ESCOLA
SUPERIOR
DE SAÚDE
POLITÉCNICO
DO PORTO

P.PORTO

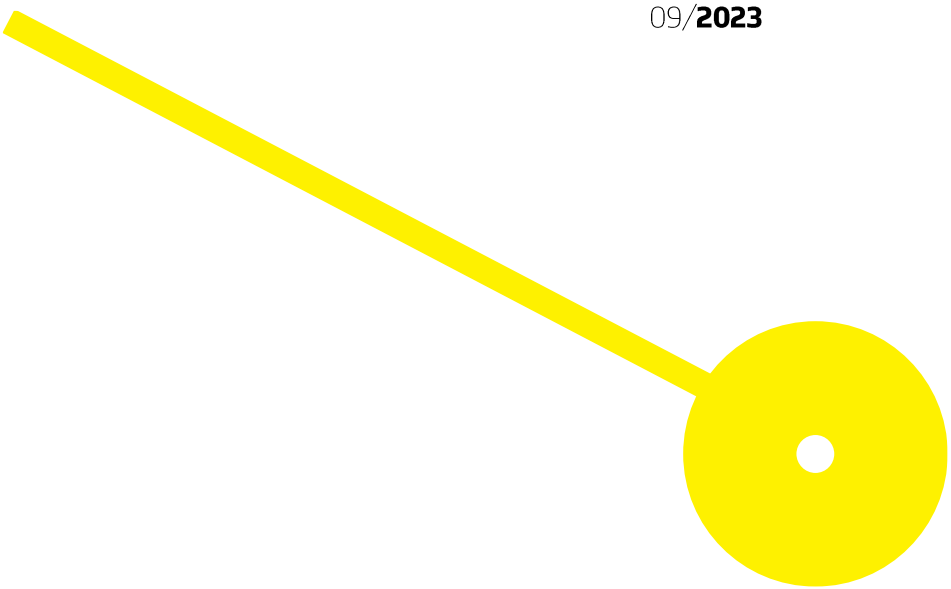
M

MESTRADO
EM BIOESTATÍSTICA E BIOINFORMÁTICA APLICADAS À SAÚDE

Application of Machine Learning Techniques for a Recommendation System in Pharmacy

Beatriz Freitas Torres

09/**2023**



ESCOLA
SUPERIOR
DE SAÚDE

Application of Machine Learning Techniques for a Recommendation System in Pharmacy

Autor

Beatriz Freitas Torres

Orientadores

Doutoramento/Alexandra Alves Oliveira/Escola Superior de Saúde - P.PORTO, LIACC e Retail Consult

Doutoramento/Brígida Mónica Faria/Escola Superior de Saúde - P.PORTO e LIACC

Doutoramento/Sandra Maria Ferreira Alves/Escola Superior de Saúde - P.PORTO

Dissertação apresentada para cumprimento dos requisitos necessários à obtenção do grau de Mestre em **Bioestatística e Bioinformática Aplicadas à Saúde** pela Escola Superior de Saúde do Instituto Politécnico do Porto.

Acknowledgements

First, I want to give special thanks to supervisor Alexandra Oliveira for her dedication, help, and patience along this journey. I would also like to thank my co-supervisors, Brígida Faria and Sandra Alves. Thank you for your guidance and support and for always being available to help. Your encouragement and motivation were essential during this period.

To my boyfriend, thanks for all the patience and encouragement through this phase. To my friends, those who helped with their knowledge and those who supported me and were always ready to encourage and advise me. I would also like to thank my family, without whose support this work would not have been possible.

To Retail Consult and everyone who made me feel very welcome, thank you for your support.

I want to express my gratitude to everyone who made this possible directly or indirectly.

Resumo

A Farmácia Comunitária tem um papel essencial na população, melhorando a qualidade de vida e minimizando os riscos associados à medicação. Em Portugal, as farmácias estão autorizadas a dispensar produtos sujeitos e não sujeitos a receita médica. Os profissionais de farmácia têm responsabilidade acrescida no aconselhamento de produtos não sujeitos a receita, com atenção às suas indicações, a possíveis interações e contraindicações. Assim, um sistema de recomendação com informação dos produtos auxiliaria o profissional no aconselhamento. Apesar de existirem alguns estudos acerca de sistemas de recomendação de medicamentos, a grande maioria é aplicada em âmbito hospitalar e a produtos sujeitos a receita médica.

Este trabalho tem como objetivos principais desenvolver uma *framework* conceptual de recomendação de produtos farmacêuticos e identificar grupos de produtos relevantes através das suas caraterísticas. Os objetivos específicos consistem em descrever sistemas de recomendação em farmácia, definir e comparar funções de distância para criar grupos clinicamente relevantes para o aconselhamento, aplicar técnicas de *machine learning* e compará-las e comunicar os resultados.

Para atingir estes objetivos foi analisado o processo do aconselhamento de produtos não sujeitos a receita médica. Foram selecionadas bases de dados públicas a incluir na *framework* conceptual e os dados obtidos foram processados. Obteve-se assim uma base de dados com 1426 produtos (medicamentos não sujeitos a receita médica, medicamentos homeopáticos e dermocosméticos) e a sua informação clínica e científica. De forma a identificar grupos de produtos relevantes foram aplicadas e avaliadas sete técnicas de *clustering* hierárquicas (*single linkage*, *complete linkage*, *average linkage*, *median linkage*, *centroid linkage*, e *ward linkage*) e não hierárquicas (K-means). Para determinar o número ideal de clusters em cada técnica e avaliar a sua validade foram utilizados dendrogramas, os índices de Calinski-Harabasz, de silhueta, de Davies-Bouldin e o método do ponto de inflexão. Para a definição da função de distância foi realizada uma consulta de especialistas de forma a que estivesse alinhada com o aconselhamento farmacêutico. Esta consulta permitiu identificar a importância das variáveis que compõem a função de distância. Os dados obtidos foram analisados no Microsoft Excel, SPSS e Python com as bibliotecas Pandas, Natural Language Toolkit (NLTK), Unidecode, Plotly, Matplotlib, NumPy, SciPy, Scikit-learn, usando o Spyder IDE.

Foram formados 22 grupos de produtos similares através do K-means, que foi a técnica mais eficaz na criação de grupos farmacologicamente homogéneos. No entanto, os grupos obtidos

não apresentaram relevância clínica suficiente para auxiliarem os profissionais no aconselhamento. Assim, foi definida uma nova função de distância enaltecendo a importância do grupo farmacoterapêutico do produto e alinhada com os resultados obtidos na consulta de especialistas. Foram formados 24 grupos de produtos similares através do K-means, que foi uma vez mais a técnica que apresentou grupos farmacologicamente homogéneos, baseados principalmente na possibilidade de uso durante a gravidez e amamentação e no grupo farmacoterapêutico. As restantes técnicas de *clustering*, técnicas não hierárquicas, não apresentaram grupos farmacologicamente homogéneos com nenhuma das funções de distância.

Palavras-chave: Sistema de recomendação, farmácia comunitária, produtos não sujeitos a receita médica, *clustering* de produtos, interações entre produtos

Abstract

Community Pharmacy (CP) plays a crucial role in the population, improving patients' quality of life and minimising medication risks. In Portugal, CPs dispense prescription and non-prescription products. Pharmacy professionals have an added responsibility when advising non-prescription products and should pay attention to self-medication and possible interactions. Therefore, a product recommendation system that incorporates relevant information about the products supports a more informed recommendation by the professional. Although there are a few studies in the area of medication RS, they are still scarce, and to the best of our knowledge, no medication RS is applied in community pharmacies in Portugal.

This work aims to develop a conceptual pharmaceutical product recommendation framework and identify relevant groups of products according to their characteristics and experts' opinions. The specific objectives consist of describing recommendation systems in pharmacy, defining and comparing distance functions capable of creating groups of similar and clinically relevant products for pharmaceutical counselling, applying machine learning techniques and comparing them, and communicating the results.

For this purpose, the background of pharmaceutical products counselling without a prescription was analysed. Public databases were selected to be included in the conceptual framework, and the data obtained was processed. Therefore, a database was obtained with 1426 products (over-the-counter medication, homoeopathic medication, and dermocosmetics) and their clinical and scientific information. In order to identify relevant groups of products, seven hierarchical (single linkage, complete linkage, average linkage, median linkage, centroid linkage, and ward linkage) and non-hierarchical (K-means) clustering techniques were applied and evaluated. Dendrograms, the Calinski-Harabasz score, silhouette score, Davies-Bouldin score and the inflexion point method were used to determine the ideal number of clusters for each technique and evaluate its validity. An experts consultation was performed to define a distance function aligned with pharmaceutical counselling. This consultation allowed the identification of the importance of the variables in the distance function definition. The resultant data was analysed in Microsoft Excel, SPSS and Python with the libraries Pandas, Natural Language Toolkit (NLTK), Unidecode, Plotly, Matplotlib, NumPy, SciPy, and Scikit-learn, using Spyder IDE.

Twenty-two groups of similar products were formed with K-means, the most effective clustering approach for forming pharmacologically homogeneous groups. However, the obtained clusters did not present enough clinical relevance to support professionals during counselling.

Consequently, a new distance function was defined, enhancing the importance of the pharmacotherapeutic group of the products and aligned with the results obtained in the experts' consultation. Twenty-four groups of similar products were formed with K-means, which was once again the technique that presented pharmacologically homogeneous groups, based mainly on safe use during pregnancy and breastfeeding and pharmacotherapeutic group. The remaining clustering techniques, non-hierarchical techniques, did not present pharmacologically homogeneous groups with any of the distance functions.

Keywords: Recommendation system; community pharmacy; non-prescription products; products clustering; products interactions.

Abbreviations and Acronyms

CP – Community Pharmacy

CNP – National Product Code

OTC – Over-the-counter

INN – International Nonproprietary Name

INFARMED – National Authority for Medicines and Health Products

DGAV – Direção-Geral da Alimentação e Veterinária

NPM – Non-Prescription Medication

SSRI – Selective Serotonin Reuptake Inhibitor

RS – Recommendation System

HRS – Health Recommendation System

SmPC – Summary of Product Characteristics

PL – Package Leaflet

WCSS – Within-Cluster Sum-of-Squares

MAH – Manufacturing Authorisation Holder

WHO – World Health Organization

ATC – Anatomical Therapeutic Chemical Code

SS – Silhouette Score

DBS – Davies-Bouldin Score

CHS – Calinski-Harabasz Score

Table of Contents

Figures

Tables

1. Introduction

Community Pharmacy (CP) in Portugal has a crucial role due to its proximity to the population. Its objective is to promote the patient's quality of life and reduce the risk associated with medication by promoting its rational use to minimise morbidity and mortality associated with them (1). This objective is fulfilled through dispensing health products and pharmaceutical advice (1).

Human resources at CP are pharmacy technicians, pharmacists and warehouse assistants. Pharmacy professionals must have a high level of knowledge, to correctly dispense medication to patients and provide all the information necessary for its rational use, reducing possible risks associated with it.

The pharmacy, in general, is divided into a public area, management office, order reception area, warehouse, laboratory and a private room for when the patient needs more privacy and to measure biochemical parameters (e.g., blood pressure, glucose, cholesterol, triglycerides) (1). Stock management depends on various factors such as demand, pharmacy location, storage space, season, and campaigns. For example, in summer, the stock of sunscreen will be much higher than in winter. Stock management is essential so there is no excess stock; however, it is also important that the pharmacy stock is sufficient for patients' needs, avoiding stock-outs. Suppliers usually deliver twice a day to guarantee all orders. The reception of orders is registered in the computer system, and the products' barcodes, CNP (National Product Code) or QR Codes are read, registering the respective quantities, expiry dates, and prices.

There is a national database with all the medicines with valid marketing authorisation in Portugal, which is maintained and updated by INFARMED, the National Authority for Medicines and Health Products in Portugal that regulates medication for human use and health products (2).

According to the *Regime Jurídico dos Medicamentos de Uso Humano*, medicines are classified, in terms of dispensing to the public, into (3):

- Medicines subject to medical prescription;
- Medicines not subject to medical prescription (over-the-counter medicines).

According to *Artigo 114º* presented in the same legislation, medicines are classified as subject to medical prescription if any of the following conditions are applicable:

- They are likely to present a danger, directly or indirectly, even when used correctly, if utilised without medical supervision;

- They are likely to present a danger, directly or indirectly, when used frequently in considerable quantities for purposes other than that for which they are intended;
- Contain substances or preparations based on these substances, whose activity or adverse reactions require further investigation;
- Intended to be administered parenterally.

When medicines do not meet these conditions, they are classified as not subject to medical prescription (3). They can be dispensed in pharmacies and parapharmacies in Portugal. Pharmacy-restricted non-prescription medicines are a sub-category of this classification that only authorises the dispensation of list non-prescription medicines exclusively in a pharmacy with established dispensing protocols (4). Some examples are ulipristal, an emergency contraceptive, several associations of acetaminophen with other active substances for migraines and pain, and ibuprofen at doses greater than 400 mg inclusive (4). In these situations, there is a need for a pharmacy professional's intervention.

There are also some exceptions, as the same International Nonproprietary Name (INN) can be dispensed as a non-prescription medication, or it can be needed to present a prescription for it differing on the dosage (5). For example, acetaminophen pills are commercialised in Portugal with two different dosages, 500 mg and 1000 mg; the first can be dispensed without a prescription, although obtaining the second requires a prescription (6).

The focus of this dissertation will be on the category of non-prescription products and their recommendation.

1.1. Non-Prescription Medical Products

In Portugal, pharmacies are authorised to dispense the following products without prescription (7):

- Over-the-counter (OTC) medication;
- Homoeopathic medication;
- Veterinary products and medication;
- Medical devices;
- Phytotherapeutic products;
- Dietary supplements;
- Dermocosmetics;

– Childcare products.

The pharmacy is usually the first place many patients seek help, which can be explained due to the fact that they have easier access to a pharmacy instead of an appointment with the doctor. Therefore the professionals have an added responsibility to refer the person for an appointment, with the doctor, if necessary or to advise one of these products as they must assess the situation and present a solution. When advising a product, they should consider possible interactions, especially with polymedicated patients.
Purchasing these products without a prescription has many benefits. However, there are also concerns about their availability (Table 1).

Table 1 *– Potential benefits and concerns of OTC (8).*

Potential benefits	Potential concerns
Increased and fast access to effective drugs	Inaccurate self-diagnosis
Fewer visits to physicians	Delay in obtaining needed therapy
Lower health care costs	Risks associated with inappropriate drug use
Improved autonomy and education of patients	The diminished role of professionals in supervising care

These products are widely used, as demonstrated by studies already carried out. A study in Spain indicates that 78.9% of adults use OTC medication (9). According to WHO, approximately 80% of people worldwide also rely on phytotherapeutic products and food supplements as therapy (10). Despite the prevalent use of OTC and dietary supplements, many consumers must be made aware of the effect on their health or the possible interactions (11).

In Portugal, the purchase of products without a medical prescription also has a high expression and a slight growth in the last years. Figures 1 and 2 demonstrate that each year there is a slight increase in the number of products sold and the percentage of these products sold in the total market. This product segment constitutes more than 30% of the total market. This high value demonstrates a demand from users for these products, therefore, advice and rational use are essential.

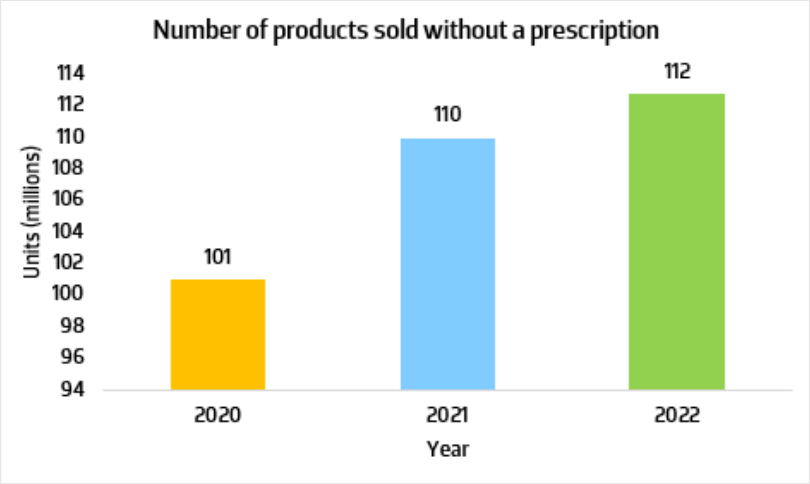


***Figure 1** – Number of products sold without a prescription* (99).

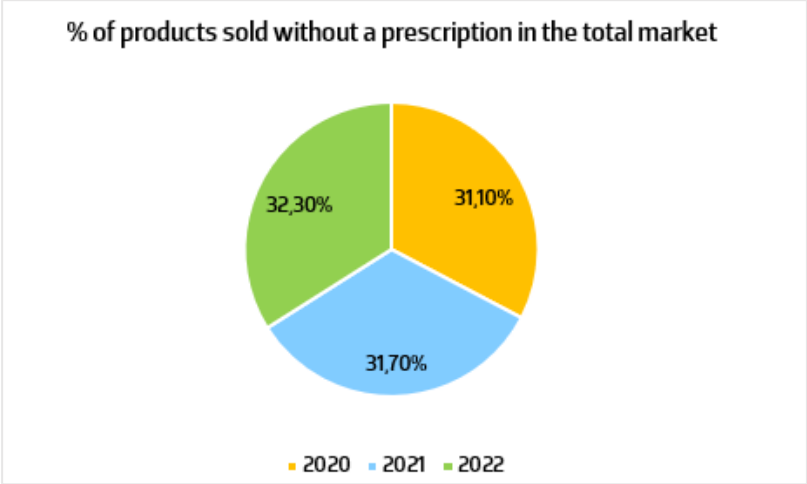


***Figure 2** – Percentage of products sold without a prescription* (99).

This dissertation will focus on OTC, homoeopathic medicines and dermocosmetics. In Portugal, OTC, homoeopathic medicines and dermocosmetics are regulated by INFARMED, whereas dietary supplements and phytotherapeutic products are regulated by DGAV (*Direção-Geral da Alimentação e Veterinária*) (2,12).

1.1.1. Over-The-Counter Medication

OTC medication or non-prescription medication (NPM) can be dispensed without a prescription in pharmacies and parapharmacies (13). These drugs are intended for short treatments for minor health conditions (1). They can be displayed in the pharmacy where the customers can see them. However, they cannot be within reach of the customers as there is the need to evaluate the situation by the pharmacy professional (14).

1.1.2. Dermocosmetics

Dermocosmetic products are defined as "any substance or preparation intended to be placed in contact with the various superficial parts of the human body, namely, the epidermis, hair and hair system, nails, lips and external genital organs, or with teeth and oral mucosa, with the purpose of, exclusively or mainly, cleaning them, perfuming them, modifying their appearance, protecting, keeping them in good condition or correcting body odours" (15).

1.1.3. Phytotherapeutic Products and Dietary Supplements

Phytotherapeutic products are defined as "any medicine that has exclusively as active substances one or more plant-derived substances" (16).
Dietary supplements are intended to supplement the diet and should never replace a varied diet, and they can be classified into three large groups (17):

- Vitamins and minerals (vitamin D, Calcium,...)
- Plants and botanical extracts (ginkgo Biloba,...)
- Other substances include fibre, probiotics, essential fatty acids, amino acids, and enzymes.

The use of dietary supplements has been the object of substantial growth internationally, especially in developed countries (18). These products exist for various situations, such as digestive problems, hypercholesterolemia, excess weight, difficulty falling asleep, strengthening the immune system, and fatigue, among others. They can also complement some medication treatments (e.g., urinary infections). On the other hand, these supplements also have adverse effects, and professionals have a significant role in advising them (19). These products can be displayed in a part of the pharmacy where customers have access (14).

1.1.4. Homoeopathic Medicine

Homoeopathic medicine is defined as "medicine obtained from substances called stocks or homoeopathic raw materials, according to a manufacturing process described in the European Pharmacopoeia or, failing that, in a pharmacopoeia officially used in a Member State, and which may contain several principles" (16).

1.2. Self-Medication, Drug Interactions and Adverse Effects

INFARMED defines self-medication as "the responsible use of NPM whenever it is intended for the relief and treatment of transient and minor health complaints, with the optional assistance or advice of a health professional" (20). It also indicates the clinical situations where self-medication can be distributed for ten systems (digestive, respiratory, cutaneous, nervous, medullary/bone, general, ocular, gynaecological, vascular, and general) (20). All the activities people perform to improve their health, prevent or reduce illness, as well as restore their health after an illness, including after hospital discharge, are considered self-medication (21). Using medicines subject to medical prescription is considered self-medication, as well as using OTC, phytotherapeutic products, or dietary supplements (14).

Although this practice has benefits, when used as recommended in the instructions, such as increased access to healthcare and improved patient autonomy, when the medication is not used rationally, it has various health risks too (22,23). According to the previous authors, the risks are inaccurate self-diagnosis and, therefore, a delay in obtaining adequate therapy, risk of drug interactions and side effects, misuse of medication, and it can lead to drug dependence and abuse. Some examples of drug misuse are the abuse of laxatives to lose weight, the use of antihistamines to sedate children or elderly people and as a medicine for insomnia, the use of superior dosage for faster relief of symptoms or a more extended use than recommended (22).

A study in Portugal estimate the prevalence of self-medication at 26,2% for urban populations and 21,5% for rural populations (24). It is more common to self-medicate between the ages of 10 to 49 years old, and when faced with a mild or moderate health problem, 46% of respondents said they consulted a doctor and 28% a pharmacist (24). According to the same study, 50% of respondents claimed to purchase medication for self-medication on the advice of a pharmacist, 30% on their initiative, 18% at the suggestion of friends or family and about 1% at the recommendation of a nurse. A second study obtained a prevalence of self-medication of 86,3%, a higher number than the previous studies (25). These results can be explained by the methods used since this study considered self-medication at any moment in the past; on the contrary, the studies conducted by Martins et al. (24) only considered that specific visit to the pharmacy and whether the patient acquired a product for self-medication. A third study was conducted later and obtained a higher prevalence (91,3%) through questionaries in pharmacies, and the respondents considered the pharmacy professionals the most credible source of medication (26). A more

recent study from another Portuguese city demonstrated that 40,7% of the population resorted to self-medication (14).

Drug interactions are defined by the modification of the pharmacological activity of a drug due to the concomitant or consecutive use of another drug, thus presenting a different activity other than the expected when administering the drugs alone (27,28). There are several risk factors related to the patient for the occurrence of these interactions, such as genetic characteristics, age, presence of diseases and most relevant, polypharmacy (27). Interactions can have positive consequences, in the case of reduced drug toxicity or synergy, thus increasing therapeutic effects; however, it can also lead to negative consequences, such as increased adverse reactions (28). They can be classified, according to severity, as mild in the case of a slight increase of the adverse effects without the need for replacement of therapy, as moderate if it causes an exacerbation of the disease and replacement of therapy is necessary, or as severe if medical treatment is necessary (29).

Drug interactions are classified as pharmacodynamic, pharmacokinetic, or pharmaceutical incompatibility (Table 2).

Table 2 *– Drug interaction classification* (27,28,30).

Pharmacodynamic interactions	Occurs between drugs with similar or opposite pharmacological effects resulting in modification of the pharmacological response through synergy or antagonism. They are, in general, predictable through knowledge of the mechanism of action of the drugs.
Pharmacokinetic interactions	Occurs when a drug alters the absorption, distribution, metabolism or excretion of another drug, resulting in a decrease or increase of its plasmatic concentration and, consequently, its pharmacological effect. These interactions are difficult to predict.
Pharmaceutical incompatibility	Chemical or physical incompatibility between the drug constituents.

NPMs are subjected to these interactions with each other and medicines which are subject to medical prescription. Some examples of interactions are: ginkgo biloba, used for cognitive function, increases the risk of haemorrhage if taken concomitantly with anticoagulants by potentiating their effect (31); St John's wort (*Hypericum perforatum*), used for depressive symptoms, decreases the plasma concentrations of antidepressants and warfarin

(anticoagulant) thus making clot formation more likely and potentiates the effect of selective serotonin reuptake inhibitors (SSRIs) that may cause serotonergic syndrome (32). There are also interactions between food and medication, such as grapefruit which interacts with almost 100 different medications, and about half can lead to severe adverse reactions (33).
Patients often do not inform their physician that they are taking NPMs, which is very important, mainly due to possible drug interactions. This fact is supported by studies that show that this happens in about 46% of cases and, more specifically, 61,7% in a study conducted in Portugal (34,35).

Adverse effects are defined as a harmful and unintended response to a drug at doses typically used in humans (28). They can range from minor problems to life-threatening events. They can occur when starting therapy with a new medicine or dietary supplement, when stopping a medication after long-term use or after the increase or decrease of the dose (36). Reactions can be classified according to their characteristics and the type of effect (Table 3).

Table 3 – Adverse effects classification (37,38).

Type of reaction	Type of effect	Characteristics	Frequency	Examples
A	Augmented	Dose-related Predictable Low mortality Related to the pharmacological action	Common	Bleeding with warfarin
B	Bizarre	Non-dose related Unpredictable High mortality Unrelated to the pharmacological action	Uncommon	Anaphylaxis
C	Chronic	Cumulative dose related Time-related	Uncommon	Adrenal suppression by corticosteroids
D	Delayed	Time-related Often dose related Occurs after some time after the use	Uncommon	Carcinogenesis
E	End of use	Occurs after drug withdrawal	Uncommon	Opioid withdrawal syndrome
F	Failure	Dose-related Unexpected failure of therapy Drugs interactions related	Common	Resistance to antimicrobial agents

Non-pharmacological measures can reduce adverse reactions, such as eating food before taking the medication to reduce the likelihood of nausea and vomiting and having a specific time interval between taking two medications, and between others (36).

When dispensing the products, it is up to the health professional, including pharmacy professionals, to educate the patient on the risks of self-medication, interactions and adverse reactions so that the patient is guided towards the rational use of the medication and can benefit from it, avoiding the risks.

1.3. Pharmacy Products Advisement/Business Rules

More and more patients are seeking help and advice from pharmacy professionals first, since it is not necessary to make an appointment to use this service (19). Thus, the professionals must have adequate training and reliable information sources. Professionals must carry out counselling with all the information necessary for adequate advice on dispensing to improve patients' quality of life and prevent possible negative results associated with medication (1).

To guarantee the best possible service to the patient, the pharmacy professional must adapt his posture and language to the patient in question to ensure that the transmitted information is well received and assimilated. Therefore, it is essential to establish simple and clear communication, and non-verbal communication is essential. For example, maintaining eye contact with the patient and having an appropriate posture demonstrates interest and receptivity to the patient and confidence.

It is necessary to invest more and more in the training of pharmacy professionals as they are the only health professionals in contact with the customers when they are searching for non-prescription products so that they can provide reliable information about them, including their risks and cautions to promote a safer use (34,39). The professional must be able to prevent, identify and solve problems and provide reliable information adequate to the patient's ability to understand so that he can use it appropriately and safely. Pharmacies should have dispensing procedures for standard product advice (40). When the patient goes to the pharmacy, he exposes his problem according to the questions that will be asked by the professional. For evaluation and decision about the therapy (providing an NPM or advising the patient to go to the doctor) for a specific problem presented by the patient, the professional must have information about (40):

- What is the existing problem;
- What are the symptoms, and how long have they been present (to determine the product's active substance according to its indication and pharmacotherapeutic group);
- If medication has already been given to solve the problem;
- What medication does the patient take for other pathologies;
- If the patient has any conditions at the moment that can make a product contraindicated (such as being pregnant or breastfeeding).

After the decision on the therapy with NPMs, there should be a conversation with the patient, and the following information should be addressed (40):

- How to take the medicine: dosage (depends on the symptoms and patient's age), how (pharmaceutical form), when;
- Duration of treatment;
- The most relevant contraindications, interactions, adverse effects, warnings and precautions;
- If applicable, what non-pharmacological measures can they adopt to complement the treatment;
- If the product requires special storage conditions, such as refrigeration;
- Place all the information in writing.

Regarding dermocosmetics, professionals must advise them according to the patient's characteristics and what he aims for using the product. Some important points to mention during the counselling, whether for aesthetic purposes or any pathologies, are how to use, duration, information about active ingredients, and the relevant actions to the patient. Due to the variety of dermocosmetics products and the rapid evolution, it is necessary for professionals to keep up to date with what exists in the market and to be able to contextualise this knowledge in practice.

1.4. Software Available in Community Pharmacy

There is some software available for CP in Portugal, and it is a valuable and indispensable tool since the entire process, from the reception of the products to the dispensing to the patients, is registered in this system. By computerising the procedures, it is possible to reduce human error, focus attention on the patient, and verify scientific information on medicines to facilitate processes and generate good counselling for the patients. With the use of technology, the pharmacy's productivity has vastly increased (41).

An example of such software is Sifarma®, implemented in about 90% of the pharmacies in Portugal (42). It is a management and service tool for CP. It assists in management processes such as stock and order management and customer service as it has scientific information that ensures the attention is focused on the user (42). When selecting a product, it is possible to check the scientific information in the software, including the therapeutic indication, which pharmacotherapeutic group it belongs to, dosage, interactions, and composition, among others. Users are registered, and new user records are created with their personal data, including their identification, date of birth, gender, address, and contact. In the user's record, it is possible to check

their purchase history. Despite having many advantages, it has some limitations since information regarding the characteristics of products is present mainly for medicines subject to medical prescription. Some OTC and dietary supplements also have this information. However, it is scarcer, and dermocosmetics may only present its list of ingredients. Another example of software used in CP is Spharm, with a minor expression than Sifarma ® (43). It has a similar operation. Although it does not have information about the medicines subject to medical prescription, it redirects to the Summary of Product Characteristics of the product and it does not present the information on the software. It does not have any information about OTC, dietary supplements and dermocosmetics.

These are general systems and are not aligned with individualised healthcare (44). One way to provide individualised healthcare is through patient-specific recommendations of products based on the individuals' characteristics (health status, objectives) and products. So, a product recommendation system that can incorporate relevant information about the product's characteristics (such as therapeutic indication, dosage, side effects and others) and that can identify possible interactions among products, as well as can classify them into similar groups could support a more informed recommendation by the healthcare provider.

1.5. Objectives

The main objectives of this dissertation are to develop a conceptual pharmaceutical product recommendation framework and identify relevant groups of non-prescription products (OTC medication, homoeopathic medication and dermocosmetics) according to their characteristics and experts' opinions. The specific objectives consist of describing recommendation systems in pharmacy, defining and comparing distance functions capable of creating groups of similar and clinically relevant products for pharmaceutical counselling, applying machine learning techniques and comparing them, and communicating the results.

1.6. Document Structure

In order to propose a solution to the problem described in Chapter 1, the state of the art was reviewed and the background of recommendation systems was presented in Chapter 2. In Section 2.1, recommendation systems are presented, the purpose of their use, the important aspects, and possible limitations. Section 2.2 describes the various recommendation system approaches and presents the advantages and limitations of each approach. Section 2.3 describes the clustering technique, the distance matrix, the calculations to obtain it, and the evaluation metrics, and Section 2.4 contains the state of the art of recommendation systems in pharmacy.

Chapter 3 describes the Methodology. Section 3.1 explains how pharmaceutical counselling of non-prescription products is conducted to understand the context better. In Section 3.2, it is presented the database selection process. Section 3.3 describes the preparation performed on the database to clean and prepare the data for the clustering analysis. In Section 3.4, are explained the machine learning algorithms that were applied. Section 3.4.1 describes how to determine if there was any influence of the expert' persional characteristics in their opinion over the criteria. Section 3.4.2 describes how to determine the optimal number of clusters and its evaluation, and section 3.5 explains how the results will be disclosed to the community.

Chapter 4 presents and discusses the results. Section 4.1 describes and discusses the results of the conceptual pharmaceutical product recommendation framework. Section 4.2 describes and provided an overview of the final database and its variables. Section 4.3 describes the results obtained in the experts' consultation. Section 4.4 describes the distance function created, and section 4.4.1 discusses the recommendation groups that were formed. Section 4.5 describes a new the distance function enhancing the importance of the pharmacotherapeutic group, and section 4.5.1 discusses the new recommendation groups formed according to the new distance function. Section 4.6 presents a discussion and the dissertation contributions.

Chapter 5 contains the conclusions and future work.

2. Recommendation Systems: Background and State of the Art

Recommendation systems are presented and explored in this chapter. Their definition and classification, the purpose of their use and the important aspects are addressed, as well as the various approaches and the advantages and limitations of each approach. Clustering, distance matrix and evaluation metrics are also described in this chapter. This chapter also contains the state of the art of recommendation systems in pharmacy.

2.1. Background of Recommendation Systems: Definitions and Classification

A recommendation system (RS) is an information filtering system that can provide item recommendations to users according to their preferences (45,46). It aims to reduce the information and cognitive load on users and make the selection process faster for users and customers by suggesting the most suitable products (45,47).

Over the past decade, RS have gained popularity and have been applied in various domains such as e-commerce, online video streaming, social media and advertising (46,48). Online information has increased more and more, thus requiring a system that can filter a large amount of data, and RS has proven to be an effective strategy (46,49). The RSs provides the user with personalised information, reducing his search time (50). Nevertheless, their application in the health domain is still limited (51). Figure 3 shows the distribution of the papers on seven RS fields from a study conducted from 2010 to 2021 (52). The same study found that RS in the fields of social networks services, tourism, healthcare and education has increased in the last few years, and its research is expanding.

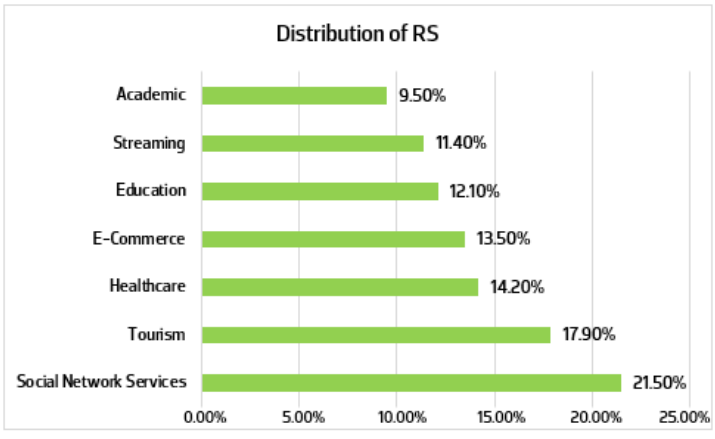


***Figure 3** – Distribution of recommendation systems (RS) fields from 2010 to 2021* (52).

Health Recommendation Systems (HRS) can support an early diagnosis, predict disease progression and optimise decision-making processes (53,54). Its objective is to improve people's

health by making personalised recommendations based on their specific information while reducing professionals' load due to the vast amount of information and products available (48,55). As it is automatic requires less expertise to use. However, it presents some challenges, especially while having sensitive information on the users, such as being the target of cyber-attacks and data exchange with web health applications (54). HRS are improving, and although there are still few, it is suggested that they can perform individualised recommendations based on the user data assisting professionals' decision-making (48).

A few aspects that need to be considered in HRS are usage context, users and items (56,57) (Table 4). The usage context describes the environment and can be divided into contextual factors and multifactorial goal setting that influence how the items are presented. The HRS can have two types of users: end-users and healthcare professionals. The items vary according to the categories of the HRS.

Table 4 *– Aspects of recommendation systems* (56,57).

	Aspect	Description
Usage context	Contextual factors	Dynamic attributes that can affect an activity (such as the optimal time to take the medication) and dynamic factors from the users (such as the emotional state).
	Multifactorial goal setting	Health goals do not follow a singular dimension, so different domain-specific criteria need to be on the item's evaluation (an option that is the "most healthy" for a user may be dangerous for another user).
Users	End-users	Healthy users or patients. There is a user profile with all relevant information.
	Healthcare professionals	Doctors, nurses, pharmacy professionals, and others.
Items		There are several categories, such as diets, physical activity/sports, recommended diagnoses, treatments/medications and medical information/sources.

Even though they present many advantages, RS also have some challenges, including cold start, data sparsity, shilling attacks and scalability problems (54,58–60).

According to the previous authors, cold starts can be divided into user cold starts and item cold starts. It occurs when a new user or product is available on the system. However, since the item

has no user interaction or reviews, the RS has insufficient information to perform and offer a personalised recommendation. The Bayes classifier is the most used to fight cold starts because it proves to be the most accurate when estimating a new item's characteristics.

The data sparsity results from the lack of or inadequate rating information. As is usually the case in a RS, a large number of items and users makes it challenging to obtain a rate of each item, producing a sparse matrix that will affect the algorithm's accuracy since most of the elements in this matrix will be zero. There are some models proposed to reduce this problem as matrix factorisation techniques and prediction methods.

A shilling attack occurs when there are fake ratings, comments and reviews to increase an item's popularity. Some algorithms are able to detect this through sentiment analysis and various statistical tests.

The scalability problems are due to the need to generate rapid results for a large-scale application as commerce constantly expands. Clustering techniques can be helpful because they reduce the data's sparsity and divide it into smaller portions, reducing the results generation speed.

Regarding HRS, there is a major concern with the user's privacy due to the sensitive information about their health and preferences (54). Patients can feel unsafe sharing their personal and medical information as they fear a leak of that information (57).

2.2. Background of Recommendation Systems: Approaches

RSs can be classified according to the approach they use to generate recommendations (Figure 4).

The first approach proposed and developed in the 1990s was collaborative filtering, and since then, RSs have been more studied and developed (52). Other commonly used approaches are content-based filtering, hybrid systems and knowledge-based filtering (55).

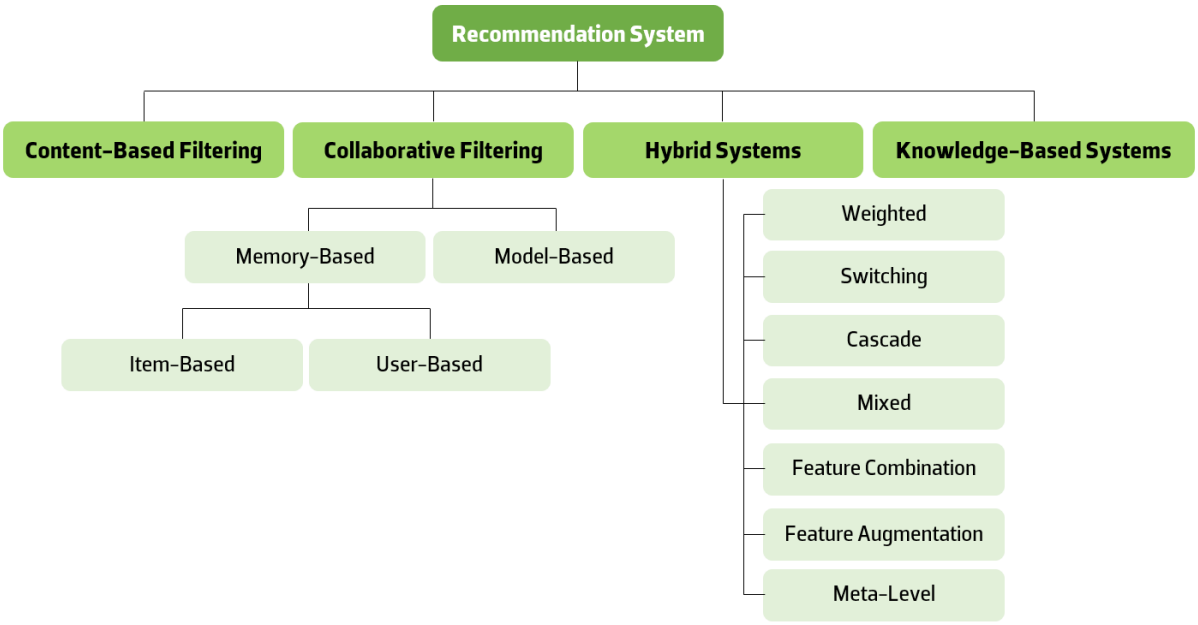


***Figure 4** – Classification of recommendation systems based on their approaches* (52,54,55).

2.2.1. Content-Based Filtering

Content-based filtering recommends similar items as the ones the user liked in the past according to the characteristics of the items (57). The item profile contains information characterising it (a set of features), and the RS recommends similar items to the ones already preferred by the user (45,55). For this approach to work, there is the need for similar products to be characterised similarly, so it is vital to have all the essential features of the items. The user profile has information on the user's preferences according to how the user rated items (50). There is no connection between users' profiles, it does not require any other user's data, and there is a connection only between items profiles and between users and respective items (59).

The recommendation occurs according to the similarity between items, which is expected to be high, and the user profile and its historical data, as demonstrated in Figure 5 (50). It uses text-mining methods, semantic analysis, neural networks, support vector machines and others (52). The similarity can be calculated by heuristic functions, such as the cosine similarity metric (58).

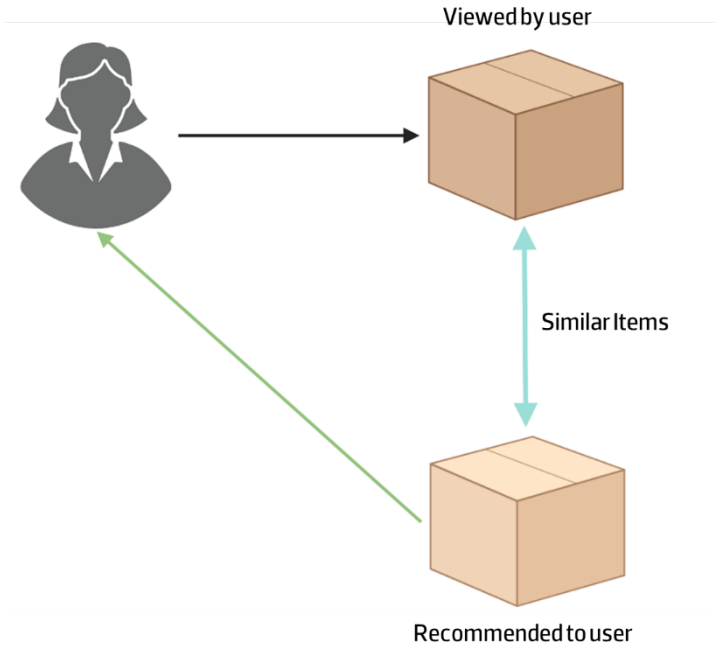


***Figure 5** – Content-based filtering.*

One example of this approach in HRSs is the recommendation of treatment for a patient's health condition that is similar to one already designed for a past treatment for that patient (57).

2.2.1. Collaborative Filtering

Collaborative filtering finds items based on past behaviour and examines similar users, also known as "people-to-people correlation", and is considered the most popular technique (58). With this previous information about the user, it identifies a neighbourhood of people that have exhibited similar behaviour and their rating activities and compares profile information to find items that will be liked by the user (Figure 6) (45,49). This approach is based on the idea that users will have similar interests now if they share the same interests in the past (57).

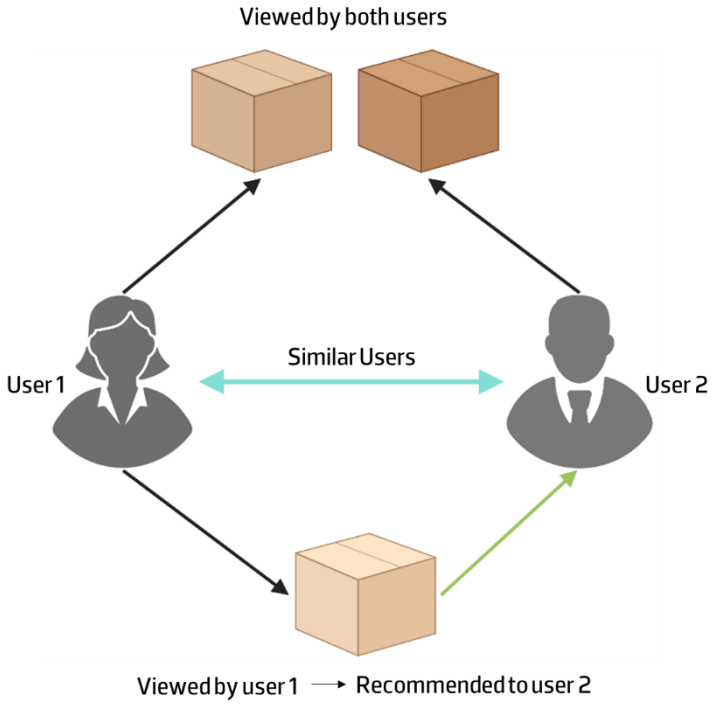


***Figure 6** – Collaborative-based filtering.*

This approach can be classified as memory-based and model-based. The memory-based approach uses a user-item matrix containing all the users and their respective ratings for the items (Figure 7) (50). Without evaluation, that cell stays empty, creating a sparse matrix (56). It can be categorised into user-based and item-based filtering, using similarity among users and items, respectively (59). User-based filtering consists of two stages. First, it locates similar users to the present user, then obtains the rates from those similar users and uses them to perform recommendations (58). The similarity can be calculated with various measures such as the mean-squared difference, Pearson correlation, Spearman correlation, cosine similarity and adjusted cosine similarity (58). The main difficulty of this technique is the amount of time consumed, and if there is not enough data, problems such as sparsity and cold start can occur (50,52). Due to these reasons, the model-based approach emerged. This model splits the rating matrix into the train data and the test data to train and evaluate the model, respectively, using data mining and machine learning techniques (50,59). The most used techniques for this are clustering, singular value decomposition, and principal component analysis (52). The accuracy of this model is dependent on the quality of the data (46).

In HRSs, in the context of a treatment's choice, this approach assumes that the treatment would be similar if the patients have similar health conditions (57).

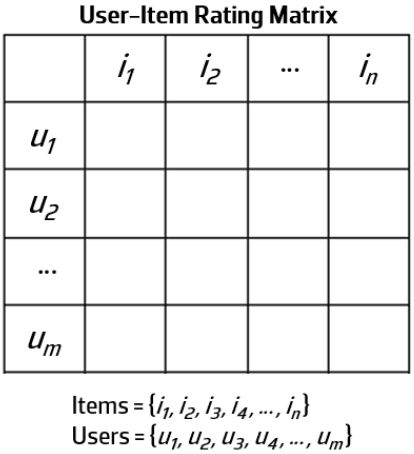


***Figure 7** – User–item rating matrix. Adapted from* (52).

2.2.2. Hybrid Recommendation Systems

The hybrid RS combines content–based filtering and collaborative filtering presented above to increase the accuracy of RS (54,55). Its main goal is to remove the limitations of individual techniques. It tries to use the advantages of the first technique to fix the disadvantages of the second that cannot recommend new items since they do not present ratings yet (cold start problem); nonetheless, content–based filtering can recommend these items since the prediction is based on the item's characteristics (45,57). This model can be divided into seven types depending on the method used to combine the filtering techniques: weighted hybridisation, switching hybridisation, cascade hybridisation, mixed hybridisation, feature combination, feature augmentation and meta–level (52). The weighted hybridisation collects the results of the combined approaches, and the weight is adjusted according to the degree to which the user's evaluation of the item and the RS's prediction (52,58). The switching hybridisation selects one of the approaches according to which one makes the best recommendation depending on the situation (58). The cascade method is a hierarchical process in which a weak approach with low priority improves the recommendation with a higher priority or more robust approach (58). When there is the need to obtain many recommendations simultaneously, mixed hybridisation is used (52). The feature combination has a main recommendation approach which can be added features by another approach for augmented data, whereas in the feature augmentation, a rating is generated, and those features are integrated into the following recommendation approach (52,58). In the meta–level method, the output from an approach is used as the other's input. It

differs from the feature augmentation since this one uses features as input on the second approach, and the meta-level method uses the entire model as input (58).

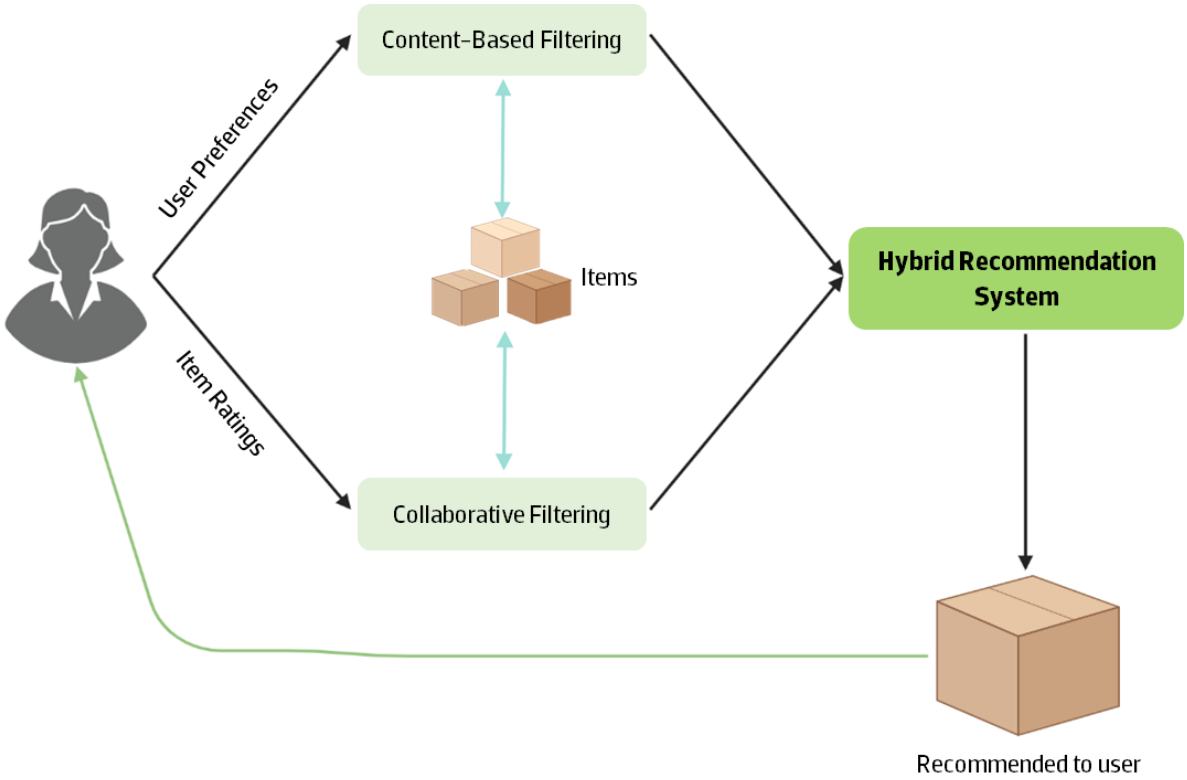


Figure 8 – Hybrid recommendation system. Adapted from (100).

2.2.3. Knowledge-Based Recommendation Systems

The knowledge-based RS finds items for the user by logical inferences (55). It is based on previously defined knowledge by experts, increasing the pertinence of the recommendations (60). This approach is valuable when a recommendation system needs to be implemented in fields where the knowledge of the other users is essential (60). In this case, the recommendation is not dependent on the users' ratings (58). The more relevant disadvantage is knowledge acquisition (constructing the rules and requirements needed) (58).

2.2.4. Advantages and Limitations of RS Approaches

All the RS approaches previously explained have advantages and limitations. Table 5 compiles that information.

***Table 5** – Advantages and Limitations of recommendation systems (RS) approaches* (52,54,58–60).

Approach	Advantages	Limitations
Content-Based Filtering	It does not experience item cold start. Users are independent of each other.	User cold start. When there is not enough information to differentiate products, the recommendations are not accurate. Overspecialisation.
Collaborative Filtering	Can recommend complex items. Helps users discover new interests. It does not require items and users analysis.	Sparsity, scalability and cold start problems. Shilling attack. Users are not independent of each other. If the neighbourhood is small, it has poor recommendations.
Hybrid RS	Combines approaches to fight their disadvantages. More accurate, effective and personalised recommendations than the others. Solves cold start and sparsity issues.	Needs more knowledge engineering. High computational complexity. Requires a large database.
Knowledge-Based RS	Minimises the cold start problem. It does not require an initial large amount of data. Independent of user's ratings, recommendations specific to the user according to the interests.	Suggestion ability is static. Requires a lot of domain knowledge.

Collaborative filtering suffers from item cold start and user cold start, however, content-based filtering only suffers from user cold start; therefore, it can recommend new items based on their information. The overspecialisation problem with content-based filtering is very important since it can only recommend items that are similar to the ones previously evaluated by the user. Consequently, there is no diversity in the items recommended, and it does not provide various

contents. On the other hand, collaborative filtering has limitations since the dataset only has a small number of the items rated (sparsity) and is not suitable for millions of users and items (scalability). The hybrid RS combines these approaches to make more accurate and personalised recommendations but has its limitations, having high computational complexity and requiring a large amount of data. The knowledge-based RS has as its main limitation the need for a lot of domain knowledge to provide accurate recommendations.
For example, content-based filtering can be found in Spotify and Internet Movie Database (IMDB). Amazon and YouTube are platforms that use collaborative filtering. Netflix uses a hybrid RS, recommending based on what similar users are watching (collaborative filtering) and on similar information between movies rated by the users (content-based filtering). Knowledge-based RSs are required when knowledge about the topic is essential for the recommendation, such as travel and education RSs.

According to a study performed between 2010 and 2020, collaborative filtering is the most used recommendation approach, representing 41.6% of the papers collected during that time (52). Nevertheless, this approach has decreased since 2014, and the hybrid RS has been increasing. The use of content-based filtering has been decreasing, being used in fields focused on text information.

2.3. Background of Recommendation Systems: Clustering, Distance Matrix and Evaluation Metrics

Clustering is an algorithmic technique that identifies clusters to describe data and clarify complicated relationships (52,61). It is an unsupervised classification and forms clusters, groups of elements that are similar to each other (high similarity) and not similar to other groups (62). Here is where the similarity and distance between the elements are important. Clustering can be divided into hierarchical and non-hierarchical, the first finds clusters using the previous ones that have formed, and the last finds all the clusters simultaneously (63). Hierarchical clustering can be divided into agglomerative (bottom-up) and divisive approaches (top-down) (64). Their difference is that the agglomerative approach begins clustering with every element in separate clusters and joins them consecutively to form larger clusters, while the divisive approach begins with every element in one cluster and divides it into smaller ones (63). K-means is a non-hierarchical algorithm used to form a specific number of clusters. A study from 2010 to 2020

revealed that clustering is the third most used recommendation technique in papers published during this time, behind text mining and neural network (52).

In order to perform clustering, there is a need for a distance matrix. The Jaccard index or Jaccard similarity coefficient can be used to calculate the similarity between sets characterised by qualitative variables through the following equation (46):

$$J\ (A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

A and B are sets. |A ∩ B| represents the number of elements that are common to both sets A and B (intersection). |A ∪ B| represents the total number of unique elements in sets A and B combined (union). The value ranges between 0 and 1, and the higher it is, the more similar the elements are (61). To perform this calculation, the elements must have their attributes represented by the values "0" in case of the absence of the attribute and "1" if it is present (57). If there are variables with a range of numbers, it is also possible to calculate their similarity with this index. A classical interval arithmetic $\mathbb{R}$ is an algebraic system formed by intervals $A = \left[\ \underline{A}, \overline{A}\ \right] \subset \mathbb{R}$ (65). The width (wid) of A is $\overline{A} - \underline{A}$, that represent respectively, the upper and lower bounds of the interval (65). According to the previous author, to calculate the Jaccard index in this case, it would be in the form $JK(a,b)$ (*K* indicates the relation to Kaucher's interval arithmetic):

$$JK\ (A,B) = \frac{wid\ (A \wedge B)}{wid(A \vee B)} \tag{2}$$

where $A \wedge B$ and $A \vee B$ are defined as:

$$A \wedge B = \left[\max\left\{\underline{A}, \underline{B}\right\}, \min\left\{\overline{A}, \overline{B}\right\}\right] \tag{3}$$

$$A \vee B = \left[\min\left\{\underline{A}, \underline{B}\right\}, \max\left\{\overline{A}, \overline{B}\right\}\right]. \tag{4}$$

The final formula to calculate the Jaccard index of an interval is:

$$JK\ (A,B) = \frac{\min\left\{\overline{A}, \overline{B}\right\} - \max\left\{\underline{A}, \underline{B}\right\}}{\max\left\{\overline{A}, \overline{B}\right\} - \min\left\{\underline{A}, \underline{B}\right\}} \tag{5}$$

To form the distance matrix, the Jaccard distance needs to be determined, which measures the dissimilarity between the elements instead of its similarity as the Jaccard index does (66). Jaccard distance is complementary to the Jaccard index and can be calculated with the following formula (66):

$$J_\delta(A,B) = 1 - J\,(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (6)$$

After creating the distance matrix, the clustering techniques can be performed. Agglomerative hierarchical clustering can have various linkage methods, such as single linkage, complete linkage, median linkage, centroid linkage, ward linkage and average linkage. A linkage method can work well on a dataset and not get good results in another dataset (67). In single linkage, also called nearest neighbour or minimum linkage, the distance between two clusters is determined by the minimum distance between their elements (68) (Figure 9). Complete linkage, also called the farthest neighbour, uses the maximum distance contrary to single linkage (68) (Figure 10). Average linkage uses the average distances between all pairs of elements from two different clusters (68) (Figure 11). Centroid linkage consists of the distance between the centroids of the elements in the clusters (67) (Figure 12). Ward linkage uses the sum of the squares (measures the deviation of an element from the mean) of the distances between all the objects in the clusters and the centroid (68,69). Median linkage merges elements based on the median distance of each of the elements in the cluster to the remaining elements (70).

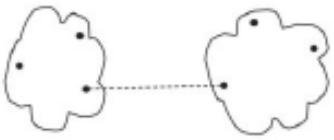


***Figure 9** – Single linkage (67).*

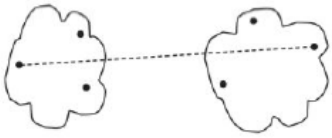


***Figure 10** – Complete linkage* (67).

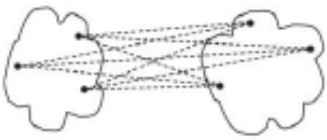


***Figure 11** – Average linkage (67).*

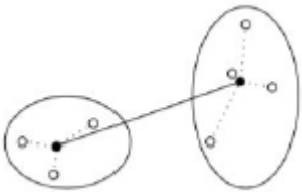


***Figure 12** – Centroid linkage (67).*

A dendrogram is a visual representation of the clusters formed; its height represents the distance between clusters, and where the distance is high, there is where the dendrogram should be cut and the clusters retrieved (67,71).

In the K-means algorithm, the number of clusters is previously chosen, and the elements are assigned to the cluster with the nearest centroid (63). Consequently, each element will be as close as possible to its cluster centre (72).

It is very important to determine the optimal number of clusters for each method and evaluate the clusters formed to decide the best approach to apply.

One of the measures that can be used is the Silhouette score (S). This score assesses how compact and separated the clusters are, it indicates how close an element is to other elements in the same cluster compared to how close it is to elements from the nearest cluster (61). The global silhouette score is defined as (73):

$$S_m = \frac{1}{\mathrm{m}} \sum_{j=1}^{m} S_j \tag{7}$$

where m is the total number of clustrer, S_j represents the silhouette of the cluster $C_j, j = 1, \dots, m$ and is defined as:

$$S_j = \frac{1}{n_j} \sum_{i:x_i \in C_j} s_i \tag{8}$$

where n_j is the number of elements in cluster C_j, x_i is an element of cluster C_j and s_i is the silhouette width of x_i defined as:

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)}. \tag{9}$$

Where a_i is the average distance of an element x_i to others within the same cluster, and b_i is the average distance of an element x_i to its closest cluster, excluding its own.

It takes values between -1 and 1. When the value is close to 1, there is a close relationship between the element and its cluster and a higher distance between the element and its closest cluster, with excellent compactness whitin the cluster and between-cluster separation (74). On the other hand, if it is close to -1, it implies that the distance between the element and its cluster is higher than the distance between the element and its closest cluster, with less compactness and separation (73). Therefore, the model is accepted if the Silhouette score is relatively high.

The Davies-Bouldin Score (DB) ranges from 0 to infinity, and it is based on the notion that good partition exhibits high inter-cluster separation and intra-cluster compactness and homogeneity. It estimates within-cluster dispersion and between-cluster separation (62). This score is defined as (73):

$$DB_m = \frac{1}{\mathrm{m}} \sum_{i=1}^{m} \; max_{j=1,\dots,m, j \neq i} \; \frac{s_i + s_j}{\mathrm{d}_{ij}} \tag{10}$$

where m is the number of clusters, s_i is the measure of dispersion of cluster C_i and d_{ij} the dissimilarity between the clusters C_i and C_j. DB_m is the average similarity between each cluster and its most similar one. The clusters must have the minimum similarity to each other. Therefore, this score should have a lower value, indicating good compactness whitin the cluster and between-cluster separation.

The Calinski-Harabasz Score (CH) compares the weighted ratio of the between-cluster sum of squares ($B(m)$) and the within-cluster sum of squares ($W(m)$), defined as (75):

$$CH(m) = \frac{N-m}{m-1} \times \frac{B(m)}{W(m)} \tag{11}$$

where N represents the total number of elements of the dataset.

If the clusters are well separated and defined, the between-cluster sum of squares would be a high value, and the within-cluster sum of squares would be low (75). Therefore, this index should present a high value in order to obtain more homogeneous clusters.

Lastly, the inflexion point method, also known as the elbow method, obtains the optimal number of clusters at the inflexion point of the line graph (76). The horizontal axis of its graphic represents the number of clusters, and the vertical axis is the within-cluster sum-of-squares (WCSS) defined as (76):

$$WCSS = \sum_{P_i \text{ in } Cluster\ 1} distance\ (P_i, C_1)^2 + \sum_{P_i \text{ in } Cluster\ 2} distance\ (P_i, C_3)^2 + \cdots + \sum_{P_i \text{ in } Cluster\ m} distance\ (P_i, C_m)^2 \tag{12}$$

where P_i represents the element i of a cluster, C_x is the centroid of cluster x and m is the number of clusters. The WCSS measures the variance within each cluster.

When there is no obvious inflexion point, it is not possible to determine the optimal number of clusters with this method. More than one of these measures should be used when deciding the optimal number of clusters and evaluating them since their performance varies according to the dataset (77).

2.4. State of the Art of Recommendation Systems in Pharmacy

The state of the art of recommendation systems in pharmacy presented was carried out using the keywords "recommendation system", "pharmacy", "pharmaceutical products", "over-the-counter", and "non-prescription products" in various combinations. The search was performed on PubMed, Web of Science and Google Scholar.

One of the fields where HRS can be applied is the pharmaceutical field in order to recommend pharmaceutical products and assist pharmacy professionals' decisions.

Bhat et al. (45) developed a hybrid recommendation system to assist in the recommendation of newly marketed pharmaceutical drugs combining item-based collaborative filtering and content-based filtering approaches. The collaborative filtering consisted of the calculation of similarities between the medication based on the users' ratings to recommend similar items to users with similar preferences; on the other hand, the content-based filtering served to analyse the medications' characteristics for them to match with the specific symptoms and health conditions. The RS was evaluated by measuring precision and recall.

Zhang et al. (72) proposed CADRE, a recommendation system for online pharmacies. The objective was to provide personalised medication recommendations. The system utilised the vector space model to represent medications and employed the k-means algorithm to cluster them. Collaborative filtering was applied, using user information and a user-medication rating matrix to identify suitable medications based on user ratings. To address the limitations of collaborative filtering (sparsity and massive data), tensor decomposition was applied. The recommended items were determined based on the ratings given by similar users. The similarity between users was calculated using Pearson distance.

Chan et al. (78) aimed to provide personalized medication recommendations using a multi-agent approach. They proposed a knowledge-based RS to obtain all the possible medications according to the client's preferences and after, provided a recommendation using a collaborative-filtering approach based on the similarity by applying the adjusted cosine similarity.

Lv et al. (79) proposed a collaborative filtering RS for pharmacies. It aimed to provide personalised recommendations of pharmaceutical products according to the users' purchase history and preferences of similar users. User-based collaborative filtering approach was used to identify similar users and present a recommendation. The RS was evaluated by measuring the precision, recall and F1 score.

HRS can assist professionals to provide treatment according to the patient's characteristics. There is a need to lower the risk when taking medication. The implementation of a medication RS that recommends a product for a specific symptom or disease while checking the individual's chronic diseases and medication and particular situations to find possible interactions and contraindications, and therefore, presenting the best product adequate for that individual would be beneficial (80). The pharmaceutical industry also benefits since there are numerous products on the market (45). The same active substance can be sold under various brand names, which can make the professionals' tasks, whether it is a doctor prescribing or a pharmacy professional dispensing an OTC medication, more complex (45). For example, acetaminophen 500 mg pills *per os* are authorised for sale without a prescription in Portugal under 38 different brand names (81). A system that could support the professionals with all the information on the medication available would be beneficial and reduce human error. The system could be personalised for each individual according to their own needs, such as the individual's age, gender, diseases, chronic medication, allergies and others.

Although there are a few studies in the area of medication RS, they are still scarce, and most of them are prepared to be applied to assist the prescription by the doctors, not the pharmacies. To the best of our knowledge, no medication RS is applied in community pharmacies in Portugal.

3. Methodology

The methodology will be based on CRISP-DM (Cross-Industry Standard for Data Mining) methodology adapted to product counselling in pharmacy (82). It consists of six stages: background analysis of pharmaceutical products counselling without a prescription, selection of data from public databases and data exploration, data preparation, modelling and evaluation, and results communication.

The dissertation proposal was submitted to the ethics committee of ESS-P.Porto and approved (process number CE0015D).

3.1. Background Analysis of Pharmaceutical Products Counselling Without a Prescription

Pharmaceutical counselling consists of three phases: clinical interview of the patient, pharmaceutical intervention, and assessing the patient's clinical outcomes (83). During the clinical interview with the patient, it must be assessed age and gender and the reason for the consultation with the five key questions, known as WWHAM, presented in Table 6. It is also essential to know if the patient has other health conditions that may interfere with the problem presented, such as diseases, pregnancy or breastfeeding, and how severe the symptoms are.

Table 6 – WWHAM questions during clinical interview (83).

W	Who is it for?
W	What are the symptoms?
H	How long ago have the symptoms started?
A	Actions taken?
M	Medication being taken?

The pharmaceutical intervention is when the professional decides the therapy (non-prescription products or non-pharmacological measures) or refers the patient for a medical appointment with the doctor. If the decision is to dispense non-prescription products, the professional must explain the posology, the duration of use, and possible side effects and evaluate possible interactions. An example of an interaction is when the patient needs a product that relieves flu symptoms, usually composed of a decongestant and an analgesic and antipyretic such as acetaminophen. It is very important to inform the patient that the product contains acetaminophen and that he cannot take this OTC medicine concomitantly with acetaminophen as it could exceed the daily dose causing severe adverse effects, such as liver injury. Interactions with supplements may also occur, such

as ginkgo biloba, which may increase the risk of bleeding when taken concomitantly with anticoagulants, as it increases their effectiveness, or St. John's wort, which on the contrary, decreases the action of antidepressants (10). Hence the analysis is fundamental. Lastly, after a week, the improvement of the symptoms must be assessed. Usually, this happens when the patient is a regular client at the pharmacy or if he comes back because the previously given product did not solve the problem as expected. Due to the wide variety of products and rapid evolution, it is important that the professional is up to date with the products existing in the market and is able to contextualise that knowledge into practice. Studies report that pharmacy professionals tend to make product recommendations based on personal experience and patient feedback (84,85). Although, when robust and reliable information about the products is provided, they make recommendations based on evidence which is essential since, according to studies carried out, counselling is relevant to reduce health problems (14,84).

3.2. Databases Overview and Selection

This work is inserted in the context of the content-based filtering and knowledge-based approaches, since only the product classification was carried out, in fact there is no information about the users, and knowledge by experts was used to increase the pertinence of the recommendations.

Data on OTC medication, homoeopathic medication, and dermocosmetic products was collected from public databases. The search for clinical and scientific information on these products was based on the knowledge acquired in pharmacology (86–88).

Information relating to OTC medication was collected from:

- Prontuário Terapêutico Online from Infarmed [1]: a database to support the prescription and dispensing of medicines. Contains information about the therapeutic goal, pharmacotherapeutic group, adverse reaction, contraindications, warnings and precautions, interactions, posology, commercial name, pharmaceutical forms, and dosages.
- Infomed[2]: a national database of medicinal products for human use. It presents similar information to Prontuário Terapêutico Online but also contains the Summary of Product Characteristics (SmPC) and the Package Leaflet (PL), which contain the previously stated

[1] https://app10.infarmed.pt/prontuario/index.php
[2] https://extranet.infarmed.pt/INFOMED-fo/

information and the ingredients list. Here there is the possibility to carry out an advanced search so that only OTC medications are displayed.

- List of authorised OTC and homoeopathic medicines from Infarmed[3].

Information relating to dermocosmetics was collected from the respective websites of the brands. Due to the large number of dermocosmetics, as new products are constantly entering the market and others are discarded, it was only used a sample of the market's products from various brands and beauty lines.

3.3. Data Preparation

Data processing was carried in Microsoft Excel and Python with the libraries Pandas, Natural Language Toolkit (NLTK), Unidecode, Plotly, Matplotlib, NumPy, SciPy, and Scikit-learn, using Spyder IDE. Initially, the entire dataset was converted into lowercase, and all accents were eliminated. A dictionary for each column was created to store the new values after pre-processing, and a dictionary named "produtos" was used to add all the products with their characteristics after the pre-processing. The columns suffer different pre-processing according to their content.

- The columns Active Substance and Pharmaceutical Form were separated by "+", Adverse Effects by Categories by ",", and Pharmacotherapeutic Group was separated by paragraph. To these four string columns, all the leading and trailing white spaces of the strings were removed.
- The text in the column Age was converted into a range. If there was no indication of the age information, the range is invalid (-1, 0). If it was higher or equal than a certain age, that number was considered the lowest in a range of 120 years (x, 120). The number was considered the upper limit (0, x) if it was lower than a certain age. It was converted into a tuple (x, y) when it had both age limits.
- The Identification Number column was converted into a string.
- The Pregnancy and Breastfeeding columns did not need any pre-processing, except for the conversion into lowercase and the accents' elimination previously stated.

[3] https://www.infarmed.pt/web/infarmed/entidades/licenciamentos/locais-de-venda-mnsrm/lista_de_mnsrm

- In the columns Indication, Interactions, Contraindications, Warnings and Precautions, all the special characters were eliminated, and each cell was tokenised by sentence and word. To eliminate irrelevant words was created a list of stopwords adequate to the data to be modeled. Each cell was tokenised by 4-grams, 3-grams, 2-grams, and 1-gram, in this order, and the pharmaceutical relevant expressions/words were saved into a new dictionary of n-grams. For example, the value corresponding to the key "Indication" of the dictionary is a list of the n-grams from the column Indication.

After pre-processing, there are columns containing a string (Pregnancy and Breastfeeding), a column with a tuple (Age), columns with a list of variables (Active Substance, Pharmaceutical Form, Adverse Effects by Categories, and Pharmacotherapeutic Group) and columns with a list of n-grams (Indication, Interactions, Contraindications, Warnings and Precautions). The descriptive statistics were addressed.

For the posterior calculation of the Jaccard Index, all the values in the products' dictionary were set to 0 (not present). Then the dictionary was filled with 1 (present) if the value was present in the cell.

3.4. Modelling

Cluster analysis, a machine learning technique, was applied to form groups of similar products according to the Jaccard distance between the values they contain. The Jaccard index was calculated for each pair of products regarding each variable, depending on whether the variable is a range of values or not, using the Equations (1) and (2), and its value was stored in a dictionary named "indices".

To define a distance function capable of creating clinically relevant groups for pharmaceutical counselling aligned with the pharmacy recommendation practice by human experts, with different weights applied to each variable, an expert consultation was conducted. During the consultation, experts were asked to rank criteria from 1 (least important) to 10 (most important) when recommending a non-prescription product for a specific problem reported by the patient. The variables were Contraindications, Warnings and Precautions, Pregnancy and Breastfeeding, Adverse Effects, Client's Age, Interactions, Pharmaceutical Form, Price, Feedback from Previous Clients, Symptoms and Duration. Table 7 demonstrates how these criteria correspond to the

variables present in the dataset in order to attribute them a weight based on the experts' consultation. It was also asked for the experts' age, years of pharmaceutical counselling experience, gender, job title and if they consider any other criteria to be important that should be addressed when recommending non-prescription products. The descriptive statistics were accessed, and statistical tests were performed in SPSS (89).

Table 7 – Correspondence between experts consultation criteria and dataset variables.

Experts Consultation Criteria	Dataset Variables
Contraindications	Contraindications
Warnings and Precautions	Warnings and Precautions
Pregnancy and Breastfeeding	Pregnancy
	Breastfeeding
Adverse Effects	Adverse Effects by Categories
Client's Age	Age
Interactions	Interactions
Pharmaceutical Form	Pharmaceutical Form
Symptoms and Duration	Active Substance
	Indication
	Pharmacotherapeutic Group

The attributed weight to each variable for Jaccard index calculation corresponded to the mean standard error.

The attribution of different weights to variables has already been tested in HRS by by Stark et al. (90). Therefore, with different weights being applied to multiple variables, the weighted Jaccard index for a pair of products (x, y) was calculated using the following equation:

$$J_w\,(x,y) = \sum_{i \in V} (J_i \times w_i) \tag{13}$$

Where V is the set of variables analysed, J_i represents the Jaccard index for the pair of products (x, y) for the variable i and w_i is the weight attributed to variable i.

A heatmap was created to visualise the final index corresponding to the sum of the individual weighted indexes. Subsequently, the Jaccard distance (J_δ) was calculated for each pair of

products through the weighted Jaccard index (J_w) using the following equation that was derived from Equation (6) (91):

$$J_\delta\ (x, y) = \sum_{i \in V} w_i\ - J_w \tag{14}$$

The proof that the weighted Jaccard is a distance is in Annex A. The distance values were saved in a matrix that comprises every product's Identification Number and its distance relative to all the other products. The diagonal distance values were set to zero to apply the clustering techniques.

Hierarchical clustering and non-hierarchical clustering were performed with the distance matrix. The hierarchical methods consisted of single linkage, complete linkage, average linkage, median linkage, centroid linkage, and ward linkage. The non-hierarchical method performed was K-means.

3.4.1. Influence of Experts' Personal Characteristics on Opinion over the Criteria

After data collection through the experts' consultation, statistical analysis was performed using SPSS (89). The characterization of the pharmacy professionals consulted, and their answers was carried out using descriptive statistics. To verify if the personal characteristics of the experts influenced their opinions it was performed the Mann-Whitney U Test, and given that the criteria do not follow a normal distribution, the Spearman correlation was calculated. It was considered a significance level of 0.05.

3.4.2. Determination and Evaluation of the Optimal Number of Clusters

Performance evaluation techniques were applied to determine and evaluate each clustering technique's optimal number of clusters. This was obtained through dendrograms with cutting point at the higher distance, the Silhouette score, the Davies-Bouldin score, and the Calinski-Harabasz score for the hierarchical clustering techniques. For the non-hierarchical clustering technique, it was used the Silhouette score, the Davies-Bouldin score, the Calinski-Harabasz score and the elbow method.

The clusters obtained were described by the relative frequencies of the variables used to calculate the Jaccard distance using SPSS (89).

3.5. Results Communication

The results communication phase will make dissertation results accessible to different audiences, such as the scientific community, stakeholders, and the general public. It is planned to publish achieved results in appropriate scientific journals.

The abstract "Application of Machine Learning Techniques for a Recommendation System in Pharmacy", related to the preliminary results, has already been accepted for oral communication at the *XXVI Congresso da Sociedade Portuguesa de Estatística*, which will be held from October 11th to 14th, 2023.

4. Results and Discussion

This section presents and discusses the obtained results. The conceptual pharmaceutical product recommendation framework created is discussed. Results related to the experts' consultation, descriptive statistics and applied statistical tests are described. The dendrograms, the results of the scores for determining the ideal number of clusters, and the groups obtained with the K-means clustering method are presented.

4.1. Conceptual Pharmaceutical Product Recommendation Framework

The population trusts and relies on the information and counselling provided by CP. It is known that there are multiple benefits from this institution. Pharmacy professionals are responsible for dispensing various products in CP in Portugal, such as prescription medication, non-prescription products, such as OTC medication, homoeopathic medication, dermocosmetics, phytotherapeutic products, dietary supplements, veterinary products, and childcare products. All of them are important items that need care and counselling. As previously stated, studies support the relevance of counselling in reducing health problems. Unfortunately, the software solutions applied in CP in Portugal do not cover all these products since most of them only have information about prescription medication and have few or none about non-prescription products.

An HRS could assist professionals in reducing the risk associated with pharmaceutical products and enhancing counselling by having all the relevant information about a product's characteristics. To the best of our knowledge, no medication RS is applied in CP in Portugal. As described before, and as Figure 4 demonstrates, RS has various approaches. In order to provide one, the first step would be product classification, and to do it, there is the need to compile all the information on the products in a database. Therefore, this dissertation aimed to create a conceptual pharmaceutical product recommendation framework. To this end, various official sources were consulted for reliable information on OTC medication, homoeopathic medication and dermocosmetics, and it was all compiled using Microsoft Excel. Each product is characterised by the variables Active Substance, Pharmaceutical Form, Indication, Age, Adverse Effects by Categories, Interactions, Contraindications, Warnings and Precautions, Pregnancy, Breastfeeding, and Pharmacotherapeutic Group. Pharmacy professionals were consulted to obtain their opinion on the importance of these criteria when recommending a non-prescription product. The professionals' personal characteristics were also obtained to verify if they influenced

their opinions. After that, different weights were applied to each variable in the distance function, according to the professionals' opinion. Hierarchical (single linkage, complete linkage, average linkage, median linkage, centroid linkage, and ward linkage) and non-hierarchical (K-means) clustering techniques were applied to form groups of similar products.

A conceptual pharmaceutical product recommendation framework was created with all the previously explained information to support the professionals when recommending a product since they have easier access to it. This framework is aligned with pharmaceutical counselling since it presents all the information about the products according to pharmacy professionals' opinions for a more informed decision by the professionals so they can provide information to the patient and minimise the risks of medication (Figure 13).

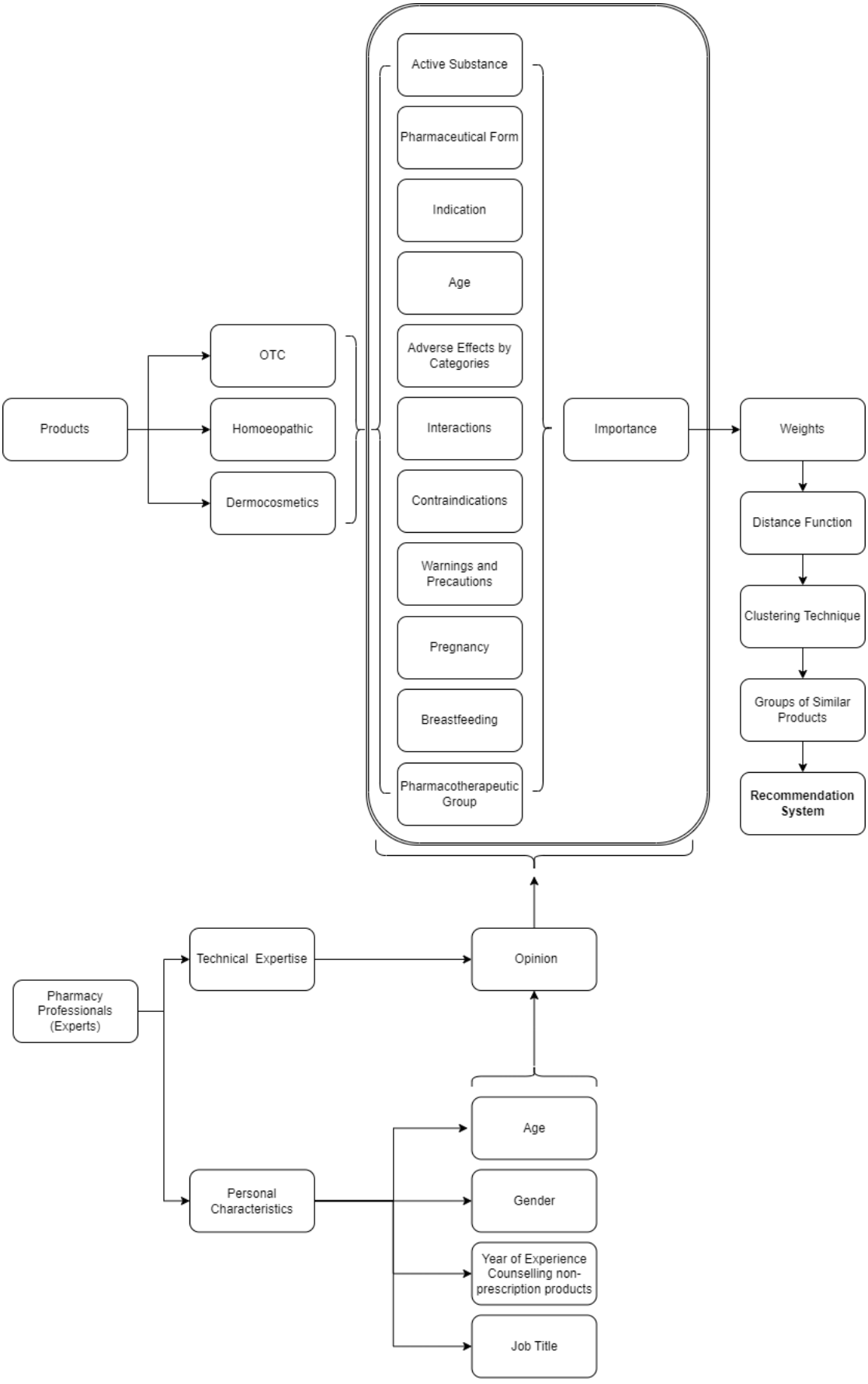


***Figure 13** – Conceptual framework for recommendation system in pharmacy.*

4.2. Database Description and Preparation

The database has information on 1426 pharmaceutical products. The majority of the database comprises OTC medication (74,5%), dermocosmetics are represented by 22,7% of the products, and 2,8% represent homoeopathic medication. Each product is characterised by 24 features, namely Identification Number, Dispense, Commercial Name, Active Substance, Pharmaceutical Form, Dosage, Packaging, Units, Quantity, Marketing Authorization Holder (MAH), Date of MAH, Indication, Age, How to Use, Adverse Effects, Adverse Effects by Categories, Interactions, Contraindications, Warnings and Precautions, Pregnancy, Breastfeeding, Ingredients List, Pharmacotherapeutic Group and ATC Code. Identification Number, Units, and Date are numeric variables, and all the others are categorical. These features represented by the columns in the dataset are described below.

Product features:

- **Identification Number**: discrete numeric variable. Each row of the dataset corresponds to a product with an identification number.
- **Dispense**: discrete variable with 3 possible values – "MNSRM", "MNSRM-EF", and "Venda Livre".
- **Commercial Name**: discrete variable with 1218 possible values – e.g., "Broncoliber", "Brufen", "Hirudoid".
- **Active Substance**: discrete variable with 317 possible values – e.g., "paracetamol", "ibuprofeno", "diclofenac".
- **Pharmaceutical Form**: discrete variable with 99 possible values – e.g., "comprimido", "cápsula", "pastilha".
- **Dosage**: discrete variable with 309 possible values – e.g., "500 mg", "200 mg", "6 mg/ml".
- **Packaging**: discrete variable with 26 possible values – e.g., "blister", "bisnaga", "saqueta".
- **Units**: discrete numeric variable indicating the number of units in a package.
- **Quantity**: discrete variable with 122 possible values – e.g., "200 ml", "100 g", "10 g".
- **Marketing Authorization Holder (MAH)**: discrete variable with 197 possible values – e.g., "Laboratórios Basi - Indústria Farmacêutica, S.A", "Bial - Portela & Cª, S.A.", "Zentiva Portugal, Lda.".
- **Date of MAH**: discrete numeric variable.

- **Indication**: discrete variable with 813 possible values – e.g., "Libertação do muco e facilita a expetoração na bronquite induzida pelo frio", "Contraceção de emergência até 72 horas após uma relação sexual não proteção ou falha de um método de contraceção", "tratamento de feridas (cortes, arranhões, escoriações, rágadas e úlceras) e queimaduras superficiais quando existe algum risco de infeção".
- **Age**: discrete variable with 38 possible values – e.g., "≥ 6 anos", "≥ 12 anos", "≥ 2 meses".
- **How to Use**: discrete variable with 985 possible values – e.g., "Aplicar uma fina camada sobre a ferida bem limpa ou zona de pele infectada, uma ou várias vezes ao dia", "uma a várias vezes ao dia até 14 dias", "1 supositório 125 mg, até 4 vezes ao dia".
- **Adverse Effects**: discrete variable with 315 possible values – e.g., "irritação epidérmica", "alteração na salivação", "ardor transitório".
- **Adverse Effects by Categories**: discrete variable with 129 possible values – e.g., "Afeções oculares", "Doenças gastrointestinais", "Doenças do sistema imunitário".
- **Interactions**: discrete variable with 201 possible values – e.g., "5-fluorouracilo", "anticoagulants", "antitússicos".
- **Contraindications**: discrete variable with 269 possible values – e.g., "Hipersensibilidade à substância ativa ou a qualquer um dos excipiente", "Insuficiência renal grave", "Insuficiência hepática".
- **Warnings and Precautions**: discrete variable with 307 possible values – e.g., "Doenças hereditárias raras", "Doentes com epilepsia", "Doentes com fenilcetonúria".
- **Pregnancy**: discrete variable with 10 possible values – e.g., "Com precaução", "Contraindicado", "Não recomendado".
- **Breastfeeding**: discrete variable with 6 possible values – e.g., "Com precaução", "Contraindicado", "Não recomendado".
- **Ingredients List**: discrete variable with 1269 possible values – e.g., "Lactose monoidratada, Estearato de magnésio", "Celulose em pó, Amido de milho", "Vaselina, Lanolina anidra, Mentol, Cânfora, Óleo essencial de terebintina".
- **Pharmacotherapeutic Group**: discrete variable with 94 possible values – e.g., "2.10. Sistema Nervoso Central. Analgésicos e antipiréticos", "3.6. Aparelho cardiovascular. Venotrópicos", "6.1.1. Aparelho digestivo. Medicamentos que atuam na boca e orofaringe. De aplicação acção tópica.".

- **ATC Code**: discrete variable with 273 possible values – e.g., "A06AD11", "R02AX03", "N02BE01".

This would be very helpful information when counselling a patient due to the fact that the products are constantly being launched into the market.

The products are identified by a specific number in order to distinguish them. The feature Dispense is also important to know if the non-prescription product is allowed to be dispensed only in a pharmacy or also in parapharmacies. The same active substance, responsible for the product's activity, can be sold under multiple commercial names, so it is essential to have both these features in the database (92). The same active substance can also have various pharmaceutical forms and dosages. The dosage can depend on the indication of the active substance. For example, acetylsalicylic acid can be used as an anticoagulant if 100 to 250 mg per day are taken, as an analgesic and antipyretic, if 500 to 1000 mg are taken per administration (93,94). Therefore, this active substance is sold in Portuguese pharmacies with dosages of 100, 150 and 500 mg (93,94). The information regarding the product's pharmaceutical form is also important since, for example, the patient may not be able to take pills or capsules and may have a preference for the form powder for oral solution. The features Indication, Age, How to Use, Adverse Effects, Interactions, Contraindications, Warnings and Precautions, Pregnancy, Breastfeeding, and Ingredients List are present in the database in order to support the professional when counselling a patient and provide him all the necessary information for safe use of non-prescription products. It is always necessary to explain to the patient which is the indication of the product recommended, given that the same product may have several indications, and it should be suitable for the age of the patient. The feature How to Use is crucial to obtain the expected outcome since, as previously explained, the dosage taken, and its frequency may differ according to the wanted results. The patient should be made aware of the most prevalent adverse effects related to the product in order to know that if any of those happen, the product is likely the cause. This feature can also exclude a product from being recommended. For example, a patient with a job that requires driving should not be taking a product that causes drowsiness or daytime sleepiness. If the patient takes any chronic medication or is under treatment for an acute condition, it is crucial to evaluate possible interactions. As well as check for all the health conditions or diseases that the patient may have, such as severe liver failure, that would contraindicate the use of a product. The Warnings and Precautions feature contains information on possible allergies (such as peanuts and soy), intolerances (such as lactose),

diseases and health conditions that may need attention or special care while the treatment is carried out. The features Pregnancy and Breastfeeding clarify if the product is safe during these periods of life. The ingredients list is present to check for any compound in case the patient has allergies or intolerances. Non-prescription medication can be classified into pharmacotherapeutic groups and ATC codes. The products are classified into 20 pharmacotherapeutic groups (and subgroups) according to their indications, which allows for a faster identification according to their objective (95). In this database, there are only represented 16 pharmacotherapeutic groups since this classification applies to medicines subject to medical prescription as well. This classification has correspondence with the ATC Codes by WHO (World Health Organization) (95). Dermocosmetics do not have this classification.

Some variables are strongly correlated (e.g., Pharmacotherapeutic Group and ATC code). Including both in the cluster analysis would be redundant and give a higher weight to this criterion. Thus, the variables with pharmacological interest in counselling were determined to define a distance function for cluster analysis. The variables used were Active Substance, Pharmaceutical Form, Indication, Age, Adverse Effects by Categories, Interactions, Contraindications, Warnings and Precautions, Pregnancy, Breastfeeding, and Pharmacotherapeutic Group.

These variables underwent pre-processing according to their content, as previously described, in order to carry out the cluster analysis. The list of stopwords was adequate to the data to be modeled. For example, the word "não" could not be a part of these list since a possible value for the column Interactions is "não foram reportadas" and the meaning would change. As an example, one product has as Indication the value "Dores de intensidade ligeira a moderada como dor de cabeça, odontalgia, dismenorreia primária. Estados de febre (com duração inferior a três dias), constipações ou gripes". After removing the stopwords and tokenize by the n-grams associated to these column, the values were "[('dor', 'ligeira', 'moderada'), ('dor', 'cabeça'), ('odontalgia'), ('dismenorreia', 'primaria'), ('febre', 'inferior', 'três', 'dias'), ('constipações'), ('gripes')]"

4.3. Variables Importance based on Experts' Consultation

In this dissertation, in the context of the content-based filtering and knowledge-based approaches, there was the need to classify pharmaceutical products. Therefore, variables from the created database were chosen according to their pharmacological interest and potential to assist professionals in counselling. The variables selected for cluster analysis were: Active

Substance, Pharmaceutical Form, Indication, Age, Adverse Effects by Categories, Interactions, Contraindications, Warnings and Precautions, Pregnancy, Breastfeeding, and Pharmacotherapeutic Group.

Since not all variables have the same importance when advising a pharmaceutical product, an experts consultation with 54 pharmacy professionals was carried out to determine their importance with the goal of applying different weights to each variable when calculating the Jaccard index. To verify if the personal characteristics of the experts influenced their opinions, it was obtained information regarding the age and gender of the professionals, years of experience counselling non-prescription products and their job titles. Table 8 has information regarding the experts' age and experience in counselling non-prescription products. The range of these variables is elevated, but the age mean is almost 30 years and nearly 7 years of experience.

***Table 8** – Pharmacy professionals' age and years of experience in counselling non-prescription products.*

	Min	**Max**	**Mean**	**STD**
Age (years)	22,00	52,00	29,48	6,44
Experience in counselling non-prescription products (years)	1,00	25,00	6,81	5,76

Legend: Min: Minimum; Max: Maximum; STD: Standard Deviation.

72,22% of the professionals were women, and 77,78% were pharmacy technicians, whereas pharmacists represented 22,22% of the experts consulted (Tables 9 and 10).

***Table 9** – Pharmacy professionals' gender.*

	%	n
Female	72,22	39
Male	28,78	15
Total	100,00	54

***Table 10** – Pharmacy professionals' job title.*

	%	n
Pharmacist	22,22	12
Pharmacy Technician	77,78	42
Total	100,00	54

The results of the importance (1 – least important to 10 – most important) given to each variable to take into consideration when recommending a non-prescription product for a specific problem reported by the patient are displayed in Table 11. The median value obtained on the criteria Feedback from Previous Clients was 7, a high value supporting the studies performed by Rutter et al. (84) and Hell et al. (85) referenced in 3.1., stating that the pharmacy professionals tend to recommend products based on patient feedback. The integration of a RS would provide more reliable information to the professional to make recommendations based on evidence. The experts' answers distribution is shown in Figures 14 and 15.

Table 11 *– Recommendation Criteria Importance for Selecting a Non-Prescription Product for a Specific Patient-Reported Problem. (1 - least important to 10 - most important).*

	Min	Max	Mean	STD	Mean STE	Median	IQR
Contraindications	3,00	10,00	8,56	1,31	6,53	9,00	2,00
Warnings and Precautions	1,00	9,00	4,41	1,48	2,98	4,00	0,00
Pregnancy and Breastfeeding	5,00	10,00	8,00	1,26	6,35	8,00	2,00
Adverse Effects	1,00	9,00	2,48	1,88	1,32	2,00	2,00
Patient's Age	1,00	10,00	5,87	1,29	4,55	6,00	1,00
Interactions	1,00	9,00	5,91	1,38	4,28	6,00	1,00
Pharmaceutical Form	1,00	7,00	2,48	1,26	1,97	2,00	1,00
Price	1,00	8,00	2,20	1,21	1,82	2,00	2,00
Feedback from Previous Clients	1,00	10,00	6,50	2,46	2,64	7,00	3,00
Symptoms and Duration	2,00	10,00	8,59	2,01	4,27	9,00	2,00

Legend: Min: Minimum; Max: Maximum; STD: Standard Deviation; Mean STE: Mean Standard Error; IQR: Interquartile Range.

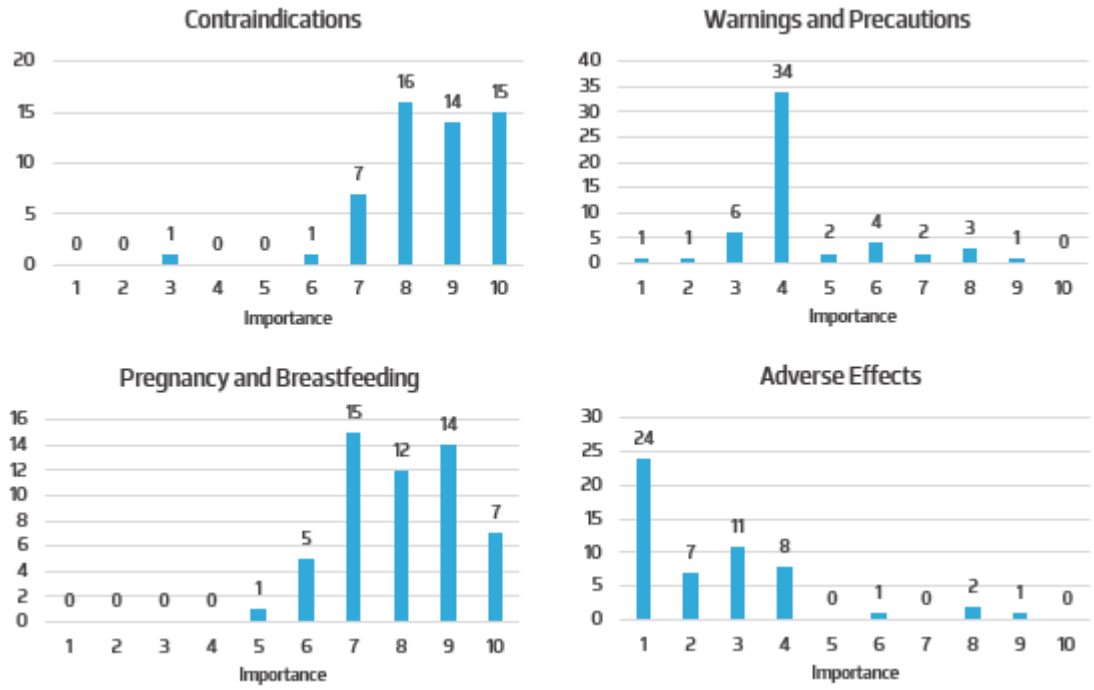


Figure 14 *– Experts answers distribution of the criteria Contraindications, Warnings and Precautions, Pregnancy and Breastfeeding. Adverse Effects.*

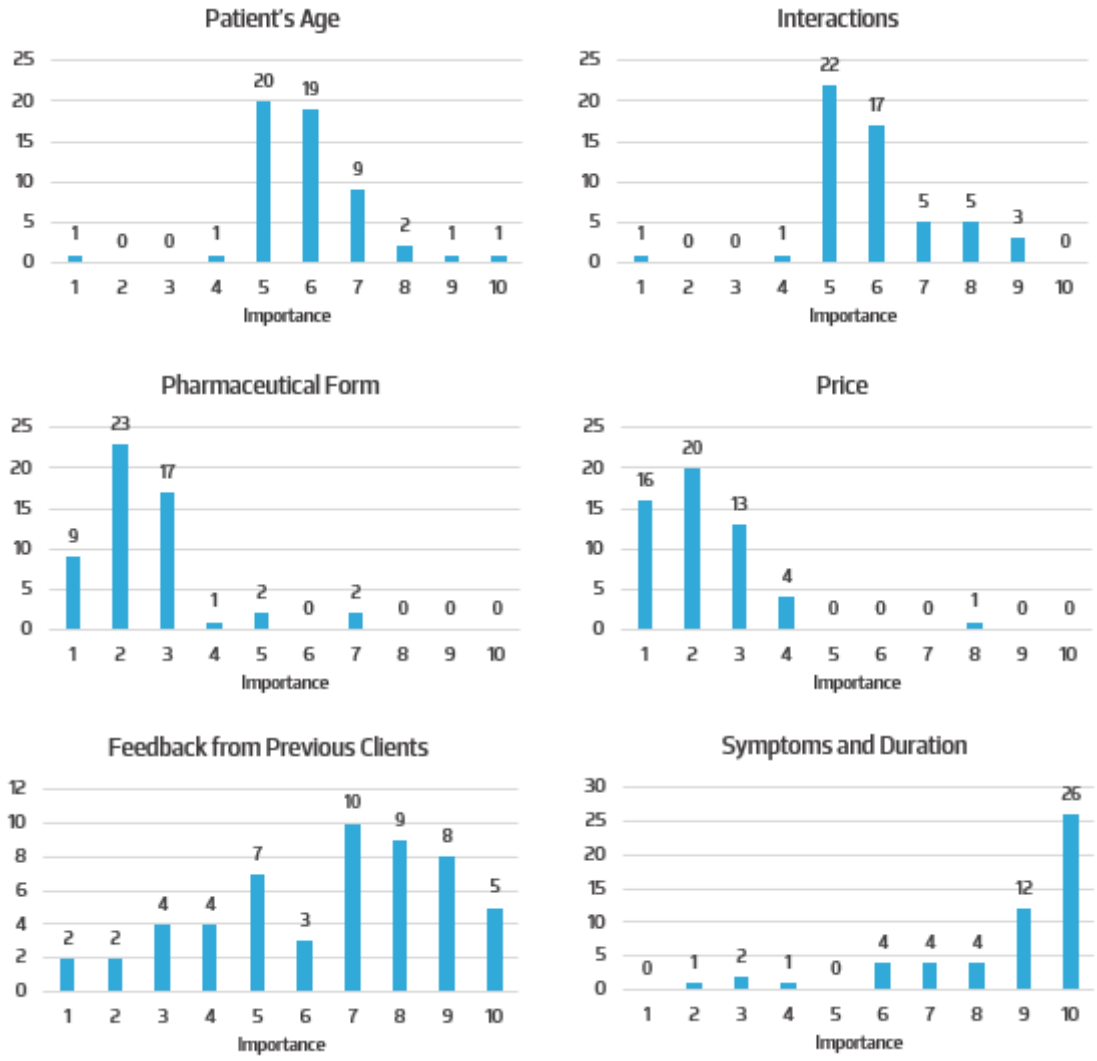


Figure 15 *– Experts answers distribution of the criteria Patient's Age, Interactions, Pharmaceutical Form, Price, Feedback from Previous Clients and Symptoms and Duration.*

There were no answers when questioned if they consider any other criteria to be important when recommending a non-prescription product.

Statistical tests were carried out in order to understand if the personal characteristics of the experts (age, years of experience, gender and job title) influenced their opinions.

The results of a non-parametric test for independent samples, the Mann-Whitney U Test, regarding the professionals' gender and job title, are displayed in Tables 12 and 13, respectively. There is no statistical evidence to affirm, with a significance level of 0,05, that there is a difference in the distribution of the importance attributed to the criteria according to the professional's gender.

Table 12 – Mann-Whitney U Test regarding the professionals' gender and the importance attributed to the criteria.

	Female			Male			P
	Count	Median	IQR	Count	Median	IQR	value[a]
Contraindications	39	8,00	2,00	15	9,00	2,00	0,639
Warnings and Precautions	39	4,00	2,00	15	4,00	1,00	0,118
Pregnancy and Breastfeeding	39	8,00	2,00	15	8,00	1,00	0,139
Adverse Effects	39	2,00	3,00	15	2,00	2,00	0,101
Patient's Age	39	6,00	2,00	15	6,00	1,00	0,968
Interactions	39	6,00	2,00	15	5,00	1,00	0,242
Pharmaceutical Form	39	2,00	1,00	15	3,00	2,00	0,395
Price	39	2,00	2,00	15	3,00	1,00	0,537
Feedback from Previous Clients	39	6,00	5,00	15	8,00	1,00	0,090
Symptoms and Duration	39	9,00	2,00	15	9,00	2,00	0,901

a. The significance level is ,05. Legend: IQR: Interquartile Range.

Regarding the results on the job title of the professionals, *p*-values lower than the significance level were obtained for the criteria Patient's Age and Interactions. These results indicate, with a significance level of 0,05, that the distribution of the importance attributed to the criteria Patient's Age and Interactions are different according to the job title. It is important to emphasize that the sample of pharmacy technicians is much higher than the sample of pharmacists. It is also important to note that there are outliers presented in these variables (Figure 15).

Table 13 – Mann-Whitney U Test regarding the professionals' job title and the importance attributed to the criteria.

	Pharmacist			Pharmacy Technician			P
	Count	Median	IQR	Count	Median	IQR	value[a]
Contraindications	12	9,00	2,00	42	9,00	2,00	0,498
Warnings and Precautions	12	4,00	0,00	42	4,00	0,00	0,904
Pregnancy and Breastfeeding	12	7,50	2,50	42	8,00	2,00	0,923
Adverse Effects	12	1,50	1,00	42	2,50	3,00	0,240
Patient's Age	12	6,00	1,50	42	6,00	1,00	0,024
Interactions	12	5,00	0,50	42	6,00	2,00	0,015
Pharmaceutical Form	12	2,50	1,00	42	2,00	1,00	0,420
Price	12	2,50	2,00	42	2,00	2,00	0,485
Feedback from Previous Clients	12	7,50	2,50	42	7,00	3,00	0,462
Symptoms and Duration	12	9,00	2,50	42	10,00	2,00	0,424

a. The significance level is ,05. Legend: IQR: Interquartile Range.

The criteria does not follow a normal distribution, according to the Shapiro-Wilk test (Table 14).

***Table 14** – Shapiro-Wilk Test Results.*

	Statistic	df	P value
Age	0,856	54	<,001
Years of Experience	0,840	54	<,001
Contraindications	0,839	54	<,001
Warnings and Precautions	0,769	54	<,001
Pregnancy and Breastfeeding	0,926	54	0,003
Adverse Effects	0,757	54	<,001
Patient's Age	0,849	54	<,001
Interactions	0,852	54	<,001
Pharmaceutical Form	0,777	54	<,001
Price	0,772	54	<,001
Feedback from Previous Clients	0,937	54	0,007
Symptoms and Duration	0,725	54	<,001

Due to the Shapiro-Wilk test results, it was used Spearman correlation coefficient to assess whether there is a correlation between the ages of professionals and their years of experience and the importance they attributed to the criteria. The results are displayed in Table 15. Regarding the correlations with age, there are three statistically significant correlations (p-value < 0,05) with the criteria Adverse Effects, Feedback from Previous Clients, and Symptoms and Duration. As the age of professionals increases, the importance attributed to the criteria Adverse Effects decreases. However, the importance given to Feedback from Previous Clients and Symptoms and Duration increases. The years of experience have a statistically significant correlation (p-value < 0,05) with the criteria Warnings and Precautions, the more experienced is the professional the less importance is given to that criterion. Despite these facts, although all previously explained correlation values obtained are statistically significant, the correlation coefficients are below 0,4, thus demonstrating a weak correlation (96).

Table 15 – Spearman correlation coefficient regarding the professionals' age and years of experience and the importance attributed to the criteria.

	Age		Years of Experience	
	Correlation Coefficient	P value[a]	Correlation Coefficient	P value[a]
Contraindications	- 0,190	0,170	- 0,157	0,256
Warnings and Precautions	- 0,254	0,064	- 0,367	0,006
Pregnancy and Breastfeeding	0,115	0,409	0,206	0,136
Adverse Effects	- 0,354	0,009	- 0,205	0,137
Patient's Age	- 0,035	0,801	- 0,029	0,838
Interactions	- 0,131	0,344	- 0,050	0,721
Pharmaceutical Form	0,054	0,699	0,116	0,402
Price	0,186	0,178	0,203	0,141
Feedback from Previous Clients	0,275	0,044	0,208	0,130
Symptoms and Duration	0,276	0,043	0,173	0,211

4.4. Distance Function

It was crucial that this distance function was aligned with the counselling process in the CP, with the importance that the professional gave to the criteria. Thus, after the statistical analysis of the results of the experts' consultation, it was possible to assign different weights to each variable that correspond to the obtained mean standard error of the importance attributed to each criteria by the professionals (Table 11). The weights were attributed according to the correspondence between the criteria present in the experts' consultation and the variables in the database, previously reported in Table 7. The higher weight was given to Contraindications (6,53), followed by Pregnancy (6,35) and Breastfeeding (6,35), Age (4,55), Interactions (4,28), Active Substance (4,27), Indication (4,27), Pharmacotherapeutic Group (4,27), Warnings and Precautions (2,98), Pharmaceutical Form (1,97) and Adverse Effects (1,32). The weighted Jaccard index for a pair of products (x, y) was calculated using Equation (13):

$$J_w\,(x,y) = J_{Active\ Substance} \times 4{,}27 + J_{Pharmaceutical\ Form} \times 1{,}97 + J_{Indication} \times 4{,}27 + J_{Age} \times 4{,}55 + J_{Adverse\ Effects} \times 1{,}32 + J_{Interactions} \times 4{,}28 + J_{Contraindications} \times 6{,}53 + J_{Warnings\ and\ Precautions} \times 2{,}98 + J_{Pregnancy} \times 6{,}35 + J_{Breastfeeding} \times 6{,}35 + J_{Pharmacotherapeutic\ Group} \times 4{,}27$$

It was created an heatmap with the final indexes (Figure 16). It is possible to visualise the diagonal that represents the similarity of a product with himself, with the higher value of similarity. There is also a group of products in the upper right part of Figure 16 that represents dermocosmetics, having a high similarity between each other; however, being dissimilar from most of the other products since the blank area (index values equal to 0) represents total dissimilarity.

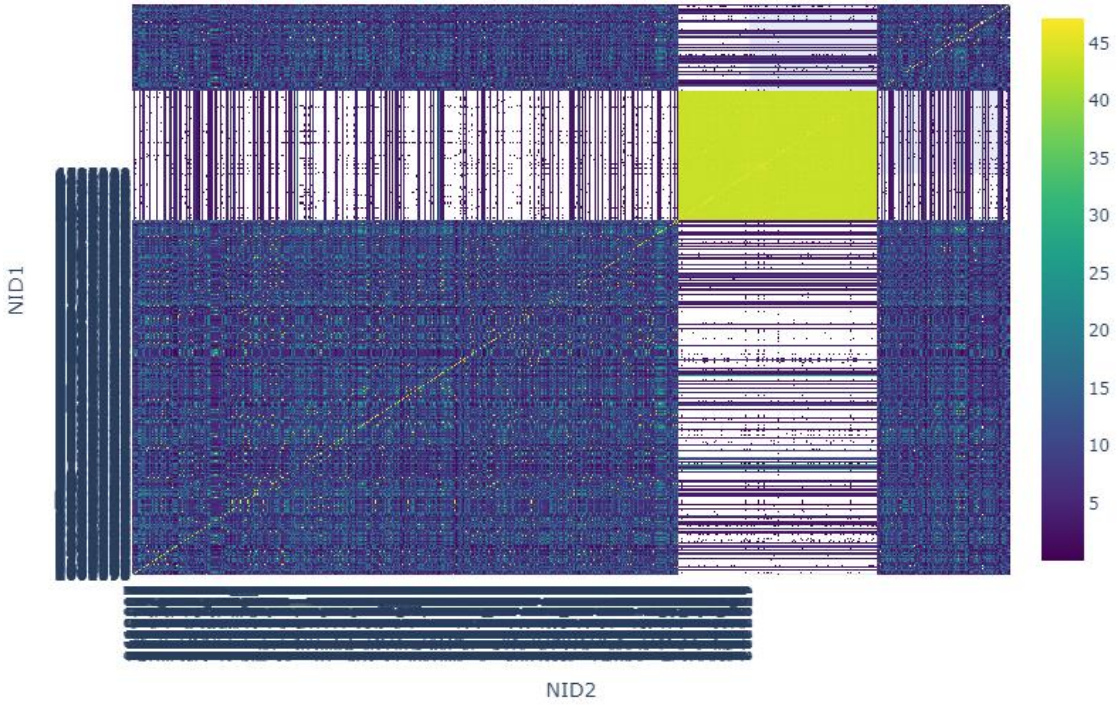


Figure 16 *– Heatmap representing the Jaccard index for each pair of products. (NID: Identification Number of the product).*

The Jaccard distance (J_δ) was calculated using Equation 14:

$$J_\delta\,(x,y) = (4{,}27 + 1{,}97 + 4{,}27 + 4{,}55 + 1{,}32 + 4{,}28 + 6{,}53 + 2{,}98 + 6{,}35 + 6{,}35 + 4{,}27) - J_w$$

$$J_\delta\,(x,y) = 47{,}14 - J_w$$

Furthermore, a distance matrix was created with all the values.

4.4.1. Recommendation Groups

Clustering is an unsupervised classification technique capable of identifying clusters to describe data. A good performance is obtained when the clusters formed different from each other, but the elements within each cluster are similar. This classification technique has been used for years in RS.

Hierarchical clustering was performed with the methods: single linkage, complete linkage, median linkage, centroid linkage, ward linkage, and average linkage; and non-hierarchical clustering, K-means. The first step was to evaluate each method's optimal number of clusters. In order to do that, stability measures were used as the Silhouette score (SS), Davies-Bouldin score (DBS), Calinski-Harabasz score (CHS) and the elbow method. It was also incorporated the domain knowledge in order to find the optimal number of clusters. A dendrogram was also obtained for each hierarchical method (Figure 17). It is possible to acknowledge that the dendrograms referent to the single, complete, median and centroid linkage methods form three clusters, and the ward and average linkage methods form two clusters. In this case, with the database of pharmaceutical products, a number of clusters so small would not have been enough to characterise the 1426 products in order for the clusters to have utility in pharmaceutical practice.

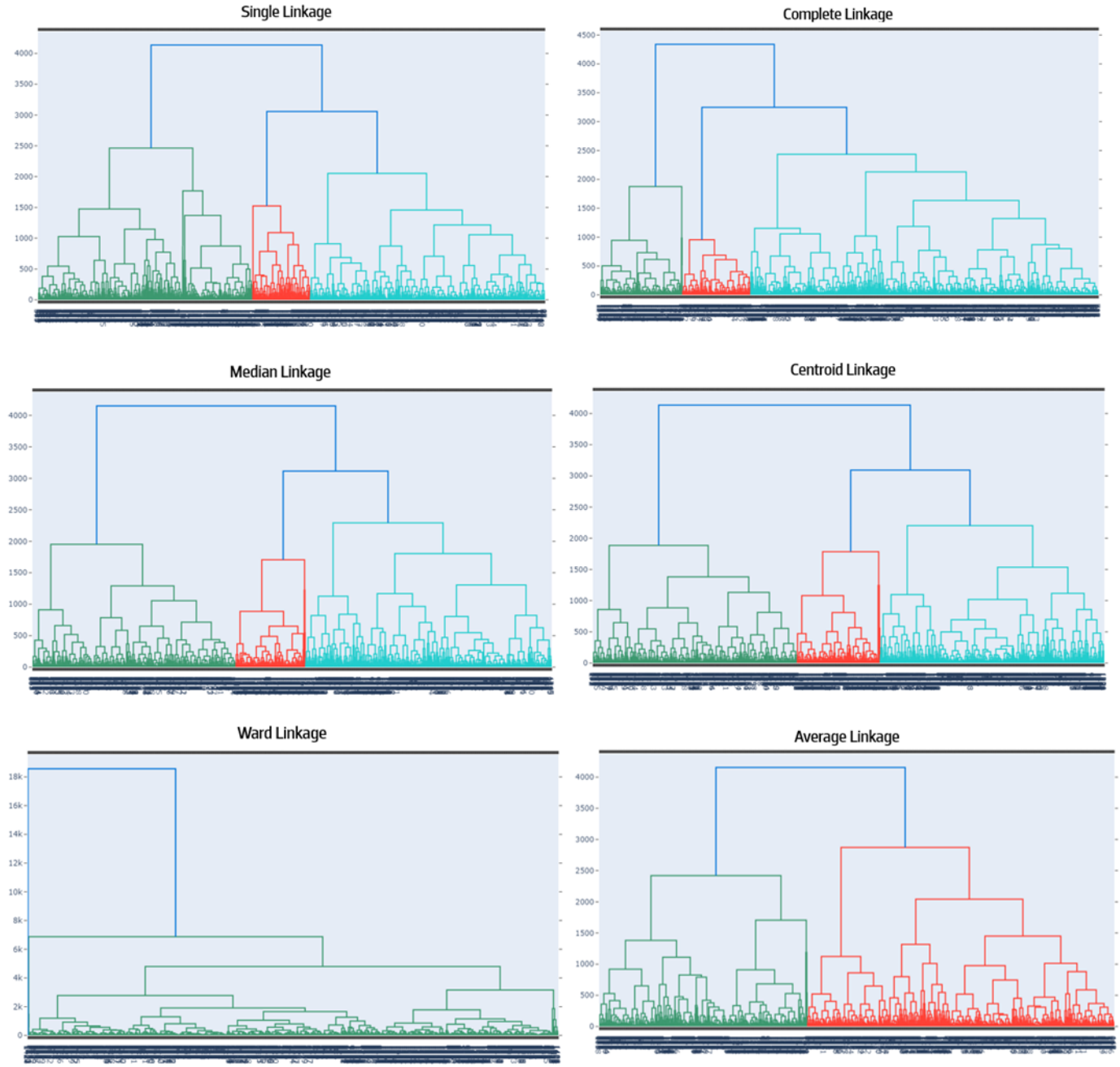


***Figure 17** – Dendrograms obtained for the hierarchical clustering methods (single linkage, complete linkage, median linkage, centroid linkage, ward linkage, and average linkage).*

Figure 18 shows the results obtained for the stability measures analysed for the method single linkage. Similar to the dendrogram, the optimal number of clusters according to these measures is also too small to have meaning in pharmaceutical practice, this also occurs for the other linkage methods. Therefore, CHS is the highest at 20 clusters, the DBS was one of the lowest at the same number and the highest SS was reached at 25 clusters, but with a very low value (0,307) since it ranges from 0 to 1. The single linkage method was performed with 20 clusters since two of the measures indicated this number as the optimal number.

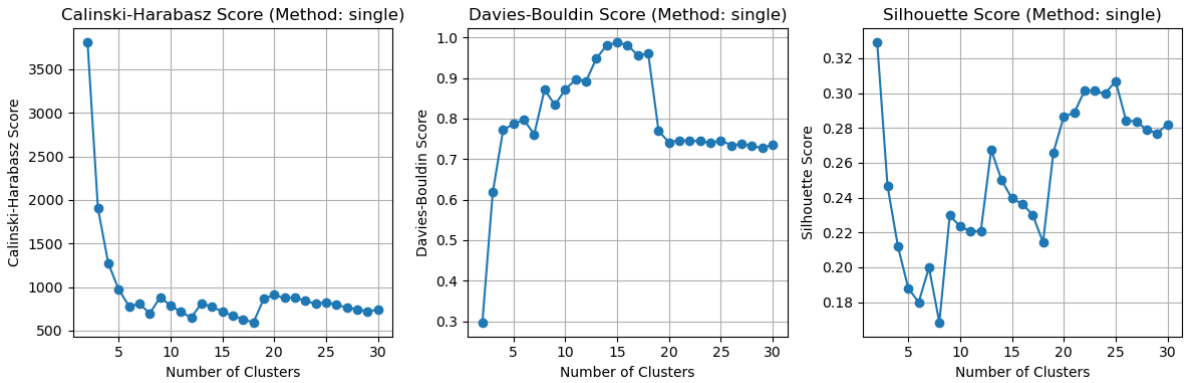


Figure 18 *– Calinski–Harabasz Score (CHS), Davies–Bouldin Score (DBS) and Silhouette Score (SS) for the single linkage method. For 20 clusters: CHS = 910, DBS = 0,741, SS = 0,287.*

In the complete linkage method, CHS peaked at 17 and 25 clusters, DBS was one of the lowest at 25 clusters, and the highest SS was reached at 30 clusters (Figure 19). The complete linkage method was performed with 25 clusters, since two measures indicate this number.

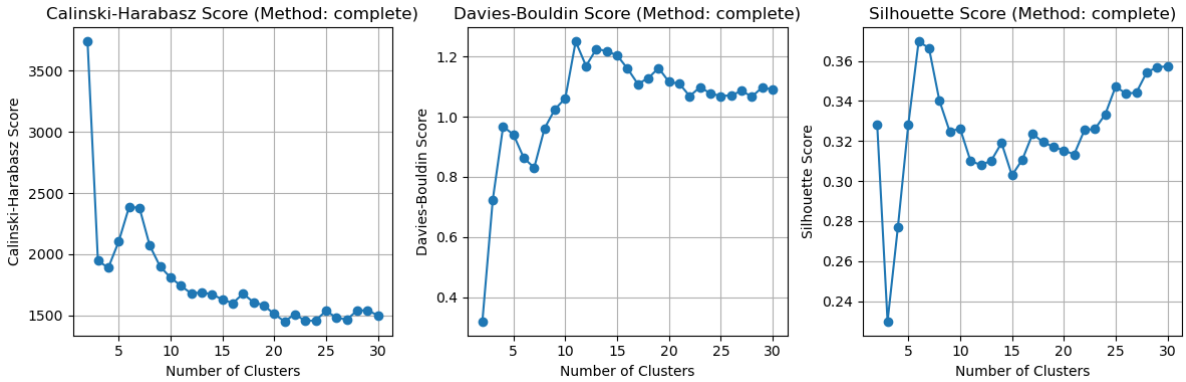


Figure 19 *– Calinski–Harabasz Score (CHS), Davies–Bouldin Score (DBS) and Silhouette Score (SS) for the complete linkage method. For 25 clusters: CHS = 1534, DBS = 1,068, SS = 0,347.*

In the median linkage method, CHS had a peak at 9 clusters, DBS had its lowest value at 18 clusters, and the highest SS was reached at 9 clusters (Figure 20). The median linkage method was performed with 18 clusters, since 9 would not represent the data in pharmaceutical practice.

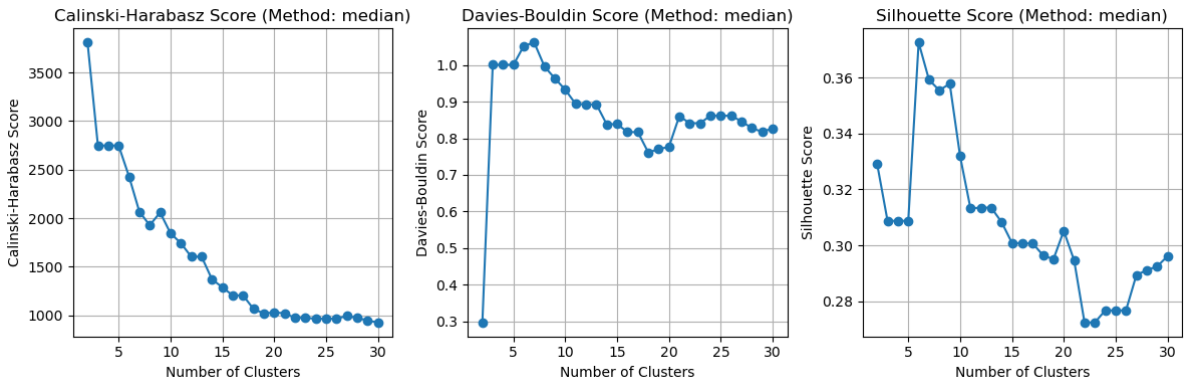


***Figure 20** – Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for the median linkage method. For 18 clusters: CHS = 1066, DBS = 0,761, SS = 0,296.*

In the centroid linkage method, CHS had a peak at 7 clusters, DBS was one of the lowest from 24 to 30 clusters, and the highest SS was reached at 7 clusters (Figure 21). The centroid linkage method was performed with a number of clusters in the range of 24 to 30, and the more homogeneous groups were obtained with 18 clusters.

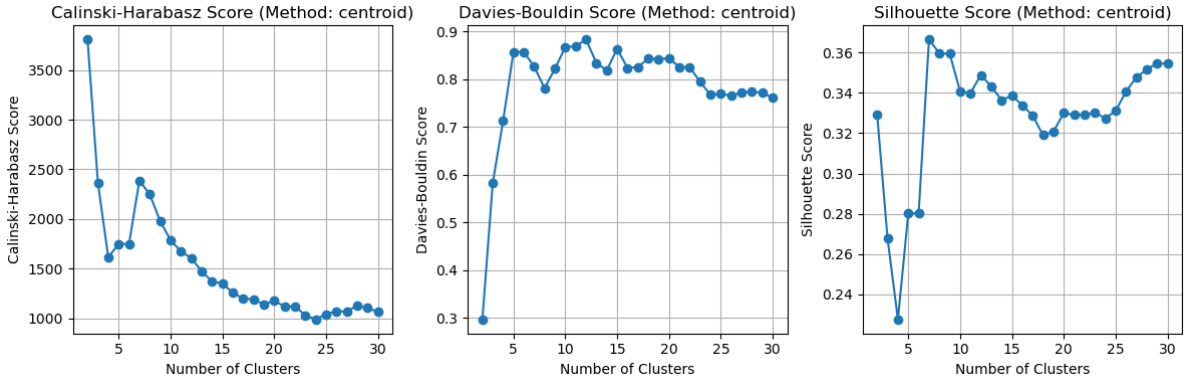


***Figure 21** – Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for the centroid linkage method. For 25 clusters: CHS = 1038, DBS = 0,770, SS = 0,331.*

In the ward linkage method, CHS had a peak at 5 clusters, DBS had its lowest value at 6 clusters, with peaks at 16 and 25, and the highest SS was reached at 5 clusters and posteriorly at 25 (Figure 22). The ward linkage method was performed with 25 clusters, due to the fact that the lower values would not have utility in pharmaceutical practice.

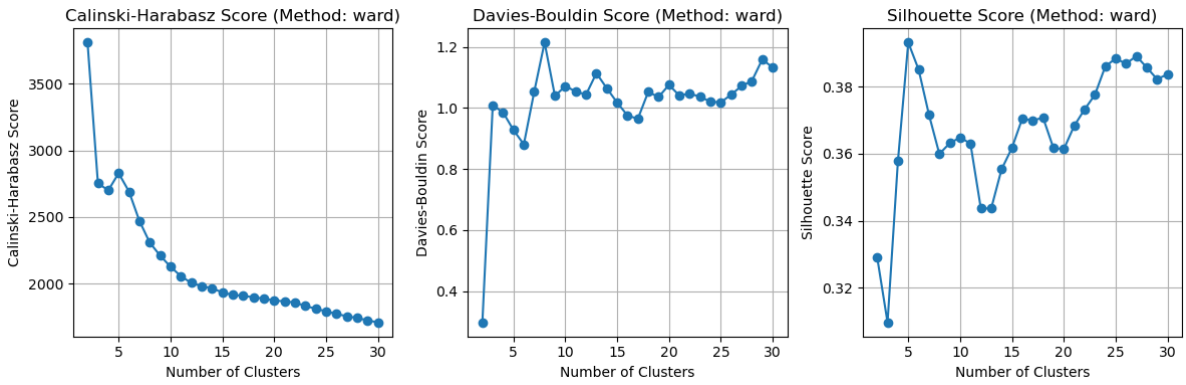


***Figure 22** – Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for the ward linkage method. For 25 clusters: CHS = 1794, DBS = 1,017, SS = 0,388.*

In the average linkage method, CHS had a peak at 6 clusters, DBS had its lowest value at 8 clusters, and the highest SS was reached at 6 clusters (Figure 23). The average linkage method was performed with various cluster numbers in order to find the most homogeneous groups since the lower values indicated as the optimal number of clusters would not have utility in pharmaceutical practice. The more homogeneous groups were obtained with 25 clusters.

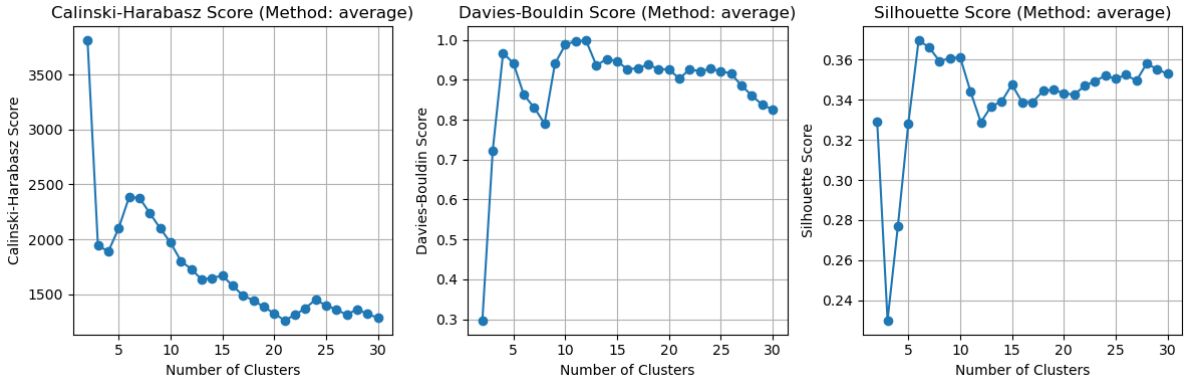


***Figure 23** – Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for the average linkage method. For 25 clusters: CHS = 1397, DBS = 0,922, SS = 0,350.*

Regarding the non-hierarchical clustering method used, K-means, CHS peaked at 5 clusters, DBS had its lowest value at 6 clusters, and the highest SS was reached at 6 clusters (Figure 24). The elbow method indicated 5 clusters through the inflexion point (Figure 24). K-means was performed with 22 clusters since this value had one of the lowest DBS and a peak in CHS.

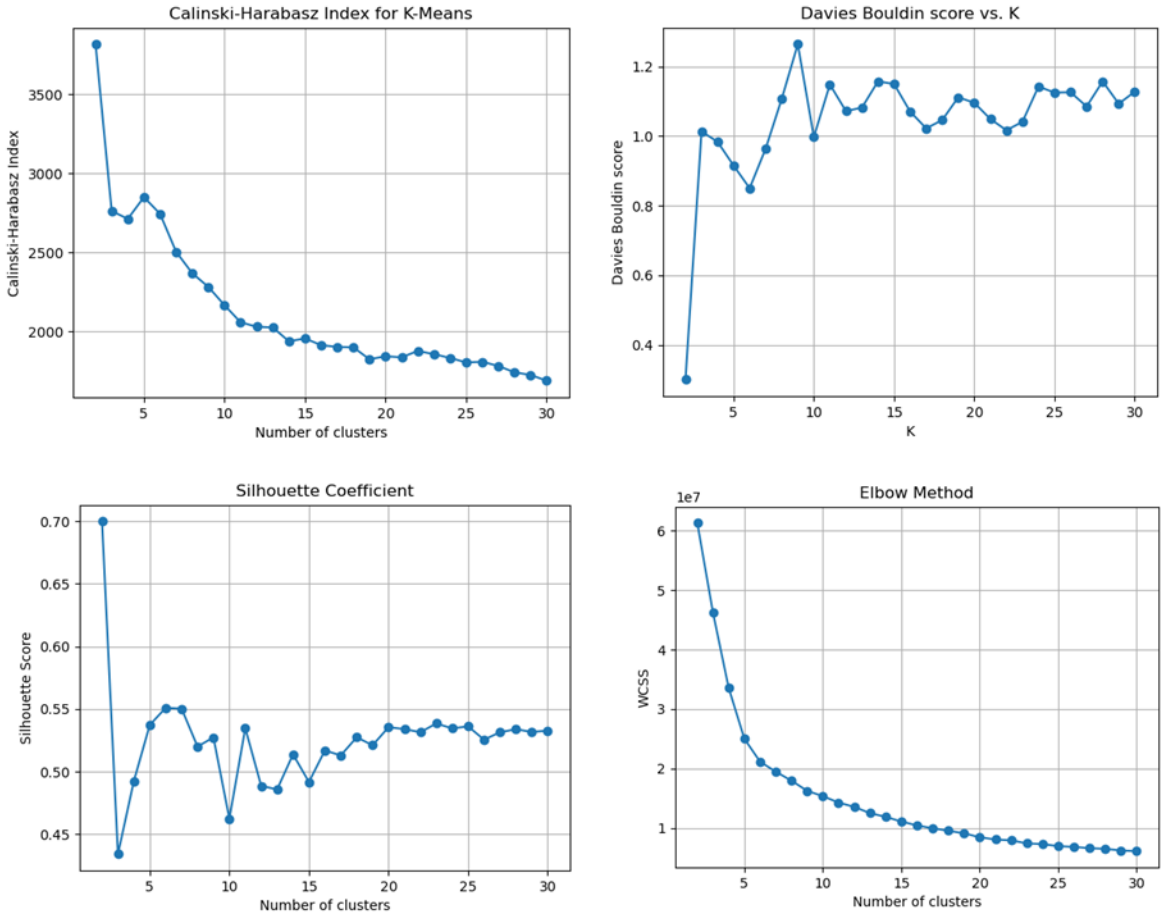


***Figure 24** – Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for K-means. For 22 clusters: CHS = 1876, DBS = 1,012, SS = 0,538.*

The relative frequencies of the formed clusters and the products that each contain in all 7 methods are shown in Figures 25 to 31. It is important to note that all methods generated a cluster with all dermocosmetics included, representing 22,7% of all products in the database, and there was no division of these products. As previously seen in the heatmap of the Jaccard indices, these products have high similarity among themselves, as well as high dissimilarity with the rest, thus predicting that this grouping would occur.

As it is possible to see in Figures 25, 27, 28 and 30, the single linkage, median linkage, centroid linkage and average linkage methods formed groups with a very small number of products, with clusters sometimes containing only 1 product. It is also observed in these cases, groups with a high number of products in which there is no homogeneity within them cluster. These methods favour the separation of outliers, however, there is no clinical reason for this to happen. Thus, these techniques are not suitable for this dataset. The complete linkage, ward linkage and K-means methods, on the other hand, obtained more homogeneous groups and the one that presented groups of products with greater utility in pharmaceutical counselling was K-Means

(Figures 26, 29 and 31). This method also presents the higher values for CHS (1876) and SS (0,538).

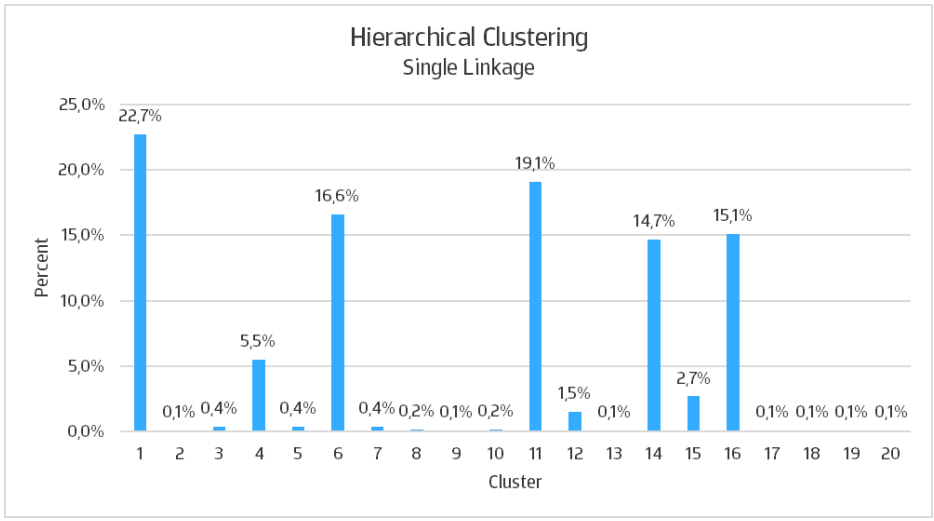


***Figure 25** – Percentage of the clusters formed with single linkage method.*

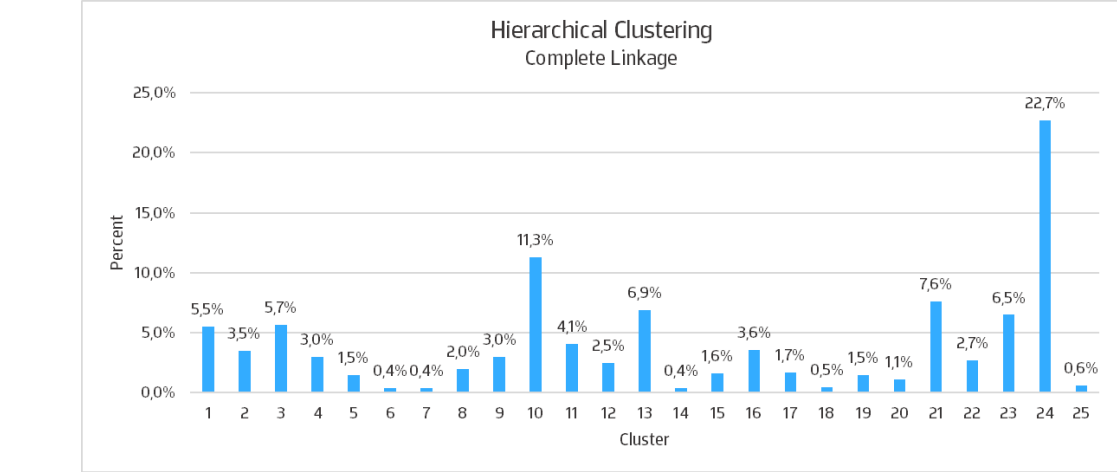


***Figure 26** – Percentage of the clusters formed with complete linkage method.*

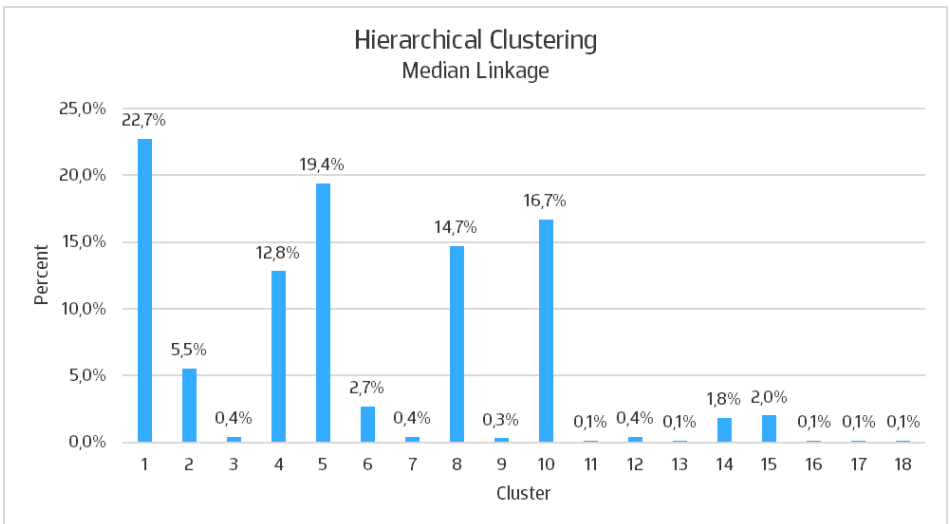


***Figure 27** – Percentage of the clusters formed with median linkage method.*

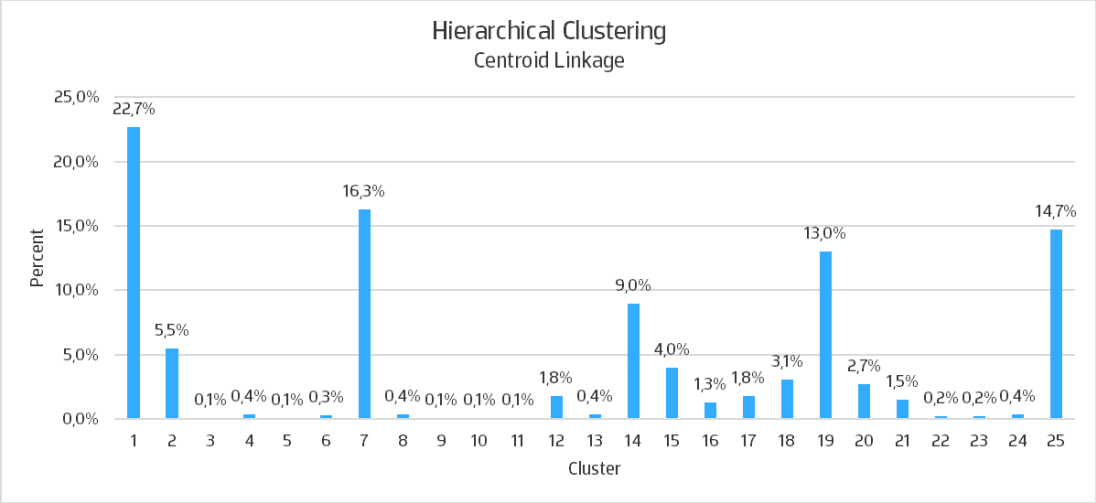


***Figure 28** – Percentage of the clusters formed with centroid linkage method.*

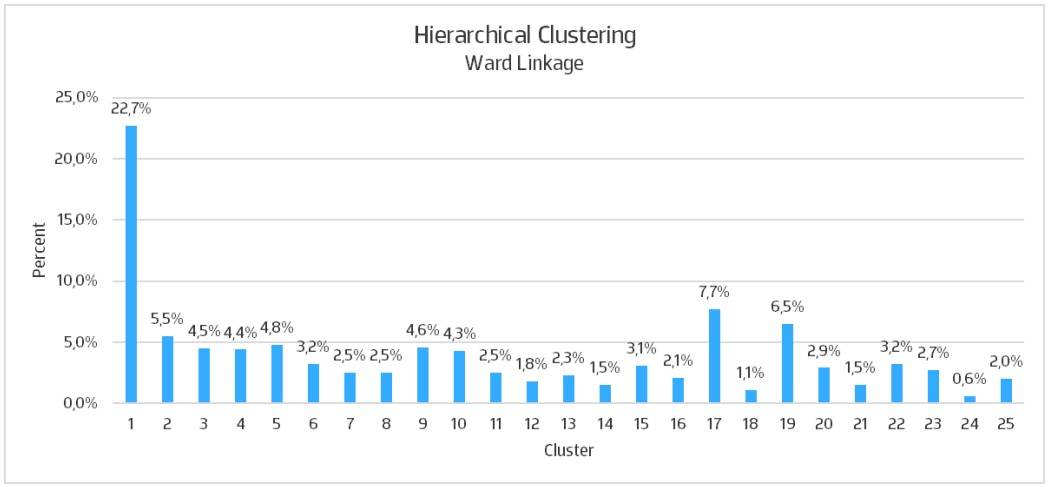


***Figure 29** – Percentage of the clusters formed with ward linkage method.*

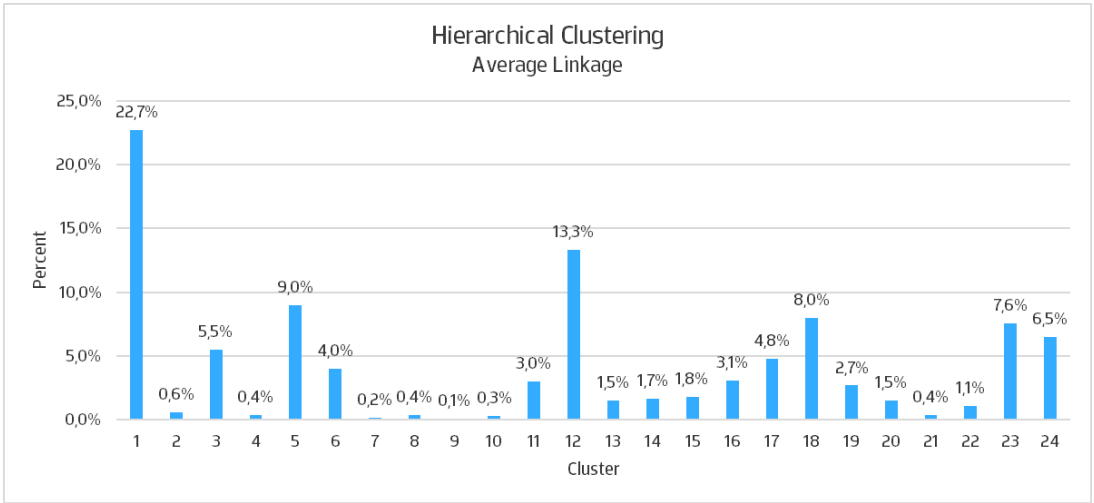


***Figure 30** – Percentage of the clusters formed with average linkage method.*

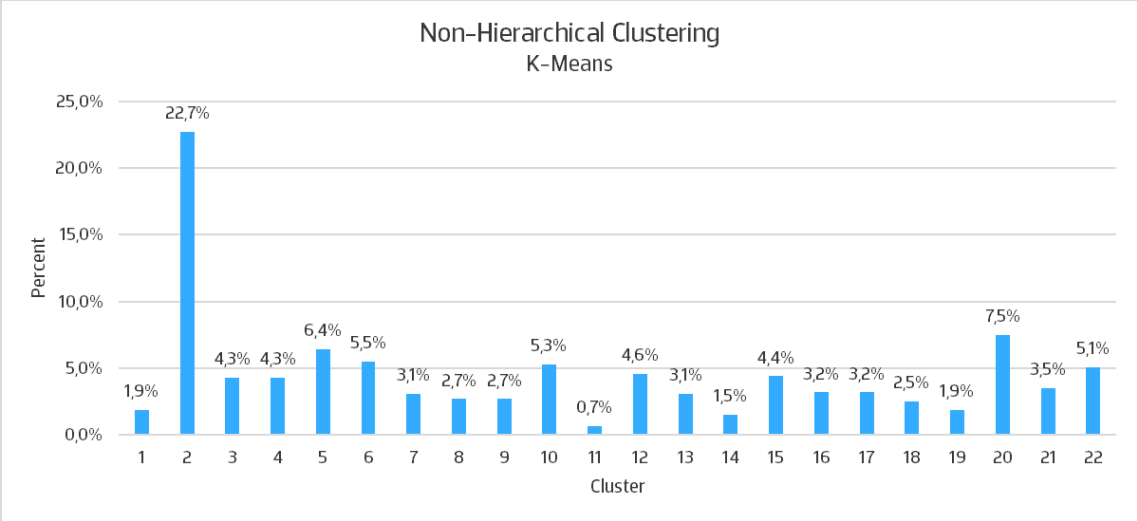


***Figure 31** – Percentage of the clusters formed with K-means.*

The relative frequency of each cluster formed by K-means was assessed. The variables with greater homogeneity within the clusters and heterogeneity between clusters were Pregnancy and Breastfeeding. Therefore, their relative frequencies are presented in Tables 16 and 17. Other variables that discriminate the cluster will be presented in bar charts for each cluster formed by K-means. The relative frequency of the remaining variables of each cluster formed with this method that do not present a defined pattern are presented in a repository due to space limitations[4]. Given the extension of the results that are not clinically relevant of the remaining six methods, they are presented in a repository for deeper analysis[4].

Regarding the variable Pregnancy, 11 of the 22 clusters formed were completely separated according to the safe use of its products during pregnancy, presenting only one value in this variable; and 18 of the 22 clusters were completely separated according to the safe use of its products during breastfeeding (Tables 16 and 17). Thus, as previously discussed, these are the most homogeneous variables within the clusters and heterogeneous variables between clusters.

[4] https://drive.google.com/drive/folders/118yRVIFI4fdXD8VyKQNF-BOv09E-99RH?usp=drive_link

Table 16 – Percentage of the variable Pregnancy present in all clusters by K-means.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
No indication	0	100,0	0	0	0	0	0	0	0	0	80,0
Can be used	0	0	0	96,7	0	100,0	0	89,7	0	1,3	0
With caution	0	0	0	0	83,5	0	0	0	0	0	0
Not recommended	85,2	0	0	0	0	0	18,2	0	0	71,1	0
Not recommended in the first trimester	11,1	0	0	0	12,1	0	81,8	5,1	0	0	0
Not recommended in the third trimester	0	0	0	3,3	0	0	0	5,1	100,0	17,1	0
Contraindicated	0	0	100,0	0	0	0	0	0	0	0	0
Contraindicated in the first trimester	3,7	0	0	0	3,3	0	0	0	0	3,9	0
Contraindicated in the third trimester	0	0	0	0	0	0	0	0	0	6,6	0
Not applicable	0	0	0	0	1,1	0	0	0	0	0	20,0

Legend: C: Cluster

Table 16 – Percentage of the variable Pregnancy present in all clusters by K-means (cont.).

	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22
No indication	0	0	0	0	0	0	0	0	0	0	0
Can be used	0	0	0	0	100,0	0	0	11,1	0	84,0	0
With caution	100,0	100,0	0	1,6	0	0	0	77,8	0	0	0
Not recommended	0	0	100,0	0	0	0	100,0	0	97,2	0	0
Not recommended in the first trimester	0	0	0	0	0	0	0	0	2,8	2,0	0
Not recommended in the third trimester	0	0	0	0	0	0	0	0	0	12,0	0
Contraindicated	0	0	0	98,4	0	100,0	0	0,0	0	0,0	100,0
Contraindicated in the first trimester	0	0	0	0	0	0	0	7,4	0	0	0
Contraindicated in the third trimester	0	0	0	0	0	0	0	0	0	0	0
Not applicable	0	0	0	0	0	0	0	3,7	0	2,0	0

Legend: C: Cluster

Table 17 – Percentage of the variable Breastfeeding present in all clusters by K-means.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
No indication	0	100,0	0	0	0	0	0	0	0	0	80,0
Can be used	0	0	0	100,0	0	100,0	0	100,0	0	0	0
With caution	0	0	0	0	93,4	0	0	0	100,0	19,7	0
Not recommended	100,0	0	0	0	0	0	100,0	0	0	80,3	0
Contraindicated	0	0	100,0	0	5,5	0	0,0	0	0	0	0
Not applicable	0	0	0	0	1,1	0	0	0	0	0	20,0

Legend: C: Cluster

***Table 17** – Percentage of the variable Breastfeeding present in all clusters by K-means (cont.).*

	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22
No indication	0	0	0	0	0	0	0	0	0	0	0
Can be used	0	0	100,0	0	100,0	0	0	0	0	100,0	0
With caution	100,0	100,0	0	0	0	0	0	88,9	0	0	0
Not recommended	0	0	0	0	0	0	100,0	0	100,0	0	0
Contraindicated	0	0	0	100,0	0	100,0	0	7,4	0	0	100,0
Not applicable	0	0	0	0	0	0	0	3,7	0	0	0

Legend: C: Cluster

Regarding the analysis of the other variables in each cluster formed with K-means, the variables that discriminate cluster 1, with few modes, are Age, Interactions, and Contraindications (Figure 32). In the Age variable, this cluster is characterised by a high percentage of the value "(-1,0)" indicating no age information was available. Interactions variable characterises the cluster by the presence of the value "não foram reportadas". The Contraindications variable presents a high percentage of "hipersensibilidade excipiente". Pregnancy and Breastfeeding describe the products of this cluster as not recommended during these periods since the higher percentage corresponds to the value "not recommended" in both variables (Tables 16 and 17). The remaining variables do not present a defined pattern, meaning the percentages are distributed among the different possible values.

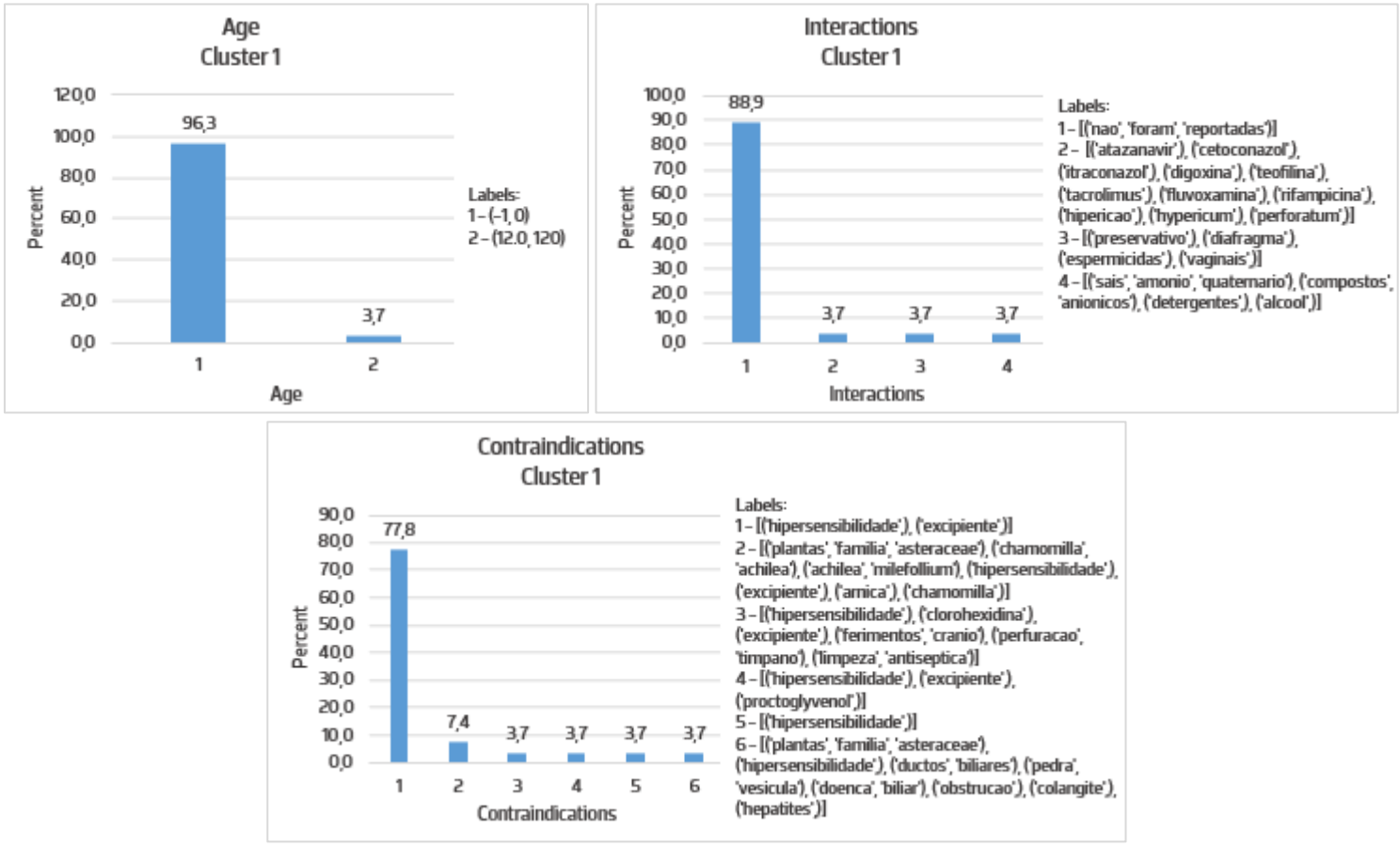


Figure 32 *– Percentage of the variables Age, Interactions, and Contraindications in cluster 1 by K-means.*

Cluster 2 is constituted only by the dermocosmetics, representing 22,7% of the database as previously discussed. These products don't contain an Active Substance, thus their value is not applicable in this case.

In cluster 3, the variable that stand out and discriminate this cluster is Warnings and Precautions (Figure 33). All the products in this cluster are contraindicated during pregnancy and breastfeeding (Tables 16 and 17).

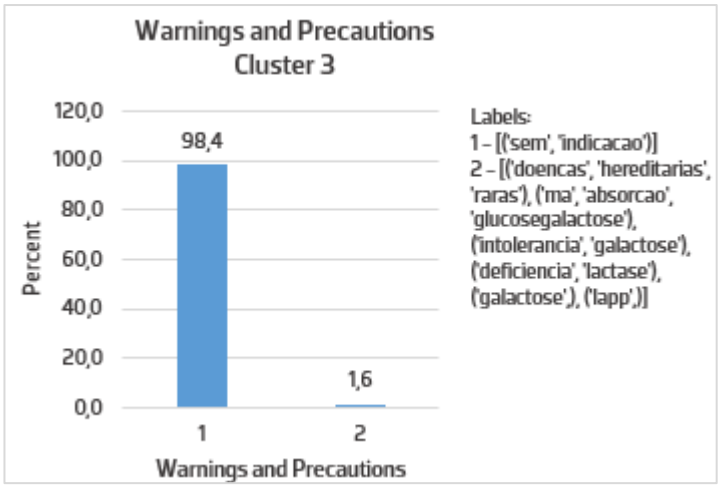


Figure 33 *– Percentage of the variable Warnings and Precautions in cluster 3 by K-means.*

Regarding the variables that discriminate cluster 4, with few modes, are Age, Contraindications, and Warnings and Precautions (Figure 34). 95,1% of the products do not present any information about the age, and the remaining products are adequate for children until 12 years old. The Contraindications variable presents a high percentage of the value "hipersensibilidade excipiente". This cluster is characterised by a high percentage of the value "sem indicação" in the variable Warnings and Precautions. During Pregnancy, 96,7% of these products can be used, and 100% can be used when Breastfeeding (Tables 16 and 17).

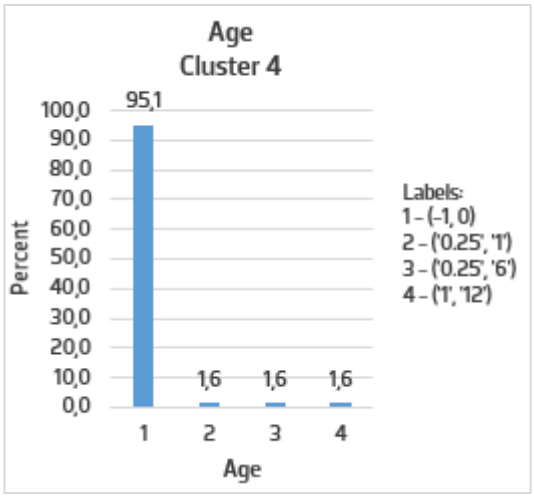


Figure 34 *– Percentage of the variables Age, Contraindications, Warnings and Precautions in cluster 4 by K-means.*

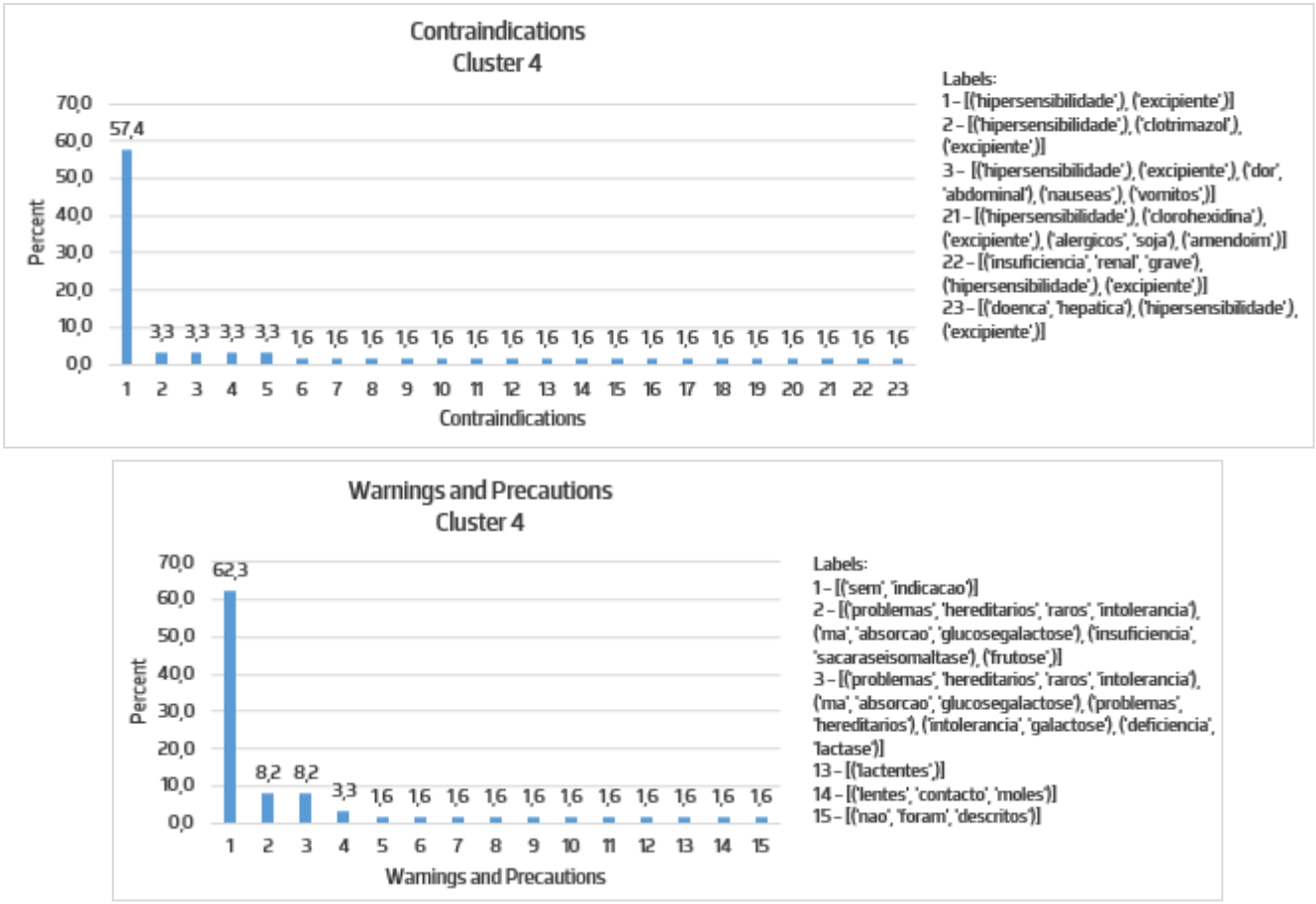


Figure 34 *– Percentage of the variables Age, Contraindications, Warnings and Precautions in cluster 4 by K-means (cont.).*

The variables that stand out and discriminate cluster 5 are Interactions and Contraindications (Figure 35). In the Interactions variable, this cluster is characterised by a high percentage of the value "não foram reportadas", which indicates there are no interactions reported for 80,2% of products within this cluster. The Contraindications variable presents a high percentage of the value "hipersensibilidade excipiente". During Pregnancy, 83,5%% of these products can be used with precaution, as well as 93,4% of them when Breastfeeding (Tables 16 and 17).

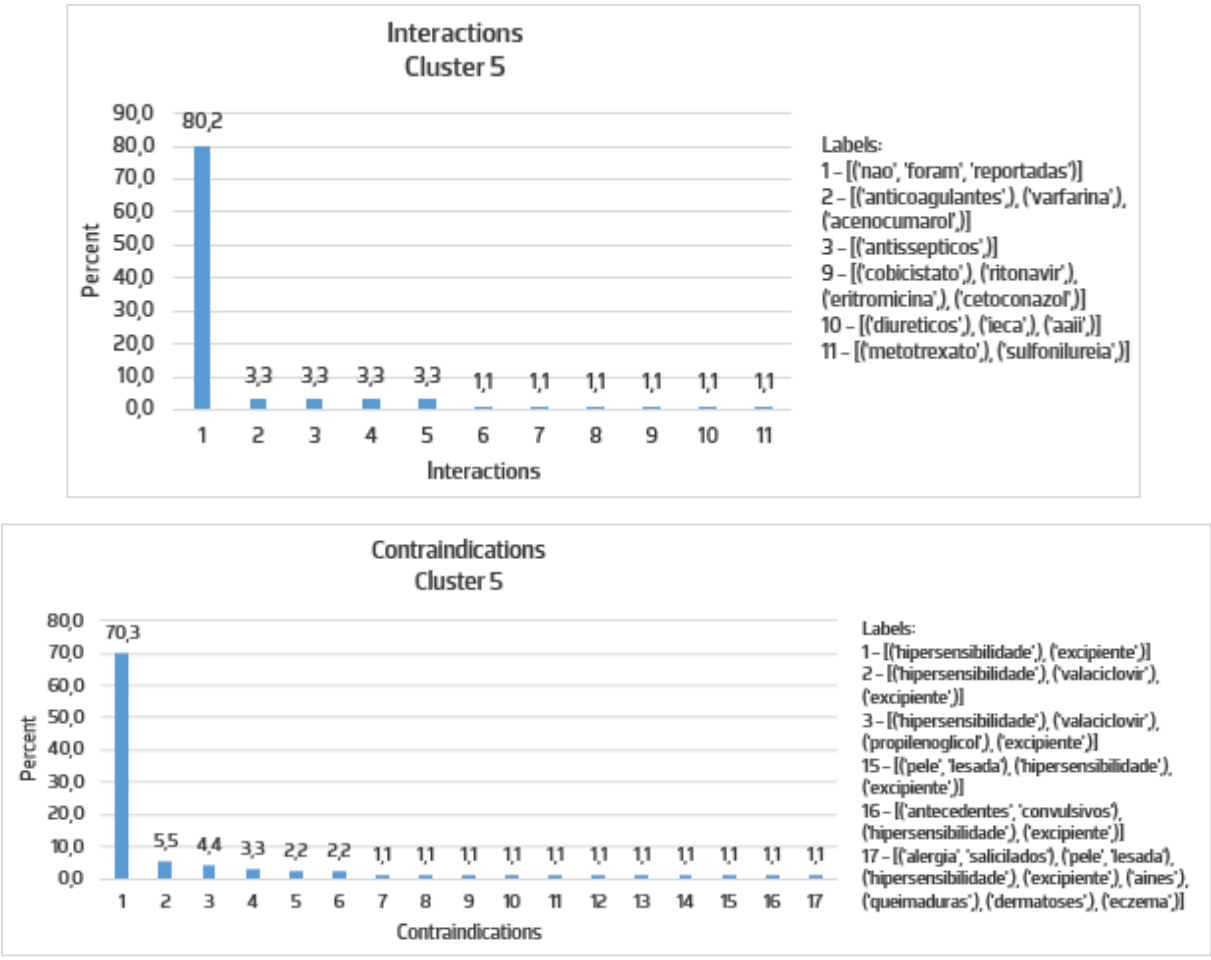


***Figure 35** – Percentage of the variables Interactions and Contraindications in cluster 5 by K-means.*

Cluster 6 is constituted only by the Active Substance "paracetamol", representing 5,5% of the database. Despite having the same active substance, these products are appropriate for various age intervals depending on their dosage and pharmaceutical form.

Cluster 7 is constituted by two Active Substances, "ambroxol" and "carbocisteína", but both of them belong to the same Pharmacotherapeutic Group, "5.2.2. aparelho respiratório. antitússicos e expetorantes. expetorantes." (Figure 36). Due to this fact, these products, like those from the previous cluster, despite having active substances for the same indications, are appropriate for various age intervals depending on their dosage and pharmaceutical form. This cluster is also characterised by different values, with high dispersion, in the variables Interactions, Contraindications and Warnings and Precautions.

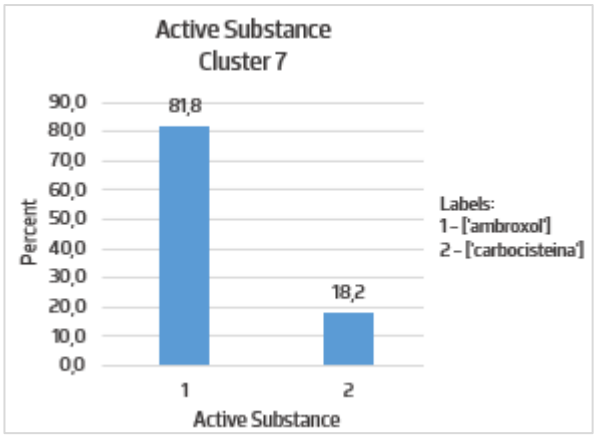


***Figure 36** – Percentage of the variable Active Substance in cluster 7 by K-means.*

Regarding the variables that discriminate cluster 8, with few modes, are Interactions, Contraindications, and Warnings and Precautions (Figure 37). This cluster is characterised by the lack of interactions reported for 100% of its products in the Interactions variable. The Contraindications variable presents a high percentage of the value "hipersensibilidade excipiente". This cluster is characterised by a high percentage of the value "sem indicação" in the variable Warnings and Precautions. During Pregnancy 89,7% of these products can be used, and 100% of them can be used when Breastfeeding (Tables 16 and 17).

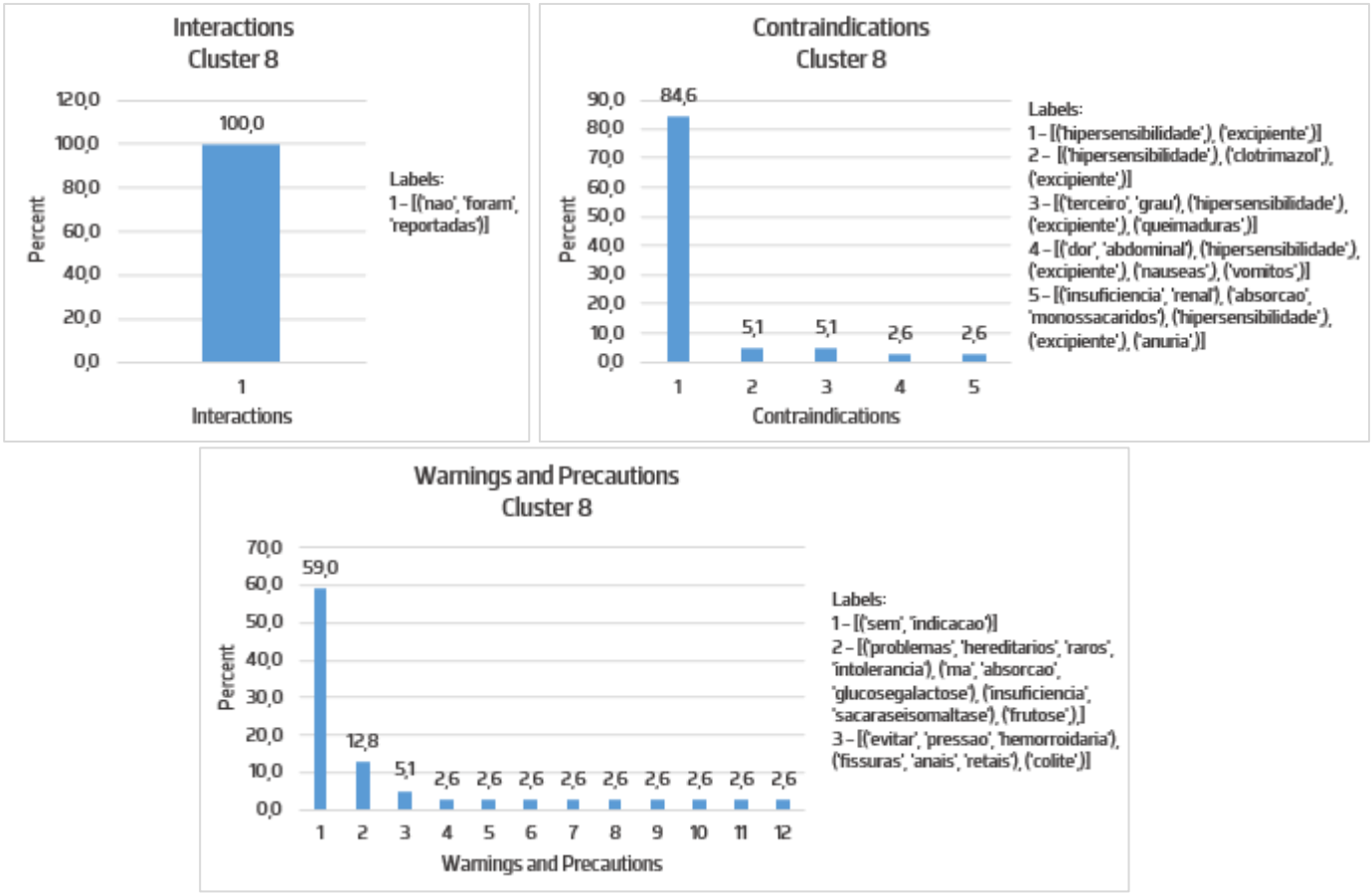


***Figure 37** – Percentage of the variables Interactions, Contraindications, and Warnings and Precautions in cluster 8 by K-means.*

Cluster 9 is constituted only by the Active Substance "diclofenac", representing 2,7% of the database. Despite having the same active substance, these products are appropriate for various age intervals (Figure 38). The variable Adverse Effects by Categories characterises this cluster since 100% of the produtcs are characterised by the value "afeções dos tecidos cutâneos e subcutâneos".

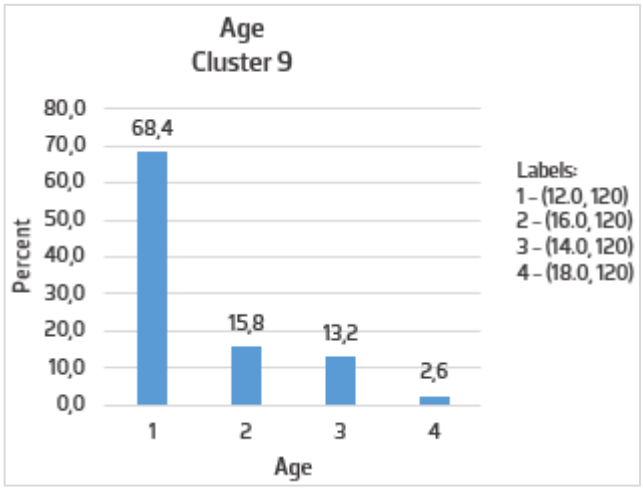


***Figure 38** – Percentage of the variable Age in cluster 9 by K-means.*

The variables that stand and discriminate cluster 10 are Pregnancy and Breastfeeding. 71,1% of the products present in this cluster are not recommended during pregnancy, and 80,3% are not recommended during breastfeeding (Tables 16 and 17). The remaining variables present high dispersion.

Cluster 11 is constituted only by 10 homoeopathic medication with 10 different Active Substances, representing 0,7% of the database, the smallest group formed by K-means (Figure 39). These products, although they are homoeopathic medication, they have 10 different indications but similar values in the variables Contraindications, Indications, Warnings and Precautions, Pregnancy and Breastfeeding.

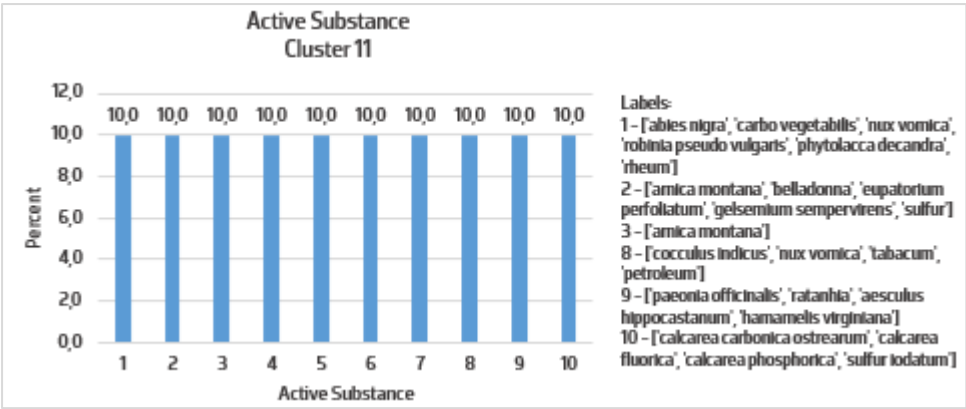


***Figure 39** – Percentage of the variable Active Substance in cluster 11 by K-means.*

Cluster 12 is characterised by the high percentage of the value "cetirizina" in the variable Active Substance (Figure 40). Although this cluster has multiple active substances, the variables Pregnancy and Breastfeeding discriminate this cluster, as stated before, since all the products can be used during these periods of life with precaution (Tables 16 and 17).

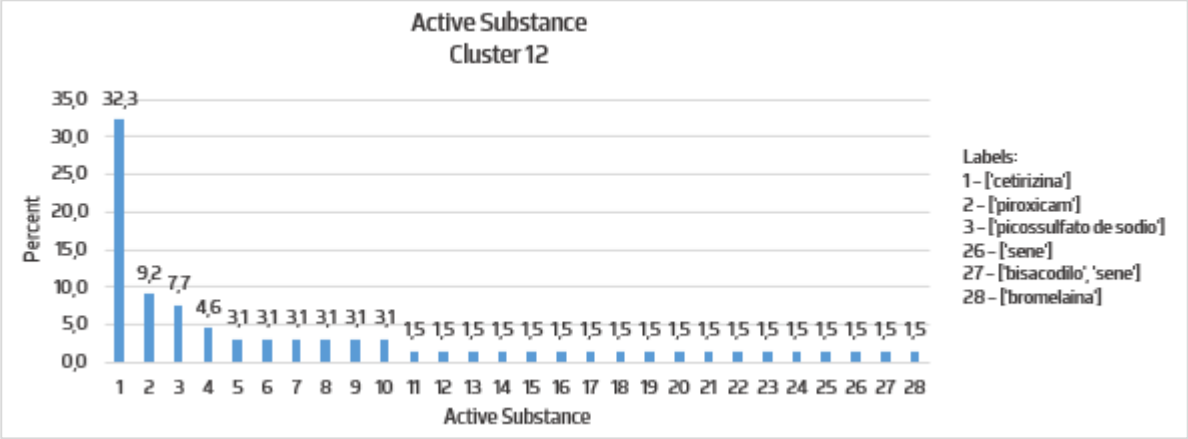


***Figure 40** – Percentage of the variable Active Substance in cluster 12 by K-means.*

Cluster 13 is constituted by two Active Substances, "acetilcisteína" and "bromexina", but both of them belong to the same Pharmacotherapeutic Group, "5.2.2. aparelho respiratório. antitússicos e expetorantes. expetorantes." (Figure 41). Due to this fact, these products, despite having active substances for the same indications, are appropriate for various intervals of age depending on its dosage and pharmaceutical form. They belong to the same pharmacotherapeutic group as the products that belong to cluster 7. In this case, they belong to two separate groups because while the products at cluster 7 are characterised by a high percentage of the value "não foram reportadas" in the Interactions variable, the products from this cluster are mainly represented by having interactions with "enoxaparina, hidroclorotiazida, litio, triamtereno, ieca, aaii, aines". These dissimilarities also happen in other variables as well, thus explaining the classification into different clusters.

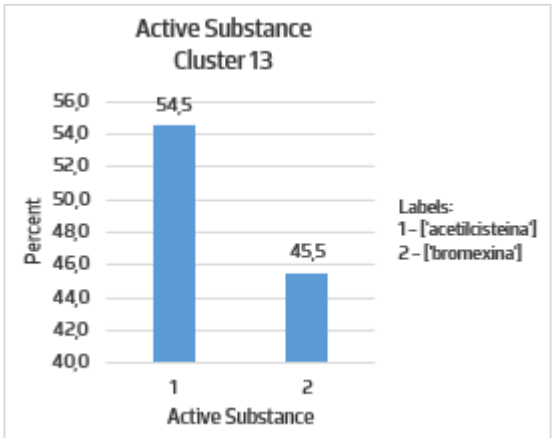


***Figure 41** – Percentage of the variable Active Substance in cluster 13 by K-means.*

Cluster 14 is constituted by two Active Substances, "etofenamato" and "ibuprofeno", but both of them belong to the same Pharmacotherapeutic Group, "9.1.10. aparelho locomotor. anti-inflamatórios não esteróides. anti-inflamatórios não esteróides para uso tópico" (Figure 42). Due to this fact, these products, despite having active substances for the same indications, are appropriate for various intervals of age depending on its dosage and pharmaceutical form.

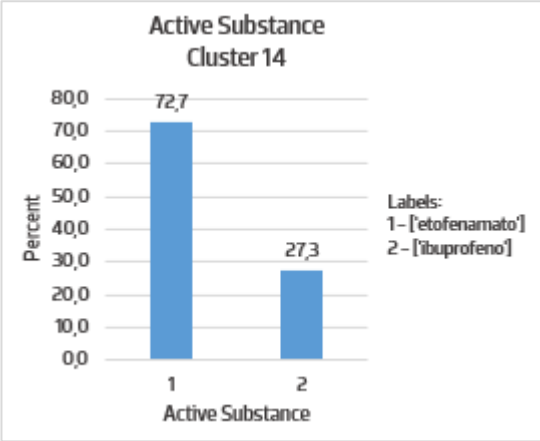


***Figure 42** – Percentage of the variable Active Substance in cluster 14 by K-means.*

The variables that discriminate cluster 15, with few modes, are Interactions and Contraindications (Figure 43). This cluster is characterised by a high percentage of the value "não foram reportadas" in the Interactions variable. The Contraindications variable presents a high percentage of "hipersensibilidade excipiente". Pregnancy and Breastfeeding describes the products of this cluster as contraindicated during these periods (Tables 16 and 17).

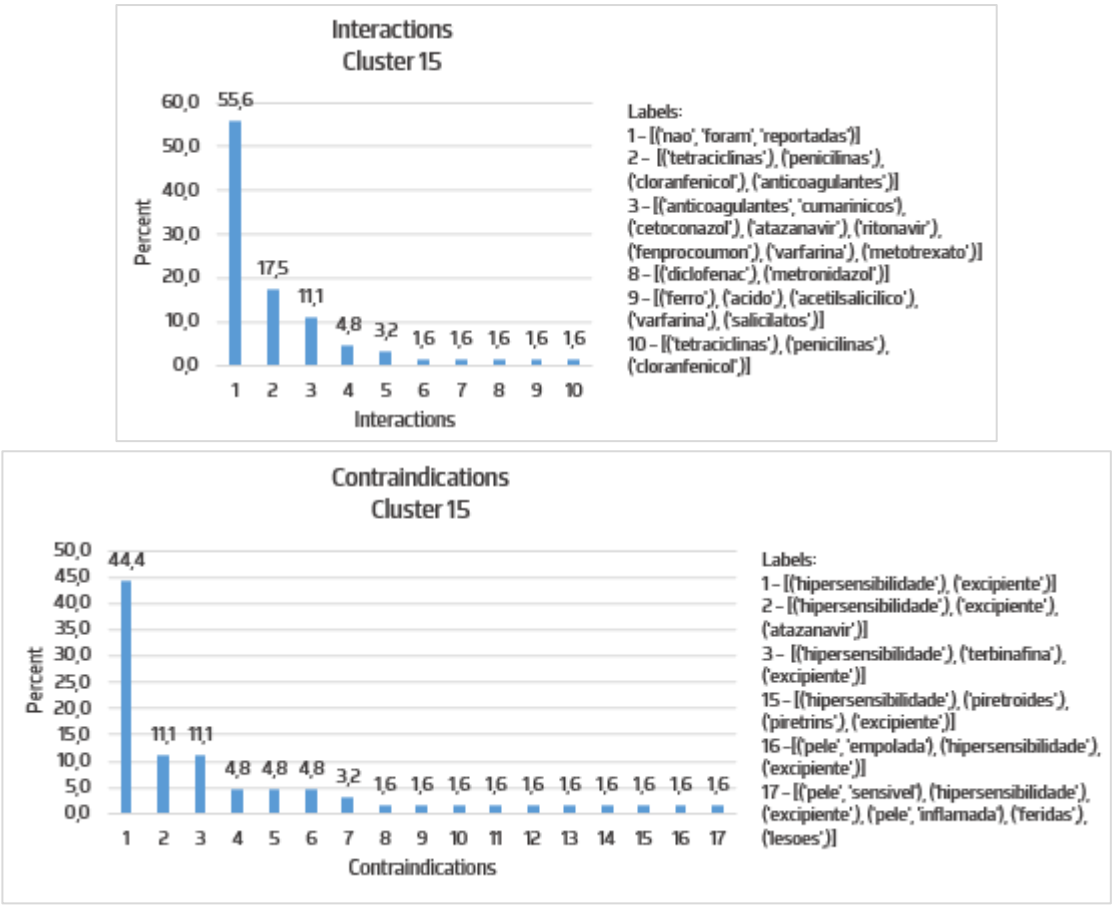


Figure 43 *– Percentage of the variables Interactions and Contraindications in cluster 15 by K-means.*

The variables that discriminate cluster 16, with few modes, are Indications and Adverse Effects by Categories (Figure 44). In the Indications variable, this cluster is characterised by a 34,4% of the value "obstipação crónica, encefalopatica porto-sistémica, pré-coma, coma, fezes moles, fissure anal, abcessos anais, pór-operatório ano-retal, hemorróidas". Although this value is not very high compared to other groups, it is important to note that the following values with a higher percentage also represent gastrointestinal diseases. This is supported by the highest value in the Pharmacotherapeutic Group since 54,3% are part of the products indicated for the digestive system, specifically laxatives (Figure 44). 71,7% of this cluster have as adverse effect "doenças gastrointestinais". All these products can be used during Pregnancy and Breastfeeding (Tables 16 and 17).

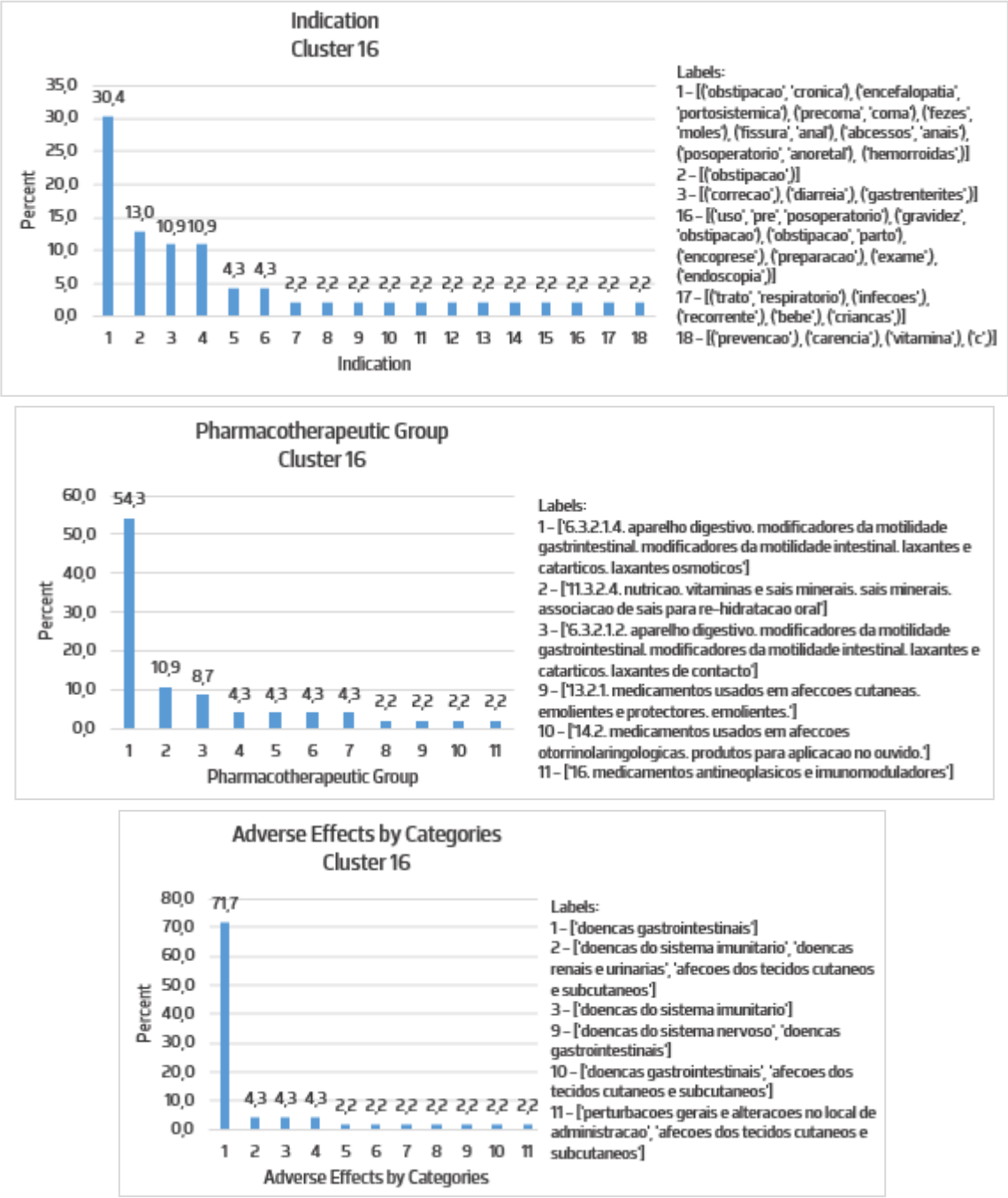


***Figure 44** – Percentage of the variables Indication, Pharmacotherapeutic Group and Adverse Effects by Categories in cluster 16 by K-means.*

Cluster 17 is constituted by two Active Substances, "naproxeno" and "ibuprofeno", but both of them belong to the same Pharmacotherapeutic Group, "9.1.3. aparelho locomotor. anti-inflamatórios não esteróides. anti-inflamatórios não esteróides. Derivados do ácido propiónico" (Figure 45). Due to this fact, these products, despite having active substances for the same indications, are appropriate for various intervals of age depending on its dosage and pharmaceutical form. The active substance "ibuprofeno" is also present in cluster 14. It is divided between the two groups since the products containing this active substance differ from each

other through the other variables, which can be explained by its different pharmaceutical forms and its ingredients list that can impact the products absorption, distribution, metabolism and elimination, creating different adverse effects, interactions, and others.

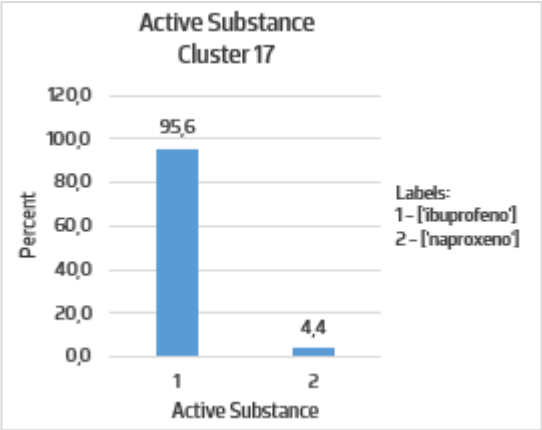


***Figure 45** – Percentage of the variable Active Substance in cluster 17 by K-means.*

Cluster 18 is constituted only by the Active Substance "nicotina", representing 2,5% of the database. Despite having the same active substance, these products can have different adverse effects, interactions and others due to the same reasons previously explained for the products in cluster 17.

The variable that discriminates cluster 19 is Age. In this variable, this cluster is characterised 100% by products that have no indication of the appropriate age to use them. The remaining variables present high dispersion.

The variable Warnings and Precautions discriminates cluster 20 (Figure 46). In this variable, this cluster is characterised by a 33,6% of the value "sem indicação". 97,2% of these products are not recommended during pregnancy, and the remaining ones are not recommended during the first trimester (Table 16). None of the products are recommended during breastfeeding (Table 17).

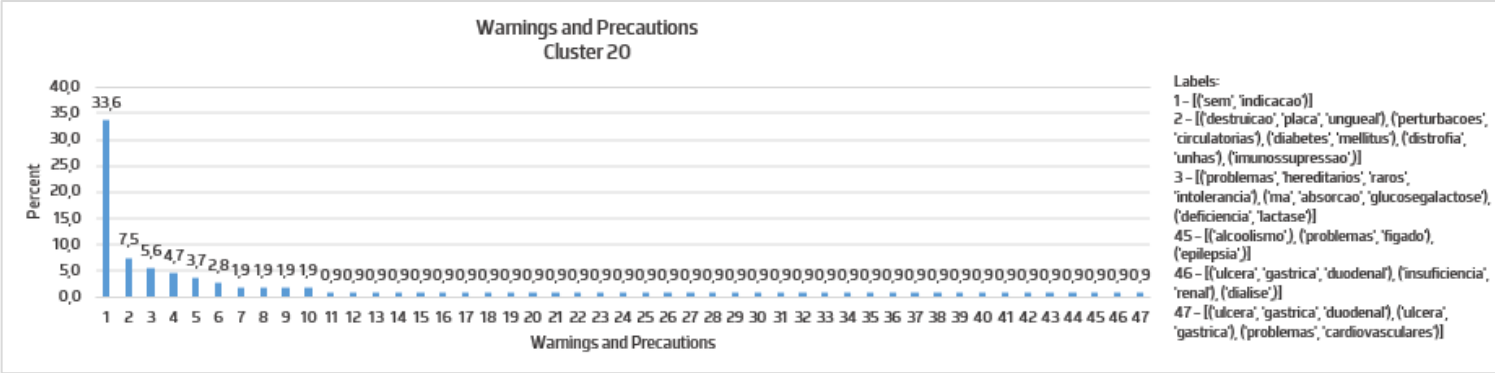
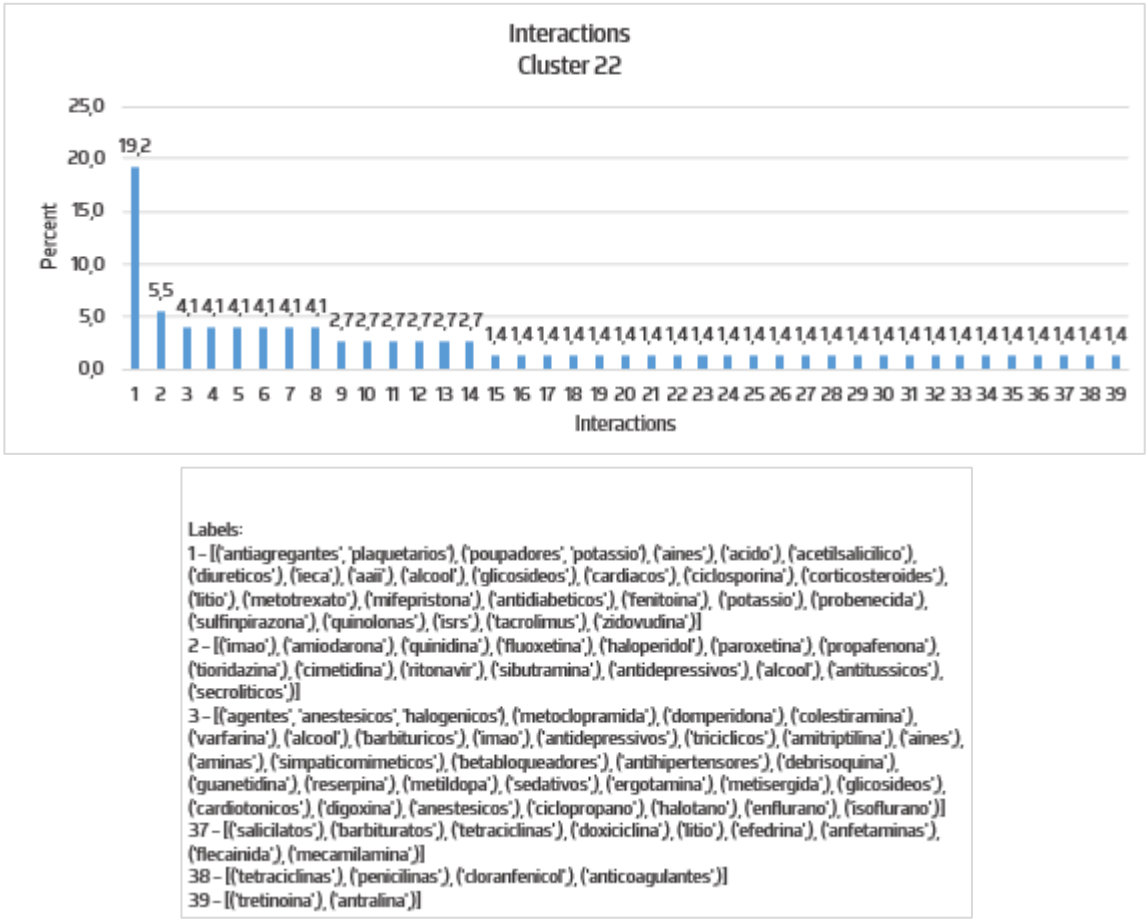


***Figure 46** – Percentage of the variable Warnings and Precautions in cluster 20 by K-means.*

The variables that stand out and discriminate cluster 21 are only Pregnancy and Breastfeeding. 84,0% of the products present in this cluster are allowed to be used during pregnancy, and 100% are allowed during breastfeeding (Tables 16 and 17).

The variable that discriminates cluster 22 is Interactions (Figure 47). It has various values but the value "antiagregantes plaquetários, poupadores potássio, aines, ácido acetilsalicílico, diuréticos, ieca, aaii, álcool, glicosídeos cardíacos, ciclosporina, corticosteroides, lítio, metotrexato, mifepristona, antidiabéticos, fenitoína, probenecida, sulfinpirazona, quinolonas, isrs, tacrolimus, zidovudina" stands out. All these products are contraindicated during Pregnancy and Breastfeeding (Tables 16 and 17).



Labels:
1 – [('antiagregantes', 'plaquetarios'), ('poupadores', 'potassio'), ('aines'), ('acido'), ('acetilsalicilico'), ('diureticos'), ('ieca'), ('aaii'), ('alcool'), ('glicosideos'), ('cardiacos'), ('ciclosporina'), ('corticosteroides'), ('litio'), ('metotrexato'), ('mifepristona'), ('antidiabeticos'), ('fenitoina'), ('potassio'), ('probenecida'), ('sulfinpirazona'), ('quinolonas'), ('isrs'), ('tacrolimus'), ('zidovudina')]
2 – [('imao'), ('amiodarona'), ('quinidina'), ('fluoxetina'), ('haloperidol'), ('paroxetina'), ('propafenona'), ('tioridazina'), ('cimetidina'), ('ritonavir'), ('sibutramina'), ('antidepressivos'), ('alcool'), ('antitussicos'), ('secroliticos')]
3 – [('agentes', 'anestesicos', 'halogenicos'), ('metoclopramida'), ('domperidona'), ('colestiramina'), ('varfarina'), ('alcool'), ('barbituricos'), ('imao'), ('antidepressivos'), ('triciclicos'), ('amitriptilina'), ('aines'), ('aminas'), ('simpaticomimeticos'), ('betabloqueadores'), ('antihipertensores'), ('debrisoquina'), ('guanetidina'), ('reserpina'), ('metildopa'), ('sedativos'), ('ergotamina'), ('metisergida'), ('glicosideos'), ('cardiotonicos'), ('digoxina'), ('anestesicos'), ('ciclopropano'), ('halotano'), ('enflurano'), ('isoflurano')]
37 – [('salicilatos'), ('barbituratos'), ('tetraciclinas'), ('doxiciclina'), ('litio'), ('efedrina'), ('anfetaminas'), ('flecainida'), ('mecamilamina')]
38 – [('tetraciclinas'), ('penicilinas'), ('cloranfenicol'), ('anticoagulantes')]
39 – [('tretinoina'), ('antralina')]

***Figure 47** – Percentage of the variable Interactions in cluster 22 by K-means.*

Clusters 6, 9 and 18 are characterised by the presence of only one active substance. Although these groups are homogeneous, with few modes or even just one in some variables, this is not the ideal classification for pharmaceutical counselling. Clusters 7, 13, 14 and 17 are characterised by the presence of only two active substances belonging to the same Pharmacotherapeutic Group. This is a better classification since it includes products with more than one Active Substance.

The clustering occurred mainly according to the variables Pregnancy and Breastfeeding, as previously discussed, the second ones with the higher weight applied. This can be explained by the highest number of values in the variable to which the higher weight was applied, Contraindications. Although Age and Interactions are following regarding the weights applied to them when calculating the Jaccard distance, few groups are characterised by homogeneity, with few modes on these variables. This may have occurred due to the same fact as the variable Contraindications: the presence of many values. On the other hand, the variables that do not characterise are Indications, Pharmaceutical Form and Adverse Effects by Categories.

The Pharmacotherapeutic Group is a classification of the products according to their indications. This is an important variable since it allows a faster identification of a product according to the patient's objective. Therefore, Table 18 presents the relative frequencies of the main group of the variable Pharmacotherapeutic Group.

***Table 18** – Percentage of the variable Pharmacotherapeutic Group present in all clusters by K-means.*

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
N/A	7,4	100,0	0	0	0	0	0	0	0	0	70,0
G1	0	0	3,2	4,8	1,1	0	0	0	0	7,9	0
G2	7,4	0	3,2	3,3	0	100,0	0	10,3	0	25,0	0
G3	11,1	0	0	3,3	8,8	0	0	0	0	0	0
G4	0	0	0	8,2	0	0	0	0	0	0	0
G5	3,7	0	0	0	1,1	0	100,0	12,9	0	5,3	0
G6	14,8	0	6,4	16,4	27,5	0	0	15,4	0	27,5	10,0
G7	0	0	3,2	0	4,4	0	0	7,7	0	1,3	0
G8	0	0	0	0	0	0	0	0	0	11,8	0
G9	7,4	0	21,0	0	4,4	0	0	0	100,0	5,3	10,0
G10	0	0	0,0	0	1,1	0	0	0	0	0	0
G11	3,7	0	1,6	19,6	0	0	0	5,2	0	0	0
G13	44,4	0	29,0	37,6	44,0	0	0	41,1	0	3,9	10,0
G14	0	0	32,3	0	3,3	0	0	2,6	0	11,8	0
G15	0	0	0	6,5	4,4	0	0	5,2	0	0	0
G16	0	0	0	0	0	0	0	0	0	0	0
G20	0	0	0	0	0	0	0	0	0	0	0

Legend: C: Cluster; N/A: Not Available; G1: 1. Anti-Infectious Agents; G2: 2. Central Nervous System; G3: 3. Cardiovascular System; G4: 4. Blood; G5: 5. Respiratory System; G6: 6. Gastrointestinal System; G7: 7. Genitourinary System; G8: 8. Hormones and Drugs to Treat Endocrine Diseases; G9: 9. Locomotor System; G10: 10. Anti-Allergic Medication; G11: 11. Nutrition; G13: 13. Drugs for Skin Disorders; G14: 14. Drugs Used in Otorhinolaryngolical Disorders; G15: 15. Drugs for Eye Disorders; G16: 16. Antineoplastic Drugs and Immune-Modulators; G20: 20. Dressing Material, Local Hemostats, Medicinal Gases and Other Products.

***Table 18** – Percentage of the variable Pharmacotherapeutic Group present in all clusters by K-means (cont.).*

	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22
N/A	0	0	0	0	0	0	0	0	4,7	0	0
G1	0	0	0	1,6	0	0	0	0	0	0	0
G2	9,2	0	0	6,3	0	0	100,0	0	10,3	12,0	19,2
G3	0	0	0	1,6	0	0	0	3,7	10,3	0	0
G4	0	0	0	0	4,3	0	0	3,7	0	0	0
G5	0	100,0	0	0	0	0	0	0	18,7	0	24,7
G6	32,1	0	0	27,0	67,3	0	0	14,8	18,6	42,0	31,4
G7	0	0	0	6,4	0	0	0	7,4	4,6	20,0	0
G8	0	0	0	0	0	0	0	0	0	0	0
G9	10,7	0	100,0	23,8	0	100,0	0	0	3,8	0	12,3
G10	33,8	0	0	0	0	0	0	0	7,5	0	2,7
G11	4,6	0	0	1,6	17,4	0	0	11,1	0	8,0	0
G13	4,5	0	0	28,6	6,5	0	0	44,4	13,1	16,0	4,1
G14	4,6	0	0	1,6	2,2	0	0	0	3,7	0	5,5
G15	0	0	0	1,6	0	0	0	11,1	2,8	2,0	0
G16	0	0	0	0	2,2	0	0	3,7	0	0	0
G20	0	0	0	0	0	0	0	0	1,9	0	0

Legend: C: Cluster; N/A: Not Available; G1: 1. Anti-Infectious Agents; G2: 2. Central Nervous System; G3: 3. Cardiovascular System; G4: 4. Blood; G5: 5. Respiratory System; G6: 6. Gastrointestinal System; G7: 7. Genitourinary System; G8: 8. Hormones and Drugs to Treat Endocrine Diseases; G9: 9. Locomotor System; G10: 10. Anti-Allergic Medication; G11: 11. Nutrition; G13: 13. Drugs for Skin Disorders; G14: 14. Drugs Used in Otorhinolaryngolical Disorders; G15: 15. Drugs for Eye Disorders; G16: 16. Antineoplastic Drugs and Immune-Modulators; G20: 20. Dressing Material, Local Hemostats, Medicinal Gases and Other Products.

It is possible to note that 8 of the 22 clusters previously described are entirely separated according to this variable. Despite that, there are clusters that belong to the same Pharamcotherapeutic Group, such as clusters 6 and 18 but were separated due to the remaining variables, mainly Pregnancy and Breastfeeding, as shown in Table 19.

***Table 19** – Clusters separation according to the products' Pharmacotherapeutic Group and their safety during Pregnancy and Breastfeeding.*

Cluster	Pharmacotherapeutic Group	Safety During Pregnancy and Breastfeeding
6	2. Nervous Central System	P/B: Can be used
18		P/B: Not recommended
7	5. Respiratory System	P/B: Not recommended
13		P/B: Can be used with caution
9	9. Locomotor System	P: Not recommended in the third trimester B: Can be used with caution
14		P: Not recommended B: Can be used
17		P/B: Contraindicated

Legend: P: Pregnancy; B: Breastfeeding.

Dermocosmetics also have been clustered in the same group without any division despite having different indications. To solve this problem, the dermocosmetics clustering may benefit if these products are in a separate database from the medication since its criteria may differ.
For the methods single linkage, complete linkage, median linkage, centroid linkage, ward linkage and average linkage, the hierarchical clustering methods, the obtained results presented high dispersion of the products within the clusters.

4.5. Distance Function Enhancing the Importance of the Pharmacotherapeutic Group

When a patient goes to a pharmacy and presents a specific problem, usually that restrains the choice of the product to one that belongs to a particular pharmacotherapeutic group. For example, if a patient has allergy symptoms, the products that can be recommended for this situation belong to the pharmacotherapeutic group "10. Medicação Antialérgica". Due to this fact, the ideal clustering with pharmaceutical products to assist professionals in counselling would occur if there was a previous separation by this variable. There are 20 pharmacotherapeutic groups. However, in this database, the products only belong to 16 different groups and dermocosmetics are not classified into these groups.
The previous clustering was not carried out according to the variable Pharmacotherapeutic Group since, initially, it was created and tested a general distance function that would not require filtering by the pharmacotherapeutic group. However, as the clusters obtained were not consistent with this variable, it was decided to increase the value of its weight in order to stand out compared to the other variables. To test this, it was attributed to the variable Pharmacotherapeutic Group a higher weight when calculating weighted Jaccard Index (weight = 20) than the previous weight attributed (weight = 4,27). For all the other variables the same weights were maintained.

The weighted Jaccard index for a pair of products (x, y) was calculated using Equation (13) as previously, but with the Pharmacotherapeutic Group's weight as 20:

$$\begin{aligned} J_w(x,y) = {} & J_{Active\ Substance} \times 4{,}27 + J_{Pharmaceutical\ Form} \times 1{,}97 + J_{Indication} \times 4{,}27 \\ & + J_{Age} \times 4{,}55 + J_{Adverse\ Effects} \times 1{,}32 + J_{Interactions} \times 4{,}28 \\ & + J_{Contraindications} \times 6{,}53 + J_{Warnings\ and\ Precautions} \times 2{,}98 \\ & + J_{Pregnancy} \times 6{,}35 + J_{Breastfeeding} \times 6{,}35 \\ & + J_{Pharmacotherapeutic\ Group} \times 20 \end{aligned}$$

It was created a heatmap with the final indexes (Figure 48).

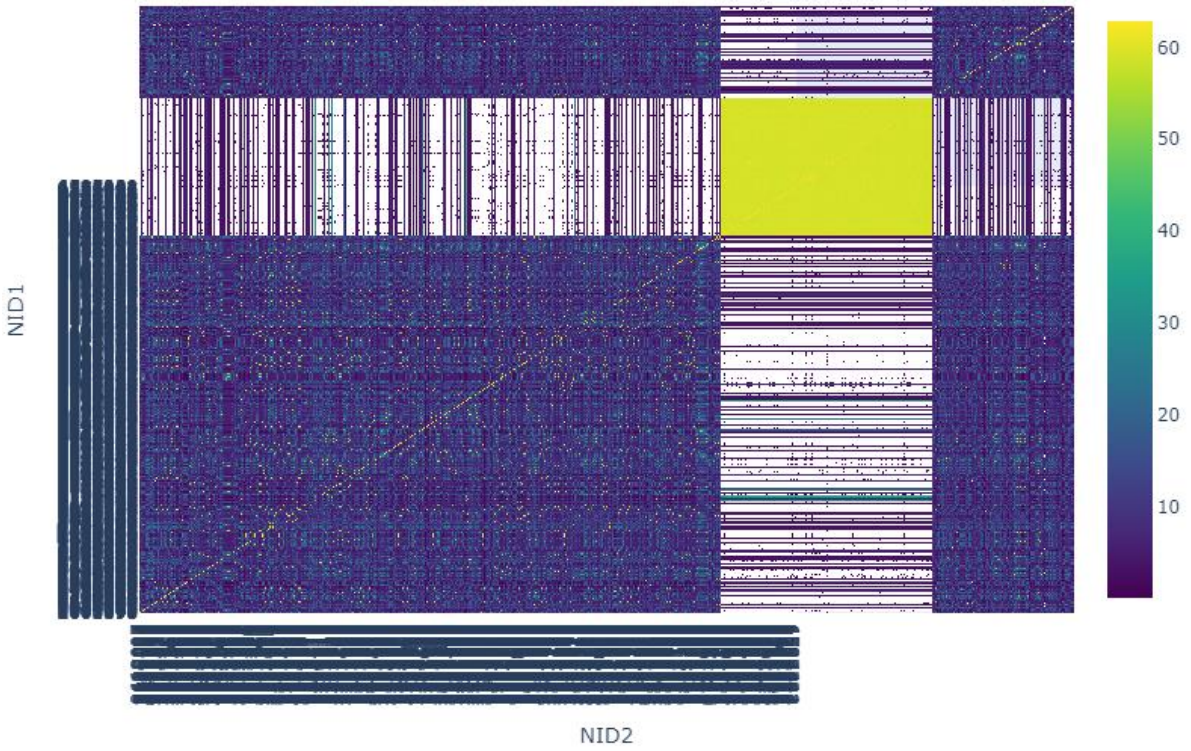


***Figure 48** – Heatmap representing the Jaccard index for each pair of products. Pharmacotherapeutic Group's weight: 20. (NID: Identification Number of the product).*

As in the previous heatmap, it is possible to visualise the diagonal representing the similarity of a product with himself, with the higher value of similarity and a group of products in the upper right part of Figure 48 representing dermocosmetics. As previously, the Jaccard distance (J_δ) was calculated through the weighted Jaccard index using Equation 14:

$$J_\delta\,(x,y) = (4{,}27 + 1{,}97 + 4{,}27 + 4{,}55 + 1{,}32 + 4{,}28 + 6{,}53 + 2{,}98 + 6{,}35 + 6{,}35 + 20) - J_w$$

$$J_\delta\,(x,y) = 62{,}87 - J_w$$

A distance matrix was created.

4.5.1. Recommendation Groups Enhancing the Importance of the Pharmacotherapeutic Group

The same clustering techniques were performed: hierarchical clustering with the methods single linkage, complete linkage, median linkage, centroid linkage, ward linkage, and average linkage; and non-hierarchical clustering, K-means. The evaluation of the optimal number of clusters for each method was performed using the same stability measures SS, DBS, and CHS and the elbow

method. It was also incorporated the domain knowledge in order to find the optimal number of clusters, and for each hierarchical method it was also obtained a dendrogram (Figure 49). The dendrograms refering to the methods single linkage, complete linkage and median linkage maintained the same number of clusters formed as in Figure 17, three clusters. Ward linkage also obtained two clusters as it had previously before the increase of the variable Pharmacotherapeutic Group weight. On the other hand, centroid linkage, which previously formed three clusters according to the dendrogram present in Figure 17, has now formed only two clusters. The inverse happened to average linkage method, which had previously formed 2 clusters, but now forms three clusters. As stated before, these numbers of clusters would not characterise these products in order for them to have utility in pharmaceutical practice.

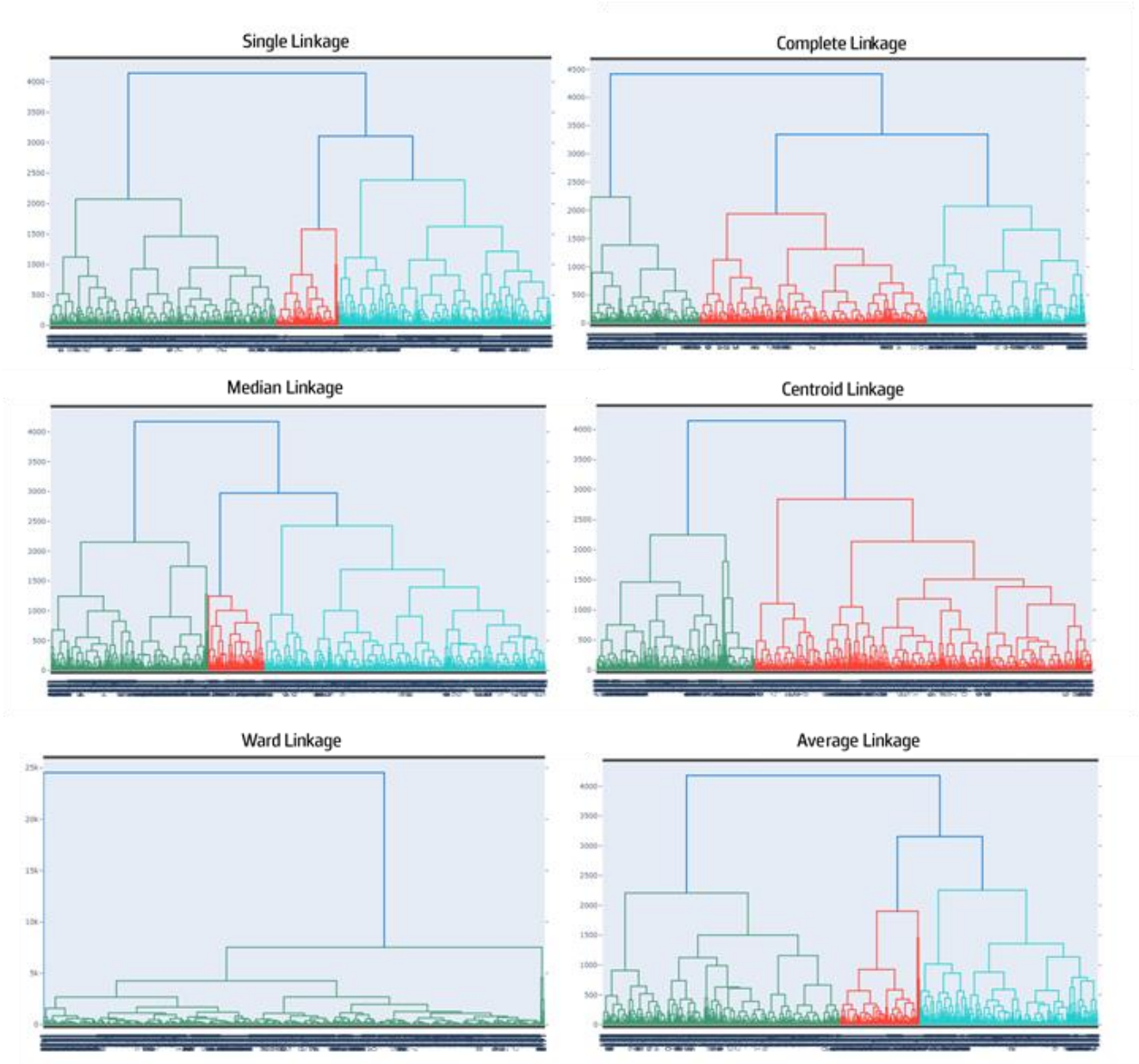


Figure 49 *– Dendrograms obtained for the hierarchical clustering methods (single linkage, complete linkage, median linkage, centroid linkage, ward linkage, and average linkage). Pharmacotherapeutic Group's weight: 20.*

Figure 50 shows the results obtained for the stability measures analysed for the method single linkage. Therefore, CHS is the highest at 17 clusters, the DBS was one of the lowest at 15 clusters, and the highest SS was reached at 28 clusters. The single linkage method was performed with 25 clusters since the groups formed when more clusters were selected, as suggested by SS, were characterised by the presence of only one active substance. As previously stated, this does not have an application in pharmaceutical counselling. Thus, were formed 25 clusters that obtained more clinically relevant groups.

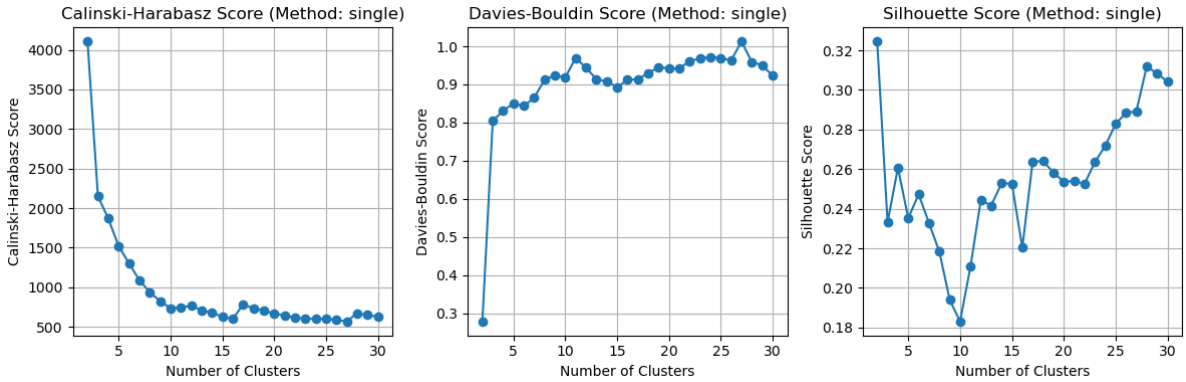


Figure 50 *– Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for the single linkage method. Pharmacotherapeutic Group's weight: 20. For 25 clusters: CHS = 620, DBS = 0,968, SS = 0,283.*

In the complete linkage method, CHS had the highest value at 10 clusters, DBS was one of the lowest at 12 clusters, and the highest SS was reached at 30 clusters, presenting some peaks at 10 and 23 (Figure 51). The complete linkage method was performed with 23 clusters, the higher and the lower numbers of clusters would have an application in pharmaceutical counselling, as explained before.

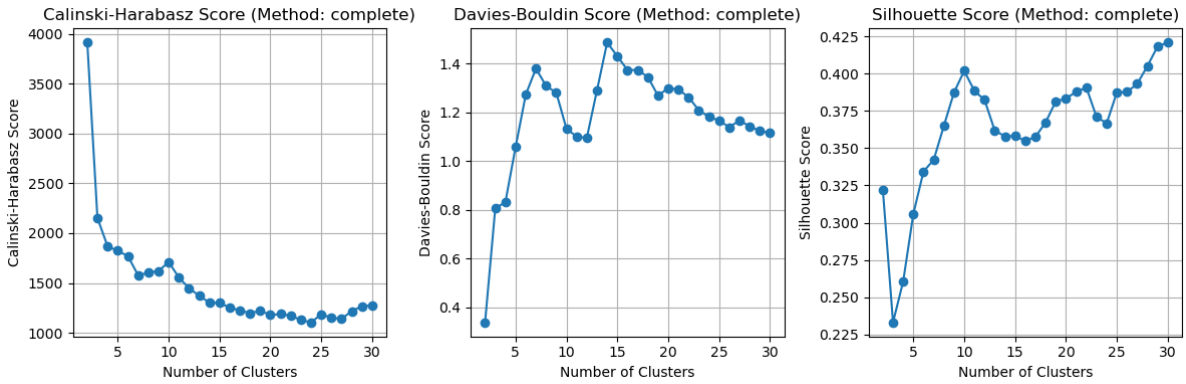


Figure 51 *– Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for the complete linkage method. Pharmacotherapeutic Group's weight: 20. For 23 clusters: CHS = 1183, DBS = 1,206, SS = 0,392.*

In the median linkage method, CHS had its highest value at 11 clusters, DBS had its lowest values in the range of 26 to 30 clusters, and the highest SS was reached at 29 clusters (Figure 52). The median linkage method was performed with 26 clusters since two measures had good outcomes for that value.

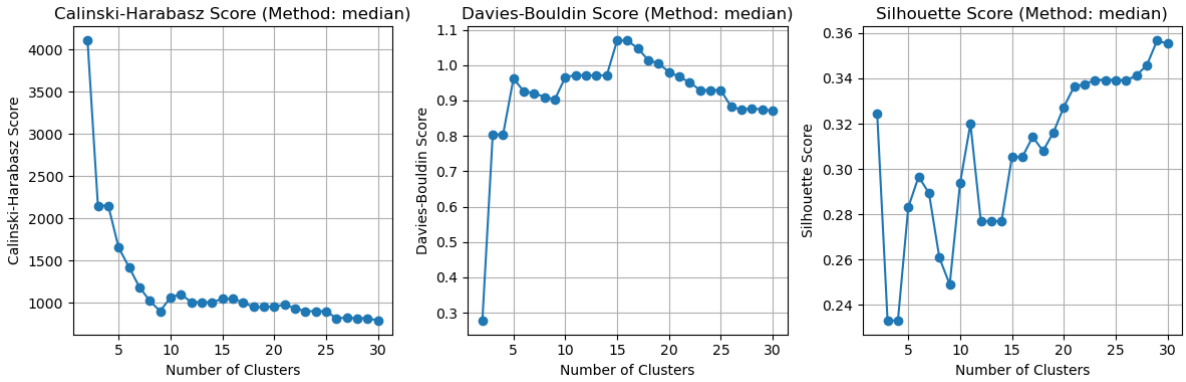


***Figure 52** – Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for the median linkage method. Pharmacotherapeutic Group's weight: 20. For 26 clusters: CHS = 802, DBS = 0,880, SS = 0,339.*

In the centroid linkage method, CHS had a peak at 16 clusters, DBS had its lower value at 23 clusters, and the highest SS was reached at 30 clusters (Figure 53). The centroid linkage method was performed with 22, which has a similar SS value, due to the same fact presented as single linkage. When more clusters were selected, as suggested by SS, they were characterised by the presence of only one active substance, which does not have an application in pharmaceutical counselling.

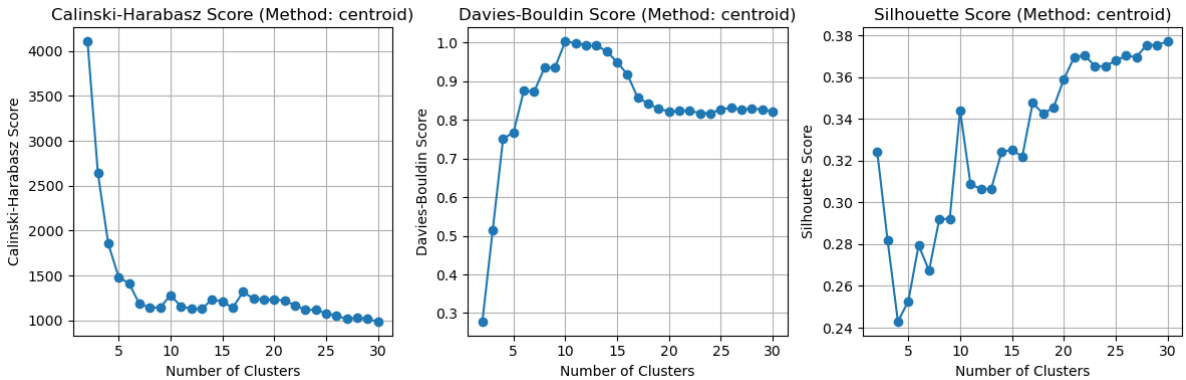


***Figure 53** – Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for the centroid linkage method. Pharmacotherapeutic Group's weight: 20. For 22 clusters: CHS = 1211, DBS = 0,824, SS = 0,372.*

In the ward linkage method, CHS presented the highest values for the lower numbers of clusters, DBS had its lowest value at 14 clusters, and the highest SS was reached at 30 clusters (Figure 54). It was tested the higher number of clusters as suggested by the SS, however, the same problem

as in the single linkage and centroid linkage emerged, and the groups formed were characterised by the presence of only one active substance. Thus, the ward linkage method was performed with 25 clusters.

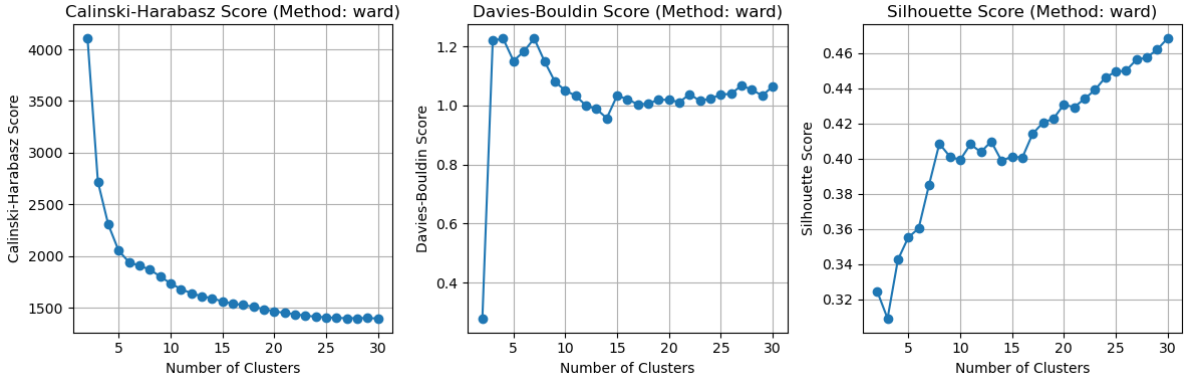


Figure 54 *– Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for the ward linkage method. Pharmacotherapeutic Group's weight: 20. For 25 clusters: CHS = 1403, DBS = 1,033, SS = 0,450.*

In the average linkage method, CHS had a peak at 9 clusters, DBS had its lowest value at 17 clusters, and the highest SS was reached at 30 clusters (Figure 55). The same problem occurred in this case as in the single linkage, centroid linkage and ward linkage methods. The groups obtained by selecting such a high cluster number were not clinically relevant to assist pharmaceutical counselling. Therefore, the average linkage method was performed with 25 clusters.

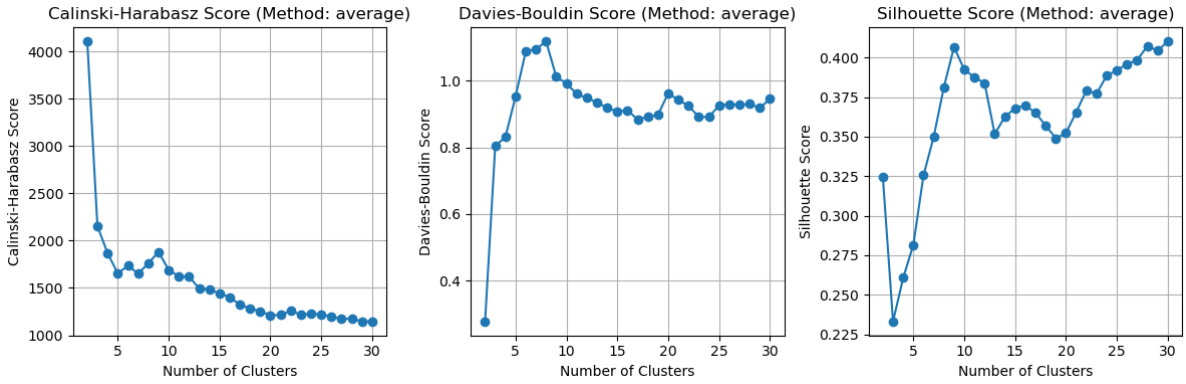


Figure 55 *– Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for the average linkage method. Pharmacotherapeutic Group's weight: 20. For 25 clusters: CHS = 1234, DBS = 0,928, SS = 0,393.*

Regarding the non-hierarchical clustering method used, K-means, CHS presented the highest values for the lower numbers of clusters, DBS had its lowest value at 17 clusters, and the highest SS was reached at 30 clusters (Figure 56). The elbow method does not present a clear inflexion point, therefore, it is not possible to determine the optimal number of clusters according to this

method (Figure 55). K-means was performed with 24 clusters since this value had one of the highest SS.

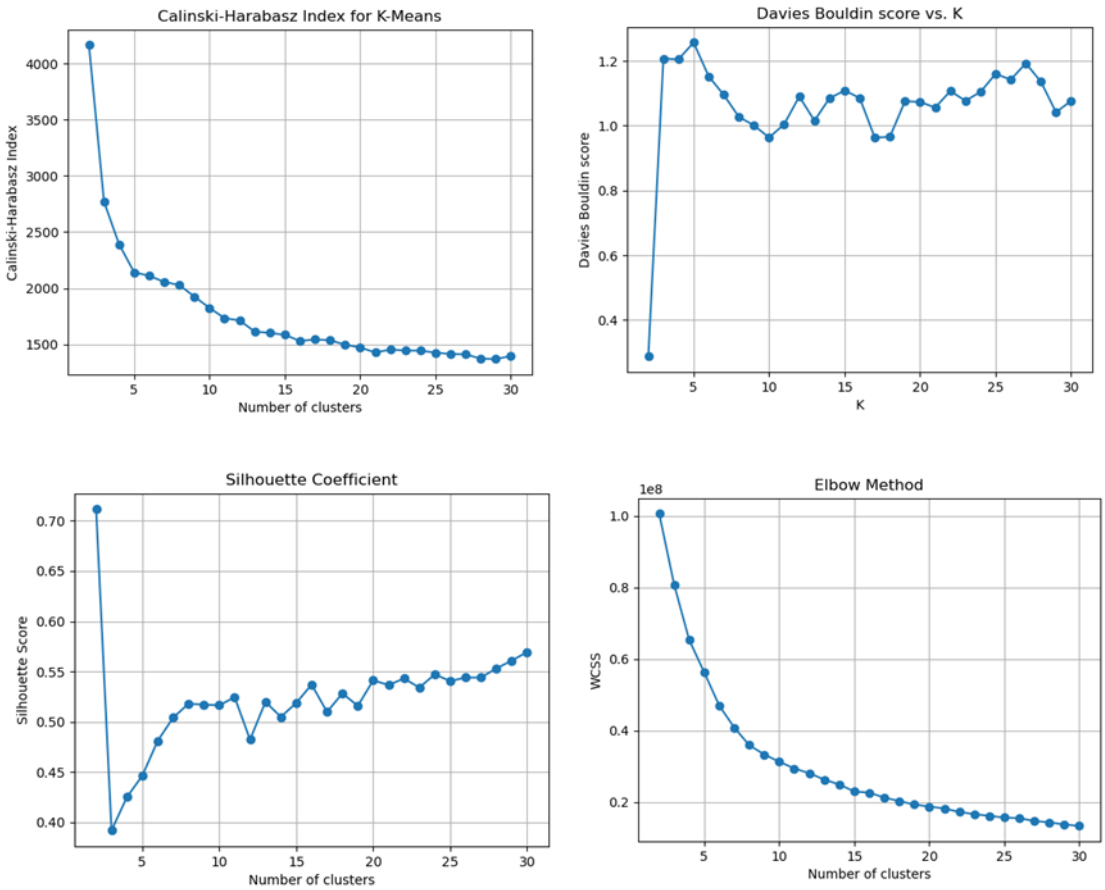


***Figure 56** – Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for K-means. Pharmacotherapeutic Group's weight: 20. For 24 clusters: CHS = 1464, DBS = 1,091, SS = 0,543.*

Table 20 contains all the results of CHS, DBS and SS for all the hierarchical and non-hierarchical methods tested and for the initial weight and the adjusted weight of the variable Pharmacotherapeutic Group. CHS were higher before except for the centroid linkage method (Figure 57). The best results of this score were obtained using K-means compared to the other methods in both weights (Figure 57). In the DBS it is verified that all the methods increased the value with the adjusted weight, obtaining a worst result than before, since the lower the value, the better (Figure 58). Single linkage presents the best DBS with the initial weight and centroid linkage with the adjusted weight (Figure 58). SS obtained before for the distance function with the initial weights according to the experts consultation results, were lower than the values obtained for all the methods with the adjusted weight, except for single linkage (Figure 59). K-means presents the higher values of SS both with the initial and the adjusted weights (Figure 59). The obtained results of SS and CHS indicate that K-means forms the most compact clusters (similar

elements within each cluster) and separated clusters, however, DBS indicates the clusters formed by this method are less homogeneous than the ones formed with the single linkage and centroid linkage for the initial and the adjusted weight, respectively.

***Table 20** – Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS) and Silhouette Score (SS) for all the clustering methods tested for the initial weight (W = 4,27) and the adjusted weight (W = 20) of the variable Pharmacotherapeutic Group.*

Method		CHS		DBS		SS	
		W = 4,27	W = 20	W = 4,27	W = 20	W = 4,27	W = 20
Single Linkage	#C	20	25	20	25	20	25
	Value	910	620	0,741	0,968	0,287	0,283
Complete Linkage	#C	25	23	25	23	25	23
	Value	1534	1183	1,068	1,206	0,347	0,392
Median Linkage	#C	18	26	18	26	18	26
	Value	1066	802	0,761	0,880	0,296	0,339
Centroid Linkage	#C	25	22	25	22	25	22
	Value	1038	1211	0,770	0,824	0,331	0,372
Ward Linkage	#C	25	25	25	25	25	25
	Value	1794	1403	1,017	1,033	0,388	0,450
Average Linkage	#C	25	25	25	25	25	25
	Value	1397	1234	0,922	0,928	0,350	0,393
K-Means	#C	22	24	22	24	22	24
	Value	1876	1464	1,012	1,091	0,538	0,543

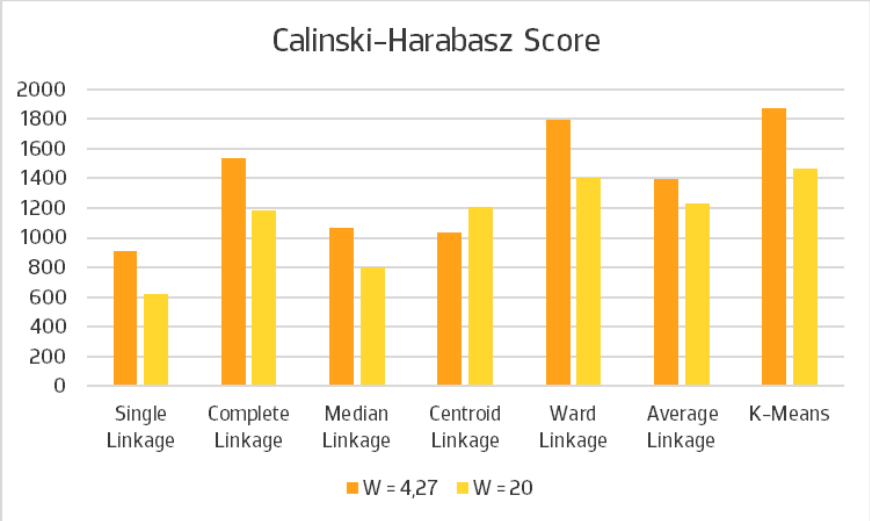


***Figure 57** – Calinski-Harabasz Score for all the clustering methods tested for the initial weight (W = 4,27) and the adjusted weight (W = 20) of the variable Pharmacotherapeutic Group.*

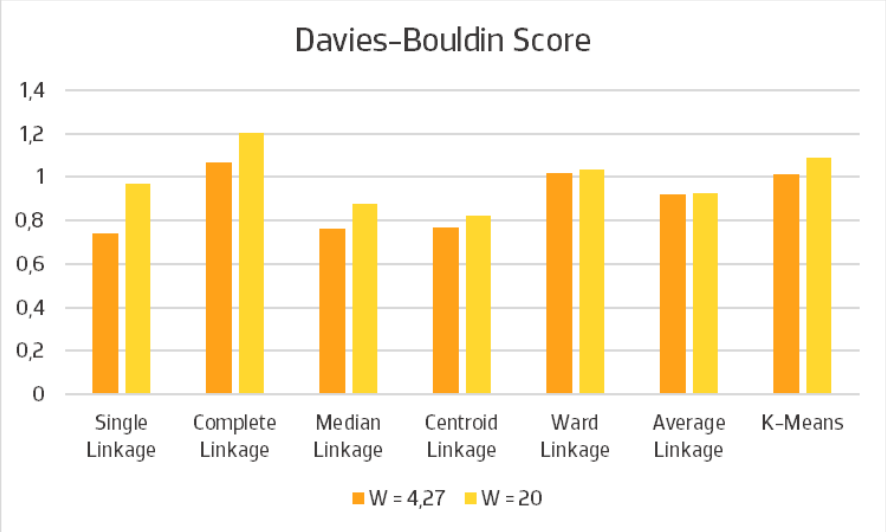


***Figure 58** – Davies-Bouldin Score for all the clustering methods tested for the initial weight (W = 4,27) and the adjusted weight (W = 20) of the variable Pharmacotherapeutic Group.*

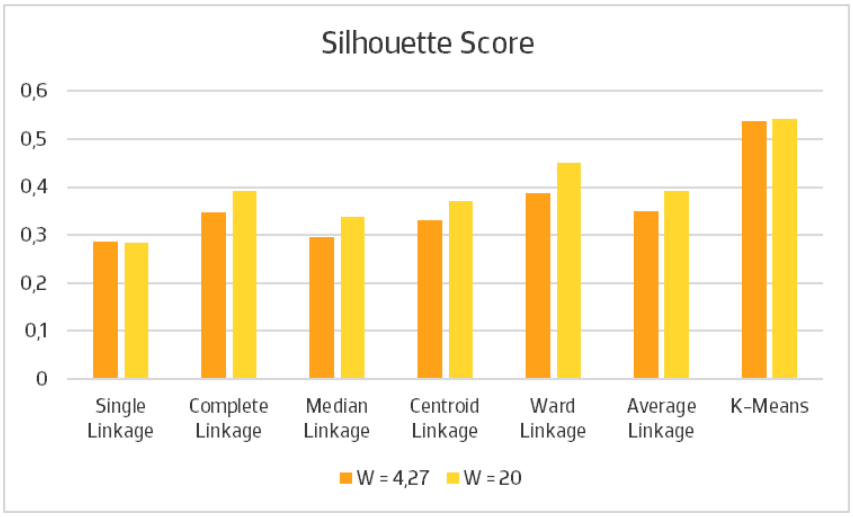


***Figure 59** – Silhouette Score for all the clustering methods tested for the initial weight (W = 4,27) and the adjusted weight (W = 20) of the variable Pharmacotherapeutic Group.*

The relative frequencies of the formed clusters and the products that each contain in all the 7 methods is shown in Figures 60 to 66. It is important to note that all methods generated a cluster with all dermocosmetics included, which represent 22,7% of all products in the database. No difference occurred when clustering these products since they do not belong to a pharmacotherapeutic group, and all have the value "sem indicação" in that variable. As previously seen in the heatmap of the Jaccard indices, these products have high similarity among

themselves, as well as high dissimilarity with the rest, thus predicting that this grouping would occur.

As happened before, it is possible to see in Figures 60, 62, 63 and 64, the single linkage, median linkage, centroid linkage and average linkage methods formed groups with a very small number of products, with clusters sometimes containing only 1 product, favouring the separation of outliers for no apparent clinical reason. Thus, these techniques are not suitable for this dataset. This provoked that groups with a high number of products were formed that had no homogeneity within clusters. The complete linkage, ward linkage and K-means methods, on the other hand, obtained more homogeneous groups and the one that presented groups of products with greater homogeneity within clusters was K-Means (Figures 61, 65 and 66). K-means obtained the highest SS value of all methods (0,543), including when compared to K-means performed before the weight of the variable Pharmacotherapeutic Group was adjusted (0,538). Although this is not a high value considering SS ranges from 0 to 1, it is the highest obtained and is consistent with the clusters formed since this method presented, once again, the more homogeneous groups with clinical relevance for pharmaceutical counselling.

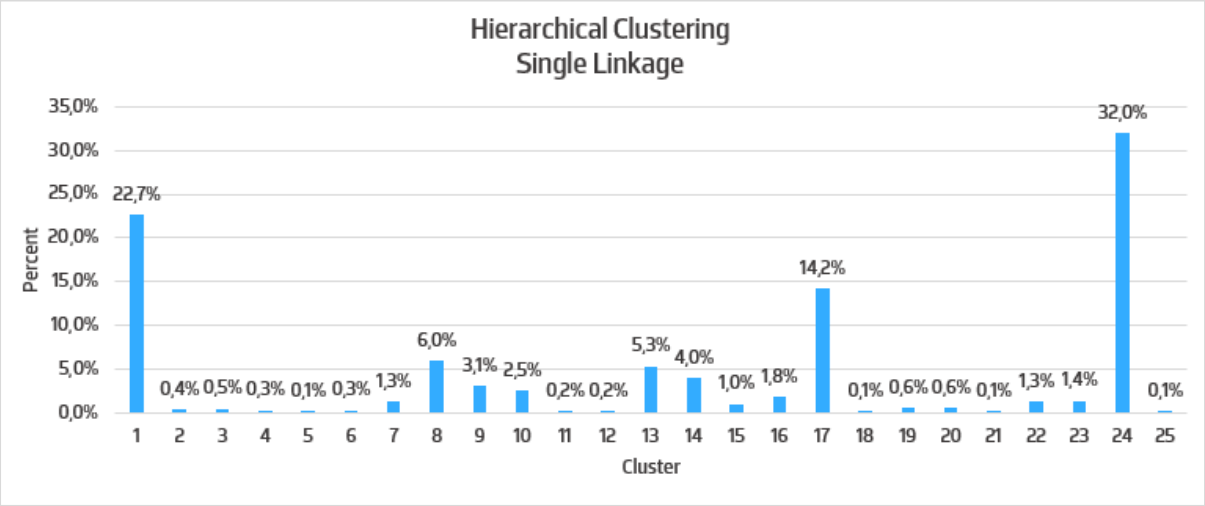


Figure 60 *– Percentage of the clusters formed with single linkage method. Pharmacotherapeutic Group's weight: 20.*

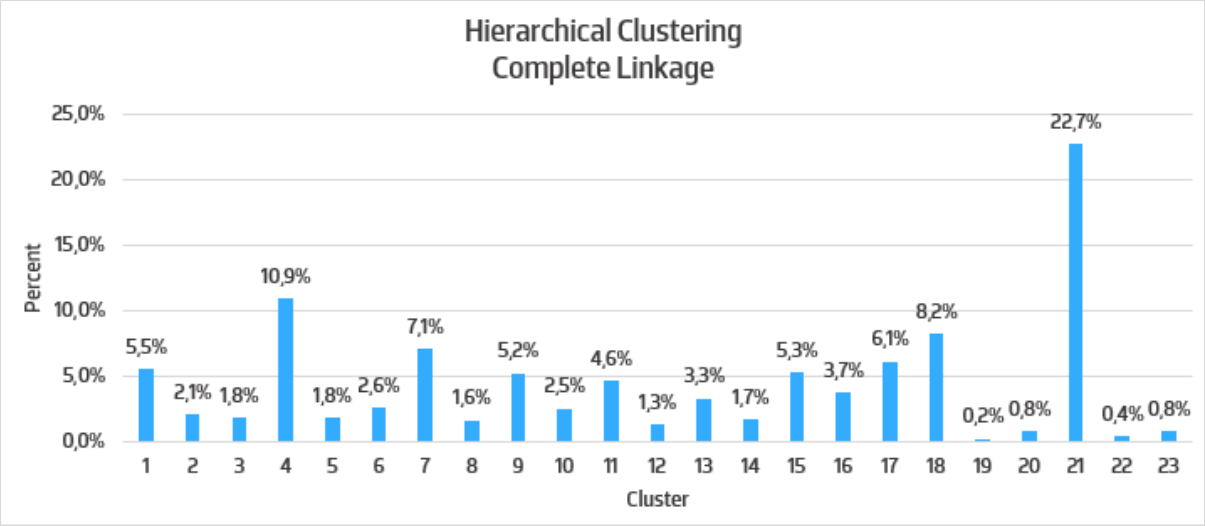


Figure 61 *– Percentage of the clusters formed with complete linkage method. Pharmacotherapeutic Group's weight: 20.*

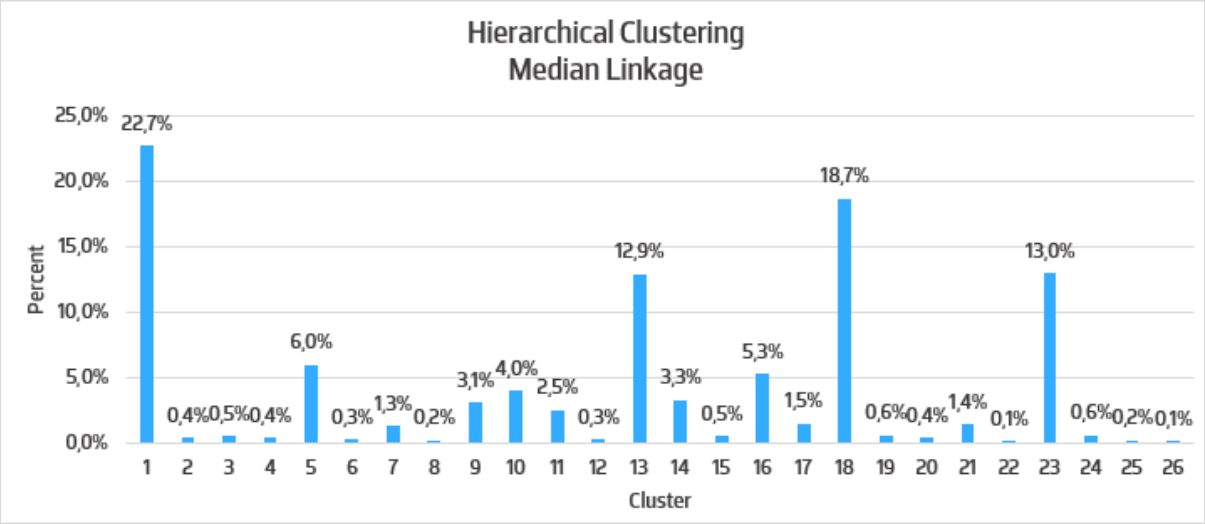


Figure 62 *– Percentage of the clusters formed with median linkage method. Pharmacotherapeutic Group's weight: 20.*

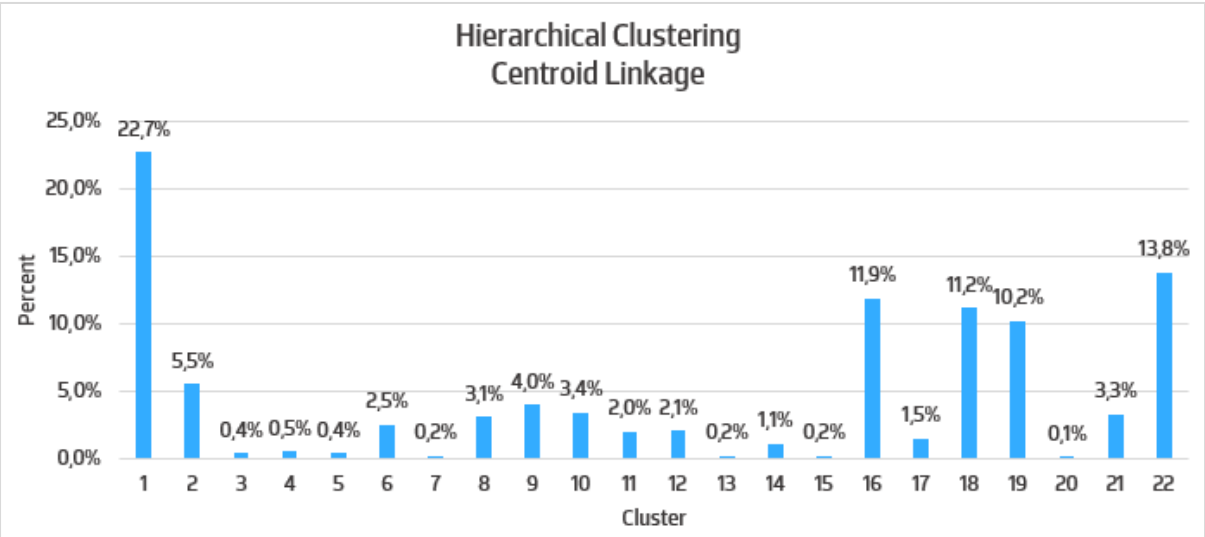


Figure 63 *– Percentage of the clusters formed with centroid linkage method. Pharmacotherapeutic Group's weight: 20.*

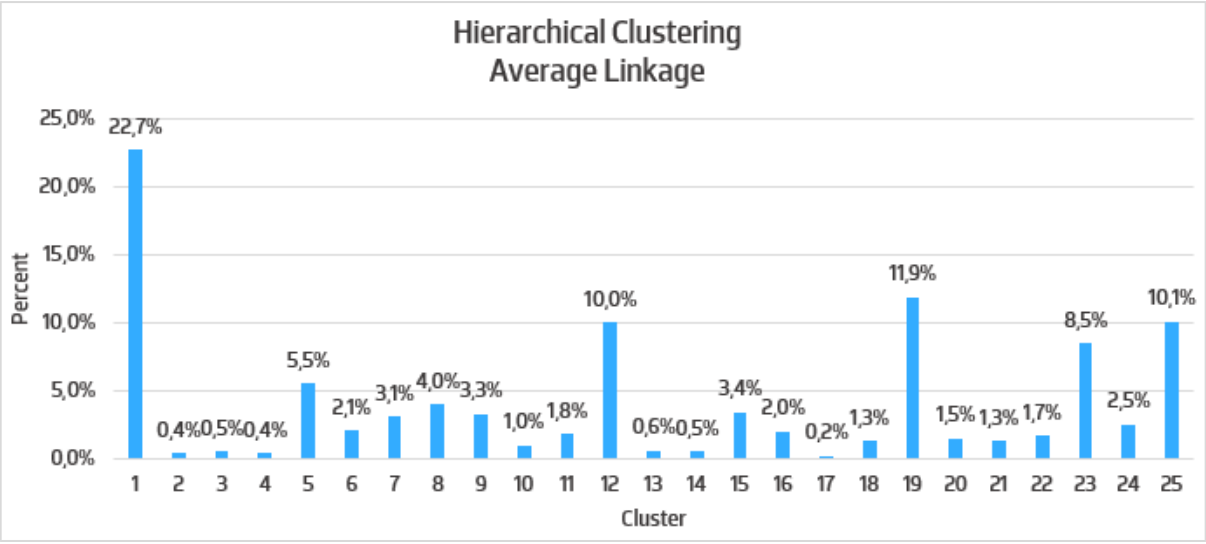


Figure 64 *– Percentage of the clusters formed with average linkage method. Pharmacotherapeutic Group's weight: 20.*

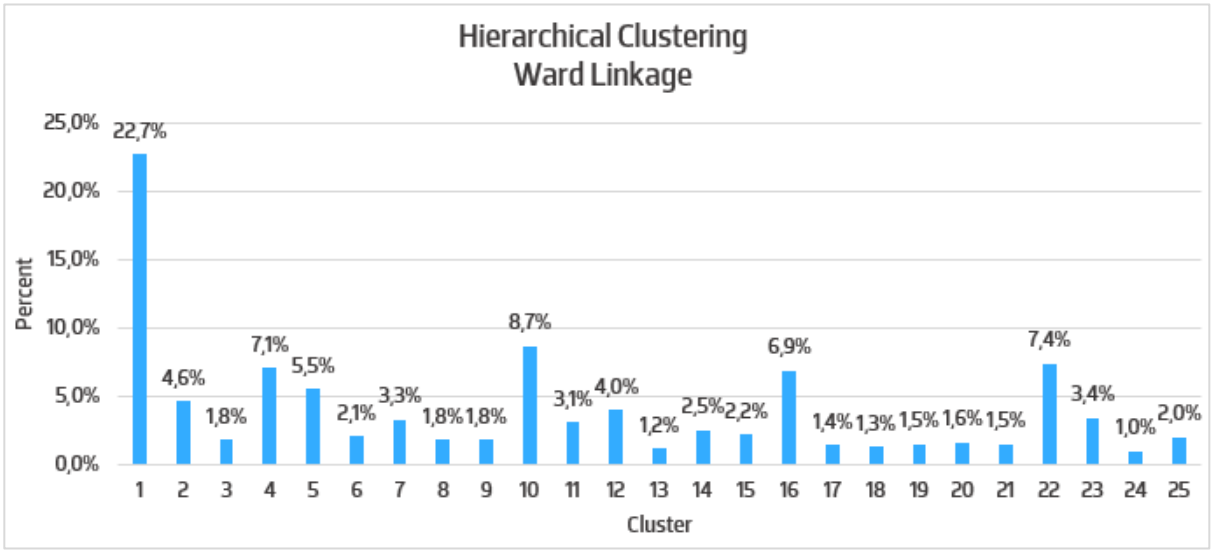


Figure 65 *– Percentage of the clusters formed with ward linkage method. Pharmacotherapeutic Group's weight: 20.*

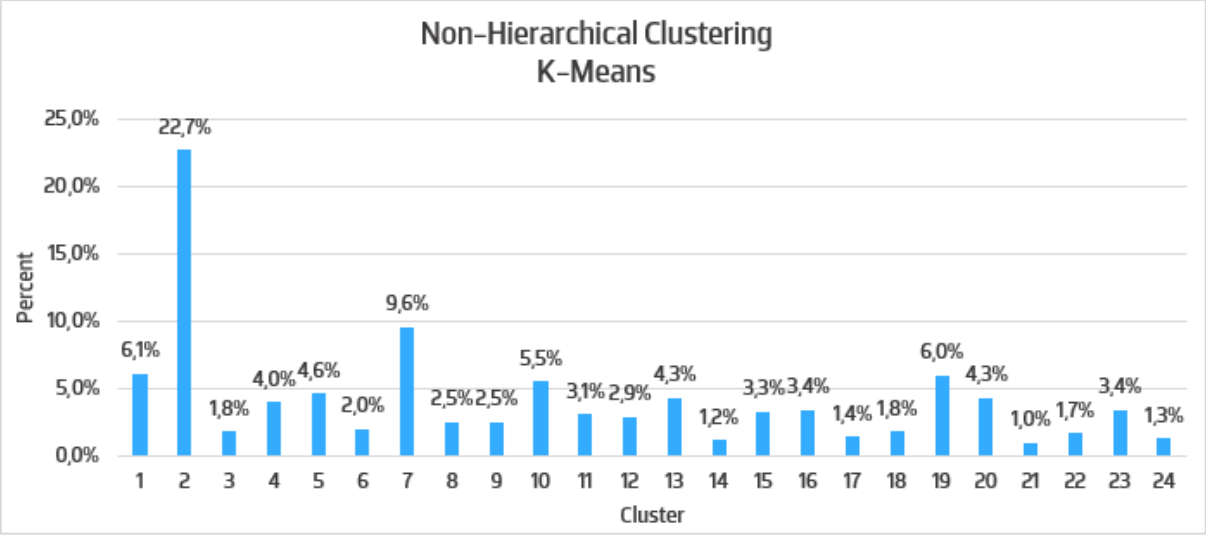


Figure 66 – Percentage *of the clusters formed with K-means. Pharmacotherapeutic Group's weight: 20.*

The relative frequencies of each cluster formed by K-means were assessed. The variables with greater homogeneity within the clusters and heterogeneity between clusters were Pregnancy and Breastfeeding, as previously. After the weight adjustment of the variable Pharmacotherapeutic Group it also presents superior homogeneity. Therefore, their relative frequencies are presented in Tables 21, 22 and 23, respectively. Other variables that discriminate the clusters will be presented in bar charts for each cluster formed by K-means.

As previously, due to space limitations, the relative frequencies of the remaining variables of each cluster formed with this method are presented in a repository[5], as well as those referring to the remaining six methods for a deeper analysis.

Regarding the variable Pregnancy, 12 of the 24 clusters formed were completely separated according to the safe use of its products during pregnancy, presenting only one value in this variable (Table 21). Before the weight adjustment, 50% of the clusters were separated by this variable, and after the adjustment, the same happened to the same percentage of clusters.

[5] https://drive.google.com/drive/folders/118yRVIFI4fdXD8VyKQNF-BOv09E-99RH?usp=drive_link

***Table 21** – Percentage of the variable Pregnancy present in all clusters by K-means. Pharmacotherapeutic Group's weight: 20.*

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
No indication	0	100,0	0	0	0	0	0	0	0	0	0	0
Can be used	0	0	100,0	0	0	0	0	0	0	100,0	0	100,0
With caution	85,1	0	0	0	0	0	0	11,4	0	0	100,0	0
Not recommended	0	0	0	36,8	0	96,4	95,6	2,9	100,0	0	0	0
Not recommended in the first trimester	12,6	0	0	63,2	0	0	2,2	0	0	0	0	0
Not recommended in the third trimester	0	0	0	0	0	3,6	0	57,1	0	0	0	0
Contraindicated	0	0	0	0	100,0	0	0	14,3	0	0	0	0
Contraindicated in the first trimester	2,3	0	0	0	0	0	1,5	0	0	0	0	0
Contraindicated in the third trimester	0	0	0	0	0	0	0	14,3	0	0	0	0
Not applicable	0	0	0	0	0	0	0,7	0	0	0	0	0

Legend: C: Cluster

***Table 21** – Percentage of the variable Pregnancy present in all clusters by K-means. Pharmacotherapeutic Group's weight: 20 (cont.).*

	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
No indication	0	47,1	0	0	0	0	0	0	0	0	0	0
Can be used	0	0	0	0	0	0	0	90,3	0	0	100,0	84,2
With caution	88,5	0	0	20,8	0	0	0	0	0	87,5	0	0
Not recommended	9,8	41,2	0	0	90	0	0	0	0	0	0	0
Not recommended in the first trimester	0	0	0	0	0	0	0	4,8	0	0	0	15,8
Not recommended in the third trimester	0	0	0	79,2	0	0	0	3,2	0	0	0	0
Contraindicated	0	0	100,0	0	0	100,0	100,0	0	100,0	0	0	0
Contraindicated in the first trimester	1,6	0	0	0	10	0	0	0	0	8,3	0	0
Contraindicated in the third trimester	0	0	0	0	0	0	0	1,6	0	0	0	0
Not applicable	0	11,8	0	0	0	0	0	0	0	4,2	0	0

Legend: C: Cluster

The safe use of the products during breastfeeding separated completely 15 of the 24 (Table 22). Before the weight adjustment, approximately 82% of the clusters were separated by this variable, and after the adjustment, the percentage decreased to 63%.

Table 22 – Percentage of the variable Breastfeeding present in all clusters by K-means. Pharmacotherapeutic Group's weight: 20.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
No indication	0	100,0	0	0	0	0	0	0	0	0	0	0
Can be used	0	0	100,0	0	0	78,6	0	22,9	0	100,0	0	100,0
With caution	93,1	0	0	0	0	0	6,6	11,4	0	0	100,0	0
Not recommended	0	0	0	100,0	0	21,4	92,7	51,4	100,0	0	0	0
Contraindicated	6,9	0	0	0	100,0	0	0	14,3	0	0	0	0
Not applicable	0	0	0	0	0	0	0,7	0	0	0	0	0

Legend: C: Cluster

Table 22 – Percentage of the variable Breastfeeding present in all clusters by K-means. Pharmacotherapeutic Group's weight: 20 (cont.).

	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
No indication	0	47,1	0	0	0	0	0	0	0	0	0	0
Can be used	0	0	0	0	0	0	0	98,4	0	0	93,8	84,2
With caution	100,0	0	0	100,0	0	0	0	0	0	87,5	6,3	0
Not recommended	0	41,2	0	0	100,0	0	0	1,6	0	0	0	15,8
Contraindicated	0	0	100,0	0	0	100,0	100,0	0	100,0	8,3	0	0
Not applicable	0	11,8	0	0	0	0	0	0	0	4,2	0	0

Legend: C: Cluster

A complete separation according to the products' main Pharmacotherapeutic Group occurred in 14 of the 24 clusters (approximately 58%) (Table 23). This is an improvement as it was expected since before the weight adjustment, this value was approximately 36%.

Table 23 – Percentage of the variable Pharmacotherapeutic Group present in all clusters by K-means. Pharmacotherapeutic Group's weight: 20.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
N/A	0	100,0	0	0	0	0	0	0	0	0	0	0
G1	1,1	0	0	0	3	0	0	0	0	0	0	0
G2	0	0	0	0	3	0	10,2	100,0	100,0	100,0	0	9,8
G3	10,3	0	0	0	0	0	10,2	0	0	0	0	0
G4	0	0	0	0	0	0	0	0	0	0	0	0
G5	1,1	0	0	100,0	0	0	8,8	0	0	0	100,0	9,7
G6	28,6	0	100,0	0	3	0	17,6	0	0	0	0	12,2
G7	3,4	0	0	0	4,5	0	5,1	0	0	0	0	26,8
G8	0	0	0	0	0	0	6,6	0	0	0	0	0
G9	0	0	0	0	1,5	100,0	2,9	0	0	0	0	0
G10	1,1	0	0	0	0	0	5,9	0	0	0	0	0
G11	0	0	0	0	1,5	0	0,7	0	0	0	0	2,4
G13	45,8	0	0	0	45,4	0	18,8	0	0	0	0	34
G14	3,4	0	0	0	37,9	0	9,5	0	0	0	0	2,4
G15	4,6	0	0	0	0	0	2,2	0	0	0	0	2,4
G16	0	0	0	0	0	0	0	0	0	0	0	0
G20	0	0	0	0	0	0	1,5	0	0	0	0	0

Legend: C: Cluster; N/A: Not Available; G1: 1. Anti-Infectious Agents; G2: 2. Central Nervous System; G3: 3. Cardiovascular System; G4: 4. Blood; G5: 5. Respiratory System; G6: 6. Gastrointestinal System; G7: 7.

Genitourinary System; G8: 8. Hormones and Drugs to Treat Endocrine Diseases; G9: 9. Locomotor System; G10: 10. Anti-Allergic Medication; G11: 11. Nutrition; G13: 13. Drugs for Skin Disorders; G14: 14. Drugs Used in Otorhinolaryngolical Disorders; G15: 15. Drugs for Eye Disorders; G16: 16. Antineoplastic Drugs and Immune-Modulators; G20: 20. Dressing Material, Local Hemostats, Medicinal Gases and Other Products.

***Table 23** – Percentage of the variable Pharmacotherapeutic Group present in all clusters by K-means. Pharmacotherapeutic Group's weight: 20 (cont.).*

	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
N/A	0	82,4	0	0	0	0	0	0	0	0	0	0
G1	9,8	0	0	0	0	0	1,2	0	0	0	4,2	0
G2	3,2	0	0	0	0	0	15,3	0	0	0	2,1	0
G3	0	0	0	0	0	0	0	0	0	4,2	4,2	0
G4	0	0	0	0	0	0	0	3,2	0	0	12,5	0
G5	0	0	0	0	0	0	21,2	1,6	0	0	0	0
G6	34,3	5,9	0	0	100,0	100,0	20,1	46,8	0	16,7	20,9	0
G7	0	0	0	0	0	0	3,5	3,2	0	8,3	0	0
G8	0	0	0	0	0	0	0	0	0	0	0	0
G9	1,6	5,9	100,0	100,0	0	0	23,5	0	100,0	0	0	0
G10	36,1	0	0	0	0	0	2,4	0	0	0	0	0
G11	4,9	0	0	0	0	0	1,2	20,9	0	12,5	29,3	0
G13	4,8	5,9	0	0	0	0	10,6	17,7	0	41,6	18,8	100,0
G14	4,9	0	0	0	0	0	0	1,6	0	0	0	0
G15	0	0	0	0	0	0	1,2	3,2	0	12,5	8,4	0
G16	0	0	0	0	0	0	0	1,6	0	4,2	0	0
G20	0	0	0	0	0	0	0	0	0	0	0	0

Legend: C: Cluster; N/A: Not Available; G1: 1. Anti-Infectious Agents; G2: 2. Central Nervous System; G3: 3. Cardiovascular System; G4: 4. Blood; G5: 5. Respiratory System; G6: 6. Gastrointestinal System; G7: 7. Genitourinary System; G8: 8. Hormones and Drugs to Treat Endocrine Diseases; G9: 9. Locomotor System; G10: 10. Anti-Allergic Medication; G11: 11. Nutrition; G13: 13. Drugs for Skin Disorders; G14: 14. Drugs Used in Otorhinolaryngolical Disorders; G15: 15. Drugs for Eye Disorders; G16: 16. Antineoplastic Drugs and Immune-Modulators; G20: 20. Dressing Material, Local Hemostats, Medicinal Gases and Other Products.

Despite the results of the variable Breastfeeding were not as successful as before, and the variable Pregnancy was equally successful, the results of the Pharmacotherapeutic Group were better than before the weight adjustment, which was its purpose.

Regarding the analysis of the other variables in each cluster formed with K-means after the new adjusted distance function, the variables that discriminate cluster 1, with few modes, are Interactions, Contraindications, and Warnings and Precautions (Figure 67). Interactions variable characterises the cluster by the presence of the value "não foram reportadas". The Contraindications variable presents a high percentage of "hipersensibilidade excipiente". 46% of

the products that belong to this cluster are characterised by having no indication regarding Warnings and Precautions.

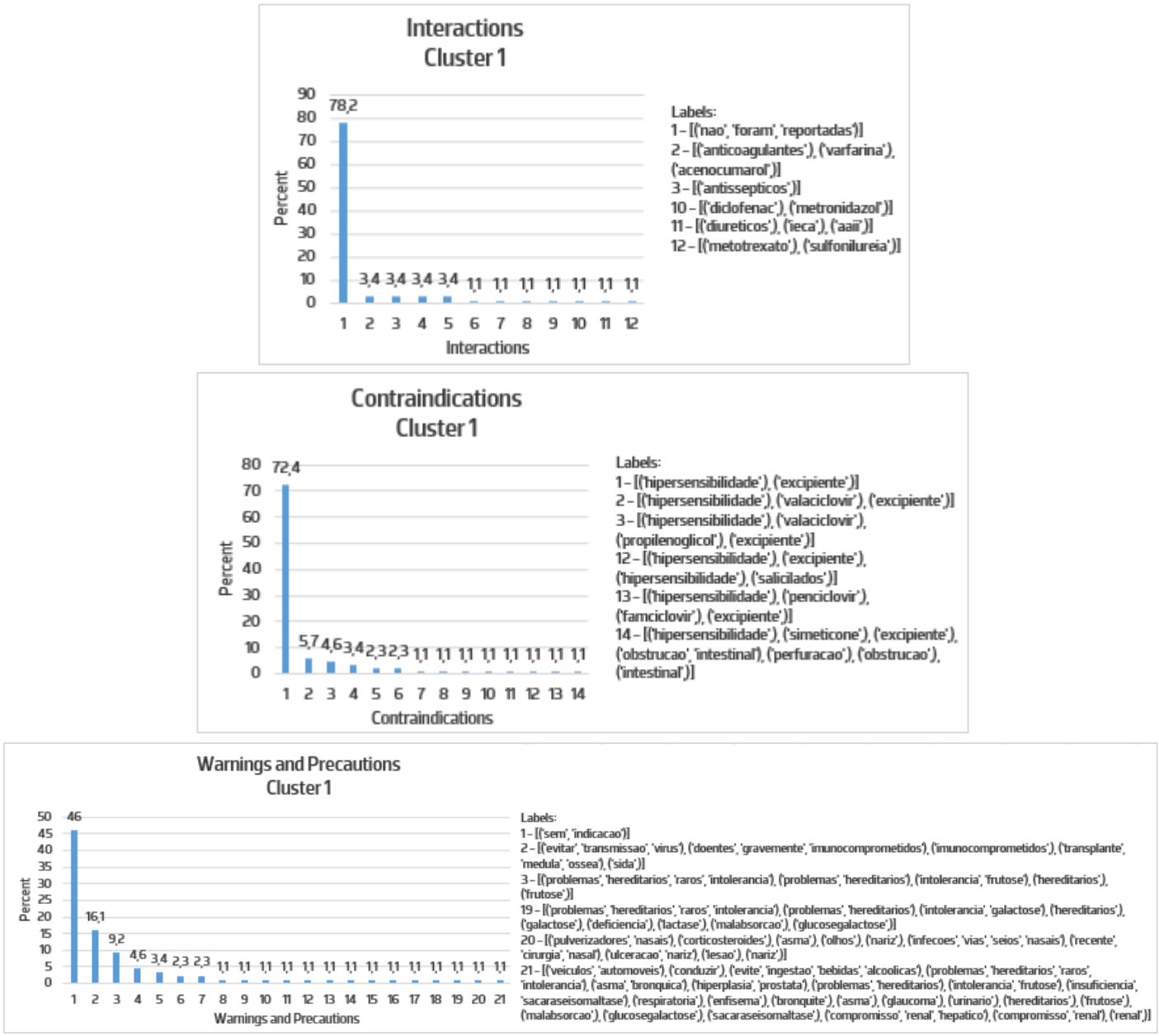


***Figure 67** – Percentage of the variables Interactions, Contraindications, and Warnings and Precautions in cluster 1 by K-means. Pharmacotherapeutic Group's weight: 20.*

Cluster 2 is constituted only by the dermocosmetics, representing 22,7% of the database as previously clustered before adjusting the Pharmacotherapeutic Group's weight. These products don't contain an Active Substance, thus its value is not applicable in this case.

The variables that discriminate cluster 3, with few modes, are Active Substance and Adverse Effects by Categories (Figure 68). Although this cluster is characterised by the presence of products of the same Pharmacotherapeutic Group (osmotic laxatives), it is composed by four different active substances (Table 23). 96% of the products can cause the same adverse effects related to the gastrointestinal system.

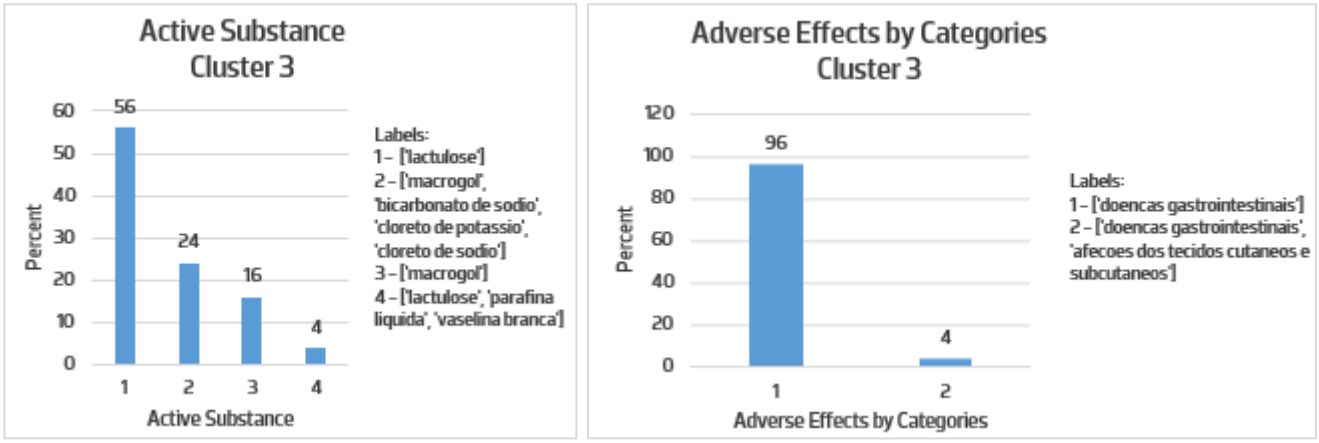


***Figure 68** – Percentage of the variables Active Substance and Adverse Effects by Categories in cluster 3 by K-means. Pharmacotherapeutic Group's weight: 20.*

The variables that discriminate cluster 4, with few modes, are Active Substance, Adverse Effects by Categories, Contraindications, and Interactions (Figure 69). In the Pharmacotherapeutic Group variable, this cluster is characterised only by the presence of products from the group "5.2.2. aparelho respiratório. antitússicos e expetorantes. expetorantes" (Table 23). All the eight values in the Active Substance variable belong to this pharmacotherapeutic group, but this cluster is characterised mainly by the presence of the active substances "ambroxol" and "carbocisteína". Before the weight adjustment, these two active substances belonged to one cluster where only the two were presented (cluster 7, as shown in Figure 36). Now other substances have been grouped due to the increase of the Pharmacotherapeutic Group's weight. Other variables of this cluster also have few modes, such as Adverse Effects by Categories, since 71,9% of this cluster is characterised by the value "doenças do sistema imunitário, doenças gastrointestinais, afeções dos tecidos cutâneos e subcutâneos". Almost half of the products have as a contraindication "úlcera gastroduodenal, hipersensibilidade excipiente", and 63,2% do not have any interactions reported.

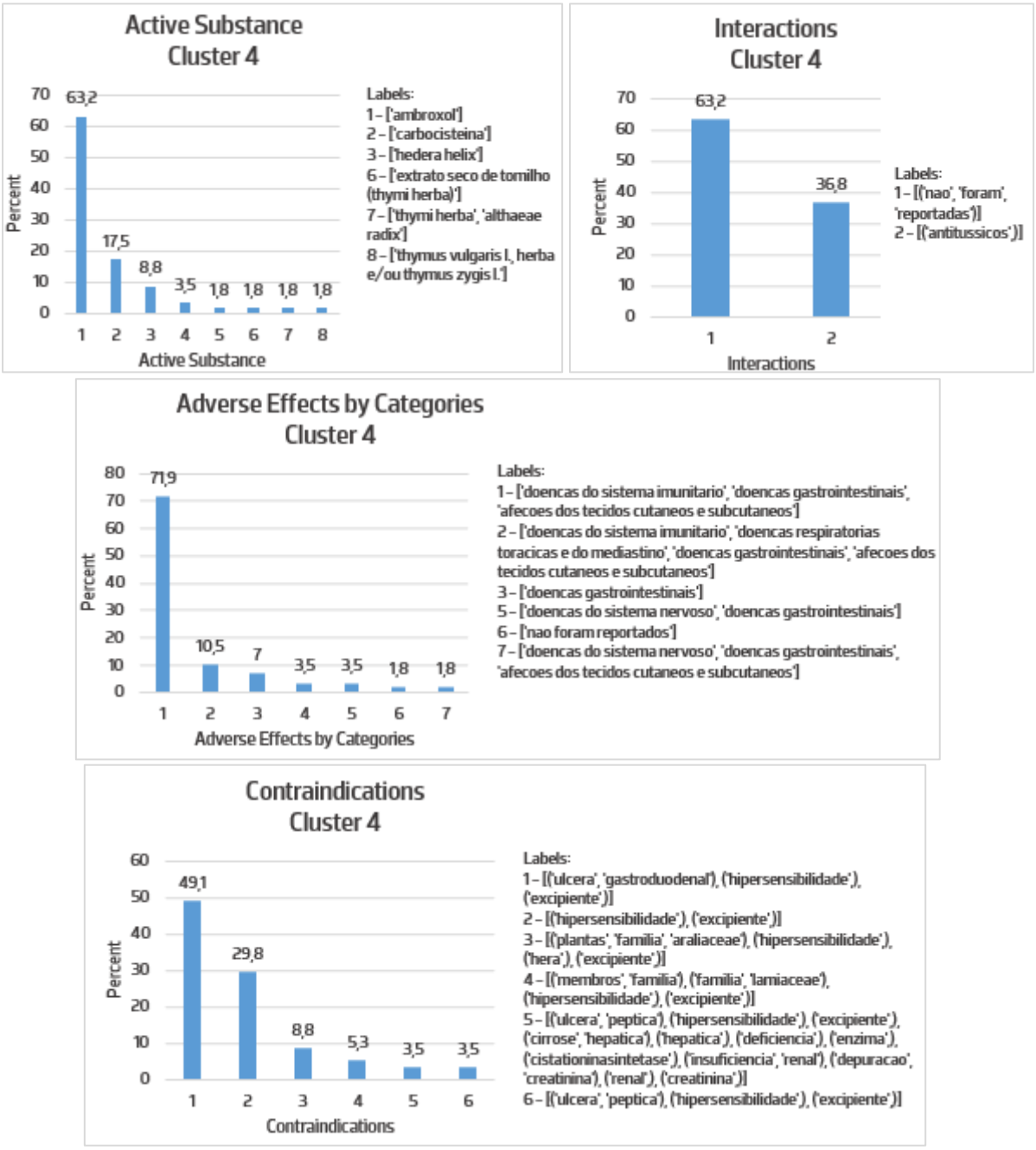


Figure 69 *- Percentage of the variables Active Substance, Adverse Effects by Categories, Contraindications, and Interactions in cluster 4 by K-means. Pharmacotherapeutic Group's weight: 20.*

The variable that discriminates cluster 5, with few modes, is Warnings and Precautions (Figure 70). In this variable, this cluster is characterised by the value "sem indicação".

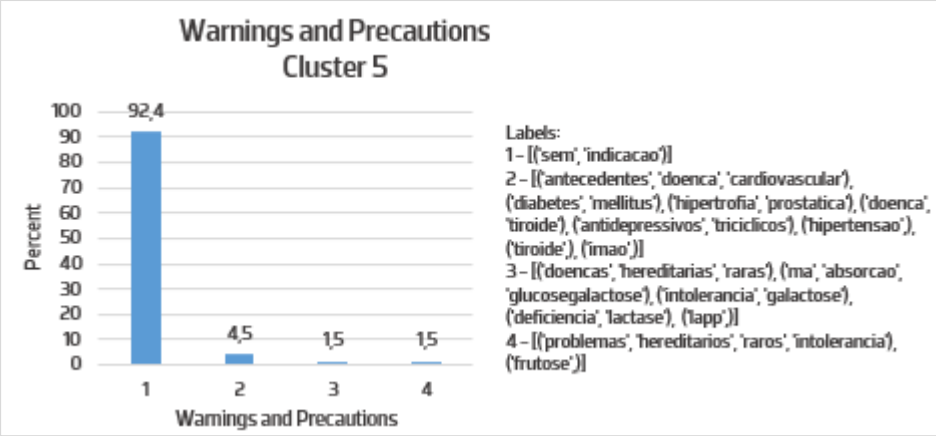


***Figure 70** – % of the variable Warnings and Precautions in cluster 5 by K-means. Pharmacotherapeutic Group's weight: 20.*

The variables that discriminate cluster 6, with few modes, are Active Substance, Pharmaceutical Form, Adverse Effects by Categories, and Warnings and Precautions (Figure 71). In the Pharmacotherapeutic Group variable, this cluster is characterised by the presence of products from the group "9.1.10. aparelho locomotor. anti-inflamatórios não esteroides. anti-inflamatórios não esteroides para uso tópico" (Table 23). All the values in the Active Substance variable belong to this pharmacotherapeutic group, but this cluster characterised mainly by the presence of the active substance "etofenamato". Other variables of this cluster also have few modes, such as Pharmaceutical Form, mainly characterised by the presence of the value "gel" and Adverse Effects by Categories presents 89,3% of the value "perturbações gerais e alterações no local de administração, afeção dos tecidos cutâneos e subcutâneos". The variable Warnings and Precautions presents approximately half of the products with no indications and half with the value "fotoalergia, contacto". These facts are consistent since this pharmacotherapeutic group is constituted by products for topical use, therefore, it is expected that all the pharmaceutical forms are meant for topical application, which can cause adverse effects related to its application on the skin tissue.

In cluster 7, although the variable Interactions presents a high value of modes, it is the variable that presents more agreement between the products (Figure 72). In this variable, the cluster is characterised by 52,6% of its products having no reported interactions.

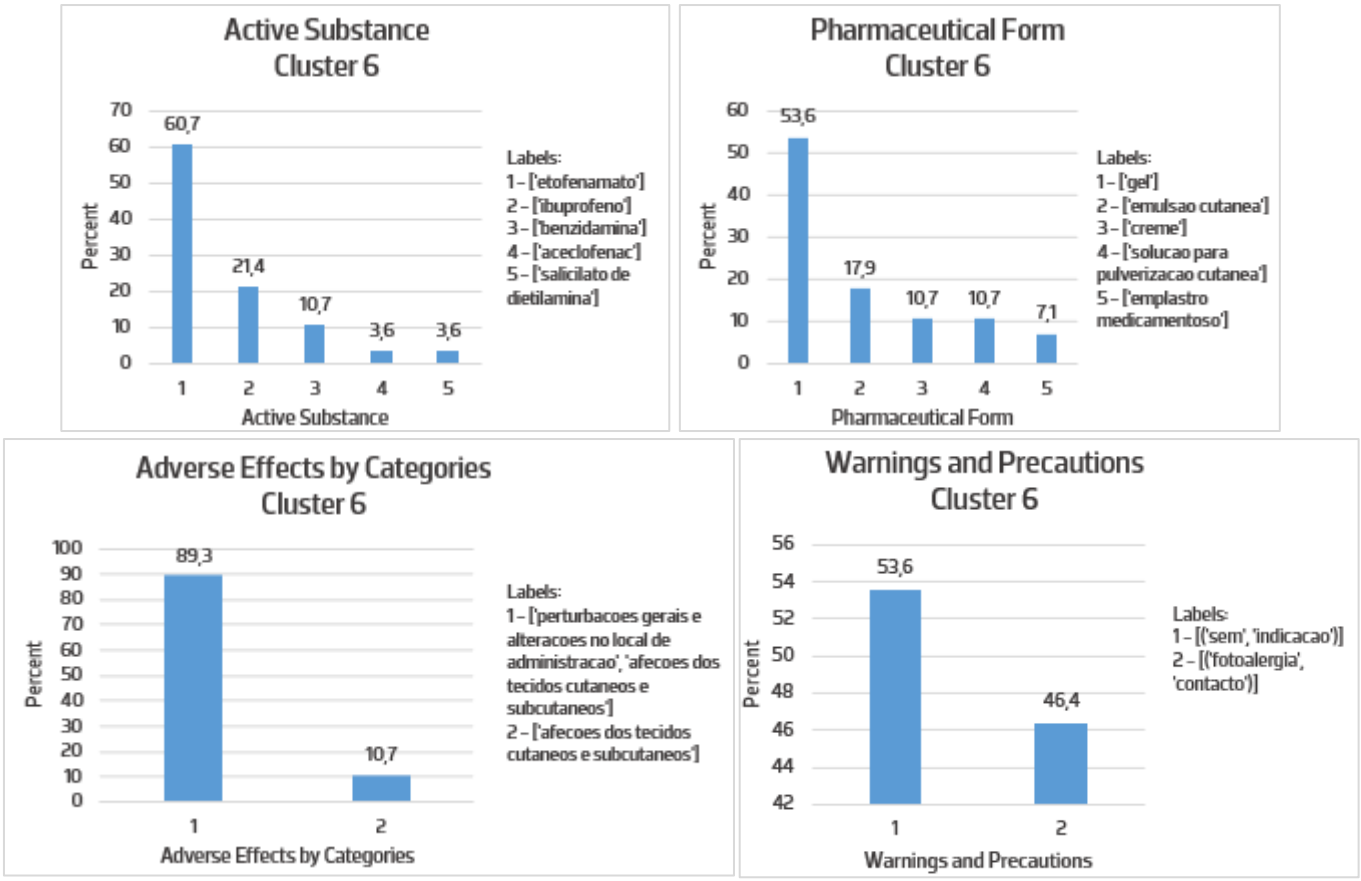


Figure 71 *– Percentage of the variables Active Substance, Pharmaceutical Form, Adverse Effects by Categories and Warnings and Precautions in cluster 6 by K-means. Pharmacotherapeutic Group's weight: 20.*

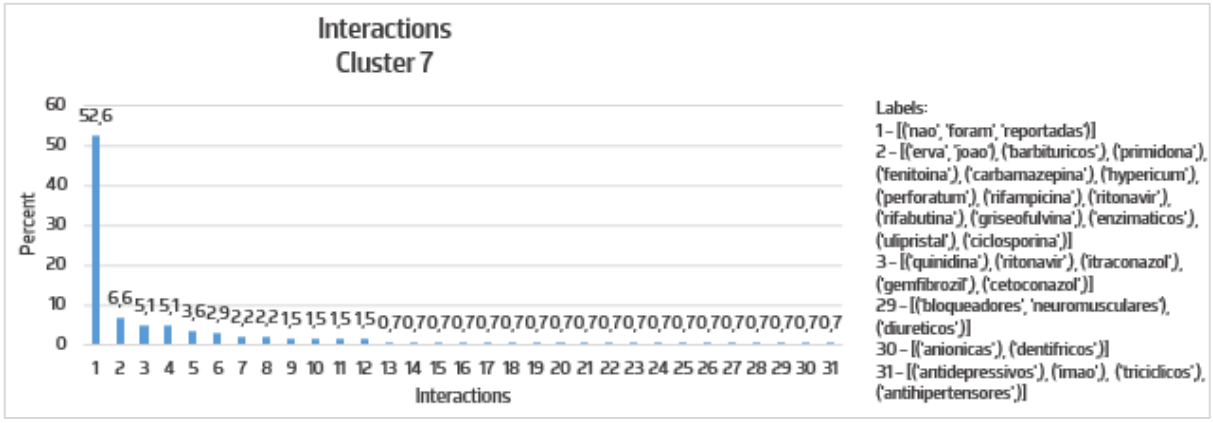


Figure 72 *– Percentage of the variable Interactions in cluster 7 by K-means. Pharmacotherapeutic Group's weight: 20.*

The variable Age discriminates cluster 8, with few modes (Figure 73). About half of the products require the patient to be older than 12 years old. In the Pharmacotherapeutic Group variable, this cluster is characterised by the presence of products from the group "2.10. sistema nervoso central. analgésicos e antipiréticos" (Table 23).

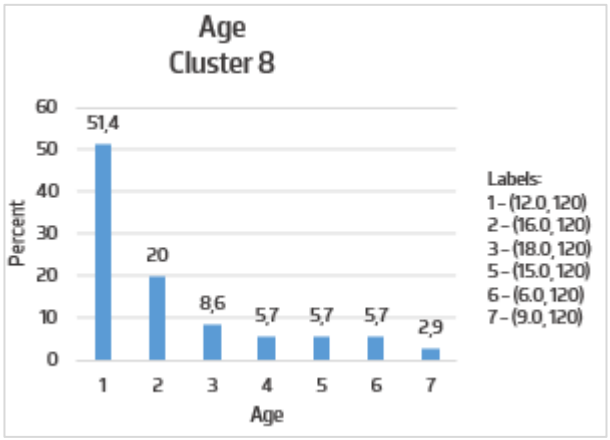


***Figure 73** – Percentage of the variable Age in cluster 8 by K-means. Pharmacotherapeutic Group's weight: 20.*

Clusters 9 and 10 contain the same products as clusters 18 and 6, respectively, before the weight adjustment. Cluster 9 is characterised by the presence of products with the Active Substance "nicotina" representing 2,5% of the database. Cluster 10 is constituted only by the Active Substance "paracetamol", representing 5,5% of the database.

Cluster 11 is constituted by two Active Substances, "acetilcisteína" and "bromexina", but both of them belong to the same pharmacotherapeutic group (Figure 74). This cluster contains the same products as cluster 13 before the weight adjustment.

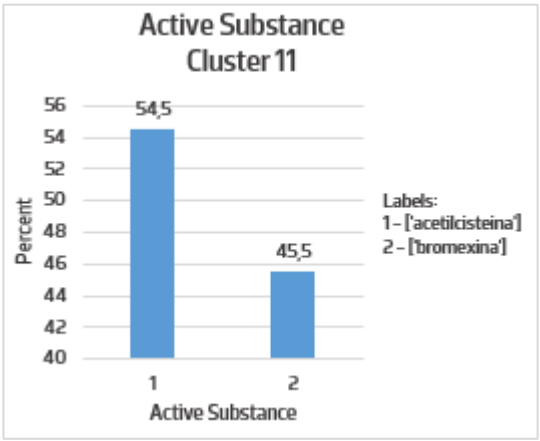


***Figure 74** – Percentage of the variable Active Substance in cluster 11 by K-means. Pharmacotherapeutic Group's weight: 20.*

the variables that discriminate cluster 12, with few modes, are Interactions, and Warnings and Precautions (Figure 75). In the Interactions variable, this cluster is characterised by the presence of the value "não foram reportadas". The cluster is also characterised by having no indication for 65,9% of its products regarding Warnings and Precautions.

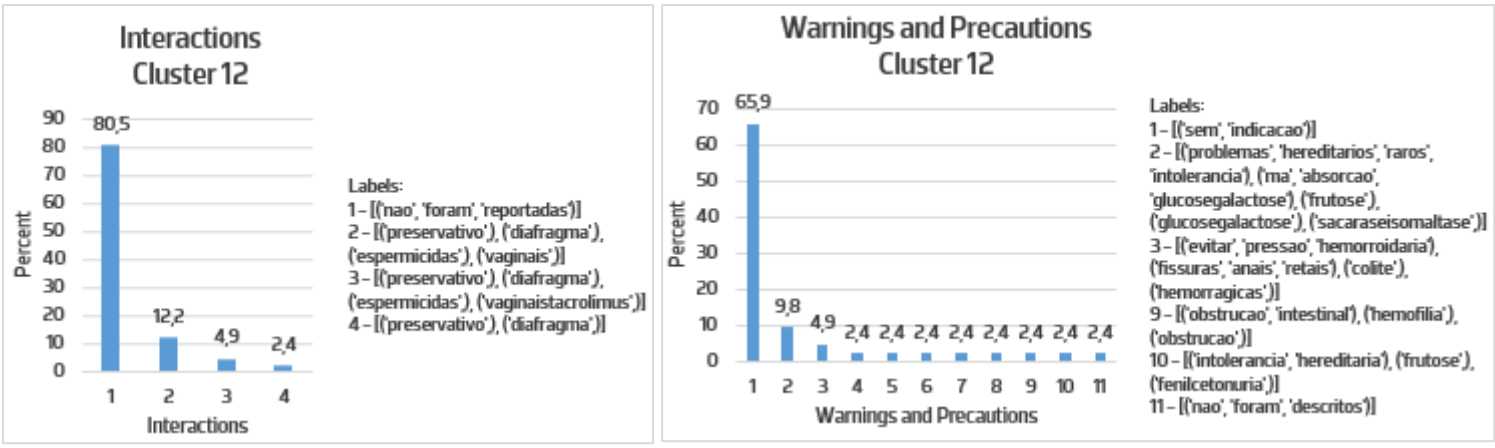


Figure 75 *– Percentage of the variables Interactions, and Warnings and Precautions in cluster 12 by K-means. Pharmacotherapeutic Group's weight: 20.*

The only variables that stand out and discriminate cluster 13 are Pregnancy and Breastfeeding. 88,5% of the products present in this cluster are allowed to be used with precaution during pregnancy, and 100% are allowed with precaution during breastfeeding (Tables 21 and 22). The remaining variables presented high dispersion.

The variables that discriminate cluster 14, with few modes, are Adverse Effects by Categories, Interactions and Contraindications (Figure 76). In the Adverse Effects by Categories variable, this cluster is characterised by the presence of the value "não foram reportados". The cluster is also characterised by having no reported interactions for 88,2 % of its products regarding the Interactions variable; and by the value "hipersensibilidade excipiente" in the Contraindications variable.

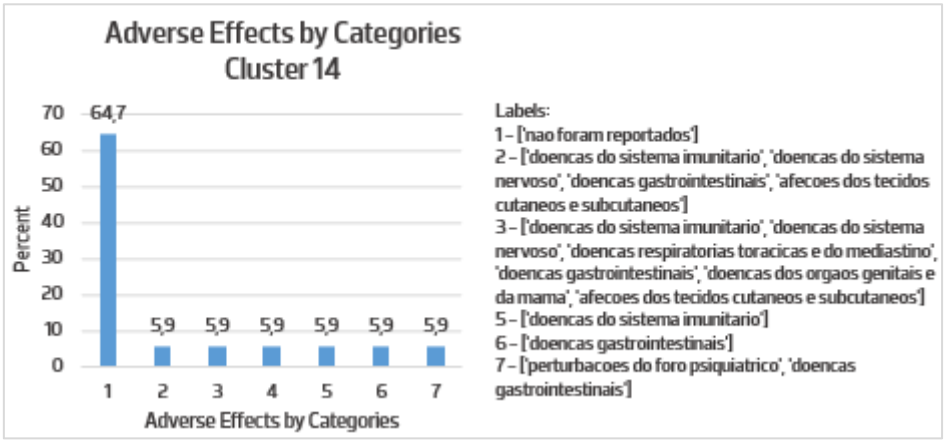


Figure 76 *– Percentage of the variables Adverse Effects by Categories, Interactions, and Contraindications in cluster 14 by K-means. Pharmacotherapeutic Group's weight: 20.*

Cluster 15 is constituted by two Active Substances, "naproxeno" and "ibuprofeno", but both of them belong to the same pharmacotherapeutic group (Figure 77). This cluster is similar to cluster 17 before the weight adjustment. It differs in the quantity of "naproxeno", which has two more units in this cluster.

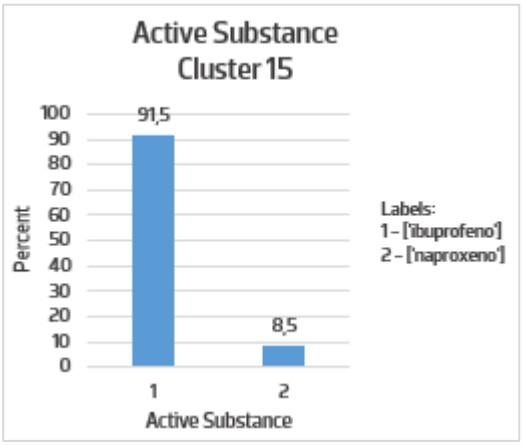


Figure 77 *– Percentage of the variable Active Substance in cluster 15 by K-means. Pharmacotherapeutic Group's weight: 20.*

The variables that discriminate cluster 16, with few modes, are Pharmaceutical Form, Adverse Effects by Categories, Interactions, Warnings and Precautions and Contraindications (Figure 78). The Pharmaceutical Form characterises the cluster by having only products to apply in the skin tissue, with the prevalent value "gel". In the Adverse Effects by Categories variable, this cluster is characterised by the presence of the value "afeções dos tecidos cutâneos e subcutâneos", which is consistent with products' pharmaceutical form, and thus, its application. The cluster is also characterised by having no reported interactions for 87,5% of its products regarding Interactions variable; by having no indication in the Warnings and Precautions variable for 83,3% of the products; and by the value "ácido acetilsalicílico, hipersensibilidade excipiente, asma, urticaria, rinite, aines" in the Contraindications variable for 77,1% of the products. All the products in this cluster belong to the Pharmacotherapeutic Group "9.1.10. aparelho locomotor. anti-inflamatórios não esteroides. anti-inflamatórios não esteroides para uso tópico", the same as the products present in cluster 6. They differ from cluster 6 through the remaining variables. Even the products with the active substance "ibuprofeno" that are divided between the two groups differ from each other through these variables, which can be explained by its different pharmaceutical forms and its ingredients list that can impact the products absorption, distribution, metabolism and elimination, creating different adverse effects, interactions, and others.

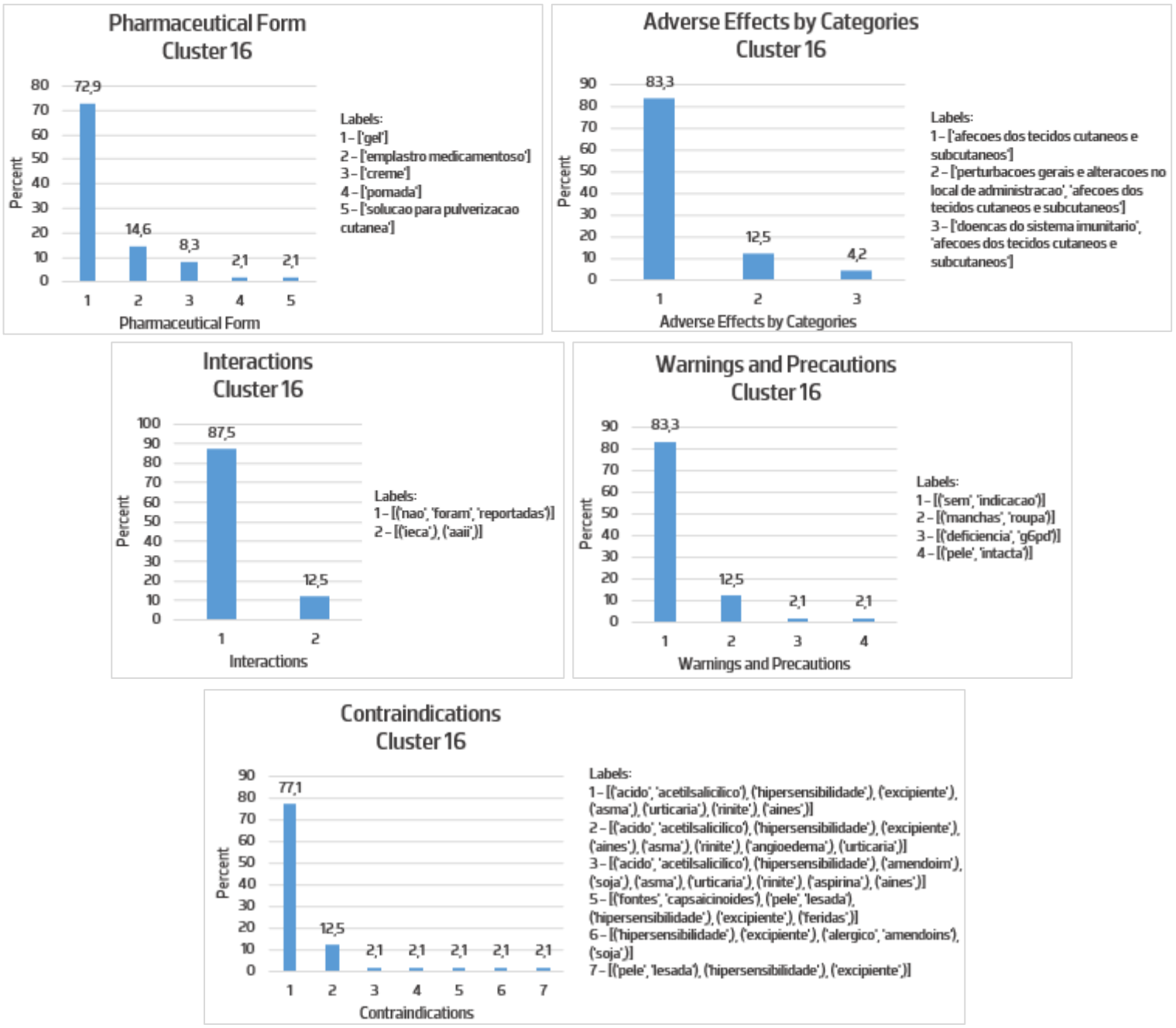


Figure 78 *– Percentage of the variables Pharmaceutical Form, Adverse Effects by Categories, Interactions, Warnings and Precautions and Contraindications in cluster 16 by K-means. Pharmacotherapeutic Group's weight: 20.*

Clusters 17 and 18 are characterised by the variable Pharmacotherapeutic Group, since both clusters contain products from the same group ("6.11. aparelho digestivo. medicamentos que atuam na boca e orofaringe de aplicação tópica") (Table 23). What leads the products to be clustered into two different clusters is the possibility of their use during the periods of pregnancy and breastfeeding. The variables Pregnancy and Breastfeeding discriminate this cluster, with few modes in these clusters. 90% of the products in cluster 17 are not recommended during pregnancy and 10% are contraindicated during the first trimester; none of the products is recommended when breastfeeding (Tables 21 and 22). On the other hand, all the products from cluster 18 are

contraindicated during all the pregnancy period as well as when breastfeeding (Tables 21 and 22). These clusters were also separated according to the variables Interactions and Contraindications. Cluster 17 has no any reported interactions, but cluster 18 interacts with various drugs. Cluster 17 presents as contraindication only the value "hipersensibilidade excipiente", whereas the 18 presents various contraindications with other drugs and diseases.

The variables that stand out and discriminate cluster 19 are Pregnancy and Breastfeeding. All the products present in this cluster are contraindicated during pregnancy and breastfeeding (Tables 21 and 22). The remaining variables presented high dispersion.

The variables that stand out and discriminate cluster 20 are also Pregnancy and Breastfeeding. 90,3% of the products present in this cluster are allowed during pregnancy, and 98,4% are as well when breastfeeding (Tables 21 and 22). The remaining variables presented once again high dispersion.

The same variables, Pregnancy and Breastfeeding, characterise cluster 21, with all its products being contraindicated during these periods. The division into this cluster or cluster 19, which also is contraindicated, occurred through the other variables; on the other hand, they do not discriminate this cluster since they have various values with dispersed percentages. One example is that cluster 19 comprises products from various Pharmacotherapeutic Groups, whereas cluster 21 only comprises products from the same Group, as seen in Table 23.

The variable that discriminates cluster 22 is Age. The Age variable characterises the cluster by only having products with no indication regarding the adequate age for their use.

The variable Age stands out and discriminates cluster 23, with few modes since 93,8% of the products do not present indication about the appropriate age for their use (Figura 79). The remaining values of this variable reveal that those products are only adequate to children under 12 years old. All the products present in this cluster are allowed during pregnancy and 93,8%% are as well when breastfeeding (Table 21 and 22).

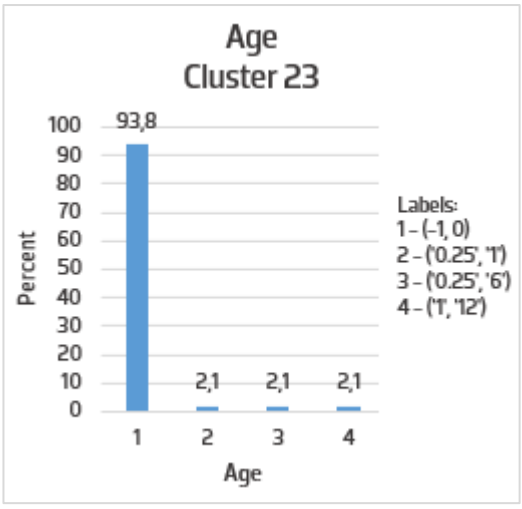


***Figure 79** – Percentage of the variable Age in cluster 23 by K-means. Pharmacotherapeutic Group's weight: 20.*

Cluster 24 is characterised by the variable Active Substance with the value "clotrimazol" and Pharmacotherapeutic Group with the value "13.1.3. medicamentos usados em afeções cutâneas. anti-infecciosos de aplicação na pele. Antifúngicos" (Figure 80).

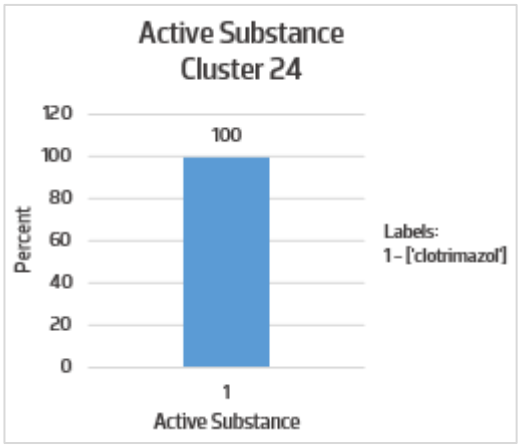


***Figure 80** – Percentage of the variable Active Substance in cluster 24 by K-means. Pharmacotherapeutic Group's weight: 20.*

It is important to note that once again clustering mainly occurred by the variables Pregnancy and Breastfeeding (Tables 21 and 22). However, after the weight adjustment, clustering also occurred accordingly to the Pharmacotherapeutic Group, which was the aim when its weight was increased in the distance function.

Clusters 2, 9, 10 and 11 were clustered as previously, before the weigh adjustment to the variable Pharmacotherapeutic Group. This can be explained by the specific values in its variables. Clusters

2, 9, 10 and 24 are characterised by the presence of only one active substance belonging to the same Pharmacotherapeutic Group.

There are clusters from the same pharmacotherapeutic group (clusters 8, 9 and 10 belong to the group "2. sistema nervoso central"; clusters 4 and 11 belong to the group "5. aparelho respiratório"; clusters 3, 17 and 18 belong to the group "6. aparelho digestivo"; and clusters 6 and 16 belong to the group "9. aparelho locomotor" , which is explained by the other variables, where the products can differ and due to the higher number of values present in some variables, as previously explained. The variables that presented less homogeneity within the clusters and heterogeneity between the clusters, with various modes and a very dispersed percentage among the possible values, were Pharmaceutical Form, Indication, and Contraindications.

K-means presented the higher values of SS and CHS both with the initial and the adjusted weights, although its DBS value was not the lowest. Since the performance of these measures varies according to the dataset, more than one measure should be used to decide the optimal number of clusters and evaluate them. Therefore, as K-means obtained the higher values of two measures, the cluster formed with this non-hierarchical clustering method, presented the best compactness within clusters and separation between clusters. After adjusting the weight in the distance function in order for the cluster to be consistent with the pharmacotherapeutic group, the results improved. However, the obtained clusters may not be clinically relevant enough to assist professionals in pharmaceutical counselling.
For the methods single linkage, complete linkage, median linkage, centroid linkage, ward linkage and average linkage, the hierarchical clustering methods, the obtained results presented high dispersion of the products within the clusters.

4.6. Discussion and Dissertation Contributions

Previous literature has used clustering techniques to cluster pharmaceutical products. Wallstrom et al. (97), used an unsupervised clustering procedure, Markov chain Monte Carlo simulation, to estimate parameters of a Bayesian clustering model. The author's objective was to cluster OTC products. The main results were the obtention of eight clusters, mainly characterised by the product's indication, such as allergy and cough treatments. On the other hand, it wasn't found any consistency regarding other information, such as the adequate age for use, for example. Zhang et al. (72), performed clustering based on the products' treatment efficacy, using vector space model

and K-means algorithm in order to build a collaborative recommendation system for an online pharmacy. The authors used Pearson distance as a similarity measure between users. It resulted in a list of the top related products presented to the users according to their symptoms. Clustering techniques have also been applied to other categories of products. Calantone et al. (98), used K-means to obtain clusters of products that are being launched in the market, according to their prices and the strategic and tactical launch decisions.

To the best of our knowledge, Jaccard distance was not used to cluster pharmaceutical products before. Nonetheless, it was previously applied to medication in order to predict its adverse effects (57). The similarity between each product was dependent only on its possible adverse effects. The attribution of different weights to the variable has already been tested in HRS. It was proposed by Stark et al. (90), a system to assist doctors in the prescription of products indicated as the treatment for migraine disease. In this case, different weights were attributed to the variables, such as disease history, blood pressure, allergies, and others. In this dissertation, the distance function was created and tested in order to be aligned with pharmaceutical counselling, according to the importance attributed to each variable when professionals are recommending a product. Thus, different weights were applied to the variables. At first, after attributing the weights resultant of the experts consultation, the results were not consistent with pharmaceutical counselling. When a patient presents a specific problem or situation, it restrains the professional's choice of the product according to the pharmacotherapeutic group that contains products that apply to that problem. For example, if a patient presents allergy symptoms, those products belong to a specific group (10. Medicação Antialérgica). With that being said, the ideal clustering would be within each pharmacotherapeutic group. In the given example, there would be clusters dividing the products within the group of products indicated to allergy symptoms. Those would be clustered according to the other variables such as interactions, contraindications, safe use during pregnancy and breastfeeding, and others. Therefore, a higher weight was given to the variable Pharmacotherapeutic Group in order for the clusters to be consistent with this variable, and then cluster the products according to the other variables. As stated before, there was an improvement in the results, however, the obtained clusters do not present enough clinical relevance to assist professionals in pharmaceutical counselling. To address this limitation, the products could be previously separated by their pharmacotherapeutic group. For example, in clusters 17 e 18 after the weight ajustment, have products from the same pharmacotherapeutic group, and the classification occurred through other variables, such as Pregnancy and Breastfeeding. The ideal

scenario would be to classify the products according to the remaining variables within each pharmacotherapeutic group. Clustering techniques are used to reduce data sparsity, and increase the generation of recommendations speed, and accuracy.

There were some limitations in this dissertation, such as the non-inclusion of all the pharmaceutical products as dietary supplements and phytotherapeutic products, due to the vast quantity of products commercialised and the time limitations to collect its information; and the non-inclusion of users classification since personal information is not publicly available and there was no access to it.

This dissertation contributes to the advance in the clustering analysis of pharmaceutical products. This was achieved through the description of recommendation systems in pharmacy and the construction of a database of non-prescription products aligned with pharmaceutical counselling in community pharmacies. A distance function was defined and tested with all the criteria verified during pharmaceutical counselling. To define the distance function, an experts' consultation was performed to obtain the pharmacy professionals' opinions on the importance of each variable present in the database when recommending a non-prescription product. Various clustering techniques were tested in order to obtain the more homogeneous groups that would be clinically relevant to assist pharmacy professionals during counselling. The abstract "Application of Machine Learning Techniques for a Recommendation System in Pharmacy", related to the preliminary results, has already been accepted for oral communication at the *XXVI Congresso da Sociedade Portuguesa de Estatística*.

5. Conclusions and Future Work

Community Pharmacy and pharmacy professionals have a crucial role due to its proximity to the population. They promote the patient's quality of life and reduce the risk associated with medication. The professionals must evaluate self-medication, drug interactions and adverse effects to help the community fulfil its health objectives. Professionals tend to make recommendations based on personal experience and patient feedback when it is not present robust and reliable information. Since counselling is crucial, the professional must be accompanied by all the necessary information. However, the current software solutions implemented in CPs in Portugal fall short, as they mainly focus on prescription medications and need more information on non-prescription products. Therefore, developing a pharmaceutical product recommendation framework presents a viable solution. Health Recommendation Systems can be applied in the pharmaceutical field in order to recommend pharmaceutical products and assist pharmacy professionals' decisions. Although there are a few studies in the area of pharmaceutical products, they are still scarce and mainly applied to assist doctors in prescribing medication. To bridge this information gap, since, to the best of our knowledge, no medication RS is applied in community pharmacies in Portugal, it was proposed the development of a conceptual pharmaceutical product recommendation framework. The resulting database includes 1426 pharmaceutical products, each with 24 essential features. Such a comprehensive database proves valuable during pharmaceutical counselling, particularly considering the continuous influx of new products into the market. With all the pertinent information consolidated in an easily accessible database, pharmacy professionals would be better equipped to provide informed recommendations and offer more effective counselling services, thus, minimising the risks associated with medication.

Machine learning techniques were applied to identify relevant product groups according to their characteristics. A distance function was defined and evaluated based on the percentage of commonalities between products to create these groups of products to assist pharmaceutical counselling. Statistical tests were carried out to understand whether there were differences in the distribution of the professionals' opinions about the importance of the criteria used during counselling according to the personal characteristics of each individual. There is no statistical evidence to affirm that there is a difference in the distribution of the importance attributed to the criteria according to the professional's gender. Regarding the job title, the distribution is different in the criteria Patient's Age and Interactions. The asymmetric sample size of pharmacists and

pharmacy technicians can explain these. Regarding the variable age, three criteria have statistically significant correlations (Adverse Effects, Feedback from Previous Clients, and Symptoms and Duration), and the years of experience have a statistically significant correlation with one criterion (Warnings and Precautions). Despite that, the obtained correlations are weak. The Jaccard distance was aligned with pharmaceutical counselling after weight was attributed to the variables according to the experts' consultation to define the importance of each variable.

Initial results with hierarchical methods did not present homogeneity within clusters, they presented a very high dispersion. Also, the methods single linkage, median linkage, centroid linkage and average linkage methods favoured the separation of outliers without an apparent clinical reason. All methods generated a cluster with all dermocosmetics included, representing 22,7% of all products in the database. K-means provided the best results, although not clinically relevant enough to assist professionals in counselling. Therefore, a higher weight was given to the variable Pharmacotherapeutic Group for the distance function to be consistent with this group. Hierarchical clustering methods presented once again dispersion between the elements of each cluster. K-means obtained more clusters consistent with the pharmacotherapeutic group and classified according to their other variables. The variables that discriminate the more the clusters formed were Pregnancy and Breastfeeding, both before and after the weight adjustment of the Pharmacotherapeutic Group. The previous clustering was not carried out according to the variable Pharmacotherapeutic Group since it initially carried out a general distance function that aimed not to require filtering by variable. Clustering techniques are able to reduce data sparsity, and increase the generation of recommendations speed, and their accuracy.

While K-means was the method that provided the more homogeneity within clusters and heterogeneity between clusters, it is possible, in future work, to refine it to obtain more valuable groups to assist in counselling. It would be essential to study this topic and test the distance function in more product categories, such as dietary supplements and phytotherapeutic products, and test them separately in dermocosmetics; apply degrees of severity to appropriate variables; perform an experts' consultation with a larger sample of professionals to obtain more robust results; make an initial separation of the products according to their pharmacotherapeutic group before applying the clustering algorithms so that the clustering is carried out according to the remaining variables, after being separated by their pharmacotherapeutic group; and include the users classification.

References

1. Santos H, Cunha I, Coelho P, Cruz P, Botelho R, Faria G, et al. Boas Práticas Farmacêuticas para a farmácia comunitária (BPF) [Internet]. 3ª Edição. Ordem dos Farmacêuticos; 2009. Available from: https://www.ordemfarmaceuticos.pt/fotos/documentos/boas_praticas_farmaceuticas_para_a_farmacia_comunitaria_2009_20853220715ab14785a01e8.pdf
2. INFARMED. Apresentação [Internet]. 2016 [cited 2023 Apr 27]. Available from: https://www.infarmed.pt/web/infarmed/apresentacao
3. Decreto-Lei n.º176 de 30 de Agosto de 2006. Publicado no. Diário da República. 2006;1ª Série.
4. INFARMED. Lista de DCI identificadas pelo Infarmed como MNSRM-EF e respetivos protocolos de dispensa [Internet]. 2023 [cited 2023 Apr 27]. Available from: https://www.infarmed.pt/web/infarmed/entidades/medicamentos-uso-humano/autorizacao-de-introducao-no-mercado/alteracoes_transferencia_titular_aim/lista_dci
5. Cooper RJ. Over-The-counter medicine abuse-a review of the literature. J Subst Use. 2013;18(2):82–107.
6. Infarmed. Prontuário Terapêutico Online - Paracetamol [Internet]. 2023 [cited 2023 Feb 24]. Available from: https://app10.infarmed.pt/prontuario/framepesactivos.php?palavra=paracetamol&x=0&y=0&rb1=0
7. Decreto-Lei n.º 307/2007 de 31 de Agosto - Regime jurídico das farmácias de oficina. Publicado no. Diário da República, 1ª série. 2007;(168):6083–91.
8. McCoul ED. Contemporary Role and Regulation of Over-the-Counter Sinonasal Medications. Otolaryngol - Head Neck Surg (United States). 2021;165(1):7–13.
9. Sánchez-Sánchez E, Fernández-Cerezo FL, Díaz-Jimenez J, Rosety-Rodriguez M, Díaz AJ, Ordonez FJ, et al. Consumption of over-the-counter drugs: Prevalence and type of drugs. Int J Environ Res Public Health. 2021;18(11).
10. Williams CT. Herbal Supplements: Precautions and Safe Use. Nurs Clin North Am [Internet]. 2021;56(1):1–21. Available from: https://doi.org/10.1016/j.cnur.2020.10.001
11. Simundic AM, Filipi P, Vrtaric A, Miler M, Nikolac Gabaj N, Kocsis A, et al. Patient's knowledge and awareness about the effect of the over-the-counter (OTC) drugs and dietary supplements on laboratory test results: A survey in 18 European countries. Clin Chem Lab Med. 2019;57(2):183–94.
12. Direção-Geral da Alimentação e Veterinária. Suplementos Alimentares [Internet]. 2023 [cited 2023 Apr 27]. Available from: https://www.dgav.pt/alimentos/conteudo/generos-alimenticios/regras-especificas-por-tipo-de-alimentos/suplementos-alimentares/
13. INFARMED I.P. Questões frequentes sobre medicamentos de dispensa exclusiva em farmácia.

2017;1–16. Available from: http://www.infarmed.pt

14. Coelho RB, Costa FA. Impact of pharmaceutical counseling in minor health problems in rural Portugal. Pharm Pract. 2014;12(4):0–0.
15. Infarmed. Decreto-Lei n.º 189/2008, de 24 de Setembro - Legislação Farmacêutica Compilada. 2008; Available from: www.infarmed.pt
16. Saúde M Da. Decreto-Lei n.º 176/2006 - Classificação de medicamentos quanto à dispensa ao público. D da Repub. 2006;6–9.
17. Martins A, Ponte A, Mousinho C, Bragança F, Hergy F, Guerra L, et al. Suplementos Alimentares: O que são e como notificar reações adversas. Bol Farm - INFARMED. 2017;21(3):1–4.
18. Mehralian G, Yousefi N, Hashemian F, Maleksabet H. Knowledge, attitude and practice of pharmacists regarding dietary supplements: A community pharmacy- based survey in Tehran. Iran J Pharm Res. 2014;13(4):1455–63.
19. Waddington F, Naunton M, Kyle G, Thomas J, Cooper G, Waddington A. A systematic review of community pharmacist therapeutic knowledge of dietary supplements. Int J Clin Pharm. 2015;37(3):439–46.
20. Infarmed. Despacho n.º 17690/2007. D da Repub. 2007;10–2.
21. Gonçalves E, Marcelo A, Vilão S, da Silva JA, Martins AP. Non-prescription medicinal products dispensed exclusively in the pharmacy: an underused access opportunity in Portugal? Drugs Ther Perspect. 2016;32(11):488–98.
22. Quintal C, Sarmento M, Raposo V. Fatores explicativos do consumo de medicamentos não sujeitos a receita médica em Portugal. Acta Farm Port [Internet]. 2015;4(1):53–66. Available from: http://www.actafarmaceuticaportuguesa.com/index.php/afp/article/view/60
23. WHO. Guidelines for the Regulatory Assessment of Medicinal Products for use in Self-Medication. 2000;
24. Martins AP, Da Costa Miranda A, Mendes Z, Soares MA, Ferreira P, Nogueira A. Self-medication in a Portuguese urban population: A prevalence study. Pharmacoepidemiol Drug Saf. 2002;11(5):409–14.
25. Peixoto J. Automedicação no Adulto. Univ Fernando Pessoa. 2008;87.
26. Martins DBS, Couto SMP do, Ribeiro MIBR, Fernandes AJG. Prevalência da automedicação na região de Bragança: a perspetiva do consumidor e do farmacêutico. Egitania Sci. 2011;(8):199–215.
27. Monteiro C, Marques FB, Ribeiro CF. Interacções medicamentosas como causa de iatrogenia evitável. Rev Port Med Geral e Fam [Internet]. 2007;23(1):63–73. Available from: http://www.rpmgf.pt/ojs/index.php?journal=rpmgf&page=article&op=view&path[]=10322
28. Oliveira RP, Jesus A. Interações Medicamentosas Potenciais em Farmácia Comunitária – Estudo Exploratório. 2022;11:12–27.

29. Almeida SM De, Gama CS, Akamine N. Prevalência e classificação de interações entre medicamentos dispensados para pacientes em terapia intensiva Prevalence and classification of drug-drug interactions in intensive care patients. Einstein. 2007;5(4):347–51.
30. Corrie K, Hardman JG. Mechanisms of drug interactions: Pharmacodynamics and pharmacokinetics. Anaesth Intensive Care Med [Internet]. 2011;12(4):156–9. Available from: http://dx.doi.org/10.1016/j.mpaic.2010.12.008
31. Juurlink DN. Drug interactions with warfarin: what clinicians need to know. 2007;177(4):3–6.
32. Boullata J. Natural health product interactions with medication. Nutr Clin Pract. 2005;20(1):33–51.
33. Bailey DG, Dresser G, Arnold JMO. Grapefruit-medication interactions: Forbidden fruit or avoidable consequences? C Can Med Assoc J. 2013;185(4):309–16.
34. Fernandes A, Palma L, Frazão F, Monteiro C. Medicamentos não sujeitos a receita médica - razões mais frequentes de seu uso. Rev Lusófona Ciências e Tecnol da Saúde. 2009;1(7):47–55.
35. Woroń J, Chrobak AA, Ślęzak D, Siwek M. Unprescribed and unnoticed: Retrospective chart review of adverse events of interactions between antidepressants and over-the-counter drugs. Front Pharmacol. 2022;13(August):1–9.
36. FDA. Finding and Learning about Side Effects (adverse reactions) [Internet]. 2022 [cited 2023 May 29]. Available from: https://www.fda.gov/drugs/information-consumers-and-patients-drugs/finding-and-learning-about-side-effects-adverse-reactions
37. Scott S, Thompson J. Adverse drug reactions. Anaesth Intensive Care Med [Internet]. 2014;15(5):245–9. Available from: http://dx.doi.org/10.1016/j.mpaic.2014.02.008
38. Zazzara MB, Palmer K, Vetrano DL, Carfì A, Graziano O. Adverse drug reactions in older adults : a narrative review of the literature. Eur Geriatr Med [Internet]. 2021;(0123456789). Available from: https://doi.org/10.1007/s41999-021-00481-9
39. Cuzzolin L, Benoni G. Safety of non-prescription medicines: Knowledge and attitudes of Italian pharmacy customers. Pharm World Sci. 2010;32(1):97–102.
40. Ordem dos Farmacêuticos. Norma específica sobre dispensa de medicamentos e produtos de saúde. Boas Práticas Farmácia Comunitária. 2018;(OF.C-N004-00):1–13.
41. Angelo LB, Christensen DB, Ferreri SP. Impact of community pharmacy automation on workflow, workload, and patient interaction. J Am Pharm Assoc [Internet]. 2005;45(2):138–44. Available from: http://dx.doi.org/10.1331/1544345053623537
42. Glintt. Sifarma [Internet]. 2022 [cited 2023 May 30]. Available from: https://www.glintt.com/pt/o-que-fazemos/ofertas/SoftwareSolutions/Paginas/Sifarma.aspx
43. SIER Group. SoftReis [Internet]. 2017 [cited 2023 May 30]. Available from: https://www.sier.pt/services/softreis/

44. Angstman KB. Individualized Health Care. Inq J Heal Care Organ Provision, Financ [Internet]. 2014 Jan 1;51:004695801456163. Available from: http://journals.sagepub.com/doi/10.1177/0046958014561637
45. Bhat S, Aishwarya K. Item-based Hybrid Recommender System for newly marketed pharmaceutical drugs. Proc 2013 Int Conf Adv Comput Commun Informatics, ICACCI 2013. 2013;2107–11.
46. Haw SC, Chew LJ, Subramaniam S. A hybrid recommender system based on data enrichment on the ontology modelling. F1000Research. 2021;10.
47. Kshour M, Ebrahimi M, Goliaee S, Tawil R. New recommender system evaluation approaches based on user selections factor. Heliyon [Internet]. 2021;7(7):e07397. Available from: https://doi.org/10.1016/j.heliyon.2021.e07397
48. Cheung KL, Durusu D, Sui X, de Vries H. How recommender systems could support and enhance computer-tailored digital health programs: A scoping review. Digit Heal. 2019;5:1–19.
49. Zhu J, Patra BG, Yaseen A. Recommender system of scholarly papers using public datasets. AMIA . Annu Symp proceedings AMIA Symp. 2021;2021:672–9.
50. Lattar H, Ben Salem A, Hajjami Ben Ghézala H, Boufares F. Health Recommender Systems: A Survey. Smart Innov Syst Technol. 2020;146(April 2021):182–91.
51. Hors-Fraile S, Rivera-Romero O, Schneider F, Fernandez-Luque L, Luna-Perejon F, Civit-Balcells A, et al. Analyzing recommender systems for health promotion using a multidisciplinary taxonomy: A scoping review. Int J Med Inform [Internet]. 2018;114(April 2017):143–55. Available from: http://dx.doi.org/10.1016/j.ijmedinf.2017.12.018
52. Ko H, Lee S, Park Y, Choi A. A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. Electronics [Internet]. 2022 Jan 3;11(1):141. Available from: https://www.mdpi.com/2079-9292/11/1/141
53. Bhimavarapu U, Chintalapudi N, Battineni G. A Fair and Safe Usage Drug Recommendation System in Medical Emergencies by a Stacked ANN. Algorithms. 2022;15(6):186.
54. Sezgin E, Özkan S. A systematic literature review on Health Recommender Systems. 2013 E-Health Bioeng Conf EHB 2013. 2013;
55. De Croon R, Van Houdt L, Htun NN, Štiglic G, Abeele V Vanden, Verbert K. Health recommender systems: Systematic review. J Med Internet Res. 2021;23(6).
56. Valdez AC, Ziefle M, Verbert K, Felfernig A, Holzinger A. Recommender systems for health informatics: State-of-the-art and future perspectives. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2016;9605 LNCS:391–414.
57. Tran TNT, Felfernig A, Trattner C, Holzinger A. Recommender systems in the healthcare domain: state-of-the-art and research issues. J Intell Inf Syst [Internet]. 2021 Aug 17;57(1):171–201.

Available from: https://link.springer.com/10.1007/s10844-020-00633-6

58. Fayyaz Z, Ebrahimian M, Nawara D, Ibrahim A, Kashef R. Recommendation systems: Algorithms, challenges, metrics, and business opportunities. Appl Sci. 2020;10(21):1–20.
59. Patel K, Patel HB. A state-of-the-art survey on recommendation system and prospective extensions. Comput Electron Agric [Internet]. 2020;178(September):105779. Available from: https://doi.org/10.1016/j.compag.2020.105779
60. Benkessirat S, Boustia N, Rezoug N. Overview of Recommendation Systems. Smart Innov Syst Technol. 2019;41:357–72.
61. Lin Z, Laska E, Siegel C. A general iterative clustering algorithm. Stat Anal Data Min. 2022;15(4):433–46.
62. Thomas JCR, Peñas MS, Mora M. New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance. Proc - Int Conf Chil Comput Sci Soc SCCC. 2013;0(1):49–53.
63. Madhulatha TS. An overview of clustering methods. J Eng. 2012;2(4):719–25.
64. Gulagiz FK, Suhap S. Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms. Int J Comput Eng Inf Technol [Internet]. 2017;9(1):6–14. Available from: www.ijceit.org
65. Bazhenov AN, Telnova AY. Generalization of Jaccard Index for Interval Data Analysis. Meas Tech [Internet]. 2023 Mar 11;65(12):882–90. Available from: https://link.springer.com/10.1007/s11018-023-02180-2
66. Shameem M-U-S, Ferdous R. An efficient k-means algorithm integrated with Jaccard distance measure for document clustering. In: 2009 First Asian Himalayas International Conference on Internet [Internet]. IEEE; 2009. p. 1–6. Available from: http://downloads.esri.com/archydro/archydro/Doc/Overview of Arc Hydro terrain preprocessing workflows.pdf%0Ahttps://doi.org/10.1016/j.jhydrol.2017.11.003%0Ahttp://sites.tufts.edu/gis/files/2013/11/Watershed-and-Drainage-Delineation-by-Pour-Point.pdf%0Awww
67. Jarman AM. Hierarchical Cluster Analysis: Comparison of Single Linkage, Complete Linkage, Average Linkage and Centroid Linkage. Int J Georg Soutern Univ [Internet]. 2020;1(1):1–13. Available from: https://www.researchgate.net/publication/339443595
68. Li Z, De Rijke M. The impact of linkage methods in hierarchical clustering for active learning to rank. SIGIR 2017 - Proc 40th Int ACM SIGIR Conf Res Dev Inf Retr. 2017;941–4.
69. Vijaya V, Sharma S, Batra N. Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering. Proc Int Conf Mach Learn Big Data, Cloud Parallel Comput Trends, Prespectives Prospect Com 2019. 2019;568–73.
70. Maroco J. Análise Estatística - Com utilização do SPSS. Analise estatistica com SPSS. 2011.
71. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview, II. Wiley Interdiscip Rev

Data Min Knowl Discov. 2017;7(6):1–16.
72. Zhang Y, Zhang D, Hassan MM, Alamri A, Peng L. CADRE: Cloud-Assisted Drug REcommendation Service for Online Pharmacies. Mob Networks Appl. 2015;20(3):348–55.
73. Theodoridis S, Koutroumbas K. Pattern Recognition. Fourth Edi. Pattern Recognition. 2009. 883–885 p.
74. Yuan C, Yang H. Research on K-Value Selection Method of K-Means Clustering Algorithm. J. 2019;2(2):226–35.
75. Simić S, Villar JR, Calvo-Rolle JL, Sekulić SR, Simić SD, Simić D. An application of a hybrid intelligent system for diagnosing primary headaches. Int J Environ Res Public Health. 2021;18(4):1–15.
76. Cui M. Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. 2020;5–8.
77. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognit. 2013;46(1):243–56.
78. Chan WN, Aung HM. Recommender System for Pharmacy Using Multi-Agent. 2018;1–6.
79. Lv Y, Kong J. Application of Collaborative Filtering Recommendation Algorithm in pharmacy system. J Phys Conf Ser. 2021;1865(4).
80. John A. Medication Recommendation System Based on Clinical Documents. Int Conf Inf Sci. 2016;180–4.
81. Infarmed. Infomed [Internet]. 2023 [cited 2023 May 20]. Available from: https://extranet.infarmed.pt/INFOMED-fo/
82. IBM. Introduction to CRISP-DM [Internet]. 2021. [cited 2022 Dec 10]. Available from: https://www.ibm.com/docs/it/spss-modeler/saas?topic=dm-crisp-help-overview
83. Veiga P, Cavaco AM, Lapão LV, Guerreiro MP. Self-medication consultations in community pharmacy: An exploratory study on teams' performance, client-reported outcomes and satisfaction. Pharm Pract (Granada). 2021;19(1):1–8.
84. Rutter P, Wadesango E. Does evidence drive pharmacist over-the-counter product recommendations? J Eval Clin Pract. 2014;20(4):425–8.
85. Hell F, Taha Y, Hinz G, Heibei S, Müller H, Knoll A. Graph convolutional neural network for a pharmacy cross-selling recommender system. Inf. 2020;11(11):1–13.
86. Brunton LL, Chabner BA, Knollmann BC. As bases farmacológicas da terapêutica de Goodman & Gilman. 12ª edição. Porto Alegre: AMGH; 2012.
87. Katzung B, Masters S, Trevor A. Basic and Clinical Pharmacology. 15ª edição. McGraw-Hill Medical; 2021.
88. Klaassen CD. Casarett & Doull's Toxicology: The Basic Science of Poisons. 5th Editio. McGraw-Hill Professional; 2001.
89. IBM Corp. Realeased 2021. IBM SPSS Statistics for Windows, Version 28.8. Armonk, NY: IBM Corp;

90. Stark B, Knahl C, Aydin M, Samarah M, Elish KO. BetterChoice: A migraine drug recommendation system based on Neo4J. 2017 2nd IEEE Int Conf Comput Intell Appl ICCIA 2017. 2017;2017-Janua:382–6.
91. Kosub S. A note on the triangle inequality for the Jaccard distance. Pattern Recognit Lett. 2016;120(1):36–8.
92. EMA. Active Substance [Internet]. 2023 [cited 2023 Jun 10]. Available from: https://www.ema.europa.eu/en/glossary/active-substance
93. INFARMED I.P. 2. Sistema Nervoso Central / 2.10. Analgésicos e antipiréticos / Ácido acetilsalicílico [Internet]. Prontuário Terapêutico Online. 2023 [cited 2023 Jun 10]. Available from: https://app10.infarmed.pt/prontuario/framepesactivos.php?palavra=aspirina&x=0&y=0&rb1=0
94. INFARMED I.P. 4. Sangue / 4.3. Anticoagulantes e antitrombóticos / 4.3.1. Anticoagulantes / 4.3.1.3. Antiagregantes plaquetários / Ácido acetilsalicílico [Internet]. Prontuário Terapêutico Online. 2023 [cited 2023 Jun 10]. Available from: https://app10.infarmed.pt/prontuario/mostra.php?flag_palavra_exacta=1&id=519&palavra=%C1cido+acetilsalic%EDlico&flag=1
95. INFARMED. Classificação farmacoterapêutica de medicamentos - Despacho n.º 4742/2014, de 21 de março. Legis Farm Compil [Internet]. 2014; Available from: https://www.infarmed.pt/documents/15786/1072289/110-AB6_Desp_4742_2014_VF.pdf
96. Schober P, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. Anesth Analg. 2018;126(5):1763–8.
97. Wallstrom GL, Hogan WR. Unsupervised Clustering of Over-the-counter Healthcare Products into Product Categories. 2008;40(6):642–8.
98. Calantone RJ, Benedetto CA. Clustering product launches by price and launch strategy. J Bus Ind Mark. 2007;22(1):4–19.
99. Health Market Research. Market Reports [Internet]. 2023 [cited 2023 Apr 27]. Available from: https://www.hmr.co.com/market-reports/
100. Bai X, Wang M, Lee I, Yang Z, Kong X, Xia F. Scientific paper recommendation: A survey. IEEE Access. 2019;7(January):9324–39.

Annex A

Proof that the weighted Jaccard is a distance

The Jaccard distance (J_δ) was calculated through the weighted Jaccard index (J_w) with the following formula:

$$J_\delta\ (A,B) = \sum_{i \in V} w_i\ -\ J_w$$

where

$$J_w\ (A,B) = \sum_{i \in V} (J_i(A,B) \times w_i)$$

and

$$J_i\ (A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where A and B are observations of a set S, V is the set of variables and w_i a positive real value.

In order for the J_δ to be a distance it must satisfy the following conditions:

1. $J_\delta(A,B)\ \geq 0$ for all $A, B \in S$ and $J_\delta(A,B) =\ 0$ if and only if A = B
2. $J_\delta(A,B) =\ J_\delta(B,A)$ for all $A, B \in S$
3. $J_\delta(A,B) \leq J_\delta(A,C) +\ J_\delta(C,B)$ for all $A, B, C \in S$

Proof:

1. $J_\delta(A,B) =\ \sum_{i \in V} w_i\ -\ \sum_{i \in V} (J_i(A,B) \times w_i) =\ \sum_{i \in V} w_i\ (1 - J_i(A,B))$

since $0 \leq J_i(A,B) \leq 1$ and w_i a positive real value then $w_i\ (1 - J_i(A,B)) \geq 0$ and therefore $J_\delta(A,B) \geq 0$.

$$J_\delta(A,B) = \sum_{i \in V} w_i \,(1 - J_i(A,B)) = 0 \Leftrightarrow \; 1 - J_i(A,B) = 0$$

$$\Leftrightarrow J_i(A,B) = 1$$

Since the Jaccard distance J_δ is known to fulfill all properties of a metric (91),

A is equal to B

2. $J_\delta(A,B) = \sum_{i \in V} w_i \; - \sum_{i \in V} (J_i(A,B) \times w_i) = \sum_{i \in V} w_i \,(1 - J_i(A,B))$

   since the Jaccard distance J_δ is known to fulfill all properties of a metric

   $$1 - J_i(A,B) = \; 1 - J_i(B,A)$$

   and therefore

   $$J_\delta(A,B) = \; J_\delta(B,A)$$

3. $J_\delta(A,B) \leq J_\delta(A,C) + \; J_\delta(C,B)$

$$J_\delta(A,C) + \; J_\delta(C,B) = \; \sum_{i \in V} w_i \left(1 - J_i(A,C)\right) \; + \sum_{i \in V} w_i \left(1 - J_i(C,B)\right)$$

$$= \sum_{i \in V} w_i \left(1 - J_i(A,C) + 1 - J_i(C,B)\right)$$

since the Jaccard distance J_δ is known to fulfill all properties of a metric

$$1 - J_i(A,C) + 1 - J_i(C,B) \geq 1 - J_i(A,B)$$

since w_i a positive real value

$$w_i \left(1 - J_i(A,C) + 1 - J_i(C,B)\right) \geq w_i \left(1 - J_i(A,B)\right)$$

$$\sum_{i \in V} w_i \left(1 - J_i(A,C) + 1 - J_i(C,B)\right) \; \geq \sum_{i \in V} w_i \left(1 - J_i(A,B)\right)$$

$$J_\delta(A,C) + \; J_\delta(C,B) \geq J_\delta(A,B)$$