Democratic and Popular Republic of Algeria
Ministry of Higher Education and Scientific Research
University of Abderrahmane Mira Bejaia
Faculty of exact sciences
Computer science department

Graduation thesis

To obtain an academic master's degree

Option : Advanced Information System

Theme

Food recommender system for cancer patients based on sentiment analysis

Presented by:
Boughafene Chayma

Evaluated by:

President	Mr Allem Khaled		U. A. Mira Bejaia
Supervisor	Dr. EL BOUHISSI BRAHAMI Houda	M.C.A	U. A. Mira Bejaia
Examiner	Mrs Khoulallen.N		U. A. Mira Bejaia

Promotion 2023 – 2024

Thanks

I thank Allah for granting me health, courage and the will to accomplish this modest work.

I would like to express my most sincere thanks to the people who gave me their help and who contributed to the development of this dissertation as well as to the success of this wonderful academic year.

I would also like to thank my supervisor Mrs. El Bouhissi for the help and assistance she provided me to enable us to successfully complete this end-of-cycle dissertation. I express my deep gratitude for her unwavering commitment to my growth and success.

I would like to express my great respect to the honorable members of the jury who agreed to evaluate this work.

I do not forget my parents for their contribution, support and patience.

Dedications

I dedicate this work as a token of my love and attachment :

To my parents (my dear father and my dear mother) thanks to them that I found the path to success and that I was able to follow long studies properly. May ALlAH protect you from evil, provide you with long life, health and happiness so that I can repay you a minimum of what I owe you.

To my dear brothers Walid and Yacin and my dear sisters Souad and Khadidja. I implore ALlAH to bring you happiness, love and that your dreams come true.

To all my family. For your love, your prayers and your encouragement which have been of great support to me during this long journey.

To all my friends. In memory of the wonderful times we've had together and the strong ties that bind us. For fear of forgetting some of them, I won't venture to name them all, but I'm sure they'll recognize themselves here. Many thanks for your support, your encouragement and your help. With all my affection and esteem, I wish you every success and happiness, both in your professional and private lives.

Chayma

Abstract

Cancer is a serious disease characterized by abnormal and irregular cell development in any part of the body, in the form of a tumor. It is considered as the second-leading cause of death in the world. Efforts to find a successful cancer therapy have led to the effective use of various treatments such as chemotherapy, and surgery to eliminate dangerous tumors. However, these treatments may affect the patient's immune system by damaging blood cells that protect the body from disease. Therefore, it is strongly recommended that cancer patients consume nutrient-rich foods to increase their strength to better cope with the side effects of treatment. In this dissertation, a hybrid food recommendation system which considers the patient's emotional state and dietary preferences and needs to help them predict the foods which can be consumed have been proposed. The approach involve to use a content-based system that filters recipes according to user needs and sentiment scores. Additionally, a rule-based sentiment analysis method was employed to identify sentiment from text reviews determining which of the foods were liked or disliked. The efficacy of this proposed approach was rigorously assessed, and the results yielded promising insights. Notably, combining content based, sentiments analysis led to a marked improvement with a precision of 97%.

Keywords: Nutrition suggestion; Cancer; Sentiment analysis; Content-based filtering.

Résumé

Le cancer est une maladie grave caractérisée par un développement cellulaire anormal et irrégulier dans n'importe quelle partie du corps, sous la forme d'une tumeur. Il est considéré comme la deuxième cause de mortalité dans le monde. Les efforts déployés pour trouver une thérapie efficace contre le cancer ont conduit à l'utilisation efficace de divers traitements tels que la chimiothérapie et la chirurgie pour éliminer les tumeurs dangereuses. Toutefois, ces traitements peuvent affecter le système immunitaire du patient en endommageant les cellules sanguines qui protègent l'organisme contre les maladies. Il est donc fortement recommandé aux patients atteints de cancer de consommer des aliments riches en nutriments afin d'augmenter leurs forces et de mieux faire face aux effets secondaires du traitement. Dans cette thèse, un système hybride de recommandation alimentaire qui prend en compte l'état émotionnel du patient et ses préférences et besoins alimentaires pour l'aider à prédire les aliments qu'il peut consommer a été proposé. L'approche consiste à utiliser un système basé sur le contenu qui filtre les recettes en fonction des besoins de l'utilisateur et des scores de sentiment. En outre, une méthode d'analyse des sentiments basée sur des règles a été employée pour identifier les sentiments à partir de critiques textuelles déterminant quels aliments ont été appréciés ou non. L'efficacité de cette approche proposée a été rigoureusement évaluée et les résultats ont donné des informations prometteuses. Notamment, la combinaison de l'analyse basée sur le contenu, analyse des sentiments a conduit à une nette amélioration avec une précision de 97%.

Keywords: Suggestions nutritionnelles; Cancer; Analyse des sentiments; Filtrage basé sur le contenu.

Abbreviations list

RS Recommendation system
CBF Content-Based Filtering
CF Collaborative filtering
HRS Hybrid recommender system
NLP Natural language processing
SA Sentiment Analysis
OM Opinion mining
FRS Food recommender system
CB Content based
ALS Alternating Least Square
BPR Bayesian Personalized Ranking
LMF Logistic Matrix Factorization
SVD Singular Value Decomposition
CapsNet Capsule Networks
PPKG Personal Preference Knowledge Graph
KGAT knowledge graph attention network
LSTM long short-term memory
NDCG Normalized Discounted Cumulative Gain
ACO Ant colony optimization
E-LSTM Enhanced Long Short-Term Memory
CSV Comma-Separated Values
EDA Exploratory data analysis
Pandas Python Data Analysis Library
Sklearn Scikit-learn

Contents

List of Figures

1
General introduction

In a world faced with multiple health challenges, chronic diseases and serious illnesses continue to have a profound impact on contemporary societies. Among these, cancer stands out as one of the main causes of morbidity and mortality on a global scale, affecting millions of people each year and highlighting the complexity of the interactions between genetics, environment and lifestyle.

Cancer is a disease in which some of the body's cells grow uncontrollably and divide to other parts of the body without stopping. Cancer cells ignore signals that would otherwise stop them dividing and influence surrounding normal cells. Cancer can start almost anywhere in the human body which is made up of trillions of cells. Normally, human cells grow and multiply to form new cells, as the body needs them. When cells grow old or became damaged they die and new cells take their place. However, sometimes this orderly process breaks down and abnormal damaged cells grow and multiply when they should not. These cells may form tumors which can be benign means do not spread or grow back and malignant tumors or cancerous tumors which means they spread and grow back. Cancer can be a genetic disease or inherited that is caused by changes to genes. It can be caused when DNA is changed exposure to environmental factors including chemicals in tobacco and smoke. It can arise because of errors that occur as cells divide. Cancer can lead to weight loss, obesity and other chronic disease. Between 2015 and 2016, the national health and nutrition examination survey demonstrated 39.8% of adults and 18.5% of youth were obese. In a recent study, it said that 900.000 adults in the United States showed a significant proportional increase between obesity and mortality risk from multiple cancers, including of the esophagus, colon and rectum pancreas, breast and uterus. It is estimated that 40 to 80% of all cancer patients will be malnourished during the course of the disease. Furthermore, malnutrition can influence treatment outcomes delay wound healing, worsen muscle function and increase the risk of post-operative complications. Is the result mainly of inadequate food intake due to a set of nutrition impact symptoms, which may result from local effects of tumor; from side effects of anticancer treatments or from infections or other complications following during the course of the disease. Malnutrition is frequent and has been reported to be present in 20- 70% of cancer patients depending on tumor entity, stage of the disease and clinical setting.

An incorrect treatment program has the capacity to have many negative consequences both for the medical practitioners and for the patients.

For medical practitioners their reputations will adversely affected and they may be accused of medical negligence, in contrast, patients have to bear the cost of unnecessary treatment and also the pain and inconvenience of having their health put at risk. Additionally in practice, there are many situations where a medical practitioner is required to not only apply his or her medical knowledge but also to consider the patient's condition, financial situation and even their personal emotional state. Generally, in treatment, some cases are more complex than others are, so they should be allocated a longer time to allow doctors discuss them thoroughly, while some cases are easy and can be treated in a regular way. Nutrition is an important part of cancer treatment. Eating the correct meals help to build the immune system and fight against disease. Food provides energy, vitamins and other essential nutrients needed by the body to function properly and sustenance for daily activities.

A healthy diet enhances body growth, promotes good mental function, boosts body beauty and promotes healthy long life. Nutrition therapy could be used to manage chronic diseases by managing the diet based on the belief that food provides vital medicine and helps to maintain a good health. In addition, a healthy food lifestyle helps to reach and maintain a healthy mind and body weight, lowers health risks, such as obesity, diabetes, hypertension and cancer, reduce effects during and after treatment. Therefore, nutrition plays a crucial role in the care pathway for cancer patients. Proper nutrition can help manage symptoms, improve quality of life, and potentially increase the effectiveness of treatments. However, the nutritional needs of patients can vary considerably depending on the type of cancer, the stage of the disease, the treatments undergone, and the general physical conditions of the patient. Malnutrition, frequently observed in cancer patients, can lead to deterioration of general condition, affect response to treatment and worsen prognoses.

Personalization of nutrition for cancer patients has become a growing area of interest. Personalized nutrition aims not only to optimize medical treatment, but also to respond to the patient's preferences and feelings, thereby improving their emotional and physical well-being. Taking into account feelings and food preferences through methods such as sentiment analysis can help design nutritional recommendations that are more accurate and better accepted by patients.

1.1 Problem statement and objective

Managing nutrition in cancer patients is a major challenge. Cancer and its treatments, such as chemotherapy and radiotherapy, can cause various side effects that affect the patient's diet and nutritional status, such as loss of appetite, changes in taste, and digestive difficulties. These symptoms make it difficult to maintain an adequate diet, essential for supporting treatment, boosting immunity and improving quality of life. However, the nutritional needs of cancer patients are highly diverse and influenced by multiple factors, including their emotions and psychological state, which can vary significantly throughout treatment.

therefore the problem posed in this project lies in the need to propose nutritional recommendations intended for cancer patients, not only to the specific physiological needs of patients, but also to their emotional states and personal preferences which are often neglected in traditional approaches.
It is in this context that our project falls, aiming to develop a food recommendation system for cancer patients, based on sentiment analysis. This system aims to closely link nutritional recommendations to patients' emotional states and dietary preferences, leveraging natural language processing techniques, providing personalized advice that supports both physical health and well-being emotional of the patient.

1.2 Work Structure

The development of this project is structured into six chapters, organized as follows:
Chapter 1: General introduction.
Chapter 2: General and Basic Concepts - This chapter will cover the key concepts necessary to understand the project, including basic notions of recommendation systems, and sentiment analysis.
Chapter 3: State of the Art - Analysis of previous work and studies relating to nutrition recommendation systems in the field of health in general and cancer specifically.
Chapter 4: Presents our solution to meet the needs presented in Chapter 3. We describe the system architecture and the implementation of the different concepts. We also define the tools used to build the system.
Chapter 5: presents the overall validation of our system, as well as the results of tests and comparisons of the different types of algorithms used. All this after having presented the technological aspects surrounding the implementation of our system as well as the description of the different interfaces of the application.
The last chapter (chapter 6) concludes this dissertation and presents some future perspectives. Finally, the thesis includes the bibliographic references used for its elaboration.

2
Fundamental concepts

2.1 *Introduction*

In a world where technology increasingly shapes the way we live and interact, recommendation systems and sentiment analysis are emerging as essential tools for personalizing user experiences and providing users with information that corresponds to their interests and this by analyzing their interactions with their information space. For cancer patients, personalizing diet based on detailed analyzes of their preferences and emotional state could significantly improve their well-being and quality of life.

In this chapter, we explore two crucial areas recommender systems and sentiment analysis. The first section is dedicated to recommender systems, we define what recommender systems are, discuss their main applications and objective, present their varied types as well as the advantages and disadvantages associated with each. In addition, we detail the key stages of their implementation. In the second section, we discuss sentiment analysis, starting with its definition, the different types and levels of analysis, the associated technical challenges and relevant application areas.

2.2 Recommendation systems

2.2.1 Definition

Recommendation systems (RS) can be defined in various ways, given the diversity of classifications proposed for these systems, which may relate to different types of data or specific approaches.

In a general way, RS is a software tool and an intelligent system that provides the user with suggestions on items or products that meet his needs or are simply likely to interest him. For example, what movie to see, what book to buy or even what music to listen, these suggestions are based on the individual's tastes by analyzing the browsing history, opinions, comments and ratings given to products and the behavior of other users [1]. A RS does not receive a direct request from the user, but must offer him new possibilities by learning his or her preferences, browsing history, opinions, comments, and ratings given to products or similar behaviors of other users. The recommendations given by a RS are pointed to support their clients in different decision-making forms, for example, what things to purchase, what music to tune in, or what online news articles to read.

Recommender frameworks are important implies for online users to manage with data over-burden and offer assistance them make superior choices. They are now one of the most popular applications of artificial intelligence, supporting information discovery on the Web[42].

2.2.2 Fields of use

Recommender systems have become important and used in several fields, most particularly healthcare. Below we mention some of the primary fields of use for recommender systems:
- E- commerce(Amazon).
- Social media (Facebook...) in this field RSs suggest pages, groups, posts, news... that might interest users.
- Music (lastFM).
- Cinema and movies (Netflix and Movielens).
- Fashion and Retail (ASOS).
- Food(Yelp).
- Education and Learning(Coursera).
- Video on demand (YouTube).
- Tourism, in this field RS's suggest destinations, activities, and accommodations to users. For example TripAdvisor.

2.2.3 Basic concepts and notations

In this section, we define some concepts and notations related to recommendation systems.

User and Items entities

The two basic entities that appear in all recommendation systems are the user and the item.

1.Users: In the context of recommender systems, a "user" is the person who uses and accesses the system, gives their opinion on various items and registers by entering his personal information (interests, age ...) and receives new recommendations from the system and receives new recommendations from the system. The set of users in the system is represented by U, where a user is $u \in U$.

2.Items: The "Item" is the general term used to designate what the system recommends to users. It is the entity which represents any element constituting a recommendation list and which corresponds to the needs of the user. Including any product likely to be sold (book, products...etc. on the sites of the e-commerce such as Amazon.com), seen (movies on online TV sites such as Netflix), listened to (music) or read (such as information in online newspapers, magazines), as well as holiday destinations, restaurants, etc. The collection of items that the system can recommend is denoted by I, where $i \in I$.

User-Item Matrix(Rating Matrix)

The user-item matrix is a fundamental structure used in recommender systems to represent user interactions or preferences towards items. of which the entire system $< u, i >$ are recorded in a sparse database called Evaluation Matrix or User-item Matrix, and it is denoted by R, where each row corresponds to the evaluations provided by a single user and a column corresponds to the evaluations given to a single item by all users[51].

Profile notion

Generally, the profile of an object is a set of characteristics that allow it to be identified or represented. Two types of profiles can be used in recommendation systems, corresponding to the two entities used in these systems : the user and the item.

1. User profile: it is a description of the user's characteristics, which may be his or her interests, demographics, or preferences expressed in the form of evaluations, etc. Several approaches for acquiring information about the user in order to build his profile exist [2];[52].

2. The item profile: it corresponds to the description of the item with a set of characteristics, also called attributes or properties, for example in a food recommendation system, items (foods) are represented by their nutrients, ingredients ..., while in a document recommendation system, attributes are keywords that describe the semantic content of the document.

Prediction

In the context of recommendation systems, "prediction" is the calculation of the probable rating (the process of estimating the ratings) that a user will assign to an item that he or she has not yet seen, evaluated or interacted with. This process is central to the functionality of recommendation systems, as it enables them to suggest items that a user is likely find useful based on their past behavior and preferences.
In general, Rating Matrices have only a few cells containing values while the others have unknown values and in the majority of cases they have a "0" inside, resulting in sparse

matrices. Therefore, the density of these matrices will not be sufficient to generate precise recommendations. Therefore, missing rating prediction methods are used to increase the density of the user-item matrix to make more powerful and relevant recommendations.
This process is central to the functionality of recommendation systems, as it enables them to suggest items that a user is likely find useful based on their past behavior and preferences.

Recommendation

Recommendation is the action of calculating a list of items (often referred to as the Top-N items) that are predicted to be most appealing to the user. This process entails scoring items based on factors such as their popularity or how well they align with the user's preferences. Contrary to prediction, which relies on assessing and forecasting user ratings, the calculation of recommendations is not based on evaluations, but rather incorporating a broader range of criteria to identify the items likely to resonate most with the user.

2.2.4 Goal of recommendation systems

The aim of a recommendation system is to provide a user with relevant resources according to their preferences. The latter thus sees reduced his research time but also receives suggestions from the system to which he would not have spontaneously paid attention. The rise of the web and its popularity have contributed in particular to the implementation of such systems as in the field of e-commerce. Recommendation systems can be seen initially as a given response to users with difficulty making a decision in the context of the use of a "classic" information research system[38].

2.2.5 Types of recommendation systems

Recommendation systems are tools and techniques aimed at presenting items likely to interest the user. According to the classic classification, there are three main types of recommender systems:
Content-based filtering, Collaborative filtering, and Hybrid recommender systems.

Content-Based Filtering (CBF)

Definition:
This type of recommendation is based on choices a user has made in the past to suggest items they will consume in the future (Suggest items similar to what a user liked in the past). For example, if a user likes science fiction movies, the system will recommend other science fiction movies[6].
Therefore, a CBF system follows two main functionalities as follows: Firstly, it compares the items not rated by the user with his, represented by all the items he has rated (such as clicks, ratings and likes ...), by calculating the similarity between them, i.e. it is an item-to-item correlation. Then, it recommends the items closest to the user's preferences.

Therefore, the content-based filtering process requires two essential constituents which are: item profiles and user profiles, because recommendations are generated based on the correlation between the profiles of these two entities[37];[39].

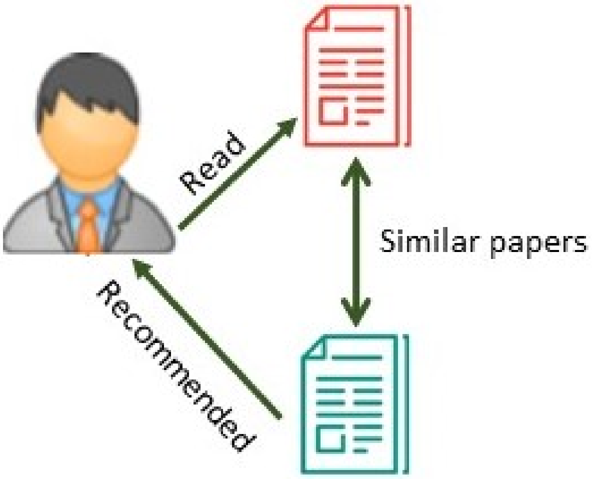


Figure 2.1: Content-Based Filtering technique.

Advantages:

✓ Recommend items similar to ones users have liked in the past.
✓ Matching between user preferences and item characteristics also works for textual data.
✓ We do not need data about other users.
✓ No need for a large community of users to be able to make.
✓ Absence of low-density problem.
✓ The possibility for users to build their own profile through exclusive ratings. In other words, CBF provides user independency.
✓ CBF recommender system gives explanation on how the recommender works (transparency).

Disadvantages:

▷ This type reduces the diversity of the recommendation item: if a user has never consumed an item with a particular set of keywords, it will never be suggested to him.
▷ Requires user profile: CBF is not effective in providing recommendations for new user. This is because the training model need the history of ratings of the user. It is necessary to have a huge number of ratings for the target user to make right predictions for him.
▷ Filtering based on the thematic criterion only, absence of other factors such as scientific quality, the target audience, the interest shown by the user, etc.
▷ New user: No history.
▷ It based entirely on item and topic scores of interest: The fewer scores, the more limited

the set of possible recommendations.
▷ The difficulty of indexing multimedia documents: The growth of multimedia documents (text, image, videos, etc.) poses the problem of taking into account the structural information of the documents to help identify relevant multimedia content.
▷ Cold start problem: A new user of the system is having difficulty expressing his profile by specifying topics that interest him.

Collaborative filtering (CF)

Definition: One of the oldest techniques used and which still remains among the simplest and most effective today is collaborative filtering. The idea of collaborative approaches is to try to predict a user's opinion on different elements. The recommendation is based on the user's previous likes and reviews and a measure of similarity with other users [6];[7]. For example, if user U_1 and user U_2 both liked item 1 and item 2, and user U_1 also liked another item 3, the system will recommend item 3 to user U_2[40].
The main steps of this approach are:
1. Many user preferences are saved;
2. A subgroup of users is identified whose preferences are similar to those of the user seeking the recommendation;
3. An average of preferences for this group is calculated;
4. The resulting preference function is used to recommend items to the user seeking the recommendation.

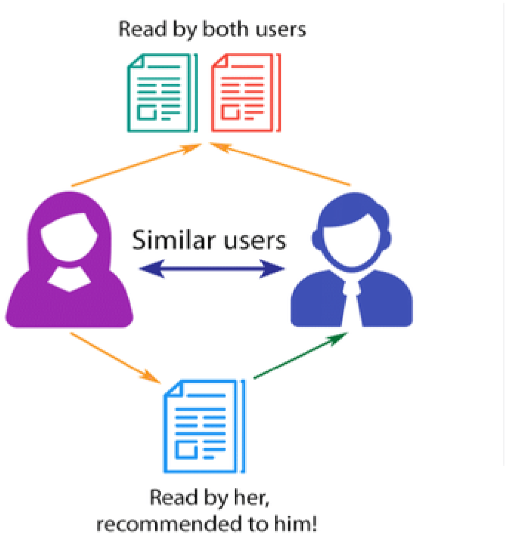


Figure 2.2: Collaborative filtering technique.

Advantages:

✓ The quality of the recommendation can be evaluated.
✓ The greater the number of users, the better the recommendation.
✓ Use the scores of other users to evaluate the usefulness of the elements.
✓ Find users or groups of users whose interests correspond to the current user.
✓ Thanks to its independence with regard to the representation of documents, collaborative filtering makes it possible to resolve the problems linked to content-based filtering, and therefore to filter any type of information (texts, images, videos).
✓ Another advantage of collaborative filtering is that users' value judgments integrate not only the thematic dimension but also other factors relating to the quality of documents such as diversity, novelty, suitability for the target audience, etc.

Disadvantages:
▷ Cold start: Collaborative filtering systems depend on user evaluations of items. Therefore, a new item cannot be recommended until no user has rated it. In recommendation systems based on collaborative filtering and content-based systems, it is impossible to predict user preferences without knowing their item evaluation histories. Therefore, new users will not receive specific recommendations before having evaluated a certain number of items.
▷ Complexity: in systems with a large number of items and users the calculation grows linearly

Hybrid systems (HRS)

Definition:
Hybrid systems are approaches that combine two or more recommendation approaches (For example,collaborative filtering and content-based filtering) to provide recommendations that are more accurate,diverse and robust.These combinations make it possible to benefit from the advantages of the approaches used and helps solve some problems faced by systems with a single approach, for example, the item-level cold start of collaborative approaches can benefit from the advantages of content-based approaches[41].

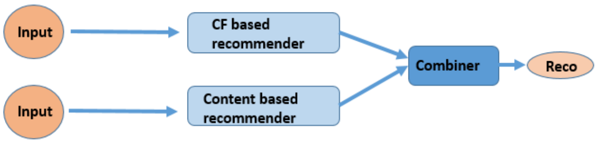


Figure 2.3: Hybrid filtering.

Advantages:

✓Reduce cold start issues.
✓Improve the accuracy of recommendations.

2.2.6 The main stages of the recommendation

Generally, a recommendation system requires three steps, as shown in the figure below:

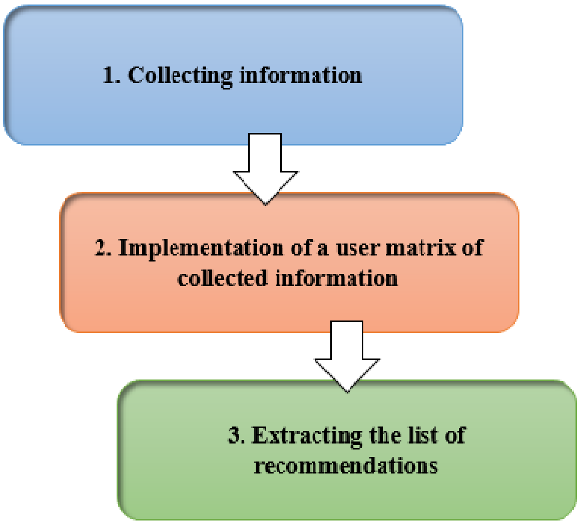

1. Collecting information
2. Implementation of a user matrix of collected information
3. Extracting the list of recommendations

Figure 2.4: Stages of the recommendation.

Collecting information

To be relevant, a recommendation system must be able to make predictions about user interests. We must therefore be able to collect a certain amount of data on them, in order to be able to build a profile for each user. There are two forms of data collection:

Collection of explicit data - Active filtering: Collection relies on the user explicitly indicating their interests to the system.
Example: Ask a user to comment, tag, rate, like or even add as favorites content (objects, articles, etc.) that interest them. We often use a rating scale ranging from one star (I don't like it at all) to five stars (I like it a lot) which are then transformed into numerical values so that they can be used by recommendation algorithms.

✓ **Advantage:** Ability to reconstruct the history of an individual and ability to avoid aggregating information that does not correspond to this single user (several people on the same station).
▷ **Disadvantage:** The information collected may contain a so-called reporting bias.

Implicit data collection - Passive filtering: It is based on an observation and analysis of user behavior carried out implicitly in the application, which embeds the recommendation system; everything is done in the ”background” (roughly without asking the user).

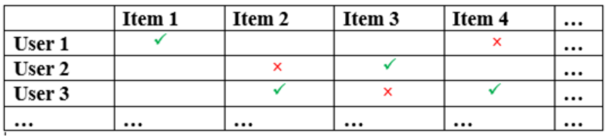
Example: Obtain the list of items that the user listened to, watched or purchased online.

✓ **Advantage:** No information is requested from users, all information is collected automatically. The data retrieved are a priori fair and do not contain reporting bias.

✓ **Disadvantage:** The data recovered is more difficult to attribute to a user and may therefore contain attribution bias (common use of the same account by several users). A user may not like certain books they purchased, or they may have purchased it for someone else

Implementation of a user matrix of collected information"user model

The implementation of a matrix called "user matrix" or "user model" including information concerning users collected during the previous stage of information collection. It can be represented as a table that contains data collected about the user associated with the products available on the website.
The table below presents a fictitious example of a binary matrix containing information such as "user u liked/did not like item i". This information can also be "purchased/did not purchase", "viewed/did not view", etc. They can also be measured on a larger number of classes: "gave 1/2/3/4/5 stars" etc. User interests generally evolve over time. The user model data should be constantly readjusted to remain consistent with the user's new interests.



	Item 1	**Item 2**	**Item 3**	**Item 4**	**...**
User 1	✓			×	**...**
User 2		×	✓		**...**
User 3		✓	×	✓	**...**
...	**...**	**...**	**...**	**...**	**...**

Figure 2.5: example of the user matrix.

Extracting the list of recommendations

To extract a list of suggestions i.e. recommendation from a user matrix, the algorithms use the notion of measuring similarity between objects or people described by the user model. The purpose of similarity is to give a value or a number (in the mathematical sense of the term) to the resemblance between two things. The stronger the resemblance, the greater the similarity value will be. Conversely, the weaker the resemblance, the smaller the similarity value will be.

2.3 Sentiments analysis:

2.3.1 Definition

Sentiment analysis also referred to as opinion mining extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, appraisal extraction, is an natural language processing approach (NLP) that identifies the feeling and emotion behind a piece of writing. The process consists of acting on a text, a sentence or a complete article, and analyzing the emotion expressed by the author. Feelings are generally classified into three types: negative, neutral and positive. The two expressions sentiment analysis (SA) or opinion mining (OM) are interchangeable. They express mutual meaning. However, some researchers have stated that OM and SA have slightly different notions: Opinion Mining extracts and analyzes people's opinion about an entity while sentiment analysis identifies the sentiment expressed in a text and then analyzes it. The goal of SA is to find opinions, identify the feelings they express, and then classify their polarity. This technique is widely used in various fields like marketing, social media monitoring, customer service, etc., to understand people's opinions and reactions[43];[45].

2.3.2 Sentiment Analysis types

There are several types of sentiment analysis. The figure below show some of them:

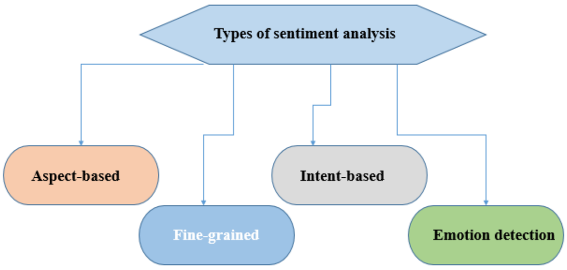


Figure 2.6: Types of sentiment analysis.

Aspect-based sentiment analysis

Aspect-based sentiment analysis focuses not just on whether the sentiment is positive or negative, but also on which aspect the sentiment is associated with (i.e. overall sentiment about a certain aspect is classified as positive, negative, or neutral). The results are more detailed, interesting and precise because this type of sentiment analysis examines in detail the information contained in a text. It is used in a wide variety of industries, including:E-commerce- Identify specific aspects of a product (quality, design, price, etc.) that are liked or disliked by customers. Healthcare- Analyze certain aspects of healthcare enterprises that are being discussed by customers such as plans of treatments, medications, quality of care, etc[3];[44].

fine-grained sentiment analysis

This type uses a lexicon approach to delve deeper into the sentiments expressed in a given text. It helps in identifying sentiments based on levels of sentiment intensity, emotions, and target. For instance, sentiments could be classified as very positive, positive, neutral, negative, and very negative. This helps in capturing more nuanced emotional responses[3];[44].

intent-based sentiment analysis

Intent-based goes beyond identifying the tone (positive, negative, or neutral) of the given text and leveraged machine learning algorithms to understand the underlying purpose of the text. It seeks to determine the intention behind texts; this could entail gauging whether the text is expressing a query, a question, feedback, a complaint, a command, or a desire. It correctly identifies the sentiment of the text along with the degree to which it is expressed, and it can improve interaction strategies[3].

Emotion detection

Most of the emotion detection systems are based on the use of sentiment lexicons or complex machine learning algorithms, this type of sentiment analysis helps identify the emotions that customers express in their comments It is more intricate than traditional sentiment analysis, It can deftly identify emotions such as anger, happiness, fear, and surprise. For example, sites like The Athletic allow readers to comment on articles, but also offer a simpler ”what do you think of this story? ” feedback option[3];[46].

2.3.3 Levels of sentiment analysis

Sentiment analyses occur at three main levels, each providing a different depth of understanding:

Document level

Known as document-level sentiment classification. In this level of analysis, sentiment is extracted from the entire review, and a whole opinion is classified based on the overall sentiment of the opinion holder[48];[47].

Sentence Level Classification

This level of analysis is closely related to the classification of subjectivity, which distinguishes between objective sentences (which presents some factual information) and subjective sentences (personal feelings, points of view, emotions, or beliefs). However, it should be noted that subjectivity is not equivalent to feeling as many objective sentences can imply opinions[4].

Aspect level

The aspect level was earlier called the feature level (feature-based opinion extraction and synthesis). It performs a finer analysis. The goal in this level consist on identifying object features that have been commented on by the opinion holder and determine whether the opinion is positive, negative, or neutral. Instead of looking at language constructs (documents, paragraphs, sentences), aspect level looks directly at the opinion itself. It is based on the idea that an opinion consists of a feeling (positive or negative) and a target (opinion)[4];[49].

2.3.4 Challenges

Sentiment analysis, also known as opinion mining, involves the study of opinions, sentiments, attitudes, and emotions expressed in text. While sentiment analysis offers valuable insights into public opinion, it also presents several challenges that can affect its accuracy and effectiveness and that need to be addressed. Here are some of the main challenges faced in sentiment analysis:

Contexte and polarity

Humans can understand the context of an interaction, but this can be an obstacle for an algorithm, where Algorithms cannot learn about contexts if they are not mentioned explicitly. One of the problems that arise from context is changes in polarity and this is the most common form of sentiment analysis, which involves classifying a piece of text as positive, negative, or neutral[5];[43]. For example, look at the following responses to a survey:
Everything about it.
Absolutely nothing!
If the question is" What did you like?" the first answer will be positive and the second negative, But if the question is "What didn't you like?" the negative in the question will make sentiment analysis change altogether. Pre-processing or post-processing will therefore be important so that the machine understands the context that may have caused certain responses. Nevertheless, it remains a difficult task.

Determine subjectivity and tone

In texts, human interactions can be subjective and objective. Objective texts do not contain explicit sentiments, whereas subjective texts do .It is difficult for algorithms to judge[5];[43]. For example, if we intend to analyze the sentiment of the following two texts:
The wallpaper is beautiful.
The wallpaper is white.
We can estimate that the feeling is positive for the first sentence and neutral for the second. All predicates (adjectives, verbs, nouns, etc.) should not be treated in the same way when analyzing the sentiment in a sentence. Here, the term "beautiful" is much more subjective than the term "white".

Identify sarcasm and irony

Detecting sarcasm and irony can be challenging; individuals might use positive words to convey negative sentiments, and without a deep understanding of the context, these sentiments can be difficult for machines to interpret accurately[5];[50]. For example, if we take the answer to the question: (Did you enjoy your experience on our site?), (Yes of course! There are no bugs!) Here, at first glance, it would seem that the answer is yes. However, we could very well see irony in it and understand the opposite. The issue here is that there are no explicit textual indicators to guide the machine in discerning or questioning the true emotion behind the statement.

Neutral posts

Another issue is neutral posts, which are not categorized. How does the algorithm handle neutral messages? What we mean by neutral does matter when we train sentiment analysis models. Since tagging data requires that tagging criteria be consistent, a good definition of the problem is necessary.

2.3.5 Sentiment analysis application

The importance of sentiment analysis exists in many fields, and a number of applications have emerged in this context. Let us briefly mention few applications:

Policy:
Before a new law is made, politicians try to get the opinion of social media users on the law.

Economy:
The customer asks for the opinion of other people who are using the product before buying it, companies can know the opinion of customers on their products or services to make changes.

Education:
Sentiment analysis helps teachers and schools take corrective action.

2.4 Conclusion

We have attempted throughout this chapter to focus on recommender systems and sentiment analysis, providing an in-depth understanding of these technologies and their potential. We have examined their definitions, areas of application, stages of RS and the different existing types of both RSs and sentiment analysis.
In the next chapter, we will present the state of the art with details on the studied papers, methodologies and tools on RSs for healthy food for cancer patients.

3
State of the art

3.1 Introduction

The advent of the internet has transformed the way of accessing information, leading to an increase in the quantity and diversity of available data due to the growing number of users. Recommender systems have gained widespread popularity as tools to help users efficiently retrieve pertinent information. In particular, food recommender systems (FRS) are of paramount importance to overcome the overwhelming abundance of information within the food domain. However, the recommendation of food is a complex domain with distinctive characteristics that pose many challenges. Despite the importance of FRS in promoting healthy eating habits and improving overall health outcomes, there have been very few systematic literature reviews conducted in this domain and specifically FRS for cancer patients. This chapter is addressed to present a systematic literature review that summarizes the current state-of-the-art in FRS.The review examines the different methods and algorithms used for recommendation, the data and how it is processed, and evaluation methods. It also presents the advantages and disadvantages of FRS.

3.2 Related works

Pawar et al [11] have proposed a hybrid food recommendation for recommending recipes tailored to cancer patients, considering user preferences and the anti-cancer properties of each recipe. The architecture consists of two main blocks, the content-based (CB) block mainly employing matrix factorization and the collaborative filtering (CF) block. The collaborative filtering approach tests Alternating Least Square (ALS), Bayesian Personalized Ranking (BPR), and Logistic Matrix Factorization (LMF), with ALS performing the best. The authors proposes an extension to the machine learning algorithm SVD++, named gSVD++, which considers both implicit and explicit feedback from users. The CB block filters recipes according to user preferences and returns a set of recipes (RL1) with preference scores calculated based on the content (ingredients, cuisine, etc.). The CF block generates a distance graph of user preferences, creates clusters of users with similar taste, and filters 15 neighbors closest to the current user to retrieve recipes that have been liked by these users. The preference scores for these recipes are calculated and scaled to generate a set of recipes (RL2). The union of recipe sets RL1 and RL2 is passed through a post-process block, which filters out any recipes already liked/disliked by the user. The resulting set is inputted to the CBM Block to get their associated health scores. The final score for a recipe is calculated using a combination of preference and health scores. The scores are then used to rank and display the recipes.

Raguvaran et al [12] have proposed a healthy and nutritious meal suggestions for cancer patients using a CapsNet model, a type of deep neural network. The three key procedures in the proposed study are Preprocessing, feature extraction, and classification. They implemented six layers in this proposed work where the first three layers are involved in the feature extraction phase, while the remaining three layers participate in the classification phase. During the feature extraction phase, the input layer receives food images and patient health details, initiating the process. The convolution layer conducts convolution to extract essential features from the food images, preparing the data for more detailed analysis. Then the primary capsule layer predicts nutritional information and probabilities of nutrients based on the extracted features. In classification phase, The Food Capsule layer uses Routing Agreement to determine likely outcomes, while Margin Loss ensures that the model learns to classify items based on the presence or absence of specific features in food images. Subsequently, the Classification Layer uses information extracted from previous layers to classify foods as beneficial or to be avoided. Finally, the Output Layer takes the results from the food classification layer and generates a final output. The authors uses Food-101 dataset and cancer patient dataset for analysis .The proposed approach achieves an accuracy of 95% with decreased loss value.

Tang et al[16] have proposed a personal preference knowledge graph model (PPKG), which adopts two modules for the recommendation. The user and item embedding modeling module, which utilizes the knowledge graph attention network architecture (KGAT) to learn item embedding and enhance recommendation precision by representing features in the knowledge graph with message-passing and update functions. The model prediction module incorporates long short-term memory (LSTM) networks to consider the influence of time on users' taste preferences. It predicts recipes based on users' historical dietary habits for more personalized recommendations beneficial for cancer prevention and treatment. To verify the effectiveness and rationality of their approach the authors created their own dataset by crawling recipe websites and extracting information from textbooks, they used rule-based methods to identify entities like diseases, symptoms, ingredients, effects and their relationships to construct a cancer knowledge graph(the identified entities and relationships were manually reviewed by medical professionals). They compare the PPKG with three other recommendation algorithms (BPRMF, CKE and CFKG) on self-created datasets showed that PPKG outperformed others in terms of recall and NDCG evaluation metrics, demonstrating its effectiveness.

Ahmad et [9] have proposed a food recommendation system that utilizes machine learning algorithms to recommend food to patients based on their health and medical conditions. This system recommends food dishes tailored to individual nutritional needs by employing the K-means clustering algorithm, the food inference algorithm, and the patient nutrition calculation algorithm. The K-means clustering algorithm is used to process the initial input of food with ingredients, compute the nutrients by components, form clusters based on the nutrients, and recommend food to the patient based on these clusters. The patient nutrition calculation algorithm is used to recommend patient nutrient data based on specific criteria such as potassium needs, calorie requirements etc. It utilizes inputs like the patient's age, gender, and blood test report to calculate and recommend suitable nutrient levels for the individual. After that the food inference algorithm use the outputs of this two algorithms as inputs and compare various parameters of the patient's health with the recommended values by the Food and Drug Administration (FDA). It then calculates the inferred nutrition of food based on FDA standards and the individual's health state to recommend food dishes that align with the person's nutritional needs. The accuracy of the recommendation system is evaluated using the confusion matrix to obtain this three metrics: Precision, Recall, and F1 score. The notable advantage of this system lies in its ability to minimize the excessive usage of medication and preventing potential side effects. Additionally, the dietary recommendation system can aid individuals in managing illnesses and avoiding the need for frequent appointments with physicians for medicines and dietary guidance.

Thongsri et al [17] have proposed a food recommender system based on collaborative filtering and the knapsack method. In the collaborative filtering segment, they computed the Pearson correlation coefficient between the ratings given by new users and old users to establish similarity based on correlation. Once similarity was determined, they calculated the prediction value for new users based on the ratings of old users. This allowed them to predict the preferences of new users based on the ratings of old users. By switching to the knapsack method, the system adopts a nuanced approach to recommend food quantities that align with individual user restrictions, including dietary considerations and health-related constraints. This method involves a strategic selection of food items, akin to solving the knapsack problem, where foods are considered based on their nutritional value (”weights”) and alignment with user preferences (”values”). The goal is to maximize the total preference value while adhering to a predetermined caloric intake, calculated from each user's basal metabolic rate (BMR). Through the application of dynamic programming, the system efficiently selects recommended menus, optimizing for user preference scores within the framework of each user's daily caloric limit. This dual-method approach ensures that recommendations are not only personalized to taste but also to nutritional needs and health goals, offering a comprehensive solution for personalized food recommendations.The evaluation results showed that users were satisfied with the system. The overall average satisfaction was 4.20, indicating a high level of satisfaction (where 5 points mean extremely satisfied, 1 point = absolutely dissatisfied).

Rehman et al [13] have proposed a food recommendation system for people suffering from common diseases. The proposed model is a cloud based food recommendation system called Diet-Right. The model utilize ant colony algorithm (ACO) that involves suggesting optimal and appropriate foods and nutrition based on the results of their pathological tests. The main objective of using ACO is to optimize the selection of foods based on users' pathological reports, thereby contributing to the control and prevention of various diet-related diseases. This approach aims at providing personalized and accurate food recommendations by leveraging the values of users' pathological reports, enhancing the relevance and effectiveness of the proposed food recommendation system. The food recommendation process involves constructing a food graph where each food is a node. Ants create local solutions by visiting nodes with the best cost relative to nutritional goals. Transition probability between nodes is calculated based on pheromone and heuristic information. Ultimately, a globally optimized solution is achieved by comparing the root mean square errors (RMSE) of different ant solutions, updating pheromone to converge towards the optimal solution. The authors constructed a database of 345 pathological test reports and their correspond-ing normal ranges. The dataset was created through a field survey, gathering pathological reports from various laboratories and verified by hospital pathologist. Additionally, a database of 3400 food items was compiled, each with 26 entries detailing common nutritional information, sourced from the official website composition of foods integrated dataset (CoFID)[?]. It is shown that parallel execution on cloud reduces convergence time by approximately 12 times. Furthermore, sufficient accuracy is attainable by increasing the number of ants.

Raguvaran et al [8] have introduced an Enhanced Long Short-Term Memory (E-LSTM) model for analyzing nutrition in food recommendation images for cancer patients. The E-LSTM method automatically suggests suitable dietary charts for cancer patients, incorporating both positive and negative nutrients for every patient. To enhance prediction and classification accuracy, a specialized layer named E-LSTM has been integrated. The proposed method consists of six layers. The input layer which takes two types of inputs: food images and that are used for classifying and generating a diet meal plan, and details about cancer patients which are varied upon the patient body condition, and the stage of cancer. Feature extraction and mapping layer which extracts features like ingredients and nutrients from the food images then maps the extracted nutrient values to predefined limits for each patient based on their cancer stage. The LSTM Layer contains multiple LSTM blocks to predict amount of each nutrient in the food images. A single LSTM block uses three gates (input, forget, output) to control flow of information. The predicted nutrient values are the outputs which passed through Softmax layer. The softmax layer is a crucial component in deep learning models used for classification tasks. It employs the softmax function, which converts predicted nutrient values to a probability distribution ensuring that they fall within the range of 0 and 1. The classification layer compares predicted nutrient values with mapped limits, and classifies nutrients as positive or negative for each patient (positive if within recommended limits, negative if exceeds limits). The output layer which is the final layer of the proposed E-LSTM model provides nutrition suggestions for patients (recommends foods with high positive nutrients and avoids foods with high negative nutrients). The authors used Food-101 [19] food image dataset for food classification, as it is publicly available. The Food-101 comprises 101 categories of food items and each category contain 1000s of images. A comparison with standard LSTM and existing food recommendation models such as DeepFood, Smart-Log, CSW-WLIFC, and Quantized DRCNN demonstrates the superior performance of the proposed E-LSTM technique in terms of F1 Score, Precision, Classification Accuracy, Recall, Training Loss, and Validation Loss.

In [18], the authors have presented a food recommendation approach, designed to suggest personalized daily meal plans for users taking into account their nutritional needs and previous food preferences. The general architecture of the system is composed of four layers, starting with the information gathering layer that captures all relevant nutrition-related information associated with the user. The user profile dataset which is focused on storing the information that characterizes the users. The intelligent system layer, which is dedicated to receive user profile information as inputs and generating recommended meal plans as outputs. This layer consists of three main components : 1) the nutritional context determination, which filters out foods that are not suitable for the current user recommendation; 2) the short-term intelligent models that generates daily meal plans using an optimization approach to maximize user preferences while ensuring nutritional requirements are met; and the long-term intelligent models that fine-tune the generated daily plans by considering weekly and monthly eating patterns for users to follow. An end-user interface that is designed to present the recommended meal plans and additional visualizations of nutritional information. This interface prioritizes the collection of user feedback on the recommendations provided. This feedback is then relayed to the information processing layer and continually and utilized on an ongoing basis in the user profiling. The objective of the presented study is then to propose a comprehensive solution intended for incorporation as the intelligent systems component within this architecture. This solution integrates the use of Multi-Criteria Decision Analysis, referred to as AHPSort, to identify the nutritional context. AHPSort is specifically used in this paper to rank food alternatives based on various criteria through pairwise comparisons, and then eliminating unsuitable foods. Additionally, it includes a short-term intelligent model based on an optimization scenario that considers both nutritional factors and individual preferences. It also includes a probabilistic approach to compute user preferences based on past common food intake, providing an alternative when the user disagrees with the initially generated menu.The dataset used in this article, comprising the list of food items, is constructed based on two popular food composition tables provided by Wander [18]. These tables contains nutritional data for 600 foods categorized into 12 groups, including calorie content and 20 different macronutrients and micronutrients. Using the proposed AHPSort method,the pre-filtering stage eliminates 32 and 40 foods detected as inappropriate for overweighted and diabetic users. A menu templates that is used in the menu recommendation is also defined as initial data. This template comprises a breakfast, a lunch, and a dinner. Once the foods are classified as appropriate an optimization scenario is used for filling the menu template presented. Moreover, an analysis of the optimization-based stage using 50 artificial user profiles indicates that it successfully achieves its goal of promoting the recommendation of foods with high consumption frequency but not recently consumed. In addition, it was explicitly confirmed that larger user profiles enhance the creation of a more varied menu composition, and that the probability-based approach performs better in such profiles. Finally, the mentioned finding globally demonstrate that the proposed approach is effective in achieving its primary objective of personalized menu delivery.

Stefanidis et al [15]focused on providing an AI-driven nutritional advisor which falls under the category of a knowledge-based recommender system for meal plans. This system takes into account user profiles such as physical characteristics, dietary choices, health conditions, preferences, etc. The general architecture of the system consists of two layers. The first one is a qualitative layer that focuses on verifying the appropriateness of individual ingredients within meals, operating as an expert system which employs fuzzy inference techniques and an ontology of rules curated by nutrition experts. The second one is a quantitative layer operates as an optimization method to generate daily meal plans. The proposed AI nutritional advisor aims to deliver daily nutritional plans (NPs) tailored to its users by analyzing their profiles. This advisor comprises two main components: the Reasoning-based Decision Support System (RDSS) and the NP generation component. The RDSS determines the suitable meals for a user accounting on their profile information, available meal options, and a qualitative rule ontology. The RDSS pre-filters meals by eliminating those that are incompatible with the user's profile and suggesting meals containing nutrients that the user needs most according to their profile. Therefore, its output is a list of acceptable meal options used to generate nutritional plans. The NP generation component constructs daily meal plans by combining appropriate meals based on quantitative rules established by a group of experts, including medical professionals, nutritionists, and physical activity specialists. More specifically, this component receives as inputs : the user profile and the list of appropriate meals recommended by the RDSS. The dataset used in this article have been validated by experts, and it was accessible at https://doi.org/10.5281/zenodo.7143234 (accessed on October 20, 2022). Hence, the framework was evaluated through three types of experiments: a small-scale experiment (200 meals) to validate the suitability of recommended meals, a large-scale experiment (3000 virtual users and 21,000 meal plans) to evaluate meal plan accuracy and variability, and a medium-scale experiment (300 virtual users) to examine the system's recommendation capacity based on user profile complexity. The user groups can be further categorized into three main categories: a.) Healthy individuals that are Adolescents, Adults and Older. b.) Individuals who are likely to require supervision from a nutrition specialist that are adults with excess weight and Athletes.c.) Individuals with health conditions that are adults with: obesity, cardiovascular disease, Type-2 Diabetes, iron deficiency anaemia and adults with a diet low in fruit and vegetables. The obtained results indicate that the system performed with high accuracy in generating suitable recommendations for macronutrients and essential food elements in daily meal plans for all user groups (92.65% and 85.86% accuracy, respectively). However, a factor limiting this accuracy was the restricted variety of meals accessible in the database for the recommender system's to use.

Lambay and Mohideen [10] have proposed a hybrid recommender system (HRS) for providing healthy diet recommendations using machine learning (ML) and big data analytics. This system incorporates natural language processing (NLP) for data pre-processing and ML algorithms for predictions and generating recommendations. In order to realize the HRS framework an algorithm named Intelligent Recommender for Healthy Diet (IR-HD) is proposed. The IR-HD algorithm uses matrix factorization and similarity measures (Euclidean distance, cosine similarity, and Pearson similarity) to predict user ratings for different foods and generate recommendations based on user preferences and food nutrition data. It aims to minimize the root mean square error (RMSE) between predicted and actual ratings. The proposed approach involves key steps such as pre-processing the data using NLP techniques, constructing a user-food matrix and a user-food ratings matrix based on the training data, computing health scores for each user-food pair based on the user's needs and the food's nutritional values, generating personalized healthy diet recommendations using the (IR-HD) algorithm. The general architecture of the proposed framework consists of three main layers. They are known as cloud layer, middle layer and application layer. The Cloud Layer is responsible for data storage and processing. This dataset is utilized for generating a user-food matrix, and includes comprehensive information on various foods, including their nutritional values and other dietary details. The middle Layer contains the underlying methods for healthy diet recommendations. Finally the application layer is the interface utilized by end users.

Rostami et al [14] presented a system called Healthy and Time-Aware Food Recommender System (HTRFS) that combines innovative elements to provide personalized and healthy food recommendations. This system utilizes time-aware collaborative filtering and a food ingredients-based model. The time-aware collaborative filtering phase uses the user-food rating matrix to estimate user-to-user similarity and recommend favorite foods to active users. The authors introduced a formula to calculate temporal similarity between users based on food ratings and the time elapsed since those ratings. In the "Food ingredients-based prediction rating" phase, a model based on food ingredients predicts user ratings by considering the nutritional information of foods. The "Preference rating prediction" process predicts user preferences based on user similarities and historical ratings. Additionally, the "Food health factor" calculation evaluates food healthiness to provide final healthy recommendations to target users. By combining these steps, the model effectively recommends healthy foods considering user preferences, food health factors, and historical ratings. The authors evaluated the model's performance against several state-of-the-art food recommendation systems using metrics like precision, recall, F1 score, AUC, and NDCG. Experimental data from Allrecipes.com and Food.com demonstrated the effectiveness of the proposed food recommender system compared to previous models.

Zioutos et al [19] in their study proposed a novel hybrid recommendation system (SHARE) that provides personalized weekly meal plans by combining collaborative filtering and content-based techniques. SHARE leverages collaborative filtering to identify users with similar dietary preferences and recommend suitable recipes accordingly. It also employs content-based filtering by analyzing recipe nutritional information and ingredients, along with a knowledge-base component that maps chronic health conditions to appropriate nutritional profiles for filtering out unsuitable recipes. The key novelty of SHARE is its dynamic adaptation feature that allows users to interactively adjust meal plan recommendations based on constraints, preferences, keyword/nutrition filters, and apply positive weightings with the system learning from these choices to refine future recommendations. The dataset used is from food.com, containing real-world recipe ratings from numerous users. It also includes assigned chronic health conditions (cancer, obesity, diabetes, etc.) to some users to simulate their health history. Experiments on food.com dataset with 40 users showed high success rates in recommending personalized and suitable recipes based on various use case scenarios involving different adaptation aspects. The system demonstrated its ability to provide accurate, dynamically adaptive meal planning aligned with user tastes and health needs, making it a promising tool for improving dietary choices and overall well-being.

3.3 Analysis and comparison

The table below represents the main characteristics of the different column approaches:
− The column " Approach " defines the author of the proposed approach.
− The column " Dataset " Indicates the set of data used to manage the system.
− The column "Evaluation of performances " Denotes how well the approach performs.
− The column "Used techniques " Defines the different methods used in this approach.
− The column " advantages " Presents the main advantages of this approach.
− The column " disadvantages "Defines the disadvantages of the approach.

After reviewing the works presented in the section above, we found out that there exists various methodologies for constructing a food recommendation system tailored to provide daily meal plan to healthy individuals and those with specific health conditions. Presently, the predominant approaches involve the use of content-based filtering, collaborative-based filtering, hybrid methods, machine learning algorithms, artificial neural networks, and deep learning approaches, optimization methods, and probabilistic techniques. The presented studies emphasize the importance of choosing appropriate algorithms or techniques to enhance the effectiveness of these models. Despite the advancements in food recommendation systems, challenges remain due to the highly contextualized and personalized nature of food

recommendations. Therefore, there is a need to investigate specialized food recommendation approaches that integrate with other types of recommendation systems to provide comprehensive and personalized solutions.
Moreover, our review show that there is a scarcity on food recommendation systems specifically designed for cancer patients. Here [12]; [8]; [11]; [16] are the presented distinct approaches to personalized dietary recommendations for cancer patients. [11] introduces a hybrid recommender system based on anti-cancer molecules in recipes, emphasizing cancer-preventing strategies and nutrition. The study in [16] offers personalized dietary recommendations using data mining techniques, tailored to individual health conditions and preferences. Hence, a personal preference knowledge graph (PPKG) recommendation system is provided recipe recommendations. The system uses the KGAT architecture and it incorporate the LSTM network to capture users' dietary habits from their historical dietary records and predict recipes which users may like based on their habit model. In [12] proposes a CapsNet model for food recommendation, achieving a 95% accuracy and reduced loss, focusing on healthy meal suggestions for cancer patients.The system aims to consider the patient's bodily state and cancer stage, such as early, medium, and severe, to suggest appropriate food items. The limitations of the system include the lack of real-world implementation and testing of the system. The research in [8] aims to propose a new approach using deep learning techniques to provide cancer patients with an effective nutrient plan. This method is an Enhanced Long-Short Term Memory (E-LSTM) classifier which is based on nutrition analysis of food images.

Furthermore, the presented approaches in the previous section explored a variety of data sources including large public food datasets, custom datasets created through surveys or web crawling, and food composition tables. Techniques employed ranged from collaborative filtering algorithms like SVD++, ALS, BPR to Content-based filtering using text/image features and neural networks (LSTM, CapsNet).
Most of the approaches using CB technique recommend items that are similar in content to the item that the user liked in the past. However, this technique is efficient only if the item can be represented as a set of features. In addition, these approaches suffer from plasticity (the ability to change the user's preferences). The CF technique matches users who shared same preferences using the ratings for items in particular domain. The majority of the approaches are based on the CF technique. For hybrid recommendation, it integrates two or more recommendation techniques to limit the weaknesses of individual ones. However, the use of RSs has exposed many challenges : data sparsity, cold start problems, fraud and privacy . Those who try to improve CF approaches only take into concern the ratings given to the products by the users, which mean that they do not include the knowledge about the active user in the recommendation process, active user's neighbors, products nor relationships between them. Some methods incorporated knowledge graphs and domain knowledge about recipes, ingredients, and diseases. Hybrid approaches combining collaborative and content-based techniques to limit the weaknesses of individual ones.were popular, along with clustering, matrix factorization, and similarity measures. Knowledge Graphs (KGAT) and domain knowledge incorporation were used in some cases. Clustering

Approach	**Dataset**	**Used techniques**	**Evaluation of performances**	**Advantages**
Pawar et al[11]	1M+ and K&N datasets	CBF, SVD++, Support vector classifier, CF: ALS, BPR, LMF	ALS outperforms from all the CF algorithms tested	Hybrid approach aims to overcome limitations of single techniques
Raguvaran et al[12]	Food-101, Cancer patient	CapsNet	Accuracy = 95%, Loss (0.025 at 20 epochs)	Improved classification accuracy due to multiple convolutional layers for feature extraction
Tang et al[16]	Self-created	KGAT, LSTM	Recall = 0.7213, NDCG = 0.08902	Considers the influence of time on users' taste preferences by modeling dietary records as sequences
Ahmad et al[9]	/	K-means clustering, food inference, patient nutrition calculation	Very proficient system	/
Thongsri et al[17]	/	CF, knapsack method	Satisfaction = 4.20 per 5	Addresses the cold-start problem by collecting initial user preferences through questionnaires
Rehman et al[13]	Created through a field survey	ACO	/	Cloud-based system provides scalability and pervasiveness
Raguvaran et al[8]	Food-101 food image, Patient dataset	E-LSTM	E-LSTM outperforms standard LSTM, Deep-Food, Smart-Log, CSW-WLIFC and Quantized DRCNN	First work producing a dedicated feature extraction layer for LSTM for this task
Rostami et al[14]	Allrecipes.com, Food.com	Time-aware CF	/	/

Table 3.1: Comparative study of related works (Part 1).

Approach	**Dataset**	**Used techniques**	**Evaluation of performances**	**Advantages**
Lambay and Mohideen[10]	/	HRS, IR-HD, NLP, similarity measures	RMSE = 0.18445	Combines the strengths of NLP techniques for preprocessing and ML algorithms for data analysis and recommendation generation
Zioutos et al [19]	Real-world recipe ratings	User-based CF using cosine similarity, CBF, Rating decay mechanism, knowledge-based RS	Success rates (80-90%)	/
Stefanidis et al[15]	https ://doi.org/ 10.5281/ zenodo. 7143234 (accessed on October 20, 2022)	Knowledge-based RS	Accuracy = 92.65% and 85.86%	Identifying strategies to help patients eat well during and after cancer treatment
Toledo et al[18]	Constructed based on two popular food composition tables provided by Wander	AHPSort, short-term intelligent model	Effective	/

Table 3.2: Comparative study of related works(Part 2).

(K-means), Matrix Factorization, and Similarity Measures (Euclidean, Cosine, Pearson) were also applied. Optimization methods like Knapsack, Ant Colony, and AHP were also used in some studies.

3.4 Conclusion

In this chapter, we have performed a comprehensive a state-of-the-art review where we presented some of the most influential works in food recommendations that employed concepts and techniques of, machine learning, deep learning, and artificial neural networks, optimization and probabilistic approaches. We thoroughly examined each article in these works, analysing the approaches proposed by the authors, evaluating the models' results, summarizing the conclusions drawn by the researchers. Each model exhibited its set of advantages and disadvantages. Each of these approaches made a valuable contribution in providing suitable daily meal plan for their user.

4
Contributions

4.1 Introduction

Nowadays, food recommendation systems and extraction of useful information in an automatic way are very important and help doctors to effectively diagnose cancer patients and build patient databases.

The number of people suffering from cancer is increasing. Accurate diagnosis at an early stage followed by appropriate subsequent treatment can reduce the risk of complications on the patient's health resulting from the cancer disease or prevent it.

In this chapter, we will present in detail our approach that we proposed and used during our project as well as its different steps for food recommendation using medical details, starting with the collection and preprocessing of data and then the construction of the proposed model.

4.2 Contribution:

Our project consists on a food recommendation system for cancer patients based on sentiment analysis, i.e. enabling people with cancer to know the foods that match their nutrient needs as well as their preferences and feelings towards different foods. To achieve this goal, several steps must be followed to achieve better results. These steps are: data collection, preprocessing, model, data training, model evaluations.

The following figure gives an overview of the proposed approach and the different stages that make it up:

Import data from FNDDS database [20], in the form of a Dataset in CSV format, preprocessing of input data (Exploration and visualization of data, Data cleaning, etc.), then selection of characteristics, construction of the model, training, recommendation and calculation of precision, evaluation of the model.

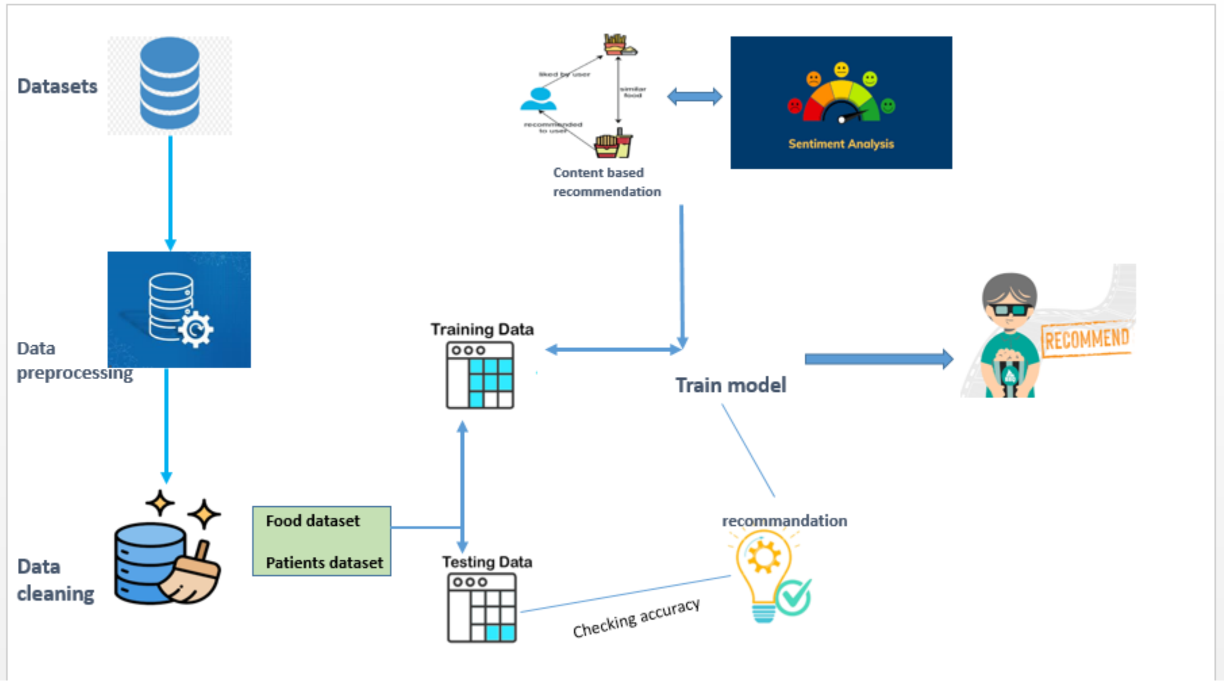


Figure 4.1: Proposed approach.

We detail each step below as follows:
The first step is devoted to data collection, which consists of actively extracting information from the source.
The second step consists on pre-processing the collected data; this step is made up of several sub-steps including: cleaning of aberrant data, analysis and visualization of the data.
The third step consists of training and testing the data, it is a very important step to achieve a good prediction result.
The last step is devoted to the construction of the model followed by a final recommendation.

4.3 Data collection

Data collection is a very important initial step in recommending foods for cancer patients. This allows for proper processing and evaluation of the chosen approach. For our approach, we have two data sets: food data set and cancer patients' data set.

Food dataset: The USDA's Food and Nutrient Database for Dietary Studies (FNDDS) is an application database created for analyzing dietary intakes from What We Eat in America (WWEIA), National Health and Nutrition Examination Survey (NHANES). It converts food and beverage portions reported in the survey into gram amounts and determines their nutrient values.
The dataset we use in our study (FNDDS 2019-2020) contains 5,624 food and beverage items (4,982 foods/642 beverages).

Cancer patients' dataset: It is a self-created dataset; contain 1000 patients with different types of cancer and 10 columns.

4.4 Data preprocessing

After data collection, the next step is preprocessing, the latter is very important to extract a perfect dataset in order to obtain quality results. This step involves cleaning, transforming the data into a format that is processed more easily and efficiently in algorithms, and integrating the raw data to prepare it for analysis. When we collect real-world data, it consists of redundant data, missing values, and outliers that will always result in a model that would not be effective for recommendation and data analysis. Therefore, with the help of data preprocessing, we can remove all these problems [21].

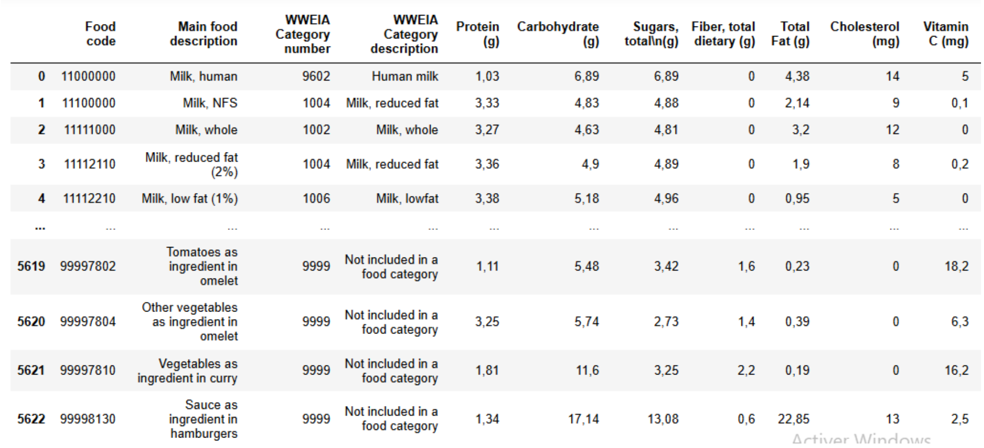
Data preprocessing includes different phases:

Data exploration and visualization:
Data exploration is considered as the first step in data analysis process. It is a crucial step in this process, because it helps identify outliers within the data that can inform subsequent data cleaning.
This process is multifaceted and involves several key steps beginning with data collection and preparation, then, exploratory data analysis (EDA) techniques are applied.
EDA involves generating summary statistics, visualizing data distributions, and identifying existing relationships between variables. As well as, data visualization is an essential part of data analysis. It is defined as the visual exploration of data, which helps to obtain and know in-depth and clear information and features about the dataset and variables. It represents data through use of common graphics, such as info-graphics, charts… It helps to grasp the underlying patterns and variations in the data.

We note that:
− The number of rows in the foods dataset is 5624 (0 to 5623 foods) and 11 columns named as follow(Food code, Main food description, WWEIA Category number, WWEIA Category description, Protein (g), Carbohydrate (g), Sugars, total (g), Fiber, total dietary (g), Total Fat (g), Cholesterol (mg), Vitamin C (mg)),
− In the patients' dataset, we have 1000 rows (patients) and 10 columns(Patient Id,

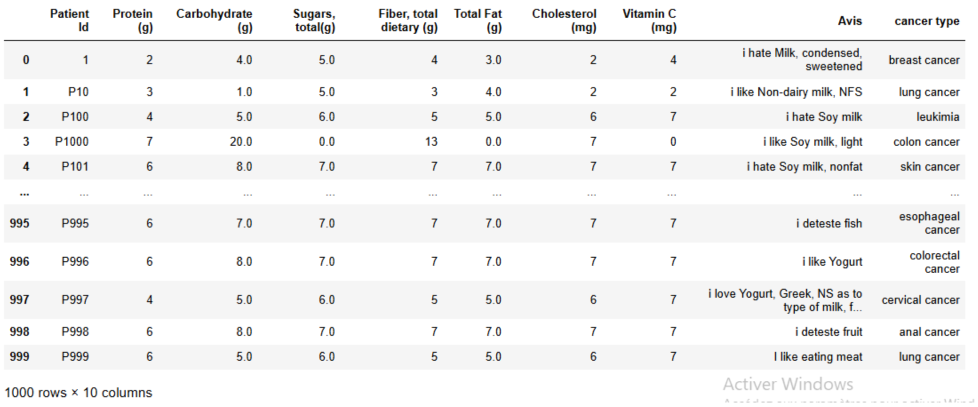
Protein (g), Carbohydrate (g), Sugars, total (g), Fiber, total dietary (g), Total Fat (g), Cholesterol (mg), Vitamin C (mg), Avis, cancer type).

The figures (4.2) and (4.3) below shows an overview of the two data sets used in our approach:



	Food code	Main food description	WWEIA Category number	WWEIA Category description	Protein (g)	Carbohydrate (g)	Sugars, total\n(g)	Fiber, total dietary (g)	Total Fat (g)	Cholesterol (mg)	Vitamin C (mg)
0	11000000	Milk, human	9602	Human milk	1,03	6,89	6,89	0	4,38	14	5
1	11100000	Milk, NFS	1004	Milk, reduced fat	3,33	4,83	4,88	0	2,14	9	0,1
2	11111000	Milk, whole	1002	Milk, whole	3,27	4,63	4,81	0	3,2	12	0
3	11112110	Milk, reduced fat (2%)	1004	Milk, reduced fat	3,36	4,9	4,89	0	1,9	8	0,2
4	11112210	Milk, low fat (1%)	1006	Milk, lowfat	3,38	5,18	4,96	0	0,95	5	0
...	...	...	...	...	...	...	...	...	...	...	...
5619	99997802	Tomatoes as ingredient in omelet	9999	Not included in a food category	1,11	5,48	3,42	1,6	0,23	0	18,2
5620	99997804	Other vegetables as ingredient in omelet	9999	Not included in a food category	3,25	5,74	2,73	1,4	0,39	0	6,3
5621	99997810	Vegetables as ingredient in curry	9999	Not included in a food category	1,81	11,6	3,25	2,2	0,19	0	16,2
5622	99998130	Sauce as ingredient in hamburgers	9999	Not included in a food category	1,34	17,14	13,08	0,6	22,85	13	2,5

Figure 4.2: Foods dataset overview.



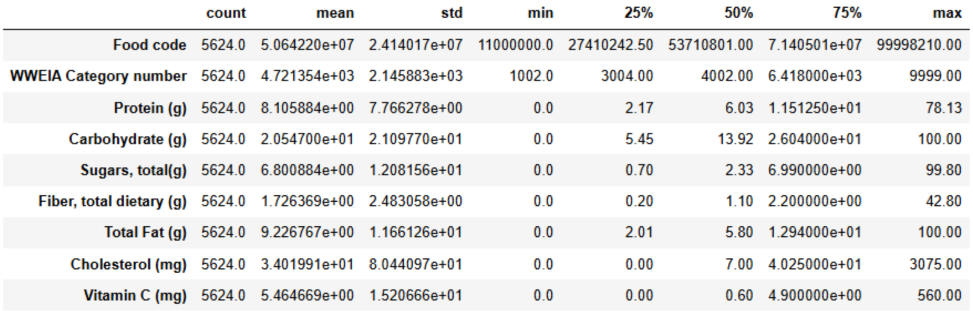
	Patient Id	Protein (g)	Carbohydrate (g)	Sugars, total(g)	Fiber, total dietary (g)	Total Fat (g)	Cholesterol (mg)	Vitamin C (mg)	Avis	cancer type
0	1	2	4.0	5.0	4	3.0	2	4	i hate Milk, condensed, sweetened	breast cancer
1	P10	3	1.0	5.0	3	4.0	2	2	i like Non-dairy milk, NFS	lung cancer
2	P100	4	5.0	6.0	5	5.0	6	7	i hate Soy milk	leukimia
3	P1000	7	20.0	0.0	13	0.0	7	0	i like Soy milk, light	colon cancer
4	P101	6	8.0	7.0	7	7.0	7	7	i hate Soy milk, nonfat	skin cancer
...	...	...	...	...	...	...	...	...	...	...
995	P995	6	7.0	7.0	7	7.0	7	7	i deteste fish	esophageal cancer
996	P996	6	8.0	7.0	7	7.0	7	7	i like Yogurt	colorectal cancer
997	P997	4	5.0	6.0	5	5.0	6	7	i love Yogurt, Greek, NS as to type of milk, f...	cervical cancer
998	P998	6	8.0	7.0	7	7.0	7	7	i deteste fruit	anal cancer
999	P999	6	5.0	6.0	5	5.0	6	7	I like eating meat	lung cancer

1000 rows × 10 columns

Figure 4.3: Overview of patients' dataset.

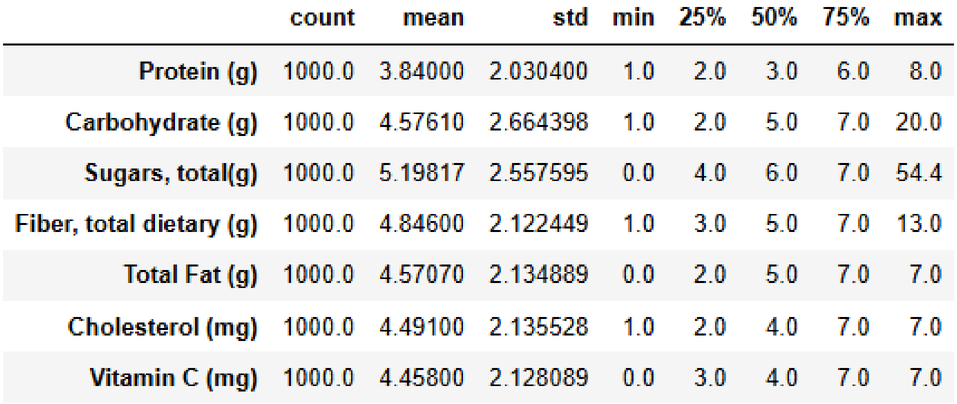
Statistical summary of the Data Frame:

This summary gives as a quick overview of the dataset. We use pandas describe method to view some basic statistical details like: count, mean, std, min, max. . . , pandas information shows column data types (feature), number of non-zero values and memory usage.



	count	mean	std	min	25%	50%	75%	max
Food code	5624.0	5.064220e+07	2.414017e+07	11000000.0	27410242.50	53710801.00	7.140501e+07	99998210.00
WWEIA Category number	5624.0	4.721354e+03	2.145883e+03	1002.0	3004.00	4002.00	6.418000e+03	9999.00
Protein (g)	5624.0	8.105884e+00	7.766278e+00	0.0	2.17	6.03	1.151250e+01	78.13
Carbohydrate (g)	5624.0	2.054700e+01	2.109770e+01	0.0	5.45	13.92	2.604000e+01	100.00
Sugars, total(g)	5624.0	6.800884e+00	1.208156e+01	0.0	0.70	2.33	6.990000e+00	99.80
Fiber, total dietary (g)	5624.0	1.726369e+00	2.483058e+00	0.0	0.20	1.10	2.200000e+00	42.80
Total Fat (g)	5624.0	9.226767e+00	1.166126e+01	0.0	2.01	5.80	1.294000e+01	100.00
Cholesterol (mg)	5624.0	3.401991e+01	8.044097e+01	0.0	0.00	7.00	4.025000e+01	3075.00
Vitamin C (mg)	5624.0	5.464669e+00	1.520666e+01	0.0	0.00	0.60	4.900000e+00	560.00

Figure 4.4: Statistical summary of the food dataset.



	count	mean	std	min	25%	50%	75%	max
Protein (g)	1000.0	3.84000	2.030400	1.0	2.0	3.0	6.0	8.0
Carbohydrate (g)	1000.0	4.57610	2.664398	1.0	2.0	5.0	7.0	20.0
Sugars, total(g)	1000.0	5.19817	2.557595	0.0	4.0	6.0	7.0	54.4
Fiber, total dietary (g)	1000.0	4.84600	2.122449	1.0	3.0	5.0	7.0	13.0
Total Fat (g)	1000.0	4.57070	2.134889	0.0	2.0	5.0	7.0	7.0
Cholesterol (mg)	1000.0	4.49100	2.135528	1.0	2.0	4.0	7.0	7.0
Vitamin C (mg)	1000.0	4.45800	2.128089	0.0	3.0	4.0	7.0	7.0

Figure 4.5: Statistical summary of the patients' dataset.

The two figures below (4.4) and (4.5) are the statistical of our data sets:

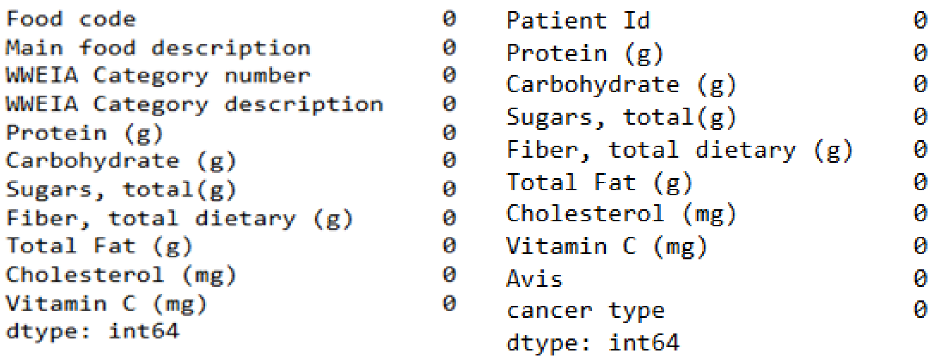
Data cleaning:
Data cleansing is a process that aims to identify and correct corrupted, inaccurate or irrelevant data. This fundamental step in data processing improves the consistency, reliability and value of data and helps strengthen the integrity and relevance of data by reducing inconsistencies, avoiding errors and enabling better, more accurate decisions.

The data cleaning process involves several steps such as:

Missing value analysis:
"Missing values" refer to missing data in a dataset. These absences can occur for a variety of reasons, such as measurement errors, data entry errors, unanswered questions in surveys, or data collection problems.

Missing values pose a challenge because most algorithms cannot handle incomplete data directly.

To detect missing values in our datasets with Python, we used the "is null() and sum()" methods from the Pandas library.



```
Food code                      0
Main food description          0
WWEIA Category number          0
WWEIA Category description     0
Protein (g)                    0
Carbohydrate (g)               0
Sugars, total(g)               0
Fiber, total dietary (g)       0
Total Fat (g)                  0
Cholesterol (mg)               0
Vitamin C (mg)                 0
dtype: int64
```

```
Patient Id                 0
Protein (g)                0
Carbohydrate (g)           0
Sugars, total(g)           0
Fiber, total dietary (g)   0
Total Fat (g)              0
Cholesterol (mg)           0
Vitamin C (mg)             0
Avis                       0
cancer type                0
dtype: int64
```

Figure 4.6: Missing values in foods and patient's datasets..

The figure 4.6 shows that there are no missing values in our datasets.

A duplicate is an entry in a dataset that is identical to one or more other entries. Duplicates can pose analysis problems and bias model results because they distort descriptive statistics and can give excessive weight to certain observations.
To detect duplicates with python we used the method (duplicated()) from the Pandas library, and we did not find any duplicates in our datasets.

Correlation matrix: The correlation matrix is a table showing the correlation coefficients between several variables. Each cell of the matrix shows the correlation between two different variables. The correlation value generally varies between -1 and 1:
- +1 indicates a perfect positive correlation (as one variable increases, the other increases as well).
- -1 indicates a perfect negative correlation (as one variable increases, the other decreases).
- 0 indicates no correlation (variables are not linearly related)[22].

The visualization of correlation matrices for the two datasets are shown in the following figures:

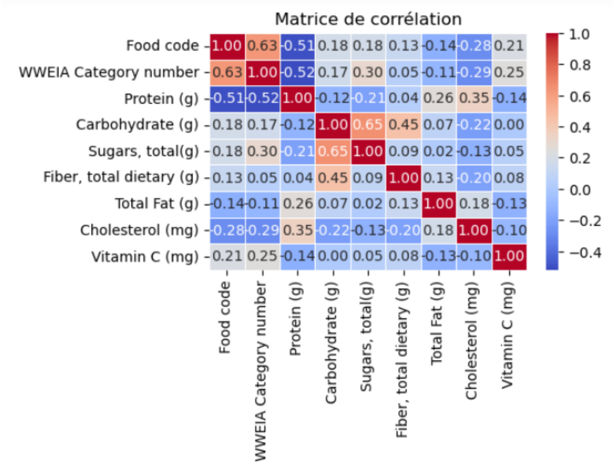


Figure 4.7: Correlation matrix of foods dataset.

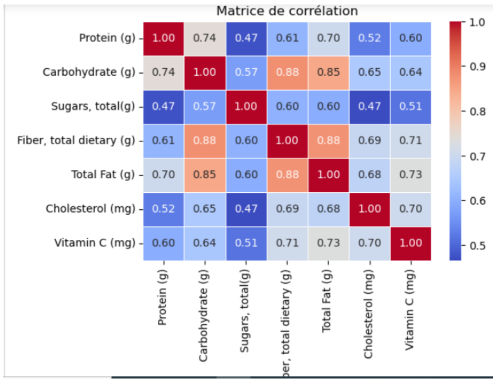


Figure 4.8: Correlation matrix of patients' dataset.

4.5 Training and Testing Data

It is important to train and test a model to achieve a good recommendation result. This method consists of dividing the data set into two parts: training part on which the model does its learning and test part on which we test the model and evaluate its performance. If a model performs better in both datasets, then the expected accuracy is better.

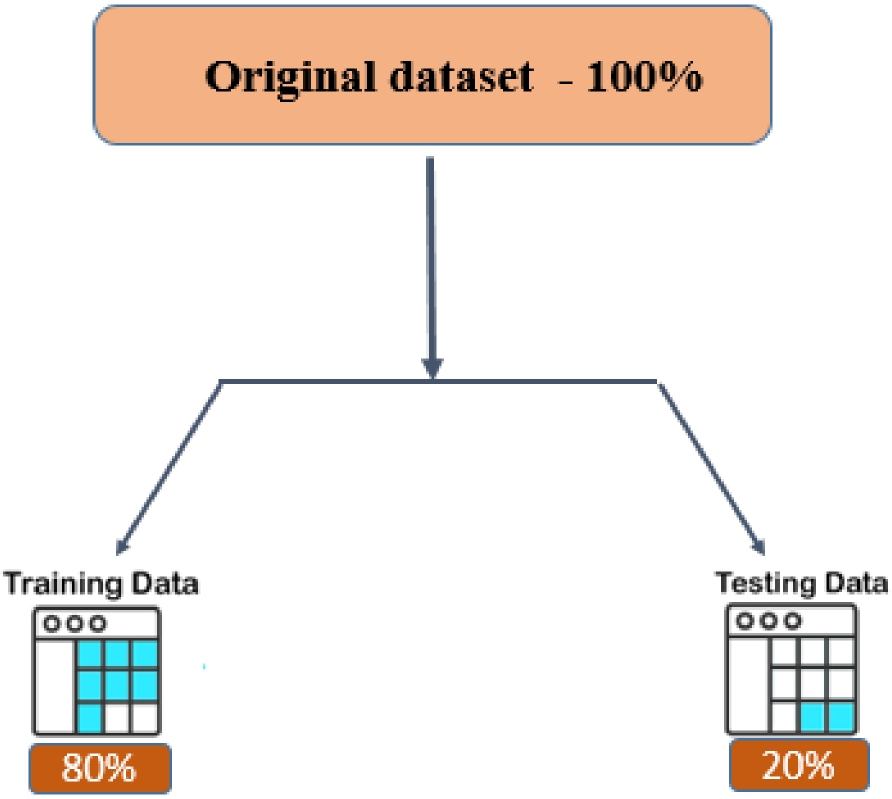


Figure 4.9: Train/test data.

4.6 Construction of the model

After data splitting, the next step is model selection to build a food recommendation system for cancer patients. This step involves the implementation of different layers and algorithms. This work consists of three main layers (input layer, selection layer, classification layer) represented as follows:

Input layer:

The first layer of the proposed approach is an input layer where inputs are introduced. Two types of entries were used: a food list containing information on different types of foods and another list containing details about cancer patients such as their nutrient requirements (protein, fiber, etc.).

The food CSV dataset and a dataset containing cancer patient details are fed to this input layer.

Selection layer:

The food list and patient details are used as inputs to extract their features, which are the amounts of nutrients for each food and each patient's requirement of those nutrients (The limit for each food nutrient is presented for each patient based of his state of health). They include key nutritional information such as protein, carbohydrates, sugars, fiber, fat, cholesterol and vitamin C.

Classification layer:

For all patients, recommended foods, foods to avoid will be classified using the Content-Based Filtering method.

Content-Based Filtering
For all cancer patients,

1. Characteristics of Items (Food) :
Food characteristics are represented by their nutritional values, such as the amounts of protein, carbohydrates, sugars, fiber, total fat, cholesterol and vitamin C.

2. User Preferences (Patients) :
◦ Patient needs are represented by their nutritional requirements and the feedback they provided, which is analyzed for positive or negative feelings.

Content-Based Filtering Steps in the approach
1. **Extraction of the Patient's Nutritional Needs :**
○ The patient's nutritional requirements are extracted from the relevant columns in the patient dataset.
2. **Filtering Foods According to Nutritional Needs :**
○ Foods whose nutritional values exceed the patient's needs are filtered out.
3. Calculation of Cosine Similarity :
○ The cosine similarity between the patient's nutritional needs and the nutritional values of foods is calculated to identify foods most similar to the patient's needs. If the cosine similarity is greater than the defined threshold then the food will be recommended otherwise it will be avoided.

Once the list of foods is recommended, sentiment analysis is used to optimize this list.

Sentiment Analysis

1. Extraction of Patient Reviews

○ Each patient can leave a text review describing their food preferences, dislikes, or general feelings toward certain foods.
○ Each food is associated with a Feeling Score : This is a value that varies from -1 (very negative) to 1 (very positive). A sentiment score close to 0 is considered neutral.

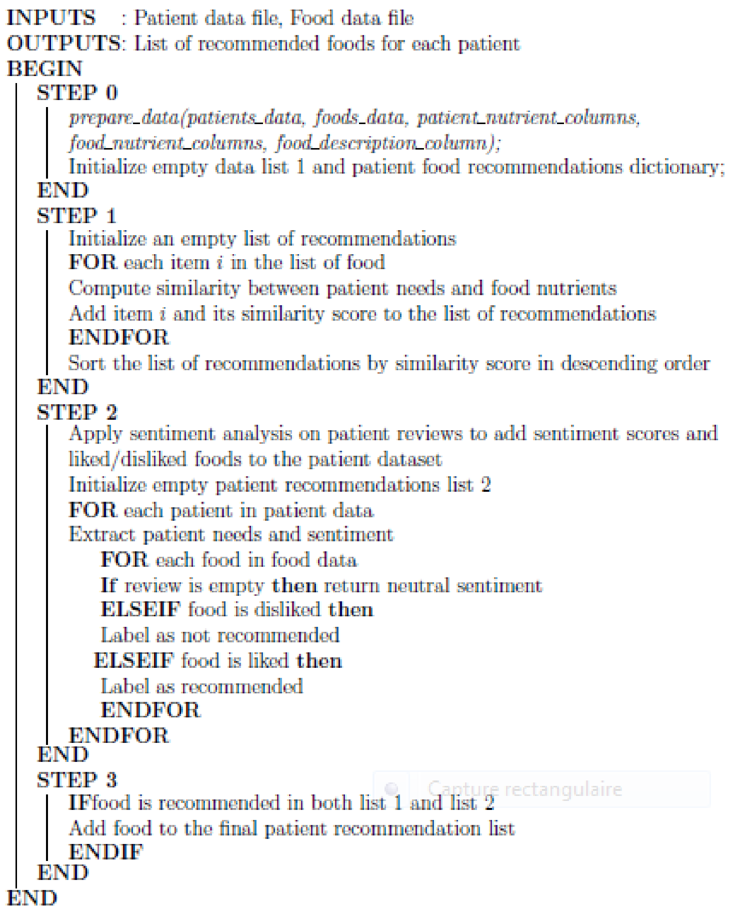
2. Identifying Liked and Disliked Foods
○ In addition to calculating the sentiment score, we look for specific words in reviews to identify foods that patients like or dislike.
○ For example, if a review says ”I hate broccoli ” or ”I love apples ”, this information is extracted for later use in recommendations.
○ Sentiment scores and lists of liked and disliked foods are added to patient data.
The food that exists in the recommendation list will be recommended according to the patient's opinion regarding this food. The method used here is a combination of word-based and rule-based sentiment analysis.

1. Word-based sentiment analysis (TextBlob):
– ‘TextBlob‘ uses lexicons and linguistic rules to determine the polarity (sentiment score) of patient reviews. Polarity varies from -1 (very negative) to +1 (very positive).

2. Rules-based sentiment analysis:
– The code scans reviews for specific words like ”hate”, ”like”, ”love” and ”prefer” to identify disliked or liked foods.
– If a review contains a food and the word ”hate”, that food is added to the list of hated foods.
– If a review contains a food and the words ”like”, ”love” or ”prefer”, this food is added

to the list of liked foods.

Here's an algorithmic which describes the sequence of the entire process.



```
INPUTS   : Patient data file, Food data file
OUTPUTS: List of recommended foods for each patient
BEGIN
  STEP 0
      prepare_data(patients_data, foods_data, patient_nutrient_columns,
      food_nutrient_columns, food_description_column);
      Initialize empty data list 1 and patient food recommendations dictionary;
  END
  STEP 1
      Initialize an empty list of recommendations
      FOR each item i in the list of food
      Compute similarity between patient needs and food nutrients
      Add item i and its similarity score to the list of recommendations
      ENDFOR
      Sort the list of recommendations by similarity score in descending order
  END
  STEP 2
      Apply sentiment analysis on patient reviews to add sentiment scores and
      liked/disliked foods to the patient dataset
      Initialize empty patient recommendations list 2
      FOR each patient in patient data
      Extract patient needs and sentiment
          FOR each food in food data
          If review is empty then return neutral sentiment
          ELSEIF food is disliked then
          Label as not recommended
         ELSEIF food is liked then
          Label as recommended
          ENDFOR
      ENDFOR
  END
  STEP 3
      IFfood is recommended in both list 1 and list 2
      Add food to the final patient recommendation list
      ENDIF
  END
END
```

Figure 4.10: Proposed approach.

4.7 Conclusion

In this chapter, we presented in detail our food recommendation approach for cancer patients, using content based filtering RS and sentiment analysis. Our approach is inspired by research relating to food recommendation which allows the use of several metrics for recommendation. In the next chapter, we will proceed to explain all aspects related to the implementation of our approach.

5

Experiment and evaluation

5.1 Introduction

The main objective of our project is to develop a system capable of providing personalized dietary recommendations for cancer patients, taking into account the specific nutritional needs of patients as well as their food preferences and aversions expressed in their opinions.

In this chapter, we present and analyze the results obtained from our recommendation model for cancer patients and the evaluation of the effectiveness of this system as well as the definition of the metrics used.

5.2 Datasets description

Dataset: A dataset is a structured collection of data in rows and columns containing information from multiple sources and can be of different types or of the same type, organized and stored together for analysis or processing.

A dataset can have a CSV, TSV extension, which can be imported with pandas functions in python.

The amount of data that a dataset can contain is deferred from one to another small (a few characteristics and 100 rows), large ((more than 1,000 characteristics and more than a million rows), the selection of features in the dataset is very essential in creating a model [23];[24].

Food dataset: The USDA's Food and Nutrient Database for Dietary Studies (FNDDS) is an application database created for analyzing dietary intakes from What We Eat in America (WWEIA), National Health and Nutrition Examination Survey (NHANES). It converts food and beverage portions reported in the survey into gram amounts and determines their nutrient values.

The dataset we use in our study (FNDDS 2019-2020) contains 5,624 lines (food and beverage items (4,982 foods/642 beverages)) and 11 columns as follow [20]:

- Food code: Unique identifier of the food.
- Main food description: name and main description of the food.
- WWEIA Category number: the category number to which the food belongs.
- WWEIA Category description:the name of the category to which the food belongs.
- Protein (g): the amount of protein in grams contained in this food.
- Carbohydrate (g): the amount of Carbohydrate in grams contained in this food.
- Sugars, total (g): the total amount of Sugars in grams contained in this food.
- Fiber, total dietary (g): the total amount of Fiber in grams contained in this food.
- Total Fat (g): the amount of Fat in grams contained in this food.
- Cholesterol (mg): the amount of Cholesterol in milligrams contained in this food.
- Vitamin C (mg): the amount of Vitamin C in milligrams contained in this food.

Patients' dataset: The user dataset used for our experiments was constructed by gathering information about cancer patient needs from different internet sites. The dataset consists of user IDs and associates the needed nutrients with their rating scores. The scale reflects the level of satisfaction expressed by each individual user towards the nutrients in question. All entries are arranged into a CSV file, wherein the first column includes the user ID, followed by columns with the nutrient name and then the given rating. We added also a column that includes the review of each used toward the foods (liked and disliked food), and a column indicating the type of cancer the patient has.

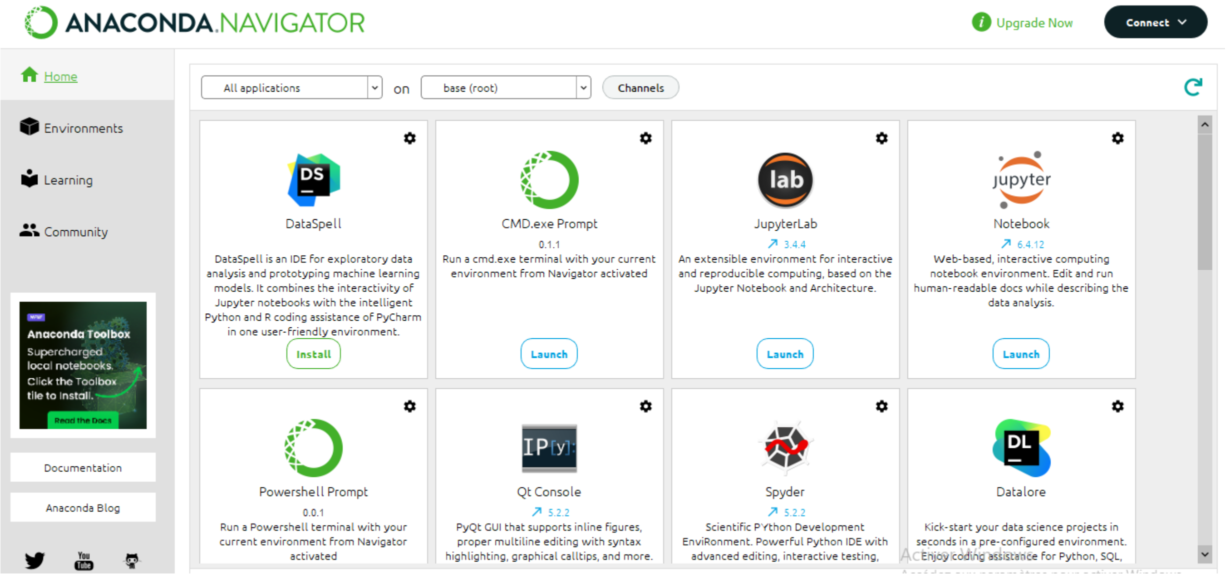
The dataset we use in our study contains 1000 lines (patients) and 10 columns represented as follow:

- Patient Id: Unique identifier of each patient.
- Protein (g): The maximum amount of Protein that patients need depending on the type of cancer they have.
- Carbohydrate (g): the maximum amount of Carbohydrate that patients need depending on the type of cancer they have.
- Sugars, total (g): The maximum amount of Sugars that patients need depending on the type of cancer they have.
- Fiber, total dietary (g): The maximum amount of Fiber, total dietary that patients need depending on the type of cancer they have.
- Total Fat (g): The maximum amount of Total Fat that patients need depending on the

type of cancer they have.
- Cholesterol (mg): The maximum amount of Cholesterol that patients need depending on the type of cancer they have.
- Vitamin C (mg): The maximum amount of Vitamin C that patients need depending on the type of cancer they have.
- Avis: patients' opinions on foods (the foods they like or dislike).

5.3 Development environment

- **Anaconda:** Anaconda is a free and open source distribution of the Python and R programming languages applied to application development, allows you to simplify the management of packages and virtual environments. It is very popular in the field of data science, machine learning(data processing and analysis, etc.), and software development due to its ease of use and integrated tools [25].



DataSpell is an IDE for exploratory data analysis and prototyping machine learning models. It combines the interactivity of Jupyter notebooks with the intelligent Python and R coding assistance of PyCharm in one user-friendly environment.

Run a cmd.exe terminal with your current environment from Navigator activated

An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Run a Powershell terminal with your current environment from Navigator activated

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing,

Kick-start your data science projects in seconds in a pre-configured environment. Enjoy coding assistance for Python, SQL,

Figure 5.1: Anaconda environment.

- **Jupyter notebook :** Jupyter Notebook is a web-based interactive development environment for notebooks, code, and data for creating and sharing documents that contain executable code, visualizations, and narrative text. Its flexible interface allows users to configure and organize workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality [29].

It is widely used in data science, education, and software development for its flexibility and ease of use.

5.4 Programming language

• **Python:** Python is a powerful, extensible, free, structured and easy to learn programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Originally developed by Guido Van Rossum in 1993, it is currently the most widely used language in the world [30].

Python's elegant syntax and dynamic typing, as well as its interpreted nature, make it ideal for scripting and rapid application development in many domains on most platforms.

Bibliothèques de Python :
• **Textblob :**
TextBlob is a python library used for processing textual data and it is free and open-source. It provides a simple API for common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. TextBlob is particularly useful for explore text's grammatical structure through linguistic annotations and extraction feature, and allows us to determine whether the input textual data has a positive, negative, or neutral tone [26]. . .

• **Pandas :**
Pandas: Acronym for Python Data Analysis Library is an open source library under the BSD license providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. This library was designed and developed primarily by Wes McKinney starting in 2008. Its installation by opening the command shell and calling the command: Pip install pandas. This allows you to easily manipulate data tables with labels of variables and individuals, these tables are called "DataFrames" (stored in CSV, TSV, XSL files, etc.). We can easily read and write these dataframes from or to a tabulated file as well as draw graphs from these DataFrames using Matplotlib [31].

• **Sklearn :**
Known as Scikit-learn is a Python library that provides a standard interface for implementing machine learning algorithms and efficient tools for data mining and data analysis. It supports various supervised and unsupervised learning algorithms. It includes other auxiliary functions that are an integral part of the ML pipeline, such as data preprocessing steps, data resampling techniques, evaluation parameters, and search interfaces to tune/optimize performance of an algorithm [32].

• **Matplotlib:**
Matplotlib is a comprehensive library used to create static, animated and interactive visu-

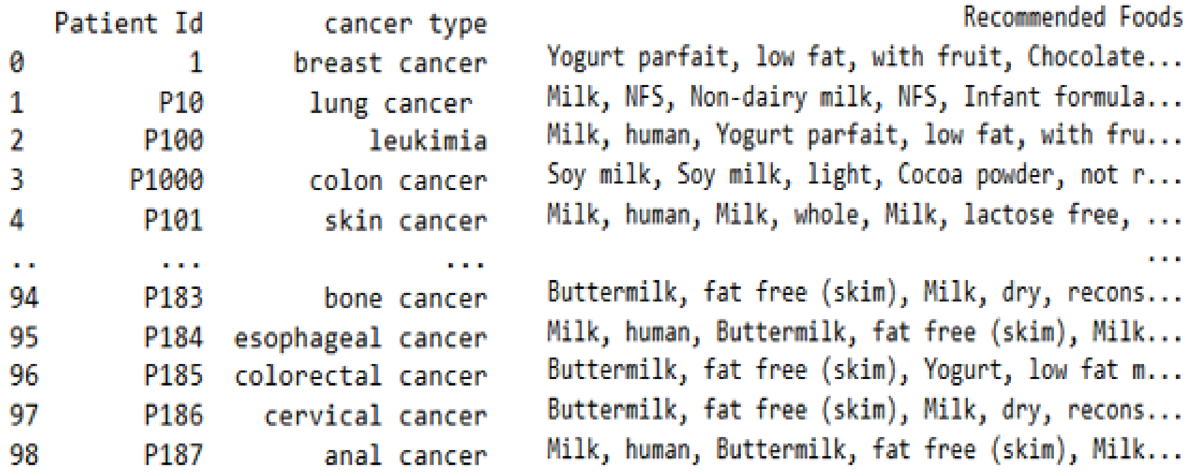
alizations in Python. Matplotlib makes easy things easy and hard things possible[33].

5.5 Results and Discussion

Our research results are a list of recommended foods for every cancer patient in the dataset, based on their nutritional needs.

This list is generated by calculating the similarity between the patient's needs (in terms of nutrients) and the quantity of these same nutrients in the foods in the food dataset. it is also based on the patients' feelings (i.e. we took into account counts patients' opinions on the foods they like or don't like) to optimize the list of recommendations.

The table below provides an overview of the dietary recommendations generated by our system for different patients with various cancer types. Each patient is identified by a unique Id and their cancer type is specified, followed by a list of recommended foods. These recommendations are based on an analysis of individual nutritional needs as well as dietary preferences and feelings toward foods that are expressed by patients.



```
   Patient Id         cancer type                                   Recommended Foods
0           1       breast cancer  Yogurt parfait, low fat, with fruit, Chocolate...
1         P10        lung cancer   Milk, NFS, Non-dairy milk, NFS, Infant formula...
2        P100            leukimia  Milk, human, Yogurt parfait, low fat, with fru...
3       P1000        colon cancer  Soy milk, Soy milk, light, Cocoa powder, not r...
4        P101         skin cancer  Milk, human, Milk, whole, Milk, lactose free, ...
..        ...                 ...                                                ...
94       P183         bone cancer  Buttermilk, fat free (skim), Milk, dry, recons...
95       P184   esophageal cancer  Milk, human, Buttermilk, fat free (skim), Milk...
96       P185   colorectal cancer  Buttermilk, fat free (skim), Yogurt, low fat m...
97       P186     cervical cancer  Buttermilk, fat free (skim), Milk, dry, recons...
98       P187         anal cancer  Milk, human, Buttermilk, fat free (skim), Milk...
```

Figure 5.2: Recommandations exemple.

5.6 Analysis of Recommendations

The dietary recommendations provided by the model show a wide variety of foods tailored to each patient's specific nutritional needs. Here are some key points to note:

1.Personalization of Recommendations:
− Recommended foods are personalized based on patients' cancer type and dietary preferences. For example, for a breast cancer patient, the model recommended low-fat yogurt with fruit, which is generally considered a healthy and beneficial option.

2. Food Diversity:
− Recommendation lists include a diversity of foods t. This diversity makes it possible to meet the different nutritional needs and dietary preferences of patients.

3. Conformity to Nutritional Needs:
− The recommended foods appear consistent with the typical nutritional needs of cancer patient.

5.7 Evaluation

After the recommendation of foods for cancer patients, an evaluation is necessary to determine the performance of the chosen approach. The model results were evaluated by analyzing a few criteria, namely the parameters of accuracy, precision and recall. Performance evaluation is a very necessary step to test model quality, to ensure the reliability of model results.

Evaluation metrics: The quality of RS algorithms can be evaluated using different types of measurement such as accuracy, precision ... The type of metrics used depends on the type of filtering technique.

In the following, we will define the metrics that we used to evaluate our system.
• **Accuracy:**

Accuracy is a metric commonly used in machine learning and statistics to evaluate the overall performance of a model. It measures the proportion of correct predictions made by the model among all the predictions it has made (the fraction of correct recommendations out of the total possible recommendations)[28]. The formula for accuracy is:

Accuracy=(True Positive+True Negative) / (TP+ TN+FP+FN)

Here is a breakdown of the components involved in this formula:

1. True Positives (TP): means the instances that were correctly predicted as positive by the model.
2. True Negatives (TN): are the instances that were correctly predicted as negative.
3. False Positive (FP): The model predicts positive but the observation is actually negative.
4. False Negative (FN): The model predicts negative but the observation is actually positive.
Accuracy provides an overall assessment of how well a model is performing in terms of making correctly both positive and negative predictions.

- **Precision:**

Precision(also known as positive predictive value) is a fundamental metric used in various fields to evaluate the performance of constructed models. This is the ratio of correctly predicted positive observations to the total predicted positive observations [35];[36].
Below is the formula to calculate the precision:

Precision = True Positives/False Positives + True Positives

Precision focuses on the quality of positive predictions made by the constructed model. It quantifies the model's ability to avoid making false positive predictions. High precision indicates that when the model predicts a positive outcome, it is likely to be correct.

- **Recall:**

Recall (also called sensitivity or true positive rate)Recall measures the model's ability to correctly identify all relevant instances in the dataset [35];[36].
It is the ratio of correctly predicted positive observations to all the observations in the actual class. It is calculated as follow:

Recall = True Positives / True Positives + False Negative

- **F1 Score:**

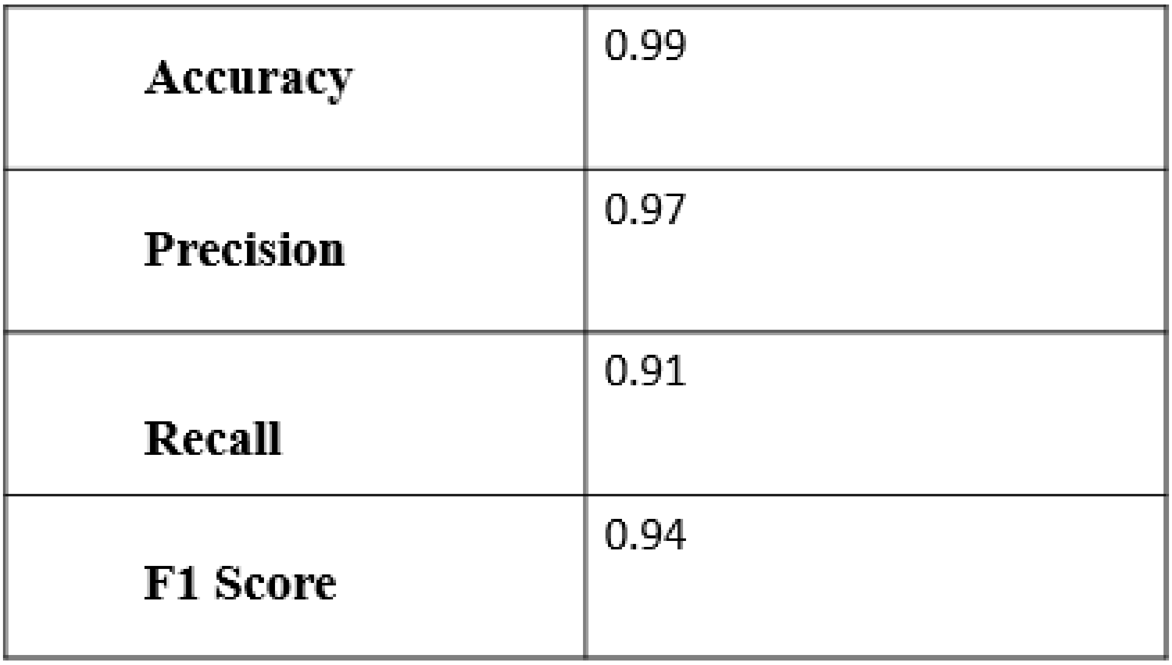
The F1 score is the harmonic mean of precision and recall. It gives a balanced measure of the two metrics and is particularly useful when you need to balance precision and recall [34].

Below the formula to calculate it:

$$F1Score = 2Precision \times Recall/Precision + Recall$$

The results of this metrics in our model are presented in the figure below:

○ Accuracy= 0.99



Accuracy	0.99
Precision	0.97
Recall	0.91
F1 Score	0.94

Figure 5.3: Evaluation metrics.

⋆ The accuracy is very high, at 99%. This means that 99% of all predictions made by the model (both positive and negative) are correct.

○ Precision= 0.97

⋆ The 97% precision indicates that of all instances classified as positive by the model, 97% were actually positive.

○ Recall (Sensitivity) = 0.91

⋆ The recall of 91% means that among all the positive instances in the dataset, the model correctly identified 91%.

○ F1 Score= 0.94

⋆ The F1 Score, which is the harmonic average of precision and recall, is 94%. This shows a good balance between precision and recall.

⋆ A high F1 Score means that the model maintains a good balance between making few false positive and false negative errors. This makes it an ideal metric for evaluating overall model performance when precision and recall are both important.

Therefore, the results obtained show an exceptional performance of the model with high metrics in all aspects. By combining high precision and recall, the model demonstrates its ability to provide precise and relevant recommendations. The figure below shows the visualization of these metrics with python:

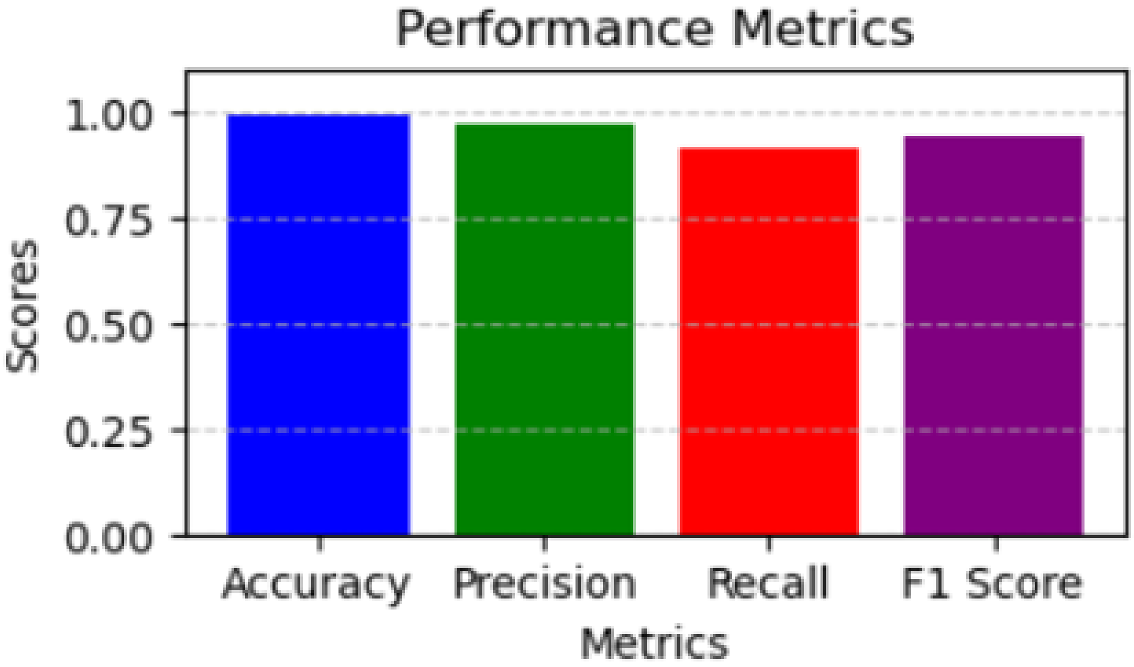


Figure 5.4: Metrics visualization.

5.8 Conclusion

In this chapter, we presented the essence of our work, which consists of creating a food recommendation system for cancer patients taking into account their emotional state towards foods. We presented the results obtained by our food recommendation system for cancer patients as well as the evaluations of the proposed approach with definitions of the Python libraries used and the evaluation metrics.
In the following chapter, we will finish our work with a general conclusion by generally summarizing our study.

6
General conclusion

This dissertation serves as an investigation into the critical field of recommending nutritional needs to cancer patients. Cancer, as a complex disease and one of the current leading cause of morbidity and mortality worldwide, presents complex challenges that extend beyond the physiological effects of the disease itself. Malnutrition is a common issue among cancer patients, often exacerbated by the side effects of treatments such as chemotherapy and radiotherapy. These nutritional deficiencies can have a significant impact treatment outcomes and patients' quality of life. Thus, a personalized diet, which considers the patient's emotional state and dietary preferences and needs, can greatly enhance the effectiveness of cancer treatment.

The main aim of this work is to provide a healthy and nutritious meal suggestions for cancer patients. Therefore, a hybrid food recommendation system which takes into account the patient nutrient needs and his sentiments toward some foods is proposed. The approach consisted of a content-based system that filtered the recipes according to user needs and returned a set of recipes, a supervised Machine Learning (Random Forest Classifier) for predicting whether a food is recommended for a patient, incorporating nutritional needs, sentiment scores and a rule-based sentiment analysis approach which consists on identifying sentiment from text reviews based on keyword presence to determine disliked and liked foods.

Due to the limited knowledge sources, the proposed approach may still be incomplete and inaccurate. Thus, our future perspectives are to improve our approach by considering more information about the users such as the impacts of different cancer stages and and on diet recommendation as well as using a large food dataset in order to improve the performance of the system. This will help to provide users with more comprehensive and healthier recipe recommendations tailored to their needs.

Carrying out this work allowed us to further enrich our capabilities, and to better understand the functioning and importance of recommendation systems. Through this study we were able to learn multiple pieces of knowledge. Thus, we consider this work the interesting start of our future challenges.

References:

[1] Ricci, F., Rokach, L., & Shapira, B. (2010). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35). Boston, MA: springer US.

[2] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, *12*, 331-370.

[3] https://www.nitorinfotech.com/blog/top-4-types-of-sentiment-analysis/

[4] Katrekar, A., & AVP, B. D. A. (2005). An introduction to sentiment analysis. *GlobalLogic Inc*, 1-6.

[5] https://monkeylearn.com/sentiment-analysis/

[6] Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems*, *111*, 180-192..

[7] Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in artificial intelligence*, *2009*(1), 421425. [8] Raguvaran, S., Anandamurugan, S., & Zubair Rahman, A. M. J. (2023). Harnessing LSTM Classifier to Suggest Nutrition Diet for Cancer Patients. *Intelligent Automation & Soft Computing*, *35*(2).

[8] Raguvaran, S., Anandamurugan, S., & Zubair Rahman, A. M. J. (2023). Harnessing LSTM Classifier to Suggest Nutrition Diet for Cancer Patients. *Intelligent Automation & Soft Computing*, *35*(2).

[9] Ahmad, M., Khan, A. U., & Sajid, M. (2023). A Diet Recommendation System for Persons with Special Dietary Requirements. *Journal of Computing & Biomedical Informatics*, *5*(01), 153-164.

[10] Lambay, M. A., & Mohideen, S. P. (2022). A Hybrid Approach Based Diet Recommendation System using ML and Big Data Analytics..

[11] Pawar, R., Gupta, S., Arora, H., Mehta, J., & Patil, A. (2021). Hybrid Food Recommender System based on cancer beating score of ingredients. *Rajendra Pawar, et. al. International Journal of Engineering Research and Applications*, 31-34.

[12Raguvaran, S., Anandamurugan, S., Anitha, E., & Rajakumareswaran, V. (2022). Nutrition-rich Food Suggestion for Cancer Patient using CapsNet. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(4), 443-448.

[13] Rehman, F., Khalid, O., Bilal, K., & Madani, S. A. (2017). Diet-right: A smart food recommendation system. *KSII Transactions on Internet and Information Systems (TIIS)*, *11*(6), 2910-2925.

[14] Rostami, M., Farrahi, V., Ahmadian, S., Jalali, S. M. J., & Oussalah, M. (2023). A novel healthy and time-aware food recommender system using attributed community detection. *Expert Systems with Applications*, *221*, 119719.

[15] Stefanidis, K., Tsatsou, D., Konstantinidis, D., Gymnopoulos, L., Daras, P., Wilson-Barnes, S., ... & Dimitropoulos, K. (2022). PROTEIN AI advisor: a knowledge-based recommendation framework using expert-validated meals for healthy diets. *Nutrients*, *14*(20), 4435..

[16] Tang, J., Huang, B., & Xie, M. (2023). Anticancer Recipe Recommendation Based on Cancer Dietary Knowledge Graph. *European Journal of Cancer Care*, *2023*(1), 8816960.

[17] Thongsri, N., Warintarawej, P., Chotkaew, S., & Saetang, W. (2022). Implementation of a personalized food recommendation system based on collaborative filtering and knapsack method. *Int. J. Electr. Comput. Eng*, *12*(1), 630-638.

[18] Toledo, R. Y., Alzahrani, A. A., & Martinez, L. (2019). A food recommender system considering nutritional information and user preferences. *IEEE Access*, *7*, 96695-96711.

[19] Zioutos, K., Kondylakis, H., & Stefanidis, K. (2023). Healthy Personalized Recipe Recommendations for Weekly Meal Planning. *Computers*, *13*(1), 1..

[20] https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds-download-databases/

[21] https://www.sciencedirect.com/topics/engineering/data-preprocessing

[22] https://www.displayr.com/what-is-a-correlation-matrix/#:~:text=A%20correlation%20matrix%20is%20a,a%20diagnostic%20for%20advanced%20analyses.

[23] https://www.databricks.com/glossary/what-is-dataset

[24] https://www.salesforce.com/fr/resources/definition/dataset/

[25] Rolon-Mérette, D., Ross, M., Rolon-Mérette, T., & Church, K. (2016). Introduction to Anaconda and Python: Installation and setup. *Quant. Methods Psychol*, *16*(5), S3-S11.

[26] Loria, S. (2018). textblob Documentation. *Release 0.15*, *2*(8), 269.

[27] Gujjar, J. P., & Kumar, H. P. (2021). Sentiment analysis: Textblob for decision making. *Int. J. Sci. Res. Eng. Trends*, *7*(2), 1097-1099.

[28] https://kobia.fr/classification-metrics-accuracy/#:~:text=L'accuracy%20est%20une%20m%C3%A9trique%20de%20performance%20qui%20%C3%A9value%20la,positifs%20et%20les%20individus%20n%C3%A9gatifs.

[29] Zastre, M. (2019, May). Jupyter notebook in CS1: An experience report. In *Proceedings of the Western Canadian Conference on Computing Education* (pp. 1-6).

[30] Python, W. (2021). Python. *Python releases for windows*, *24*..

[31] McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing*, *14*(9), 1-9.

[32] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

[33] Ari, N., & Ustazhanov, M. (2014, September). Matplotlib in python. In 2014 11th International Conference on Electronics, Computer and Computation (ICECCO) (pp. 1-6). IEEE.

[34] https://kobia.fr/classification-metrics-f1-score/

[35] https://builtin.com/data-science/precision-and-recall

[36] Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 79-91).

[37] Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, *16*(3), 261-273.

[38] Melville, P., & Sindhwani, V. (2010). Recommender systems. *Encyclopedia of machine learning*, *1*, 829-838.

[39] Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization* (pp. 325-341). Berlin, Heidelberg: Springer Berlin Heidelberg.

[40] Hameed, M. A., Al Jadaan, O., & Ramachandram, S. (2012). Collaborative filtering based recommendation system: A survey. *International Journal on Computer Science and Engineering*, *4*(5), 859.

[41] Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent data analysis*, *21*(6), 1487-1524.

[42] Sharma, L., & Gera, A. (2013). A survey of recommendation system: Research challenges. *International Journal of Engineering Trends and Technology (IJETT)*, *4*(5), 1989-1992.

[43] Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, *30*(4), 330-338.

[44] Tang, F., Fu, L., Yao, B., & Xu, W. (2019). Aspect based fine-grained sentiment analysis for online reviews. *Information Sciences*, *488*, 190-204.

[45] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731-5780.

[46] Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, *11*(1), 81.

[47] Behdenna, S., Barigou, F., & Belalem, G. (2018). Document level sentiment analysis: a survey. *EAI endorsed transactions on context-aware systems and applications*, *4*(13), e2-e2.

[48] Behdenna, S., Barigou, F., & Belalem, G. (2016). Sentiment analysis at document level. In *Smart Trends in Information Technology and Computer Communications: First International Conference, SmartCom 2016, Jaipur, India, August 6–7, 2016, Revised Selected Papers 1* (pp. 159-168). Springer Singapore.

[49] Schouten, K., & Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE transactions on knowledge and data engineering*, *28*(3), 813-830.

[50] Mohammad, S. M. (2017). Challenges in sentiment analysis. *A practical guide to sentiment analysis*, 61-83.

[51] Shi, Y., Larson, M., & Hanjalic, A. (2014). Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, *47*(1), 1-45.

[52] Dyhia, D., & Celia, M. (2019). *Definition d'un profil utilisateur pour un système de recommandation en recherche d'information* (Doctoral dissertation, Université Mouloud Mammeri).