regression	1.3431	1.0477

3.7 Comparative Analysis

3.7.1 Accuracy Metrics

- **Root Mean Squared Error (RMSE)**

Our evaluation revealed that the Hybrid model once again excelled with RMSE values between **1.2945** when α is 0.8 and **1.3328** when α is 0.1.

This range highlights that the Hybrid model consistently minimizes the prediction errors across various scenarios, thereby demonstrating its robustness and flexibility. The lower RMSE values reflect the model's ability to deliver precise recommendations with fewer large deviations from actual ratings, which is critical for maintaining a positive user experience in an e-commerce setting.

The SVD model, with an RMSE of **1.2954**, performed almost on par with the lower bound of the Hybrid model's performance. This similarity underscores the effectiveness of SVD in capturing the underlying patterns in user preferences.

However, the slightly broader range in the Hybrid model's RMSE indicates that it can adapt to different configurations while maintaining a high accuracy level.

The k-NN (k-Nearest Neighbors) model showed an RMSE of **1.3191**, which is higher than both the Hybrid and SVD models. This suggests that while k-NN is effective in identifying similar users or items based on a proximity measure, it may not capture the latent factors as effectively as SVD.

The higher RMSE reflects its moderate performance, indicating that k-NN may introduce more significant errors, particularly when the dataset is sparse or when user preferences are not well represented by simple proximity measures.

- **Mean Absolute Error (MAE)**

The Hybrid model achieved the lowest MAE values, ranging from **1.0408** α is 0.8 to **1.0405** α is 0.1. This consistency indicates minimal error variance across different α values, showcasing the model's capability to provide uniformly accurate recommendations regardless of the specific balance between collaborative and content-based components.

The low MAE values reflect the Hybrid model's proficiency in making accurate predictions with minimal deviation from the true ratings, thereby enhancing its reliability as a recommender system.

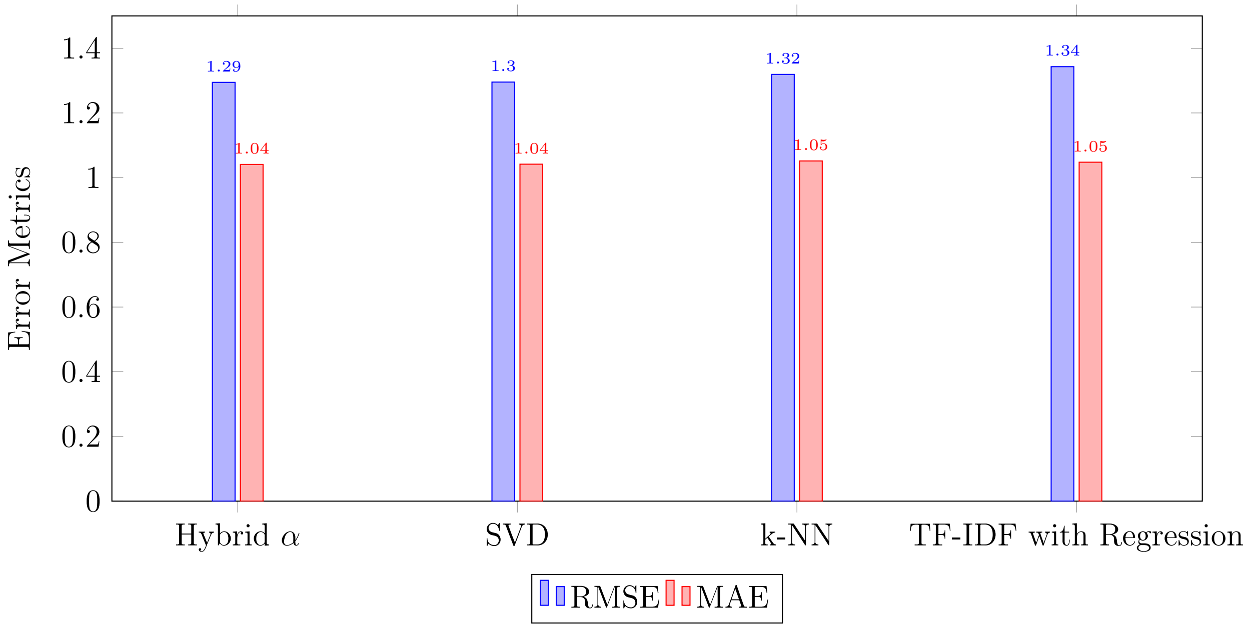
Following closely, the SVD model recorded an MAE of **1.0416**. This proximity to the Hybrid model's performance emphasizes SVD's effectiveness in minimizing prediction errors by capturing the latent factors that drive user preferences. However, the Hybrid model's slight advantage suggests that integrating content-based features can further refine the accuracy of recommendations, particularly in cases where user-item interactions are not solely determined by past behavior but also by content characteristics.

The k-NN model and the TF-IDF with Regression model showed slightly higher MAE values of **1.0516** and **1.0477**, respectively.

These higher values indicate that these models tend to deviate more from the actual ratings on average, reflecting less precise predictions.

- For k-NN, the higher MAE suggests that the model is less effective than SVD or the Hybrid model in capturing complex user-item relationships.

We could summarize the different results obtained in table 3.1 in a histogram :



3.7.2 Addressing Key Challenges

Our approach effectively tackles several challenges inherent in recommender systems, including the cold start problem, serendipity, and data sparsity.

3.7.2.1 Cold Start Problem

The cold start problem arises when a new user or item has insufficient data, making it challenging to generate accurate recommendations. The hybrid model mitigates this issue by leveraging TF-IDF's content-based filtering, which does not rely on user history.

By analyzing the textual features of items, the model can recommend relevant products to new users or suggest new items based on their content. This makes the hybrid model particularly effective in scenarios with many new users or products.

3.7.2.2 Enhancing Serendipity

The hybrid model, by combining SVD and TF-IDF, can suggest products that users may not have explicitly searched for but are likely to find engaging.

SVD captures latent preferences through collaborative filtering, while TF-IDF introduces novelty by recommending items with similar content but less obvious relevance. This dual approach broadens the range of recommendations, increasing the likelihood of serendipitous discoveries.

3.7.2.3 Handling Data Sparsity

Data sparsity which represent a common issue in recommender systems where user-item interaction data is sparse, leading to challenges in making accurate predictions. The hybrid model effectively addresses this by combining collaborative and content-based filtering. More precisely, SVD efficiently deals with sparse matrices by decomposing them into latent factors, capturing the underlying structure despite missing values. While TF-IDF complements this by leveraging available textual data, ensuring that the model can generate recommendations even in sparse environments. This hybrid approach makes the model resilient to data sparsity, improving its robustness and reliability.

3.7.3 The Impact of α on The Model's Performance

The parameter α is pivotal in tuning the performance of our recommendation system, our approach allows us to balance the strengths of collaborative filtering through SVD and content-based filtering via TF-IDF, catering to various recommendation scenarios.

Figure 3.4 illustrates how RMSE and MAE metrics vary across different α values.

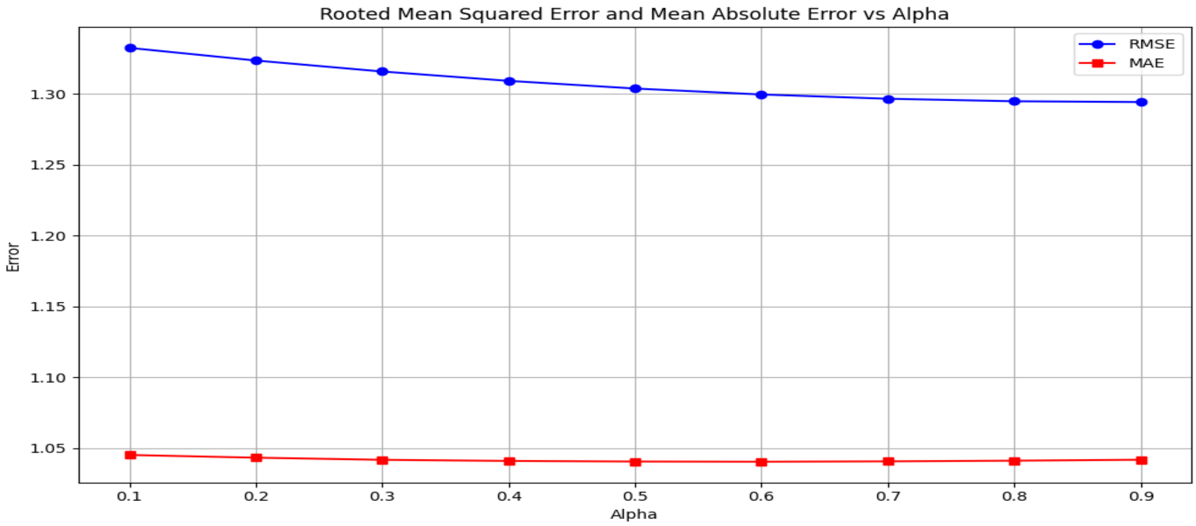


Figure 3.5: Performance Metrics for Different α Values

3.7.3.1 Effects of Higher α Values

When α is increased, the model relies more on SVD, resulting in several notable effects. Firstly, higher values of α typically enhance accuracy metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), indicating improved precision in recommendations due to SVD's adeptness at capturing latent factors in user-item interactions.

However, this heightened emphasis on SVD introduces trade-offs: it exposes the model to the cold start problem by increasing reliance on historical data, thereby hindering recommendations for new users or items lacking sufficient interaction history. Moreover, as α rises, the model tends to reduce serendipitous recommendations, focusing more on established patterns from past interactions rather than exploring diverse options. Furthermore, in scenarios with sparse data, a high α can lead to overfitting, as the model overly relies on limited interactions, compromising its ability to generalize to unseen data.

3.7.3.2 Effects of Lower α Values

Lowering α diminishes the influence of SVD and elevates the importance of the content-based model using TF-IDF, resulting in several specific outcomes. Firstly, with a reduced α, the model relies more on TF-IDF, which uses item features rather than user interaction history. This adjustment proves advantageous in scenarios where item descriptions alone suffice to generate pertinent recommendations, addressing the cold start problem effectively.

Additionally, a lower α enhances the model's capacity to introduce users to novel and unexpected items by emphasizing content features. This approach can heighten the novelty of recommendations, thereby enhancing user engagement. However, decreasing α may compromise the accuiracy of the model because it relies less on SVD's collaborative filtering capabilities, which excel in capturing intricate user-item interaction patterns.

3.7.3.3 Balancing α for Optimal Performance

Fine-tuning α enables optimal model performance across different contexts by balancing the strengths of SVD and TF-IDF. Opting for moderate α values ensures a harmonious blend of leveraging SVD for precise predictions and using TF-IDF to manage cold start challenges and promote serendipity.

Moreover, dynamically adjusting α based on specific data and circumstances can further optimize the performance. Lower α values are advantageous when handling new users or items, while higher values are more effective when there is sufficient user interaction data for robust collaborative filtering.

In summary, α is a critical parameter that influences the balance between the accuracy of recommendations and the model's ability to handle challenges such as the cold start problem, serendipity, and data sparsity. By carefully tuning α, we can create a recommendation system that is both versatile and capable of meeting diverse user needs and conditions.

3.8 Conclusion

In conclusion, our hybrid model has shown remarkable performance on real-world data, outperforming traditional approaches such as SVD, k-NN, and TF-IDF with Regression in key metrics like Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). This model's strength lies in its ability to dynamically balance collaborative filtering with content-based recommendations, optimizing its effectiveness across various scenarios.

This dual approach enhances recommendation accuracy and user satisfaction, making it a powerful tool for e-commerce platforms.

The model addresses critical challenges inherent in recommendation systems. It effectively tackles the cold start problem by leveraging both interaction data and textual content, ensuring new users and items receive relevant recommendations. By integrating TF-IDF, the system enriches recommendations with diverse and unexpected items, enhancing user engagement and satisfaction. Moreover, SVD's capability to extract latent factors from sparse data enables the model to make accurate predictions even with limited user-item interactions, addressing the common issue of data sparsity.

As a comprehensive recommendation tool, our model provides a robust solution that improves accuracy and effectively handles the complexities of sparse data and new user/item scenarios. Its ability to adapt and optimize across different situations makes it invaluable for enhancing the user experience and driving engagement on e-commerce platforms.

Future work could explore further tuning of the hybrid model's parameters and integrating additional data sources to enhance its predictive power and user experience.

Conclusion and Future Perspectives

In the ever-evolving landscape of e-commerce, the challenge of providing personalized and accurate product recommendations remains paramount. Traditional recommendation approaches often struggle with issues such as data sparsity, the cold start problem, and the need for serendipitous discoveries to keep users engaged. Addressing these challenges is crucial for enhancing user satisfaction, increasing engagement, and ultimately driving sales in a competitive online marketplace.

Recognizing these challenges, our research embarked on developing a robust recommendation system capable of overcoming the limitations of traditional methods. We aimed to create a model that could effectively handle sparse data, provide relevant suggestions to new users, and about new items, and enhance the discovery of unexpected yet interesting products. Through a comprehensive investigation of various techniques, we settled on a hybrid model that integrates Singular Value Decomposition (SVD) and Term Frequency-Inverse Document Frequency (TF-IDF) in our recommendation system.

The hybrid model we developed combines the strengths of collaborative filtering and content-based recommendation approaches. SVD, a powerful collaborative filtering technique, excels at uncovering latent factors from user-item interaction data, even when such data is sparse. It captures the hidden relationships between users and products, allowing for accurate predictions of user preferences. On the other hand, TF-IDF leverages the textual content associated with items, to find similarities between products, enabling the model to make recommendations even in the absence of sufficient user interaction data.

During the experimental phase, we tested our hybrid model using real-world data from Amazon beauty products. The results were compelling, demonstrating superior performance across key accuracy metrics, including Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Our hybrid model consistently outperformed traditional models such as SVD, k-NN, and TF-IDF with Regression, highlighting its robustness and effectiveness.

One of the key advantages of our hybrid model is its ability to dynamically adjust the weighting between collaborative and content-based components. This flexibility allows the model to optimize its performance across different scenarios, catering to a variety of user behaviors and preferences. It provides a practical solution for improving recommendation accuracy and user satisfaction, making it a valuable tool for e-commerce recommender systems.

The model's efficacy in addressing the cold start problem is particularly noteworthy. By analyzing the textual features of new items and leveraging the content-based capabilities of TF-IDF, the model can recommend relevant products to new users or suggest new items based on their content. This capability ensures that even new users or items can be effectively integrated into the recommendation process, maintaining the system's effectiveness in dynamic environments.

Furthermore, the hybrid model enhances serendipity by combining the comprehensive user preference capture of SVD with the novelty-driven suggestions of TF-IDF. This dual approach broadens the range of recommendations, increasing the likelihood of users discovering unexpected and engaging products that align with their interests.

The success of our hybrid model underscores the potential of blending collaborative and content-based approaches to create a more personalized and accurate recommendation system. It highlights the model's ability to capture nuanced relationships between items and user preferences, significantly enhancing the overall quality of recommendations. The comprehensive evaluation on real-world data showcases the model's scalability, accuracy, and coverage, demonstrating its robustness and capacity to handle diverse user preferences and extensive product catalogs.

In summary, our hybrid model offers a sophisticated yet practical solution for enhancing user experience, driving engagement, and increasing revenue for businesses in the competitive online marketplace.

And finally, for us as students, this experience that we went through in carrying out this work brought us several values, delving in the e-commerce sector and the importance of AI in it, and for the application of our knowledge in the field of machine learning for added values and more performance in our field of research, also for the discipline in the context of carrying out this modest work, giving us more opportunities for future works and successes.

Looking forward, several avenues for further research can be explored to enhance the hybrid model:

- **Model Parameter Optimization**: by conducting experiments with different similarity thresholds and weighting schemes for collaborative and content-based components.

  For example, adjusting the threshold for collaborative filtering to include more or fewer similar users or products, and fine-tuning the weights assigned to different features in content-based filtering, such as product attributes or user preferences.

- **Integration of Additional Data Sources**: by incorporating user ratings from social media platforms like Facebook or Twitter to supplement existing ratings on the e-commerce platform.

  Contextual data like user location or time of day could be used to personalize recommendations further, such as suggesting winter clothing items to users in colder regions during winter months.

- **In-depth User Experience Evaluation**: by conducting surveys or interviews with users to gather feedback on their satisfaction with recommended products and overall experience with the recommender system. Analyzing click-through rates, conversion rates, and user engagement metrics to assess the effectiveness of recommendations and identify areas for improvement.

- **Integration of Deep Learning Techniques**: exploring the use of deep learning techniques like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to analyze unstructured data such as product descriptions, reviews, or images.

  For example, using CNNs to extract features from product images and RNNs to process textual data, leading to more accurate and personalized recommendations based on visual and textual similarities between products.

- **Exploration of Different Forms of Hybridization**:the success of our hybrid model opens avenues for future research to explore different methods of hybridization. For instance, investigating cascade methods that sequentially combine outputs from individual recommendation algorithms could potentially enhance recommendation accuracy by leveraging complementary strengths. Ensemble techniques, such as blending predictions from diverse algorithms using stacking models, offer another promising approach to improve recommendation quality by aggregating diverse perspectives.

These avenues of research offer promising perspectives to extend and improve our hybrid recommendation model, paving the way for more efficient and personalized recommendation systems for e-commerce.

Bibliography

[1] Hussam Abdulla, Asim Eltahir, Saleh Alwahaishi, Khalifa Saghair, Jan Platos, and Vaclav Snasel. Chatbots development using natural language processing: A review. pages 122–128, 07 2022.

[2] Nikhil Agnihotri. Classification of machine learning algorithms. *Engineers Garage*. Accessed on May 16, 2024.

[3] Kamal Al-Barznji and Atanas Atanassov. Collaborative filtering techniques for generating recommendations on big data. 10 2017.

[4] Hael Al-bashiri, Mansoor Abdulhak, Awanis Romli, and Fadhl Hujainah. Collaborative filtering recommender system: Overview and challenges. *Advanced Science Letters*, 23:9045–9049, 09 2017.

[5] Hael Al-bashiri, Mansoor Abdulhak, Awanis Romli, and Fadhl Hujainah. Collaborative filtering recommender system: Overview and challenges. *Advanced Science Letters*, 23:9045–9049, 09 2017.

[6] Nourah Al-Rossais. Improving cold start stereotype-based recommendation using deep learning. *IEEE Access*, PP:1–1, 01 2023.

[7] Alain clapaud. Machine learning : décryptage d'une technologie qui monte. `https://www.journaldunet.com/iot/1152606-machine-learning-decryptage-d-une-technologie-qui-monte/`, 2015. Consulted May 16,2024.

[8] Pegah Alamdari, Nima Navimipour, Mehdi Hosseinzadeh, Ali Safaei, and Aso Darwesh. A systematic study on the recommender systems in the e-commerce. *IEEE Access*, PP:1–1, 06 2020.

[9] Constantin Aliferis and Gyorgy Simon. *Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI*, pages 477–524. 03 2024.

[10] Amer Alnuaimi and Tasnim Albaldawi. An overview of machine learning classification techniques. *BIO Web of Conferences*, 97:00133, 04 2024.

[11] Steven Alter. System interaction theory: Describing interactions between work systems. *Communications of the Association for Information Systems*, 42, 02 2018.

[12] Tomás Alves, Daniel Gonçalves, Sandra Gama, and Joana Henriques-Calado. Examining user preferences based on personality factors in graphical user interface design. pages 1–9, 11 2022.

[13] A. Anjum, A. Akram, and T. Mahmood. Hybrid recommender system for movie recommendations using deep learning. *Applied Soft Computing*, 74:372–381, 2019.

[14] Payam Bahrani, Behrouz Minaei-Bidgoli, Hamid Parvin, Mitra Mirzarezaee, and Ahmad Keshavarz. A new improved knn-based recommender system. *The Journal of Supercomputing*, 80(1):800–834, 2024.

[15] Vimala Balakrishnan, Zhongliang Shi, Chuan Law, Regine Lim, Lee Teh, and Yue Fan. A deep learning approach in predicting products' sentiment ratings: a comparative analysis. *The Journal of Supercomputing*, 78, 04 2022.

[16] Daniel Berrar. *Cross-Validation.* 01 2018.

[17] Yanelys Betancourt and Sergio Ilarri. Use of text mining techniques for recommender systems. In *ICEIS (1)*, pages 780–787, 2020.

[18] Aditya Bhardwaj. Movie recommendation system using svd (letterboxd). *IJARCCE*, 12, 10 2023.

[19] Steven Brunton and J. Kutz. *Singular Value Decomposition (SVD)*, pages 3–46. 02 2019.

[20] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12:331–370, 2002.

[21] Robin Burke, Alexander Felfernig, and Mehmet Göker. Recommender systems: An overview. *Ai Magazine*, 32:13–18, 09 2011.

[22] Erion Çano and Maurizio Morisio. Hybrid recommender systems: A systematic literature review. *Intelligent data analysis*, 21(6):1487–1524, 2017.

[23] Jie Chen, Cai Chen, and Yi Liang. Optimized tf-idf algorithm with the adaptive weight of position of word. 01 2016.

[24] X. Chen, L. Wang, C. Zhang, and X. Xu. An improved weighted hybrid recommender system based on user preferences and similarities. *Complexity*, 2022:1869762, 2022.

[25] Prakhar Dixit. Sentiment analysis: Building from the ground up, May 9, 2020. Accessed: 2024-06-12.

[26] Tripti Dodiya. Using term frequency - inverse document frequency to find the relevance of words in gujarati language. *International Journal for Research in Applied Science and Engineering Technology*, 9:378–381, 04 2021.

[27] Tomislav Duricic, Dominik Kowald, Emanuel Lacic, and Elisabeth Lex. Beyond-accuracy: a review on diversity, serendipity, and fairness in recommender systems based on graph neural networks. *Frontiers in Big Data*, 6, 12 2023.

[28] Antonella Falini. A review on the selection criteria for the truncated svd in data science applications. *Journal of Computational Mathematics and Data Science*, 5:100064, 10 2022.

[29] Xiang Gao, Zhiliang Zhu, Xue Hao, and Hai Yu. An effective collaborative filtering algorithm based on adjusted user-item rating matrix. *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*, pages 693–696, 2017.

[30] Subhasish Ghosh, Nazmun Nahar, Mohammad Wahab, Munmun Biswas, Mohammad Hossain, and Karl Andersson. *Recommendation System for E-commerce Using Alternating Least Squares (ALS) on Apache Spark*, pages 880–893. 02 2021.

[31] Elliot Gould, Hannah Fraser, Timothy Parker, Simon Griffith, Peter Vesk, Fiona Fidler, Daniel Hamilton, Robin Abbey-Lee, Jessica Abbott, Luis Aguirre, Carles Alcaraz, Irith Aloni, Drew Altschul, Kunal Arekar, Jeff Atkins, Joe Atkinson, Chris Baker, Meghan Barrett, and Rachel Zitomer. Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology. 10 2023.

[32] Keisha Harmandini and Kemas L. Analysis of tf-idf and tf-rf feature extraction on product review sentiment. *Sinkron*, 8:929–937, 03 2024.

[33] Arsalan Hasanvand. An introduction to singular value decomposition. 04 2023.

[34] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[35] Alain Hertz, Tsvi Kuflik, and Noa Tuval. Estimating serendipity in content-based recommender systems. 06 2023.

[36] Tianzhen Hong, Zhe Wang, Xuan Luo, and Wanni Zhang. State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, 212:109831, 2020.

[37] Yujiao Hu, Xiaolin Gui, Xinyue Hu, Cong Zeng, and Youqi Wu. Content-based personalized dating recommendation system. *Journal of Physics: Conference Series*, 1314:012104, 10 2019.

[38] IBM - Intelligence Artificielle. `https://www.ibm.com/fr-fr/topics/artificial-intelligence`. Consulté le 14 mai 2024.

[39] F.O. Isinkaye, Y.O. Folajimi, and B.A. Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3):261–273, 2015.

[40] Folasade Isinkaye. How does scalability affect collaborative filtering?, 08 2021.

[41] Margarita Išoraitė and Neringa Miniotienė. Electronic commerce: Theory and practice. *IJBE (Integrated Journal of Business and Economics)*, 2:73, 06 2018.

[42] Sonal Jain, Harshita Khangarot, and Shivank Singh. *Journal Recommendation System Using Content-Based Filtering: IC3 2018*, pages 99–108. 01 2019.

[43] Vipin Jain, Bindoo Malviya, and Satyendra Arya. An overview of electronic commerce (e-commerce). *Journal of Contemporary Issues in Business and Government*, 27:665–670, 05 2021.

[44] Aariz Faizan Javed and Syed Abdullah Ashraf. Novelty in recommender systems for effective personalization in e-commerce and retail. *Journal of Informatics Education and Research*, 3(2):488, 2023.

[45] Esra Kahya-Özyirmidokuz. Mining unstructured turkish economy news articles. *Procedia Economics and Finance*, 16, 05 2014.

[46] W. Kang and J. Kim. Context-aware hybrid recommender system for tourist attractions. *Information Sciences*, 478:469–482, 2019.

[47] Manami Kawasaki and Takashi Hasuike. A recommendation system by collaborative filtering including information and characteristics on users and items. pages 1–8, 11 2017.

[48] Hassan A. Khalil. Towards optimizing hybrid movie recommender systems. *Journal of Movie Recommender Systems*, 1(1):1–10, 2024.

[49] Muhammad Khan, Wiyada Kumam, Poom Kumam, Muhammad Riaz, and khalid naeem. Information measures for pythagorean m-polar fuzzy sets and their applications to robotics and movie recommender system. *AIMS Mathematics*, 8:10357–10378, 02 2023.

[50] Talha Khan, Rehan Sadiq, Zeeshan Shahid, Md Alam, and Mazliham Su'ud. Sentiment analysis using support vector machine and random forest. *Journal of Informatics and Web Engineering*, 3:67–75, 02 2024.

[51] Shristi Khanal, Prasad P.W.C, Abeer Alsadoon, and Angelika Maag. A systematic review: machine learning based recommendation systems for e-learning. *Education and Information Technologies*, 25:1–30, 07 2020.

[52] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: Recommendation models, techniques, and application fields. *Electronics*, 11(1):141, 2022.

[53] Natalia Koupanou. How snorkel, a semi-supervised learning technique, solved invoice accounting at tide, Apr 22, 2020. Accessed: 2024-06-11.

[54] Akshay Kulkarni, Adarsha Shivananda, Anoosh Kulkarni, and V Krishnan. *Content-Based Recommender Systems*, pages 63–87. 11 2022.

[55] Larousse - Intelligence Artificielle. `https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257`. Consulté le 14 mai 2024.

[56] George lawton. What is regression in machine learning?, 30 Aug 2023. Accessed: 2024-06-12.

[57] Legito, Fegie Wattimena, Yulianto Rofi'i, and Munawir. E-commerce product recommendation system using case-based reasoning (cbr) and k-means clustering. *International Journal Software Engineering and Computer Science (IJSECS)*, 3:162–173, 08 2023.

[58] Lei Lei, Yue Tan, Kan Zheng, Shiwen Liu, Kuan Zhang, and Xuemin Shen. Deep reinforcement learning for autonomous internet of things: Model, applications and challenges. *IEEE Communications Surveys and Tutorials*, PP:1–1, 04 2020.

[59] X. Li, M. Yu, and J. Yu. Hybrid collaborative filtering with feature combination for cold-start recommendation. *Information Sciences*, 568:75–88, 2021.

[60] Yunfei Li and Shichao Yin. User cold start recommendation system based on hofstede cultural theory. *International Journal of Web Services Research*, 20:1–17, 01 2023.

[61] Ping Liu, Mengchu Xie, Jing Bian, Huishan Li, and Liangliang Song. A hybrid pso–svm model based on safety risk prediction for the design process in metro station construction. *International Journal of Environmental Research and Public Health*, 17:1714, 03 2020.

[62] Nguyen Luong Vuong, Trinh Vo, and Tri-Hai Nguyen. Adaptive knn-based extended collaborative filtering recommendation services. *Big Data and Cognitive Computing*, 7:106, 05 2023.

[63] Tingjun Mao. Real estate price prediction based on linear regression and machine learning scenarios. *BCP Business Management*, 38:400–409, 03 2023.

[64] Rachana Mehta and Keyur Rana. A review on matrix factorization techniques in recommender systems. In *2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*, pages 269–274. IEEE, 2017.

[65] Ravita Mishra and Sheetal Rathi. *Efficient and Scalable Job Recommender System Using Collaborative Filtering*, pages 842–856. 05 2020.

[66] ml concepts.com. Overfitting and underfitting in machine learning (ml) concepts. `https://www.linkedin.com/pulse/overfitting-underfitting-machine-learning-ml-concepts-com/`. Accessed on May 16, 2024.

[67] Weilong Mo, Xiaoshu Luo, Yexiu Zhong, and Wenjie Jiang. Image recognition using convolutional neural network combined with ensemble learning algorithm. *Journal of Physics: Conference Series*, 1237:022026, 06 2019.

[68] Christian Moewes and Rudolf Kruse. Fuzzy control for knowledge-based interpolation. pages 91–101, 01 2012.

[69] Marwa Mohamed, Mohamed Khafagy, and Mohamed Ibrahim. Recommender systems challenges and solutions survey. 02 2019.

[70] Saeed Mohsen, Ahmed Elkaseer, and Steffen Scholz. Human activity recognition using k-nearest neighbor machine learning algorithm. 09 2021.

[71] D. Monti and et al. Explainable recommender systems: A survey and new perspectives. *ACM Transactions on Information Systems*, 39(1):1–33, 2021.

[72] Abdallah Moubayed, Mohammadnoor Injadat, Ali Nassif, Hanan Lutfiyya, and Abdallah Shami. E-learning: Challenges and research opportunities using machine learning and data analytics. *IEEE Access*, PP:1–1, 07 2018.

[73] Balakumar Muniandi, Apeksha Garg, and Eric Howard. Adaptive content recommendation systems for digital marketing platforms: A deep learning approach. page 2024, 04 2024.

[74] Ismail Muraina, Edward Aiyegbusi, and Solomon Abam. Decision tree algorithm use in predicting students' academic performance in advanced programming course. *International Journal of Higher Education Pedagogies*, 3:13–23, 01 2023.

[75] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *Empirical Methods in Natural Language Processing (EMNLP)*, 2019. Retrieved from `https://arxiv.org/abs/1905.12614`.

[76] Daniel Nilson. What is dimensionality reduction? `https://www.unite.ai/what-is-dimensionality-reduction/`, October 20 2020. Consulted June 12,2024.

[77] Gold Okorie, Zainab Egieya, Uneku Ikwue, Chioma Udeh, Ejuma Adaga, Obinna DaraOjimba, and Osato Oriekhoe. Leveraging big data for personalized marketing campaigns: A review. *International Journal of Management Entrepreneurship Research*, 6:216–242, 02 2024.

[78] Dagobert Pakpahan, Veronika Siallagan, and Saut Siregar. Classification of e-commerce product descriptions with the tf-idf and svm methods. *sinkron*, 8:2130–2137, 10 2023.

[79] Manish Panwar, Jayesh Jogi, Mahesh Mankar, Mohamed Alhassan, and Shreyas Kulkarni. Detection of spam email. *American Journal of Innovation in Science and Engineering*, 1:18–21, 12 2022.

[80] Tulasi K Paradarami, Nathaniel D Bastian, and Jennifer L Wightman. A hybrid recommender system using artificial neural networks. *Expert Systems with Applications*, 83:300–313, 2017.

[81] Ronakkumar Patel and Priyank Thakkar. *Addressing Item Cold Start Problem in Collaborative Filtering-Based Recommender Systems Using Auxiliary Information*, pages 133–142. 10 2022.

[82] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*, pages 325–341. Springer, 2007.

[83] Yanni Ping, Yang Li, and Jiaxin Zhu. Beyond accuracy measures: the effect of diversity, novelty and serendipity in recommender systems on user engagement. *Electronic Commerce Research*, 02 2024.

[84] Sandy Putra and Anjar Wanto. Analysis of artificial neural network accuracy using backpropagation algorithm in predicting process (forecasting). *International Journal Of Information System Technology (IJISTECH)*, 1:34–42, 11 2017.

[85] Luyl Da Quach, Anh Quynh, Khang Nguyen, and An Nguyen. Using the term frequency-inverse document frequency for the problem of identifying shrimp diseases with state description text. *International Journal of Advanced Computer Science and Applications*, 14:2023, 05 2023.

[86] Behnam Rahdari, Peter Brusilovsky, and Branislav Kveton. Towards increasing the coverage of interactive recommendations. *The International FLAIRS Conference Proceedings*, 35, 05 2022.

[87] Md. Mijanur Rahman, Ismat Shama, Siamur Rahman, and Rahmatullah Nabil. Hybrid recommendation system to solve cold start problem. *Journal of Theoretical and Applied Information Technology*, 100, 06 2022.

[88] Bushra Ramzan, Imran Bajwa, Noreen Jamil, Riaz Ulamin, Shabana Ramzan, Farhaan Mirza, and Nadeem Sarwar. An intelligent data analysis for recommendation systems using machine learning. *Scientific Programming*, 2019:1–20, 10 2019.

[89] Miguel Rebelo, Duarte Coelho, Ivo Pereira, and Fábio Fernandes. A new cascade-hybrid recommender system approach for the retail market. pages 371–380, 01 2022.

[90] Deepjyoti Roy and Mala Dutta. A systematic review and research perspective on recommender systems. *Journal of Big Data*, 9(1):59, 2022.

[91] N. S. Safa, N. Mustapha, and M. Mohd. An enhanced hybrid recommender system based on deep learning for e-learning applications. *IEEE Access*, 9:122844–122854, 2021.

[92] Haidara Saleh and Jamil Layous. *Machine Learning -Regression.* PhD thesis, 01 2022.

[93] A. L. Samuel. Some studies in machine learning using the game of checkers. ii—recent progress. *IBM Journal of Research and Development*, 11(6):601–617, 1967.

[94] PANKTI SATRA. Amazon beauty product recommendation dataset, 2021. Retrieved from `https://www.kaggle.com/datasets/satrapankti/amazon-beauty-product-recommendation`.

[95] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer, 2007.

[96] Samah Senbel. *Fast and Memory-Efficient TFIDF Calculation for Text Analysis of Large Datasets*, pages 557–563. 07 2021.

[97] Thiago Silveira, Min Zhang, Xiao Lin, Yiqun Liu, and Shaoping Ma. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10:813–831, 2019.

[98] Avinash Sonule, Hemant Jagtap, and Vikas Mendhe. Weighted hybrid recommendation system. *INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS*, page 402, 03 2024.

[99] George Stalidis, Iphigenia Karaveli, Kostas Diamantaras, Marina Delianidi, Konstantinos Christantonis, Dimitrios Tektonidis, Alkiviadis Katsalis, and Mike Salampasis. Recommendation systems for e-shopping: Review of techniques for retail and sustainable marketing. *Sustainability*, 15:16151, 11 2023.

[100] Craig Starbuck. *Predictive Modeling*, pages 239–260. 03 2023.

[101] Ja-Hwung Su, Chu-Yu Chin, Yi-Wen Liao, Hsiao-Chuan Yang, Vincent Tseng, and Sun-Yuan Hsieh. A personalized music recommender system using user contents, music contents and preference ratings. *Vietnam Journal of Computer Science*, 7, 11 2019.

[102] Manish Suyal and Sanjay Sharma. A review on analysis of k-means clustering machine learning algorithm based on unsupervised learning. *Journal of Artificial Intelligence and Systems*, 6:85–95, 01 2024.

[103] Manish Suyal and Sanjay Sharma. A review on analysis of k-means clustering machine learning algorithm based on unsupervised learning. *Journal of Artificial Intelligence and Systems*, 6:85–95, 01 2024.

[104] Farah Tawfiq, Abdul Monem Rahma, and Hala Abdul wahab. Recommendation systems for e-commerce systems an overview. *Journal of Physics: Conference Series*, 1897:012024, 05 2021.

[105] Human Tsattabhayya. All about supervised learning, 2023. Accessed: 2024-06-12.

[106] Andreu Vall, Matthias Dorfer, Hamid Eghbalzadeh, Markus Schedl, Keki Burjorjee, and Gerhard Widmer. Feature-combination hybrid recommender systems for automated music playlist continuation. *User Modeling and User-Adapted Interaction*, 29:527–572, 04 2019.

[107] Iosif Viktoratos and Athanasios Tsadiras. A machine learning approach for solving the frozen user cold-start problem in personalized mobile advertising systems. *Algorithms*, 15:72, 02 2022.

[108] Xiang Wan, Anuj Kumar, and Xitong Li. How do product recommendations help consumers search? evidence from a field experiment. *Management Science*, 10 2023.

[109] Jianfang Wang, Pengfei Han, Yanling Miao, and Fengming Zhang. A collaborative filtering algorithm based on svd and trust factor. 01 2019.

[110] Indra Waspada, Nurdin Bahtiar, Panji Wirawan, and Bagus Awan. Performance analysis of isolation forest algorithm in fraud detection of credit card transactions. *Khazanah Informatika : Jurnal Ilmu Komputer dan Informatika*, 6, 10 2020.

[111] Naina Yadav, Rajesh Kumar, Anil Singh, and Sukomal Pal. *Diversity in Recommendation System: A Cluster Based Approach*, pages 113–122. 01 2021.

[112] R. Yu, T. Zhou, Q. Li, L. Li, and X. Li. Hybrid recommender system for music streaming services based on contextual and collaborative filtering. *IEEE Transactions on Broadcasting*, 67(3):506–516, 2020.

[113] Hongli Yuan and Alexander Hernandez. User cold start problem in recommendation systems: A systematic review. 12 2023.

[114] Gisela Yunanda, Dade Nurjanah, and Selly Meliana. Recommendation system from microsoft news data using tf-idf and cosine similarity methods. *Building of Informatics, Technology and Science (BITS)*, 4, 06 2022.

[115] Sobia Zahra, Mustansar Ali Ghazanfar, Asra Khalid, Muhammad Awais Azam, Usman Naeem, and Adam Prugel-Bennett. Novel centroid selection approaches for kmeans-clustering based recommender systems. *Information sciences*, 320:156–189, 2015.

[116] Shichao Zhang. Challenges in knn classification. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 01 2021.

[117] Z. Zhang, R. Cai, Y. Zhang, W. Zhang, and Y. Chen. Personalized recommendation via cascade hybrid collaborative filtering with social influence. *Information Sciences*, 507:228–244, 2020.

[118] Zhi-Peng Zhang, Yasuo Kudo, Tetsuya Murai, and Yong-Gong Ren. Enhancing recommendation accuracy of item-based collaborative filtering via item-variance weighting. *Applied Sciences*, 9(9):1928, 2019.

[119] Kai Zhou, Lan Luo, and Tianlin Chen. *The Influence of Online Media on College Students' Self-identity in Mobile Learning Environment*, pages 1195–1203. 06 2023.

[120] Hanane Zitouni. On solving cold start problem in recommender systems using web of data. pages 1–8, 10 2022.

Abstract

The e-commerce sector has experienced significant growth, leading industry players to implement recommendation systems aimed at enhancing user satisfaction and boosting revenue through personalized suggestions. As a result, hybrid filtering techniques combining collaborative filtering (CF) and content-based methods have emerged as a promising approach to capture user preferences effectively

Our study presents a novel hybrid recommendation system for e-commerce that combines content-based filtering using inverse document frequency-term frequency (IDF-TF) with collaborative filtering through singular value decomposition (SVD). By integrating linear regression and leveraging the α parameter, we effectively balance between content-based and collaborative filtering to address cold start issues and data sparsity. This approach enhances recommendation accuracy and quality, as demonstrated by our experiments on real e-commerce data, focusing on scalability, accuracy, and coverage.

Keywords: Recommendation system, Machine Learning, SVD algorithm, TF-IDF, Hybrid.

Résumé

Le secteur du commerce électronique a connu une croissance significative, conduisant les acteurs du secteur à mettre en œuvre des systèmes de recommandation visant à améliorer la satisfaction des utilisateurs et à augmenter les revenus grâce à des suggestions personnalisées. En conséquence, les techniques de filtrage hybrides combinant le filtrage collaboratif (CF) et les méthodes basées sur le contenu sont apparues comme une approche prometteuse pour capturer efficacement les préférences des utilisateurs.

Notre étude présente un nouveau système de recommandation hybride pour le commerce électronique qui combine un filtrage basé sur le contenu utilisant la fréquence inverse des termes de fréquence des documents (IDF-TF) avec un filtrage collaboratif par décomposition en valeurs singulières (SVD). En intégrant la régression linéaire et en exploitant le paramètre α, nous équilibrons efficacement le filtrage basé sur le contenu et le filtrage collaboratif pour résoudre les problèmes de démarrage à froid et la rareté des données. Cette approche améliore la précision et la qualité des recommandations, comme le démontrent nos expériences sur des données réelles de commerce électronique, en mettant l'accent sur l'évolutivité, la précision et la couverture.

Mots-clés: Système de recommendation, apprentissage automatique, algorithme SVD, TF-IDF, Hybride.