NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

Improving relationship channels in the pharmaceutical industry with machine learning

Is it possible to develop a model(s) based on machine learning, capable to suggest a mix of channels for each physician individually?

Reinaldo dos Santos Barros

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Information Management

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Improving relationship channels in the pharmaceutical industry with machine learning

Is it possible to develop a model(s) based on machine learning, capable to suggest a mix of channels for each physician individually?

by

Reinaldo dos Santos Barros

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence.

Supervised by

Carina Isabel Andrade Albuquerque, PhD, NOVA Information Management School

João Pedro Martins Ribeiro da Fonseca, PhD, NOVA Information Management School

De

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

13 de dezembro de 2024

Reinaldo dos Santos Barros

ABSTRACT

The pharmaceutical industry is crucial in modern society, helping people live longer and better through scientific research and modern medications. In this industry, there is a concept that face-to-face interactions between a sales representative and a physician are more effective and efficient in promoting products. It is important to note that in this industry, the target customers are physicians, and all the marketing efforts are focused on convincing them to prescribe their medications. This study explores whether machine learning can help determine the most effective combination of communication channels. These channels may include webinars, email engagement, social media, or in-person interactions. Can machine learning models suggest the optimal combination of physician communication channels? This study uses three datasets. The first one provides interactions among physicians and different communication channels. The second one shares the prescription data from physicians, while the third one offers basic demographic information from physicians. The study discovered that machine learning could enhance CRM planning by applying digital channels to improve communication with physicians. It identified the classifier chain as the simplest and most effective model for predicting the mix of communication channels. Different machine learning methods and algorithms were applied to understand their use from different perspectives. Other algorithms that could be utilized in future research are also highlighted, revealing alternative approaches to addressing the issue to propose a combination of communication channels.

KEYWORDS

Machine Learning; Supervised learning; Classifier Chain; Binary Relevance; Communication Channels

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CSV	Comma-Separated Values
ETL	Extract, Transform and Load
F2F	Face to Face
FN	False Negative
FP	False Positive
HCP	Health Care Practitioner
KPIs	Key Performance Indicators
MCM	Multi-Channel Marketing
ROS	Random Over Sampling
TN	True Negative
TP	True Positive

1. INTRODUCTION

Pharmaceutical research and development require significant investments, and the development of a new drug carries high expectations of a return on this investment (Nord, 2011). Additionally, introducing a new medication to the market faces regulatory barriers, which partially hinder its widespread dissemination. (Costa-Font et al., 2015; Dewick & Miozzo, 2002).

Those barriers drive pharmaceutical companies to use different approaches to communicating with physicians. Most companies in this industry use their sales force to convince physicians to prescribe their new products and promote new drugs, making it one of the most expensive activities in the pharmaceutical sector (Mantrala et al., 2010).

1.1. CONTEXTUALIZATION

Companies in this industry have attempted to optimize these channels in various ways. They segment by product, physician profiles, region, and specialty, always seeking the best use of the sales force, expense, and highly specialized team (Lopes, 2012; Manchanda & Chintagunta, 2004). The digitization of the pharmaceutical industry has led to the significant use of digital channels, applications, websites, webinars, and social networks to disseminate new medicines (Fecha, 2017; Hole et al., 2021) and starts to play an essential role in supporting the sales force team. For example, before visiting a physician, one sales representative can share a recorded webinar to present drug details or even share an article posted on LinkedIn.

In this new approach to engaging with physicians, we face an additional challenge: effectively utilizing these channels. This will enable us to target the sales force toward the right physicians, optimizing their costs (Alloghani et al., 2018).

Another critical element in this matter is the distribution of free samples by the sales force to persuade physicians to prescribe the medication, which leads to increased costs for this relationship channel (Khazzaka, 2019).

1.2. STUDY GOALS

The study of (Lopes, 2012) shows that implementing relationship marketing as a coherent process can improve physician loyalty. Five out of ten variables, such as free samples and sales force visits, had a significant impact. In contrast, other variables, such as salesperson knowledge and communication skills, had a negligible impact from the physician's perspective. The perspective of (Fecha, 2017) argues that using digital channels, such as email marketing, has a positive impact on consumer attitudes and provides a better return on investment. Additionally, (Parekh et al., 2016) study demonstrates the patient-centric perspective of pharmaceutical industry digitalization, utilizing social media to showcase the advantages of their new drugs. The research by (Fernandes, 2022) indicates that the pharmaceutical industry has embraced digitalization in the wake of the COVID-19 pandemic, resulting in new

approaches to engaging with patients and establishing a hybrid (digital and face-to-face) work model.

Those studies examine the advantages of various communication channels, analyzing them individually or by specific clusters (brand, region) and promoting their systematic implementation.

This study aims to answer the following question:

- **Is it feasible to create machine learning models to recommend a customized mix of communication channels for each physician?**

1.3. RELEVANCE OF THE STUDY

The diffusion of new drugs is primarily determined by pharmaceutical companies' strategies, government policies, and medical professionals' behavior. Exposure to scientific facts and marketing communication can sway physicians' behavior, and the more exposure alternatives there are, the easier it will be to change this behavior.

The most relevant objective of this study is to apply machine learning to suggest the best combination of communication channels to accelerate the diffusion of new drugs and medications. This would enable physicians to access modern medicines and directly affect the patient's well-being.Introduction sample text Introduction sample text Introduction sample text Introduction sample text Introduction sample text Introduction sample text Introduction sample text Introduction sample text Introduction sample text Introduction sample text.

2. LITERATURE REVIEW

This chapter presents the literature review that formed the foundation for this work. Three major contents are presented:

1) Domains of this Study.
2) How do pharmaceutical companies and other industries manage purchasing behavior and relationship channels?
3) Which Machine Learning techniques are used in different studies?

This study will cover two different domains: relationship channel prediction, and Multilabel classifier.

2.1. HCPS RELATIONSHIP CHANNELS

HCP is the definition of a health care practitioner. This acronym is used not only for doctors but also for nurses, supervisors, and all healthcare professionals. In this study, whenever we use the expression HCP, it is exclusively related to the physician.

2.1.1. DIGITAL RELATIONSHIP CHANNELS

HCPs' relationship behavior is heavily influenced by their culture, community, and background (Nair et al., 2010). Also, the variety set of new technologies that are being developed every day, complemented by new ways to communicate, search for new content, present ideas, share thoughts, drive, read, and relate among communities, has a considerable influence on physician relationship behavior (Parekh et al., 2016).

Based on this perspective, every physician has their technology preference, and following these choices can positively influence how to present or "sell" a new brand. Suppose that one specific physician spends most of your online time on Facebook. The impact of presenting a new drug on Facebook will be greater than on YouTube. Otherwise, a different physician in the same specialty prefers that someone make a presentation of this new drug to a small group of physicians (Manchanda & Chintagunta, 2004).

"*Digital media is an essential part of life. In the pharmaceutical industry, digital marketing is replacing traditional marketing strategies*". This statement from (Jawaid & Ahmed, 2018) study reinforces the paradigm change in the pharmaceutical industry.

The article of (Anthony Jnr & Abbas Petersen, 2021) introduces the concept of a Virtual Enterprise, which was raised after the COVID-19 pandemic. Digitalization has become imperative in medical consultations, leading to the emergence of the digital ecosystem.

A survey conducted by researchers (Jawaid & Ahmed, 2018) in Pakistan revealed that to stay updated on medical information, physicians preferred the below digital channels:

- 32.31% of physicians prefer to read medical websites.
- 20.3% of participants use mobile applications.
- 9.05% use WhatsApp/SMS.
- 8.21% use e-detailing.
- 6.12% use webinars/webcasts.
- 5.15% use tele-detailing.

Although the use of WhatsApp for medical purposes is brief (7 minutes per week on average), it is the most frequently used platform for medical-related purposes (7.5 ± 12.6 times per week). Among the different platforms, webinars/webcasts are used for the longest duration per week (31.67 ± 10.39 minutes), followed by mobile applications (26.5 ± 7.6 minutes) and informative health websites (10 ± 4.2 minutes).

All the elements presented in this chapter emphasize the importance of digital channels in communication strategy. Furthermore, physicians are using digital tools to communicate and acquire knowledge about products and their peers, and companies in the pharmaceutical industry cannot neglect them.

2.1.2. SALES FORCE AND DIGITAL CHANNELS

The pharmaceutical companies do not know each physician's values, myths, and lifestyle. Even so, all kinds of cultural slogans to influence or change the values of the HCPs, like “ethical standard” and “fighting disease first” are used. Those arguments or impacted phrases are still an essential piece of marketing communication strategy (Lopes, 2012).

Management-oriented models are utilized to assess the significance of the sales force in comparison to other marketing channels. These models help to allocate sales efforts among customer segments, geographic areas, and products, determining the optimal size of the sales force and efficiently managing territories. All management aspects of sales force are designed to improve the efficiency of sales visits. However, they also bring added complexity and cost. By gaining a better understanding of the independent variables in Sales force and the various relationship channels, pharmaceutical companies can gain insights to enhance their interactions with customers (Parsons & Abeele, 1981).

The (Mantrala et al., 2010) study conceptualizes the five-layered “complex” of sales force management problem, illustrated on Figure 1.

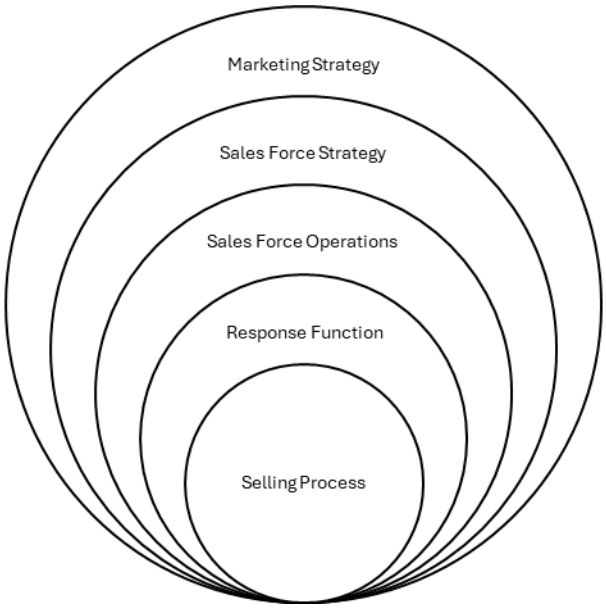

Figure 1 - five-layer complex - adapted from (Mantrala et al.,2010)

In the marketing strategy, the relationship between marketing and sales is revealed to define go-to-market strategies. This relationship becomes even more convoluted when different channels emerge from digitalization (Mantrala et al., 2010). The strategy for the sales force addresses the challenges of determining the size and structure of the sales team, as well as methods for keeping the team motivated through compensation and training. Sales force operations expose the complexity of defining and implementing sales territory design and call planning. The response function describes how interactive the sales efforts should be and the results from these efforts in a way that calibrates the sales force organization. Finally, the selling process includes the tasks of sales representative needs to complete to persuade the physician.

The research from (Mantrala et al., 2010) examines the impact of various digital channels on prescription behavior and how they can be used together to have a positive influence. It reinforces the importance of managing relationship channels and domains to communicate more effectively, thereby helping to reduce the high cost and complexity of sales force organizations.

2.2. MACHINE LEARNING

Behind the Machine Learning concept, there are Data mining and Predictive Analytics definitions. Combining the process of discovering useful patterns, and trends, and the process of extracting information from large datasets to predict and estimate future outcomes respectively (Larose & Larose, 2015a).

Machine learning is the computerized process of discovering patterns and trends from a large dataset, predicting future outcomes, and primarily using these results to adjust future predictions. Using the same patterns and trends always drives your results to the same estimation. Otherwise, enriching that information with results from machine learning can enhance them (Larose & Larose, 2015a).

Machine learning is a field of computer science in which a machine is equipped to analyze data and use statistical methods to produce output within a specific range of information or knowledge (Maimon & Rokach, 2010).

2.2.1. MACHINE LEARNING IN MARKETING

When considering the life cycle of a product or service sale, we are often influenced by the final "touchpoint" or interaction before the purchase. As a result, we tend to attribute significant importance to this particular contact channel in the buying process. However, we cannot fail to consider that before this last move, the customer searched the web, watched a video, signed up for a lecture, received an e-mail, visited a store (physical or online) to learn about the product, and may have gone through infinite different paths (or channels) before purchasing the product.

The marketing department faces a significant challenge regarding brand and product disclosure costs. Concentrating solely on a specific channel may offer cost savings but not yield the desired efficiency. Conversely, employing multiple channels for all customers can be excessively expensive. Machine learning became part of marketers' vocabulary every time that marketing mix is presented or planned, essentially when reports, dashboards, and the huge dataset are explored inside the organization (Duarte et al., 2022).

2.2.2. MULTILABEL PROBLEM APPROACHES

The key challenge in Multilabel prediction is identifying the correlation among them, followed by a reasonable engine to rank them. It is common sense to follow a simple categorization, divide the problem into single labels, or use adapted algorithms to reach the objective of predicting a Multilabel problem.

The most used algorithms for Multilabel prediction are highlighted by the research conducted by (Zhang & Zhou, 2014), and grouped by adaptive algorithms and problem transformation methods.

The following chapters explore the main algorithms used in different studies to help determine which ones better suit this thesis's purpose.

2.2.3. MULTILABEL ALGORITHM ADAPTATION METHODS

2.2.3.1. MULTI-LABEL K-NEAREST NEIGHBOR

The Multilabel k-nearest Neighbor uses k-nearest neighbor techniques to handle Multilabel data, making predictions and getting the labeling information of the neighbors through the maximum a posteriori (MAP) rules (Zhang & Zhou, 2014).

The nearest neighbor rule is a well-known and nonparametric decision procedure for machine learning and data mining tasks. It has been considered one of the most effective algorithms in machine learning and one of the top ten methods in data mining (Garcia et al., 2012; Wu et al., 2008). In traditional supervised learning, a sample x is assigned to the class of the nearest training instance using certain distance metrics. The voting k-nearest neighbor rule (*k*-NN), with k > 1, is a generalization of the NN approach where the predicted class of x is set as equal to the class represented by a majority of its k-nearest neighbors in the training set (Kanj et al., 2016).

2.2.3.2. MULTI-LABEL DECISION TREE

The Multi-Label Decision Tree algorithm aims to handle multi-label data using decision tree techniques. It utilizes an information gain criterion based on multi-label entropy to construct the decision tree in a recursive manner (Zhang & Zhou, 2014). Hierarchical multi-label classification, which deals with instances belonging to multiple classes organized in a hierarchy, has several decision tree induction approaches. An empirical study of their use in functional genomics has been presented in an article by (Vens et al., 2008). One suggestion from (Vens et al., 2008) considers building a separate tree for each class in the hierarchy. Each of these trees is a single-label binary classification tree.

2.2.3.3. COLLECTIVE MULTI-LABEL CLASSIFIER

One algorithm that addresses multi-label data is the Collective Multi-Label Classifier, which utilizes the maximum entropy principle. In this approach, correlations among labels are taken into account by encoding them as constraints that the resulting distribution must satisfy (Zhang & Zhou, 2014).

(Kong et al., 2011) describe multi-label collective classification problem as a way to predict the label sets of a group of related instances simultaneously in the label set space, the Figure 2 demonstrates the power set of all labels where x_i denotes the i-th instance, $\{Y_i^j\}$ is the set of labels assigned to x_i and instances directly linked by an edge are related.

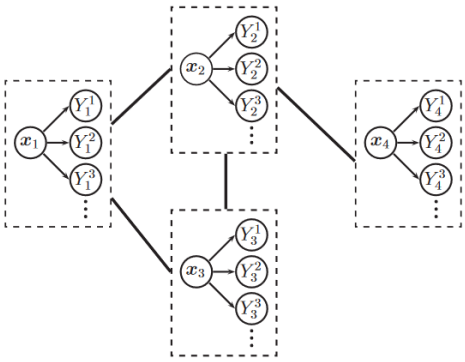

Figure 2 - Collective Multilabel classification (Kong et al., 2011)

2.2.4. MULTILABEL PROBLEM TRANSFORMATION METHODS

2.2.4.1. RANDOM K-LABELSETS

Random k-Labelsets is an algorithm that addresses the Multilabel learning problem by generating an ensemble of multi-class classification models. Each model in the ensemble focuses on a random subset of the label space Y and applies the Label Powerset (LP) technique to induce a multi-class classifier. The LP approach treats each unique combination of labels in the training set as a distinct class value in a single-label classification task. While LP is a straightforward and effective method, it may face computational and performance challenges in scenarios with many labels and few training examples. This is especially true when many labels are associated with very few training instances (Tsoumakas et al., 2011).

2.2.4.2. CALIBRATED LABEL RANKING

Calibrated Label Ranking is an algorithm that transforms the Multilabel learning problem into the label ranking problem, where labels are ranked using pairwise comparison techniques (Zhang & Zhou, 2014).

Label ranking addresses the challenge of learning a mapping from instances to rankings over a predetermined set of labels. Existing approaches to label ranking implicitly operate on an underlying utility scale that lacks a natural zero point, resulting in a lack of calibration. To address this issue, (Fürnkranz et al., 2008) proposes an extension of label ranking that incorporates calibration, greatly expanding the expressive power of these methods. This extension introduces a new technique for extending the common learning by pairwise comparison approach to the Multilabel scenario, which was previously not amenable to pairwise decomposition.

Table 1 - pairwise comparison among labels, adapted from (Fürnkranz et al., 2008) illustrates the pairwise comparison among labels to support the calibration ranking after a binary relevance of each label.

Table 1 - pairwise comparison among labels, adapted from (Fürnkranz et al., 2008)

Labels	λ_1	λ_2	λ_3	λ_4	λ_5	Rank
λ_1		$P\lambda_1\lambda_2$	$P\lambda_1\lambda_3$	$P\lambda_1\lambda_3$	$P\lambda_1\lambda_3$	$\sum P\lambda_1\lambda_j$
λ_2	$P\lambda_2\lambda_1$		$P\lambda_2\lambda_3$	$P\lambda_2\lambda_4$	$P\lambda_2\lambda_5$	$\sum P\lambda_2\lambda_j$
λ_3	$P\lambda_3\lambda_1$	$P\lambda_3\lambda_2$		$P\lambda_3\lambda_4$	$P\lambda_3\lambda_5$	$\sum P\lambda_3\lambda_j$
λ_4	$P\lambda_4\lambda_1$	$P\lambda_4\lambda_2$	$P\lambda_4\lambda_3$		$P\lambda_4\lambda_5$	$\sum P\lambda_4\lambda_j$
λ_5	$P\lambda_5\lambda_1$	$P\lambda_5\lambda_2$	$P\lambda_5\lambda_3$	$P\lambda_5\lambda_4$		$\sum P\lambda_5\lambda_j$

The main idea of the CLR is to identify the dividing point between the relevant and irrelevant labels in the data sample, ranking the labels through their peer-to-peer comparison. The equation 1 below defines the breakpoint.

This breakpoint is defined as follows:

$$\lambda_{i1} > \ldots > \lambda_{ij} > \lambda_0 > \lambda_{ij+1} > \ldots > \lambda_{ic}$$

$$\text{Relevant} = \{\lambda_{i1}, \ldots, \lambda_{ij}\} \,/\, \text{Not relevant} = \{\lambda_{ij+1}, \ldots, \lambda_{ic}\}$$

Equation 1 – Ranking equation (Fürnkranz et al., 2008)

Where λ_n defines a label and λ_0 it's the zero-point.

2.2.4.3. BINARY RELEVANCE

Binary Relevance is an algorithm that addresses the Multilabel learning problem by breaking it down into "n" different binary classification problems, each of which corresponds to a specific prediction in the label space. The current approach has some limitations because it does not consider the correlation among the labels. To overcome this weakness, several modifications to the Binary Relevance algorithm have been introduced in the last ten years to account for correlations between the labels (Zhang et al., 2018).

The Figure 3 - Multilabel dataset (Pfahringer et al., 2021) describes what a Multilabel dataset looks like and the Figure 4 - Two-class datasets independent binary classifiers (Pfahringer et al., 2021) how the same dataset can be split into different binary problems to solve them independently (Pfahringer et al., 2021). The goal of a model is to predict all label $\hat{y}_j \vee j = 1, \ldots, L$ for a given test instance $\mathscr{x}$.

$\boldsymbol{X}$	Y_1	Y_2	Y_3	Y_4
$\boldsymbol{x}^{(1)}$	0	1	1	0
$\boldsymbol{x}^{(2)}$	1	0	0	0
$\boldsymbol{x}^{(3)}$	0	1	0	0
$\boldsymbol{x}^{(4)}$	1	0	0	1
$\boldsymbol{x}^{(5)}$	0	0	0	1
$\tilde{\boldsymbol{x}}$	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$

Figure 3 - Multilabel dataset (Pfahringer et al., 2021)

$\boldsymbol{X}$	Y_1	$\boldsymbol{X}$	Y_2	$\boldsymbol{X}$	Y_3	$\boldsymbol{X}$	Y_4
$\boldsymbol{x}^{(1)}$	0	$\boldsymbol{x}^{(1)}$	1	$\boldsymbol{x}^{(1)}$	1	$\boldsymbol{x}^{(1)}$	0
$\boldsymbol{x}^{(2)}$	1	$\boldsymbol{x}^{(2)}$	0	$\boldsymbol{x}^{(2)}$	0	$\boldsymbol{x}^{(2)}$	0
$\boldsymbol{x}^{(3)}$	0	$\boldsymbol{x}^{(3)}$	1	$\boldsymbol{x}^{(3)}$	0	$\boldsymbol{x}^{(3)}$	0
$\boldsymbol{x}^{(4)}$	1	$\boldsymbol{x}^{(4)}$	0	$\boldsymbol{x}^{(4)}$	0	$\boldsymbol{x}^{(4)}$	1
$\boldsymbol{x}^{(5)}$	0	$\boldsymbol{x}^{(5)}$	0	$\boldsymbol{x}^{(5)}$	0	$\boldsymbol{x}^{(5)}$	1
$\tilde{\boldsymbol{x}}$	$\hat{y}_1$	$\tilde{\boldsymbol{x}}$	$\hat{y}_2$	$\tilde{\boldsymbol{x}}$	$\hat{y}_3$	$\tilde{\boldsymbol{x}}$	$\hat{y}_4$

Figure 4 - Two-class datasets independent binary classifiers (Pfahringer et al., 2021)

Binary relevance is widely used in supervised machine learning to classify the genre of movies on some platforms, given their relative independence. According to (Kumar et al., 2023) separating each genus as a different label in each instance of data allows for a secure identification of the multiple types of genera.

2.2.4.4. ONE VS REST

The term “one-vs-rest” describes a generic classification paradigm that uses binary classification algorithms for multi-class classification and involves tackling the multiclass problem as multiple, one-per-class “binary” classification ones (Vogiatzis et al., 2021). The Multilabel nature of this thesis and the interdependencies among labels allow us to use this method.

Also known as one-vs-all, this strategy involves fitting a separate classifier for each class. Each classifier distinguishes one class from all other classes. This approach is computationally efficient, requiring only a subset of classifiers, and is highly interpretable. One can gain insights into the class by examining the classifier corresponding to each class. As the most used strategy for multiclass classification, it serves as a reliable default choice (scikit-learn OneVsRestClassifier internet, 2024).

2.2.4.5. CLASSIFIER CHAIN

The Classifier Chain is an algorithm that transforms the Multilabel learning problem into a chain of binary classification problems, where each subsequent binary classifier in the chain is built upon the predictions of the previous classifier (Zhang & Zhou, 2014).

(Pfahringer et al., 2021) describe Classifier Chains as a connected binary classifier in a 'chain', such that the output prediction of one classifier is appended as a feature to the input of all subsequent classifiers. This method obtains improved performance over the binary relevance approach. Figure 5 shows the dataset transformation corresponds to a sequential and aleatory chain of labels, one dataset per node/label.

$\boldsymbol{X}$	Y_1
$\boldsymbol{x}^{(1)}$	0
$\boldsymbol{x}^{(2)}$	1
$\boldsymbol{x}^{(3)}$	0
$\boldsymbol{x}^{(4)}$	1
$\boldsymbol{x}^{(5)}$	0
$\tilde{\boldsymbol{x}}$	$\widehat{y}_1$

$\boldsymbol{X}$	Y_1	Y_2
$\boldsymbol{x}^{(1)}$	0	1
$\boldsymbol{x}^{(2)}$	1	0
$\boldsymbol{x}^{(3)}$	0	1
$\boldsymbol{x}^{(4)}$	1	0
$\boldsymbol{x}^{(5)}$	0	0
$\tilde{\boldsymbol{x}}$	$\widehat{y}_1$	$\widehat{y}_2$

$\boldsymbol{X}$	Y_1	Y_2	Y_3
$\boldsymbol{x}^{(1)}$	0	1	1
$\boldsymbol{x}^{(2)}$	1	0	0
$\boldsymbol{x}^{(3)}$	0	1	0
$\boldsymbol{x}^{(4)}$	1	0	0
$\boldsymbol{x}^{(5)}$	0	0	0
$\tilde{\boldsymbol{x}}$	$\widehat{y}_1$	$\widehat{y}_2$	$\widehat{y}_3$

$\boldsymbol{X}$	Y_1	Y_3	Y_3	Y_4
$\boldsymbol{x}^{(1)}$	0	1	1	0
$\boldsymbol{x}^{(2)}$	1	0	0	0
$\boldsymbol{x}^{(3)}$	0	1	0	0
$\boldsymbol{x}^{(4)}$	1	0	0	1
$\boldsymbol{x}^{(5)}$	0	0	0	1
$\tilde{\boldsymbol{x}}$	$\widehat{y}_1$	$\widehat{y}_2$	$\widehat{y}_3$	$\widehat{y}_4$

Figure 5 - chain classifier dataset (Pfahringer et al., 2021)

Classifier Chain takes advantage of correlation among labels, but due to the random nature of label ordering, it could yield worse results than an independent model.

The Classifier Chain is the meta-estimator (an estimator that uses an inner estimator) that implements a more advanced strategy. The ensemble of binary classifiers is used as a chain, where the prediction of one classifier in the chain is used as a feature to train the next classifier on a new label. Therefore, these additional features allow each chain to exploit correlations among the labels (scikit-learn classifier chain internet, 2024).

An important consideration in this algorithm is the order in which the labels are processed. It is crucial to disregard randomness and, using business knowledge, determine the appropriate order for adding the labels to the chain.

There are two options to consider. One is to establish the most used labels. Another is to prioritize labels based on preferences using a relationship strategy. For instance, if the strategy involves favoring digital channels, we may consider face-to-face interactions a lower priority in the classification process or exclude them from the set of labels altogether.

2.2.5. MACHINE LEARNING DOMAIN

This thesis is based on those three concepts. The relationship channel interactions, prescription volume, and physician basic information to predict the best relationship channel mix by physician.

The ability to assign a value to a channel, considering a specific individual, poses a challenge. However, this becomes increasingly feasible and easy to achieve with the application of machine learning combined with a considerable volume of data (Kumar et al., 2020; Ling et al., 2019; Nan & Hu, 2022; Punia et al., 2020).

This thesis aims to demonstrate the feasibility of predicting the optimal channel mix for each physician using machine learning classifiers. This thesis will use three different algorithms to evaluate and confirm this analysis. 1) binary relevance, 2) Calibrated Label Ranking (One vs. rest), and 3) Classifier Chains.

The decision to use those algorithms is based on their simplicity and similarity to the real world. We must also understand the evolution from binary relevance to classifier chain, passing through the one vs. rest method.

2.2.5.1. TRAIN AND TEST DATASET

The training and test datasets are critical for any model in machine learning. The bigger the training data, the better the classifier. On the other hand, the bigger the test data, the better the estimate of the classifier's performance on unseen data (Duarte et al., 2022). A random technique was applied to split datasets.

In the machine learning domain, there is a discussion about applying static training and testing datasets to complex models where data evolves or transforms (Tan et al., 2021). However, this thesis keeps the static aspect of the training and test bases, applying a correction to the imbalance between variables and labels in the training base through the Random Oversampling technique.

All the datasets are unbalanced, with varying proportions for each label, so we used the Random Oversampling technique. This technique involves replicating (cloning) the labels with smaller quantities proportionally until the rows of the training dataset reach the quantity of the majority of the inputs.

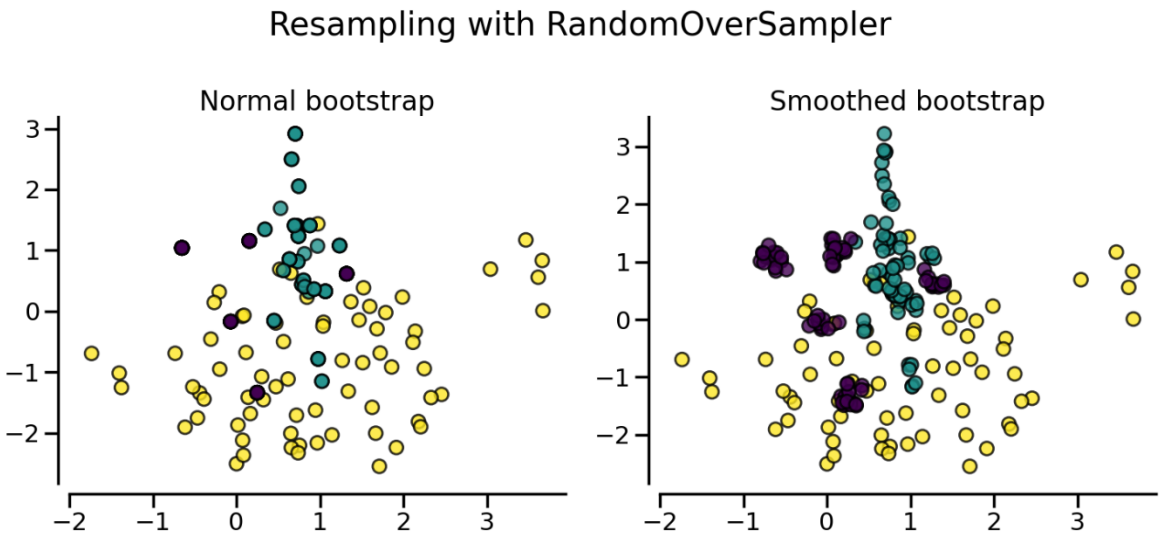

Figure 6 - Resampling with Random Oversampling (*Over-Sampling Doc Internet*, 2024)

Random Oversampling enlarges the density of the minority label until it reaches a balanced training dataset. Random oversampling applies to binary or multi-class data, and due to the Multilabel nature of this thesis, it was necessary to convert the Multilabel to multi-class data (Maimon & Rokach, 2010).

2.2.5.2. VARIABLES SELECTION

Feature selection is a critical aspect of the Machine Learning process, and the goal of this thesis is even more important because it deals with 53 different features.

This thesis uses recursive feature elimination with cross-validation (RFECV) to select features. The number of features selected is tuned automatically by fitting an RFE selector on the different cross-validation splits. The performance of the RFE selector is evaluated using a scorer for a different number of selected features and aggregated together. Finally, the scores are averaged across folds, and the number of features selected is set to the number of features that maximize the cross-validation score.

This thesis used the same algorithm to predict the results and to fit an RFE algorithm to select the best variables. The exception was the Neural Networks algorithm, which does not apply to selecting features. To remedy this situation, this thesis adopted Logistic Regression to select features.

Considering the nature of estimators, this thesis uses the “F1 Score” as a scorer to select features. The F1 score is a harmonic means of precision and recall, where the best is near 1 and the worst is near 0. Both recall and precision contributions are the same. The relative contributions of precision and recall to the F1 score are equal. Equation 2 - F1 score formula presents the F1 formula.

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}$$

Equation 2 - F1 score formula (scikit-learn f1_score metrics internet, 2024)

Where:

- TP - True Positives
- FN - False Negatives, and
- FP - False Positives.

2.2.5.3. CONFUSION MATRIX

In the Machine Learning domain, a confusion matrix is a table that allows visualization of the performance of a classification algorithm, as shown in Figure 7. This contingency table is also called an error matrix. Each row of the array represents instances of a classification, while each column represents instances of the actual class. The name comes from the easiest way to see confusion between two classes: comparing the prediction and current class.

		True values	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

Figure 7 – Confusion Matrix representation adaptation from (Larose & Larose, 2015)

Where:

- Prediction matches the current class:
  - TP (True Positive)
  - TN (True Negative)
- Prediction does not match the current class
  - FP (False Positive)
  - FN (False Negative)

2.2.5.4. EVALUATION SCORES

The precision, recall, and F1-measure are widely used in single-label classification evaluation, which is also applicable for multi-label classification by using two averaging methods named micro and macro (Tsoumakas et al., 2010).

F1

As used in Variable Selection this thesis calculates F1 score to support prediction. “Support beyond binary targets is achieved by treating multiclass and Multilabel data as a collection of binary problems, one for each label.” (scikit-learn f1_score metrics internet, 2024).

ACCURACY

The accuracy score calculates the percentage of events correctly identified (positive or negative) out of all events. This measure is good when the dataset is balanced, and True Positive and True Negative have the same weight (Larose & Larose, 2015).

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

Equation 3 - Accuracy formula (scikit-learn Metrics and scoring internet, 2024)

Where $\boldsymbol{y}$ is the true value and $\hat{\boldsymbol{y}}$ is the prediction.

ERROR RATE

The proportion of overall events incorrectly identified (positive or negative). This measure is good when the dataset is balanced; consequently, False Positives and False Negatives have the same weight (Larose & Larose, 2015).

$$\text{Error rate} = 1 - Accuracy$$

Equation 4 - Error rate formula adapted from (scikit-learn Metrics and scoring internet, 2024)

PRECISION

The precision is intuitively the classifier's ability not to label a negative sample as positive. This is a good strategy when the consequences of incorrectly identifying something as positive are costly (Larose & Larose, 2015).

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

Equation 5 - Precision formula adapted from (Larose & Larose, 2015)

The best value is 1, and the worst value is 0.

RECALL

The recall is intuitively the classifier's ability to find all the positive samples. This is an effective approach when the consequences of missing something important are very costly. (Larose & Larose, 2015).

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

Equation 6 - recall formula adapted from (Larose & Larose, 2015)

The best value is 1 and the worst value is 0.

DUMMYCLASSIFIER

The DummyClassifier defines a baseline for comparison with the classification models used in this thesis. All strategies used in this method ignore the variables to make predictions, using only the labels to randomly and statistically suggest results for the model. Considering that the datasets are not balanced, this thesis used the “most frequent” generation strategy, thus allowing the inclusion of both results in the label set (Pfahringer et al., 2021).

3. METHODOLOGY

This work will apply a Design Science Research Methodology. (Peffers et al., 2020) - to define the concept, build the prototype, evaluate the result, and develop a machine learning model to answer those questions (an artifact).

Additionally, it will use quantitative/positivism methods, respectively, to emphasize quantitative data and positivism philosophy (Hameed, 2020).

2.3. DSRM – DESIGN SCIENCE RESEARCH METHODOLOGY

Because this work will deliver an artifact – a data analytics solution – the best standard process to support the development is the principle of design science research (Peffers et al., 2007).

DSRM is based on an interactive process where we create a concept, define its objective, build an IT system, present the results, and evaluate them. This cycle can be repeated until we have a satisfactory answer to our concept (Gledson et al., 2024).

Figure 8 – DSRM general process – adapted from (Peffers et al., 2020) process model

The interactive nature of DSRM helped our research to evolve the solution during its development, and based on the process model I mapped the deliverables like below.

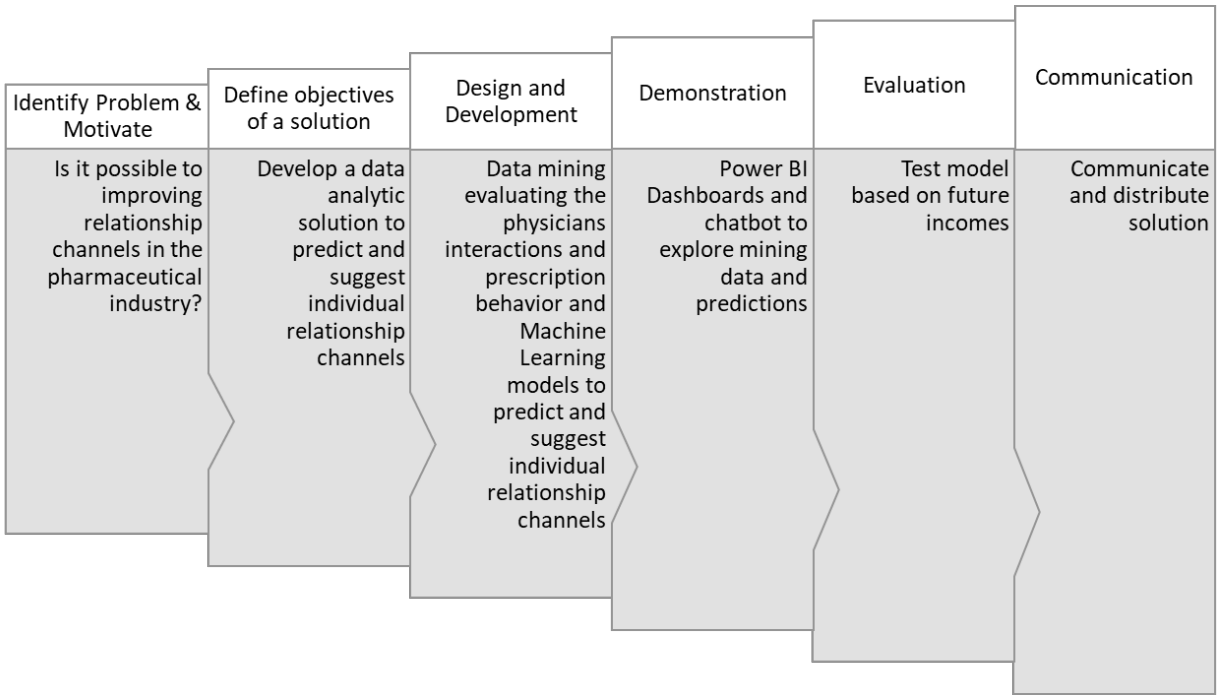

Figure 9 – DSRM process deliverable - adapted from process model (Peffers et al., 2007)

2.4. ITERATIVE PROCESS

The iterative nature of this work suggests using the AGILE methodology to support development, including build and evaluation.

This thesis appropriates the development cycle of AGILE METHODOLOGY to develop the algorithms to support the analysis and evaluation of this work. The Figure 10 - DSRM and Iterative process - adapted from (Peffers et al., 2007; Scrum.org, 2023) methods illustrates the iterative process over the DSRM method.

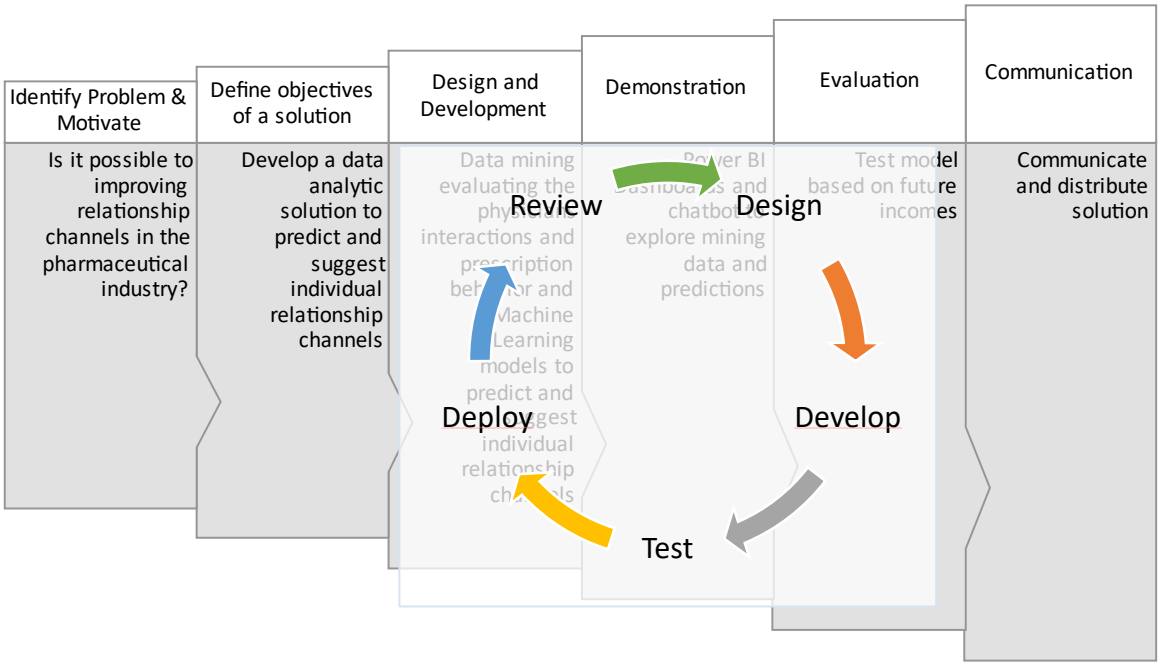

Figure 10 - DSRM and Iterative process - adapted from (Peffers et al., 2007; Scrum.org, 2023) methods

4. DEVELOPMENT

This study aims to evaluate the effectiveness of utilizing machine learning to convey product benefits in the pharmaceutical industry through optimal communication channels. This section details the framework, data organization, and the models and techniques employed to support this study.

2.5. FRAMEWORK

The tools used in this thesis were based on Microsoft technology. The list of tools is provided below:

- SQL Server – database technology to store all data used in this work.
- Azure Synapse – mainly used as ETL technology to prepare all datasets used in this study.
- Visual Studio Code – coding technology to develop Python algorithms used in this study.
- Power BI – reporting technology to present KPIs and scores to analyze and evaluate the objective of this study.

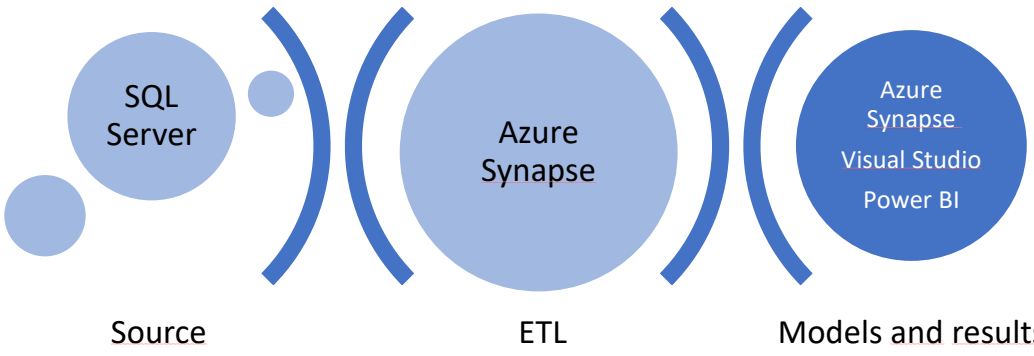

Figure 11 - tools framework (own authorship)

2.6. DATA UNDERSTANDING

The data sets used in this study include qualified communication actions by channel, prescriptions, and physician´s demographic data from Brazil. Qualified reach communication action considers the relevance of the interaction; for example, clicking on a link in an email is more important than sending the email to a Physician.

To simplify the study and reduce the data volume, this study aggregates all data of 2019. This thesis removes the product and brand dimensions for prescription data and aggregates the number of prescriptions by market. For communication interaction, this study aggregates all qualified actions by Channels. For example, for e-mail mass interaction, this study considers only the actions “clicked the email” and “Opened the e-mail.” For Demographic data, this study considers the position at the end of 2019.

The market definition is a combination of products among proprietary brands and their competitors, where the marketing team selects the major competitor brands. This thesis did not provide any critical validation for those market combinations.

In the below datasets, all data were loaded non-normalized in Azure Blob Storage in delta parquet format and CSV file. All data is already cleaned, and no transformation was needed.

- **HCPs (customersfeatures)** – Features of HCPs – demographic data.
- **HCPsPrescriptions (salesfeatures)** – Features of HCP – prescriptions number by market.
- **HCPsInteractions (relationshiplabels)** – Labels of HCPs – channels interactions and date.

2.6.1. HCPS DEMOGRAPHIC (FEATURES)

HCPS demographic data includes three classes of information.

1. Influence in the medical community – Clinical investigator and speaker are the features that define whether a physician is an investigator in the Research and Development of new drug and whether a physician is a speaker at conferences and seminars.
2. Gender – Physician's gender
3. Specialty – Identifies which specialization a doctor can have; it is common for a doctor to be a general practitioner and have an additional specialty.

The features dataset does not include sensitive or private customer information. It also contains foreign keys (HCPID) to identify physicians in prescription features data and communication channels interaction labels data.

Variables definition:

- **clinical_investigator** - Define if a physician is an investigator, usually participating on scientific investigation for new drugs.
- **gender** - Define the physician´s gender. Domain of gender:
  1. Female
  2. Male
  3. Other
- **speaker** - Define if a physician is also a speaker in congress or special events.
- **SPECIALTIES** (Anesthesiology, Cardiovascular Disease, Gastroenterology, etc.) - A set of specialties to which one physician can belong. A physician can be a Cardiologist or a General Medicine doctor.

Except for gender, all other features are binary classification variables, where 0 (zero) means that the physician is not included, and 1 (one) is included. Table 2 shows some descriptive statistics for the variables gender, clinical investigator, speakers, and specialties.

Table 2 – statistics of gender, clinical investigator, speakers and specialties features

Variables	unique	mean	std	min	max
gender	3	2.775	0.574	1	3
clinical_investigator	2	0.001	0.025	0	1
speaker	2	0.005	0.072	0	1

Variables	unique	mean	std	min	max
Anesthesiology	2	0.049	0.216	0	1
Cardiovascular Disease	2	0.067	0.250	0	1
Dermatology	2	0.070	0.255	0	1
Gastroenterology	2	0.023	0.149	0	1
General Medicine	2	0.436	0.496	0	1
Hematology	2	0.010	0.099	0	1
Nephrology	2	0.012	0.111	0	1
Neurology	2	0.033	0.180	0	1
Nutrition	2	0.004	0.065	0	1
Obstetrics Gynaecology	2	0.071	0.257	0	1
Oncology	2	0.023	0.150	0	1
Ophthalmology	2	0.011	0.105	0	1
Orthopedic Trauma	2	0.019	0.136	0	1
Other Specialty	2	0.147	0.354	0	1
Otolaryngology	2	0.048	0.214	0	1
Pediatrics	2	0.206	0.404	0	1
Proctologys	2	0.004	0.064	0	1

Psychiatrys	2	0.064	0.244	0	1
Pulmonarys	2	0.028	0.165	0	1
Rheumatologys	2	0.016	0.127	0	1
Surgerys	2	0.088	0.283	0	1
Urologys	2	0.037	0.188	0	1
Allergy_Immunology	2	0.000	0.000	0	0
Endocrinology_Diabetes_Metabolism	2	0.000	0.007	0	1
Geriatrics_Gerontology	2	0.000	0.000	0	0

The total amount of physicians in all datasets is 133230. The highlight of this dataset is the number of General Medicines, most physicians are part of the General Medicines specialty, followed by pediatric and other unspecified specialties. The frequency of each category of physicians by gender, clinical investigator, speakers and specialties is presented in Table 3.

Table 3 – quantity of physician by gender, clinical investigator, speakers and specialties

	Categories		
Variable	**1**	**2**	**3**
gender	10380	9192	113658

	Categories	
Variable	**0**	**1**
clinical_investigator	133149	81
speaker	132544	686
Allergy_Immunology	133230	0
Anesthesiology	126672	6558
Cardiovascular Disease	124290	8940
Dermatology	123938	9292
Endocrinology_Diabetes_Metabolism	133224	6
Gastroenterology	130217	3013
General Medicine	75190	58040
Geriatrics_Gerontology	133230	0
Hematology	131912	1318
Nephrology	131581	1649
Neurology	128772	4458
Nutrition	132670	560
Obstetrics Gynaecology	123754	9476
Oncology	130171	3059
Ophthalmology	131735	1495
Orthopedic Trauma	130702	2528
Other Specialty	113641	19589
Otolaryngology	126824	6406
Pediatrics	105836	27394

Proctologys	132684	546
Psychiatrys	124747	8483
Pulmonarys	129477	3753
Rheumatologys	131043	2187
Surgerys	121540	11690
Urologys	128317	4913

2.6.2. HCPs Prescriptions (Features)

Another set of features is prescription. Prescription is not the sale of one specific product, but the number of products consumed by one user. This data can be captured in drug store databases or, in some countries, when physicians prescribe medicines electronically.

Prescription data are provided by product and physician, but this thesis aggregates on the market and physician levels. One market includes products from different competitors' laboratories, and the objective is to create market share based on the Marketing team's perspective.

This dataset includes total prescriptions by market and physicians of 2019. It also includes prescriptions for different laboratories and brands. This dataset is captured in pharmacies and is not exhaustive, not covering the entire universe of prescriptions. Table 4 describes the statistics of dataset prescriptions for each market.

Table 4 – statistics of HCPsPrescriptions dataset

Variables	mean	std	min	25%	50%	75%	max
R03-TRELEGY BRONCO	0.242	2.085	0	0	0	0	169
J05-A.HERPETICOS	0.392	1.359	0	0	0	0	111
G04-HPB	1.246	7.946	0	0	0	0	628
N06-BUPROPIONA	5.275	22.398	0	0	0	2	795
R03-ASMA RESGATE	5.983	26.403	0	0	0	3	2024
D10-RETINOIDES	0.427	2.911	0	0	0	0	262
D11-TARGET ALOPECIA	0.672	3.277	0	0	0	0	420
J01-AMOXICILINA	10.509	30.627	0	0	2	8	2189
J01-AMOXICILINA CLAVUL	19.004	52.853	0	1	4	13	1632
R01-CORT.NASAIS	6.181	24.212	0	0	1	3	2126
D07-DERMATITE	1.478	6.613	0	0	0	1	480
D10-ACNE COMBINACAO	1.532	9.053	0	0	0	0	615
J01-ZINNAT TOTAL	17.559	36.937	0	2	6	18	1212
R03-ICS LABA ASMA	1.720	10.014	0	0	0	1	919
R03-A.ASMATICOS CORT.I	6.378	29.800	0	0	0	3	2571
G04-HPB TNS	1.246	7.946	0	0	0	0	628
N06-ANTDEPRESSAO-SSRI	18.968	74.295	0	0	2	8	2832
R03-ICS LABA ASMA SRT	1.720	10.014	0	0	0	1	919

N03-LAMOTRIGINAS	3.452	21.119	0	0	0	1	2457
D06-TARGET BACTROBAN	2.183	9.209	0	0	0	1	616
R03-TARGET TRELEGY	1.962	11.575	0	0	0	1	986
J01-ZINNAT PREMIUM	6.412	23.631	0	0	1	4	2342
R03-TARGET VANISTO	0.242	2.085	0	0	0	0	169
R03-TARGET ANORO	0.090	0.911	0	0	0	0	70

Market definition

The Table 5 - Volume of prescription by market, presents the quantity of prescription. The market concept is the combination of different brands selected by the marketing team. One market in this dataset includes proprietary brands and their main competitors. For example, J01-AMOXICILLIN_CLAVUL includes the amoxicillin antibiotic from different competitors but also the brand CLAVULIN.

Table 5 - Volume of prescription by market

market	quantity
D06-TARGET BACTROBAN	290875
D07-DERMATITE	196928
D10-ACNE COMBINACAO	204129
D10-RETINOIDES	56858
D11-TARGET ALOPECIA	89478
G04-HPB	166038
G04-HPB TNS	166038
J01-AMOXICILINA	1400114
J01-AMOXICILINA CLAVUL	2531909
J01-ZINNAT PREMIUM	854295
J01-ZINNAT TOTAL	2339386
J05-A.HERPETICOS	52208
N03-LAMOTRIGINAS	459926
N06-ANTDEPRESSAO-SSRI	2527056
N06-BUPROPIONA	702809
R01-CORT.NASAIS	823498
R03-A.ASMATICOS CORT.I	849745
R03-ASMA RESGATE	797049
R03-ICS LABA ASMA	229203
R03-ICS LABA ASMA SRT	229203
R03-TARGET ANORO	11933
R03-TARGET TRELEGY	261442
R03-TARGET VANISTO	32239
R03-TRELEGY BRONCO	32239

2.6.3. HCPSINTERACTIONS (LABELS)

This dataset represents the labels for this study. This thesis aims to assess the feasibility of recommending a combination of communication channels based on the history of interactions. All labels are binary classifications, where 1 (one) means that one physician interacted with the channel and 0 (zero) means that the physician had no interaction with the channel.

This dataset includes all qualified reach interactions among channels and physicians. One interaction means one communication event or action between the organization and its customers.

Labels definition:

- **e-Mail_1to1** – the physician had and e-mail one to one interaction – e-mail sent by sales force team - for example, open the e-mail sent or click on a link inside the e-mail.
- **MEETINGS_F2F_Meetings** – the physician participated in a face-to-face meeting among doctors and the sales team
- **e-Mail_Mass** – the physician had and e-mail one to one interaction, for example, open the e-mail sent or click on a link inside the e-mail.
- **Webinar_Webinar** – the physician participated in a webinar promoted by marketing team
- **F2F_f2f** – the physician was visited by a sales force member
- **Web Site_PORTAL** – the physician navigates and identify themselves in business website
- **F2F_PHONE_CALL** – the physician received a call from a sales force member
- **MEETINGS_REMOTE_MEETING** – the physician participated in a virtual meeting among doctors and the sales team

In the same way as other datasets, this thesis aggregates the interactions by channel and physician in 2019. This dataset shows which channels were used to interact or share information with one physician Table 6 – statistics of HCPSInteraction dataset.

Table 6 – statistics of HCPSInteraction dataset

labels	mean	std	min	25%	50%	75%	max
e-Mail_1to1	0.187	0.390	0	0	0	0	1
MEETINGS_F2F_Meetings	0.061	0.239	0	0	0	0	1

e-Mail_Mass	0.852	0.355	0	1	1	1	1
Webinar_Webinar	0.004	0.061	0	0	0	0	1
F2F_f2f	0.419	0.493	0	0	0	1	1
Web Site_PORTAL	0.014	0.119	0	0	0	0	1
F2F_PHONE_CALL	0.025	0.157	0	0	0	0	1
MEETINGS_REMOTE_MEETIN G	0.099	0.299	0	0	0	0	1

The Table 7 – quantity of physicians by channel, illustrates the volume of interactions by channel, and the highlights are the channel's e-mail mass and face-to-face with a higher volume of interactions. Considering the volume of face-to-face interaction, one assumption is that this channel was a priority in 2019. This kind of channel is highly specialized and extremely qualified for communication with physicians.

Table 7 – quantity of physicians by channel

	true value	
channel	**0**	**1**
F2F_PHONE_CALL	129862	3368
F2F_f2f	77431	55799
MEETINGS_F2F_Meetings	125134	8096
MEETINGS_REMOTE_MEETING	120013	13217
Web Site_PORTAL	131330	1900
Webinar_Webinar	132731	499
e-Mail_1to1	108312	24918
e-Mail_Mass	19747	113483

2.7. MODELS

All program codes for this thesis were developed using Python and the Scikit Learn and Imbalanced Learn library. Scikit-learn (scikit-learn internet, 2024) and imbalanced-learn (imbalanced-learn internet, 2024) are open-source machine learning libraries that supports both supervised and unsupervised learning. Scikit-learn provides tools for model fitting, data preprocessing and model selection, and imbalanced-learn, model evaluation and classification with imbalanced classes.

2.7.1. TRAIN AND TEST DATASET

The dataset was split into train and test datasets, considering a proportion of 70% and 30%, respectively for training data and testing data. The division of data into training and testing sets was carried out in two stages, each serving a different purpose. Firstly, it was done to choose the best set of variables using recursive feature elimination with cross-validation (RFECV) for the "Binary Relevance" model. Secondly, it was done to provide training and testing data for all the models used in this thesis.

2.7.1.1. OVER SAMPLING DATASET

To minimize the unbalance problem in those datasets, this thesis adopted the "Random Over Sampling" algorithm (RandomOverSampler library internet, 2024). For example, in the label set, the channel “email_1to1” has a proportion of 81% to 19%, respectively, to have interaction and not have interaction; on the other hand, the channel “email_mass” has a proportion of 15% to 85% in the opposite meaning. The main reason to use the ROS algorithm instead of SMOTE, ADASYN, or another method was your simplicity in explaining.

The ROS algorithm has the limitation of being applicable only to label sets with a single label (numeric, binary, or multiclass). This limitation was circumvented by transforming the Multilabel set of this thesis into a multiclass label. The solution was to concatenate all the labels into a single multiclass label, as shown in the Figure 12 – Labels concatenated – result extracted from python execution. Thus, it was possible to use the same algorithm to complete the training dataset in all methods, including Multilabel classifier chain model.

```
                                MC_label
0       0.0#0.0#0.0#1.0#1.0#0.0#1.0#0.0
1       0.0#0.0#0.0#0.0#1.0#0.0#0.0#0.0
```

Figure 12 – Labels concatenated – result extracted from python execution

After concatenating and addressing the imbalance in the datasets, the labels returned to their original form (Multilabel), splitting the concatenated multiclass label as presented in Table 8 - concatenated labels converted to Multilabel set. The order of concatenation is crucial for converting to Multilabel, representing the relationship channels.

Table 8 - concatenated labels converted to Multilabel set

	labels							
	F2F_PHONE_CALL	F2F_f2f	MEETINGS_F2F_Meetings	MEETINGS_REMOTE_MEETIN	Web Site_PORTAL	Webinar_Webinar	e-Mail_1to1	e-Mail_Mass
0	0	0	0	1	1	0	1	0
1	0	0	0	0	1	0	0	0

2.7.2. VARIABLES SELECTION

This thesis uses recursive feature elimination with cross-validation (RFECV) to select features. The scorer adopted was “F1-Score”.

2.7.3. ALGORITHMS

The goal of this thesis is to predict a mix of relationship channels to a physician, typically a multi-label scenario. To support this objective, this thesis uses the “Binary Relevance,” “Classifier Chain,” and “One vs Rest” models to predict and capture results. In this chapter, the thesis will present the algorithms used and the parameters approach for each model.

2.7.3.1. BINARY RELEVANCE

This study used “Binary Relevance” to predict each label individually. To enable a broader view of the possible outcomes, we applied the methods of “Logistic Regression,” “Neural Network,” “Decision Tree,” and “Random Forest.”

Logistic regression

This research uses 500000 as parameter to max_iter (Maximum number of iterations taken for the solvers to converge) and all other parameters with their default values, including the L2 penalty, where the classification space is low and gives less freedom to the LogisticRegression module. Max interaction was increased to fit the requirement of the data available.

Neural Network

This thesis uses the parameters max interaction (max_iter) with 400, activation as “Logistic,” learning rate (learning_rate_init) equal to 10% (0.01), hidden layer (hidden_layer_sizes) equal to 10 layers, and all other parameters with their default values for MLPClassifier module (scikit-learn Neural Network internet, 2024). Max_iter was increased to fit the requirement of data availability.

Decision Tree

This study uses the parameter random state equal to 0 and all other parameters with their default values to DecisionTreeClassifier model (scikit-learn DecisionTreeClassifier internet, 2024).

Random Forest

This research uses the parameter number of estimators (n_estimators) equal to 10 and all other parameters with their default values for RandomForestClassifier module (RandomOverSampler library internet, 2024).

2.7.3.2. MULTI-LABEL CLASSIFIER CHAIN

This thesis utilizes the ClassifierChain module (scikit-learn ClassifierChain internet, 2024). The base_estimator used is the same one as that used for "Binary Relevance," which includes Logistic Regression, Neural Network, Decision Tree, and Random Forest. The parameters used for each of them were the same as those adopted in “Binary Relevance.” This approach allows this thesis to compare the scores based on the same algorithms.

2.7.3.3. ONE VS REST CLASSIFIER

This thesis uses the module OneVsRestClassifier (scikit-learn OneVsRestClassifier internet, 2024) with the same estimator (base_estimator) used for Binary Relevance, Logistic Regression, Neural Network, Decision Tree, and Random Forest. The parameters applied for each of them were the same adopted in “Binary Relevance”, this approach allows this thesis to compare the scores based on the same algorithms.

5. RESULTS AND DISCUSSION

One CSV file was created to summarize each model's scores and test prediction results. Based on these results, it created a set of charts and pivots in Power BI comparing the different models presented in the Annexes of this thesis. The scores adopted in this thesis were F1, Accuracy, Error rate, Precision, and Recall, but all analyses were based on the F1 score.

Looking at the metrics results, even applying random over-sampling, the unbalanced dataset has a huge impact on “Binary Relevance” methods, for example, the channel “Web Site_PORTAL” has a distribution of 99% (no interaction) to 1% (interaction) and got a F1 score of 2,10%, on other hand, the channel “F2F_f2f” has a more equity distribution of 58% (no interaction) to 42% (interaction) and got a F1 score of 64,29%. This may have originated because of the sampling strategy that uses “not majority” where resample all classes except the majority.

Another highlight is One vs. Rest, in which the F1 Score was the best of four estimators; The exception was the decision tree. The Classifier Chain performed better with an F1 score of 67.10% compared to 66.92%.

The F1 score for the 'Dummy Classifier' is 47.18%. Considering the metrics obtained in the Multilabel models, as shown in the table below, this thesis considered it sufficient to proceed with the parameters of each model, but this is one of the enhancements mentioned in the recommendation for future works.

Table 9 - Summary metrics

Model	Algorithm	Channel	Accuracy	Error rate	F1score	Precision	Recall
Dummy Classifier			**54.28%**	45.7'1%	47.18%	43.65%	51.32%
Classifier Chain	Decision Tree	Multilabel (All)	42.53%	57.47%	67.10%	66.79%	**67.46%**
	Logistic Regression		50.46%	49.54%	61.48%	**75.83%**	56.66%
	Neural Network		55.67%	44.33%	66.08%	**77.51%**	65.36%
	Random Forest		55.27%	44.73%	66.11%	**75.67%**	63.64%
One vs Rest	Decision Tree	OneVsRest (All)	36.81%	63.19%	66.92%	**67.23%**	66.66%
	Logistic Regression		49.23%	50.77%	66.80%	**75.10%**	63.55%
	Neural Network		53.62%	46.38%	68.82%	**75.08%**	70.47%
	Random Forest		52.10%	47.90%	69.85%	**73.67%**	70.06%

Binary Relevance	Decision Tree	e-Mail_1to1	**76.87%**	23.13%	38.64%	38.48%	38.80%
		e-Mail_Mass	74.90%	25.10%	**85.14%**	86.00%	84.31%
		F2F_f2f	**70.18%**	29.82%	64.29%	64.27%	64.31%
		F2F_PHONE_CALL	**95.32%**	4.68%	10.27%	10.00%	10.55%
		MEETINGS_F2F_Meetings	**89.29%**	10.71%	16.34%	16.02%	16.67%
		MEETINGS_REMOTE_MEETING	**84.54%**	15.46%	24.49%	23.79%	25.24%
		Web Site_PORTAL	**96.73%**	3.27%	2.10%	1.83%	2.46%
		Webinar_Webinar	**99.14%**	0.86%	3.37%	2.96%	3.92%
	Logistic Regression	e-Mail_1to1	73.10%	26.90%	52.06%	39.11%	**77.85%**
		e-Mail_Mass	58.63%	41.37%	70.40%	**90.33%**	57.67%
		F2F_f2f	**78.64%**	21.36%	73.44%	76.33%	70.77%
		F2F_PHONE_CALL	**83.11%**	16.89%	19.41%	11.04%	80.18%
		MEETINGS_F2F_Meetings	72.87%	27.13%	25.27%	15.27%	**73.09%**
		MEETINGS_REMOTE_MEETING	**76.77%**	23.23%	37.13%	25.39%	69.07%
		Web Site_PORTAL	57.43%	42.57%	4.18%	2.16%	**65.32%**
		Webinar_Webinar	**80.03%**	19.97%	2.97%	1.51%	79.74%
	Neural Network	e-Mail_1to1	74.83%	25.17%	54.35%	41.19%	**79.88%**
		e-Mail_Mass	59.60%	40.40%	71.21%	**90.80%**	58.58%
		F2F_f2f	78.62%	21.38%	75.48%	72.41%	**78.82%**
		F2F_PHONE_CALL	**81.77%**	18.23%	18.08%	10.20%	79.29%
		MEETINGS_F2F_Meetings	73.27%	26.73%	25.75%	15.60%	**73.88%**
		MEETINGS_REMOTE_MEETING	**73.79%**	26.21%	35.84%	23.68%	73.70%
		Web Site_PORTAL	**62.08%**	37.92%	4.02%	2.08%	55.81%
		Webinar_Webinar	**89.74%**	10.26%	4.16%	2.16%	58.17%

	Random Forest	e-Mail_1to1	**81.63%**	18.37%	44.43%	51.37%	39.13%
		e-Mail_Mass	79.72%	20.28%	88.42%	86.22%	**90.73%**
		F2F_f2f	**77.26%**	22.74%	71.51%	74.95%	68.38%
		F2F_PHONE_CALL	**97.17%**	2.83%	3.90%	13.94%	2.27%
		MEETINGS_F2F_Meetings	**92.87%**	7.13%	10.54%	24.74%	6.70%
		MEETINGS_REMOTE_MEETING	**89.19%**	10.81%	25.78%	40.56%	18.89%
		Web Site_PORTAL	**98.32%**	1.68%	0.59%	1.83%	0.35%
		Webinar_Webinar	**99.50%**	0.50%	0.00%	0.00%	0.00%

6. CONCLUSIONS AND FUTURE RESEARCH

In this thesis, we observed that applying machine learning to suggest communication relationship channels for physicians is possible and feasible. Initiating a partnership with a doctor through past experiences (learning) can be effective in ensuring doctors' loyalty to future prescriptions.

The model developed in this thesis confirms that regardless of the algorithm or method applied—"Binary Relevance," "Classifier Chain," or "One vs Rest"—there is a positive outcome in identifying the doctor's prescriptive behavior and demographic information to suggest a known communication approach through relationship channels.

From the point of view of the methods and algorithms used, "Classifier Chain" is the most appropriate model, mainly due to its easy comprehension and flexibility in applying organizational communication strategies. Thus, when we leave the less strategic channels to the end of the classification chain, we influence the model to follow a desired pattern. Another aspect that favors using the "Classifier Chain" model is that we can influence it simply by removing some channels or actions from the database. For example, we can remove all face-to-face actions, creating a purely digital scenario.

In any case, the use of the 'One vs Rest' model is feasible and can also properly suggest a set of communication channels to start a relationship with a new physician. However, it is more challenging to explain or convince that a multi-class view, involving a combination of communication channels, can be used to propose a mix of channels.

The large volume of available data was both a benefit and a torment; the benefits are related to the greater possibility of learning. However, the processing cost of this data is very high. This limitation forced us to aggregate the information to feasible levels for testing and evaluating the application of machine learning.

The study faced a significant challenge due to the insufficient literature on Multilabel application models. While there were numerous articles evaluating different models, very few books offered detailed content on Multilabel classification.

There is potential for further research in multiple areas. However, We suggest five possible options.

1) Apply the models to the available data levels—by monthly period, product, molecule, or brand, for example. This method will enable us to make specific and time-sensitive predictions.
2) Apply the models to different countries and compare the results in similar markets.

3) Apply other algorithms, such as 'k-Nearest Neighbor' or 'Random k-labelsets', to expand the comparison of algorithms and different models.
4) Apply different parameters for the different algorithms and compare their results, allowing for detailed model calibration.

Take advantage of applying the models and enable a study on the segmentation of doctors using the obtained results. This includes classifying doctors according to their relationship profiles, creating personas, and determining their media preferences.

BIBLIOGRAPHICAL REFERENCES

Alloghani, M., Al-Jumeily, D., Hussain, A., Aljaaf, A. J., Mustafina, J., & Petrov, E. (2018). Healthcare Services Innovations Based on the State of the Art Technology Trend Industry 4.0. *2018 11th International Conference on Developments in eSystems Engineering (DeSE)*, 64–70. https://doi.org/10.1109/DeSE.2018.00016

Anthony Jnr, B., & Abbas Petersen, S. (2021). Examining the digitalisation of virtual enterprises amidst the COVID-19 pandemic: A systematic and meta-analysis. *Enterprise Information Systems*, *15*(5), 617–650. https://doi.org/10.1080/17517575.2020.1829075

Chornous, G., Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, Fareniuk, Y., & Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. (2021). Marketing mix modeling for pharmaceutical companies on the basis of data science technologies. *Access Journal - Access to Science, Business, Innovation in the Digital Economy*, *2*(3), 274–289. https://doi.org/10.46656/access.2021.2.3(6)

Costa-Font, J., McGuire, A., & Varol, N. (2015). Regulation effects on the adoption of new medicines. *Empirical Economics*, *49*(3), 1101–1121. https://doi.org/10.1007/s00181-014-0903-x

David, Z. (2021). *Digital Marketing Models Frameworks and tools for digital audits, planning and strategy*. https://www.academia.edu/37148488/Digital_Marketing_Models_Frameworks_and_tools_for_digital_audits_planning_and_strategy

Dewick, P., & Miozzo, M. (2002). Sustainable technologies and the innovation–regulation paradox. *Futures*, *34*(9), 823–840. https://doi.org/10.1016/S0016-3287(02)00029-0

Duarte, V., Zuniga-Jara, S., & Contreras, S. (2022). *Machine Learning and Marketing: A Literature Review.* (SSRN Scholarly Paper 4006436). https://doi.org/10.2139/ssrn.4006436

Fecha, P. M. S. (2017, December 30). *The Return of the Investment of the Digital Channels in Pharmaceutical Industry—ProQuest*. https://www.proquest.com/openview/6339912f9bbbb00bd593f079c4e8f81e

Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., & Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine Learning*, *73*(2), 133–153. https://doi.org/10.1007/s10994-008-5064-8

Garcia, S., Derrac, J., Cano, J., & Herrera, F. (2012). Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(3), 417–435. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2011.142

Gledson, B., Rogage, K., Thompson, A., & Ponton, H. (2024). Reporting on the Development of a Web-Based Prototype Dashboard for Construction Design Managers, Achieved through Design Science Research Methodology (DSRM). *Buildings*, *14*(2), Article 2. https://doi.org/10.3390/buildings14020335

Hameed, H. (2020). *Quantitative and qualitative research methods: Considerations and issues in qualitative research*. https://doi.org/10.62338/pw6mmp62

Hole, G., Hole, A. S., & McFalone-Shaw, I. (2021). Digitalization in pharmaceutical industry: What to focus on under the digital implementation process? *International Journal of Pharmaceutics: X*, *3*, 100095. https://doi.org/10.1016/j.ijpx.2021.100095

Huey, C. (2020). *The New Multichannel, Integrated Marketing: 29 Trends for Creating a Multichannel, Integrated Campaign to Boost Your Profits Now* (Kindle Edition.). Media Specialists.

Jawaid, M., & Ahmed, S. J. (2018). Pharmaceutical Digital Marketing and Its Impact on Healthcare Physicians of Pakistan: A National Survey. *Cureus*, *10*(6). https://doi.org/10.7759/cureus.2789

Kanj, S., Abdallah, F., Denœux, T., & Tout, K. (2016). Editing training data for multi-label classification with the k-nearest neighbor rule. *Pattern Analysis and Applications*, *19*(1), 145–161. https://doi.org/10.1007/s10044-015-0452-8

Khanna, V., Ahuja, R., & Popli, H. (2020). Role of Artificial Intelligence in Pharmaceutical Marketing: A Comprehensive Review. *Journal of Advanced Scientific Research*, *11*(03), Article 03.

Khazzaka, M. (2019). Pharmaceutical marketing strategies' influence on physicians' prescribing pattern in Lebanon: Ethics, gifts, and samples. *BMC Health Services Research*, *19*(1), 80. https://doi.org/10.1186/s12913-019-3887-6

Kong, X., Shi, X., & Yu, P. S. (2011). Multi-Label Collective Classification. In *Proceedings of the 2011 SIAM International Conference on Data Mining (SDM)* (pp. 618–629). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972818.53

Kumar, A., Shankar, R., & Aljohani, N. R. (2020). A big data driven framework for demand-driven forecasting with effects of marketing-mix variables. *Industrial Marketing Management*, *90*, 493–507. https://doi.org/10.1016/j.indmarman.2019.05.003

Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics Wiley*. (Kindle). Wiley.

Ling, C., Zhang, T., & Chen, Y. (2019). Customer Purchase Intent Prediction Under Online Multi-Channel Promotion: A Feature-Combined Deep Learning Framework. *IEEE Access*, *7*, 112963–112976. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2935121

Lopes, A. S. D. (2012). *A prática do marketing de relacionamento na conquista da lealdade do cliente médico cearense pela indústria farmacêutica [Dissertação Universidade do Ceará]. Repositório UFC Brasil*. http://www.repositorio.ufc.br/handle/riufc/42875

Maimon, O., & Rokach, L. (Eds.). (2010). *Data Mining and Knowledge Discovery Handbook*. Springer US. https://doi.org/10.1007/978-0-387-09823-4

Manchanda, P., & Chintagunta, P. K. (2004). Responsiveness of Physician Prescription Behavior to Salesforce Effort: An Individual Level Analysis. *Marketing Letters*, *15*(2), 129–145. https://doi.org/10.1023/B:MARK.0000047389.93584.09

Mantrala, M. K., Albers, S., Caldieraro, F., Jensen, O., Joseph, K., Krafft, M., Narasimhan, C., Gopalakrishna, S., Zoltners, A., Lal, R., & Lodish, L. (2010). Sales force modeling: State of the field and research agenda. *Marketing Letters*, *21*(3), 255–272. https://doi.org/10.1007/s11002-010-9111-4

Nair, H. S., Manchanda, P., & Bhatia, T. (2010). Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders. *Journal of Marketing Research*, *47*(5), 883–895. https://doi.org/10.1509/jmkr.47.5.883

Nan, H., & Hu, M. (2022). Corporate Marketing Strategy Analysis with Machine Learning Algorithms. *Wireless Communications and Mobile Computing*, *2022*, e9450020. https://doi.org/10.1155/2022/9450020

Nord, L. (2011). R&D Investment Link to Profitability: A Pharmaceutical Industry Evaluation. *Undergraduate Economic Review*, *8*(1). https://digitalcommons.iwu.edu/uer/vol8/iss1/6

Over-sampling—Version 0.12.3 documentation. (2024, January 1). [Documentation]. Over-Sampling — Version 0.12.3 Documentation. https://imbalanced-learn.org/stable/over_sampling.html

Parekh, D., Kapupara, Dr. P., & Shah, K. (2016). Digital Pharmaceutical Marketing: A Review. *Research Journal of Pharmacy and Technology*, *9*, 108. https://doi.org/10.5958/0974-360X.2016.00017.2

Parsons, L. J., & Abeele, P. V. (1981). Analysis of Sales Call Effectiveness. *Journal of Marketing Research*, *18*(1), 107–113. https://doi.org/10.1177/002224378101800113

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2020). *Design Science Research Process: A Model for Producing and Presenting Information*

Systems Research (arXiv:2006.02763). arXiv. https://doi.org/10.48550/arXiv.2006.02763

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Perry, J. E., Cox, D., & Cox, A. D. (2014). Trust and Transparency: Patient Perceptions of Physicians' Financial Relationships with Pharmaceutical Companies. *Journal of Law, Medicine & Ethics*, *42*(4), 475–491. https://doi.org/10.1111/jlme.12169

Pfahringer, B., Holmes, G., & Frank, E. (2021). Classifier Chains: A Review and Perspectives. *Journal of Artificial Intelligence Research*, *70*, 683–718. https://doi.org/10.1613/jair.1.12376

Punia, S., Nikolopoulos, K., Singh, S. P., Madaan, J. K., & Litsiou, K. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. *International Journal of Production Research*, *58*(16), 4964–4979. https://doi.org/10.1080/00207543.2020.1735666

RandomOverSampler library internet. (2024, January 1). *RandomOverSampler library* [Reference]. RandomOverSampler Library Internet. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html

scikit-learn classifier chain internet. (2024, September 10). *Scikit-learn Multilabel classification using a classifier chain* [Documentation]. Scikit-Learn Classifier Chain Library. https://scikit-learn/stable/auto_examples/multioutput/plot_classifier_chain_yeast.html

scikit-learn DecisionTreeClassifier internet. (2024, September 7). *DecisionTreeClassifier library* [Documentation]. DecisionTreeClassifier Library. https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

scikit-learn f1_score metrics internet. (2024, September 10). *Scikit-learn f1_score metrics* [Documentatin]. Scikit-Learn F1_score Metrics. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

scikit-learn internet. (2024, January 1). *Scikit-learn internet* [Getting Started]. Scikit-Learn Internet. https://scikit-learn/stable/getting_started.html

scikit-learn Metrics and scoring internet. (2024, September 10). *Metrics and scoring: Quantifying the quality of predictions* [Scikit-learn Metrics and scoring]. Scikit-Learn Metrics and Scoring Internet. https://scikit-learn/stable/modules/model_evaluation.html

scikit-learn OneVsRestClassifier internet. (2024, September 7). *OneVsRestClassifier library* [Documentation]. Scikit-Learn OneVsRestClassifier Library. https://scikit-learn/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

Scrum.org, P. (2023, January 1). *What is Scrum?* Scrum.Org. https://www.scrum.org/resources/what-is-scrum

Syam, N., & Sharma, A. (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. *Industrial Marketing Management*, *69*, 135–146. https://doi.org/10.1016/j.indmarman.2017.12.019

Tan, J., Yang, J., Wu, S., Chen, G., & Zhao, J. (2021). *A critical look at the current train/test split in machine learning* (arXiv:2106.04525). arXiv. http://arxiv.org/abs/2106.04525

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining Multi-label Data. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 667–685). Springer US. https://doi.org/10.1007/978-0-387-09823-4_34

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2011). Random k-Labelsets for Multilabel Classification. *IEEE Transactions on Knowledge and Data Engineering*, *23*(7), 1079–1089. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2010.164

Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning*, *73*(2), 185–214. https://doi.org/10.1007/s10994-008-5077-3

Vogiatzis, A., Chalkiadakis, G., Moirogiorgou, K., & Zervakis, M. (2021). A Novel One-vs-Rest Classification Framework for Mutually Supported Decisions by Independent Parallel Classifiers. *2021 IEEE International Conference on Imaging Systems and Techniques (IST)*, 1–6. https://doi.org/10.1109/IST50367.2021.9651468

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

Zhang, M.-L., Li, Y.-K., Liu, X.-Y., & Geng, X. (2018). Binary relevance for multi-label learning: An overview. *Frontiers of Computer Science*, *12*(2), 191–202. https://doi.org/10.1007/s11704-017-7031-7

Zhang, M.-L., & Zhou, Z.-H. (2014). A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *26*(8), 1819–1837. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2013.39

APPENDIX A

The Github repository of codes and files used in this thesis is https://github.com/m20210498/thesis.

Model	Shape	Accuracy	Error rate	F1score	Precision	Recall
Decision Tree	**(39969, 1)-(28049, 1)**	**85.87%**	**14.13%**	**30.58%**	**30.42%**	**30.78%**
e-Mail_1to1	(39969, 1)-(30726, 1)	76.87%	23.13%	38.64%	38.48%	38.80%
e-Mail_Mass	(39969, 1)-(29938, 1)	74.90%	25.10%	85.14%	86.00%	84.31%
F2F_f2f	(39969, 1)-(28049, 1)	70.18%	29.82%	64.29%	64.27%	64.31%
F2F_PHONE_CALL	(39969, 1)-(38099, 1)	95.32%	4.68%	10.27%	10.00%	10.55%
MEETINGS_F2F_Meetings	(39969, 1)-(35688, 1)	89.29%	10.71%	16.34%	16.02%	16.67%
MEETINGS_REMOTE_MEETING	(39969, 1)-(33791, 1)	84.54%	15.46%	24.49%	23.79%	25.24%
Web Site_GSKPRO	(39969, 1)-(38663, 1)	96.73%	3.27%	2.10%	1.83%	2.46%
Webinar_Webinar	(39969, 1)-(39625, 1)	99.14%	0.86%	3.37%	2.96%	3.92%
DT_Chain	**(39969, 8)**	**42.53%**	**57.47%**	**67.10%**	**66.79%**	**67.46%**
DT - Multilabel (All)	(39969, 8)	42.53%	57.47%	67.10%	66.79%	67.46%
DT_OneVsRest	**(39969, 8)**	**36.81%**	**63.19%**	**66.92%**	**67.23%**	**66.66%**
DT - OneVsRest (All)	(39969, 8)	36.81%	63.19%	66.92%	67.23%	66.66%
Logistic Regression	**(39969, 1)-(22955, 1)**	**72.57%**	**27.43%**	**35.61%**	**32.64%**	**71.71%**
e-Mail_1to1	(39969, 1)-(29216, 1)	73.10%	26.90%	52.06%	39.11%	77.85%
e-Mail_Mass	(39969, 1)-(23432, 1)	58.63%	41.37%	70.40%	90.33%	57.67%
F2F_f2f	(39969, 1)-(31431, 1)	78.64%	21.36%	73.44%	76.33%	70.77%
F2F_PHONE_CALL	(39969, 1)-(33220, 1)	83.11%	16.89%	19.41%	11.04%	80.18%
MEETINGS_F2F_Meetings	(39969, 1)-(29126, 1)	72.87%	27.13%	25.27%	15.27%	73.09%
MEETINGS_REMOTE_MEETING	(39969, 1)-(30685, 1)	76.77%	23.23%	37.13%	25.39%	69.07%
Web Site_GSKPRO	(39969, 1)-(22955, 1)	57.43%	42.57%	4.18%	2.16%	65.32%
Webinar_Webinar	(39969, 1)-(31987, 1)	80.03%	19.97%	2.97%	1.51%	79.74%
LR_Chain	**(39969, 8)**	**50.46%**	**49.54%**	**61.48%**	**75.83%**	**56.66%**
LR - Multilabel (All)	(39969, 8)	50.46%	49.54%	61.48%	75.83%	56.66%
LR_OneVsRest	**(39969, 8)**	**49.23%**	**50.77%**	**66.80%**	**75.10%**	**63.55%**
LR - OneVsRest (All)	(39969, 8)	49.23%	50.77%	66.80%	75.10%	63.55%
Neural Network	**(39969, 1)-(23821, 1)**	**74.21%**	**25.79%**	**36.11%**	**32.26%**	**69.77%**
e-Mail_1to1	(39969, 1)-(29907, 1)	74.83%	25.17%	54.35%	41.19%	79.88%
e-Mail_Mass	(39969, 1)-(23821, 1)	59.60%	40.40%	71.21%	90.80%	58.58%
F2F_f2f	(39969, 1)-(31425, 1)	78.62%	21.38%	75.48%	72.41%	78.82%
F2F_PHONE_CALL	(39969, 1)-(32683, 1)	81.77%	18.23%	18.08%	10.20%	79.29%
MEETINGS_F2F_Meetings	(39969, 1)-(29285, 1)	73.27%	26.73%	25.75%	15.60%	73.88%
Total	**(39969, 1)-(22955, 1)**	**74.32%**	**25.68%**	**39.92%**	**41.08%**	**53.21%**

Figure 13 - matrix of scores between models and labels

Customer_segment	Sum of e-Mail_1to1	Sum of e-Mail_Mass	Sum of F2F_f2f	Sum of F2F_PHONE_CALL
	3927	78103	10431	2149
High value Customer	12421	46323	31325	2692
Lost Customers	5237	75608	14082	2777
Low Value Customers	8648	86464	23704	3892
Medium Value Customer	12697	76500	33498	4101
Top Customers	37241	73754	74830	4463
Total	**80171**	**436752**	**187870**	**20074**

Figure 14 - quantity of interaction by HCP segmentation

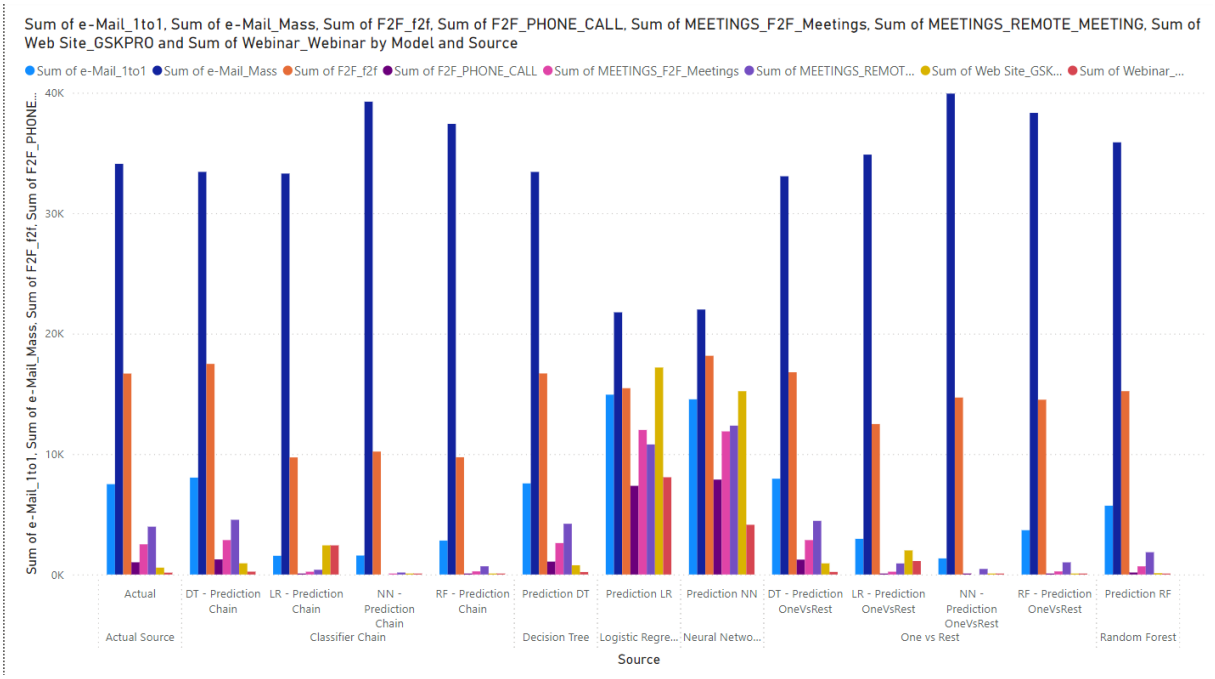

Figure 15 - bar chart od prediction for each model

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa