Technische Universiteit Eindhoven University of Technology
Master of Science in
Industrial and Applied Mathematics

POLITECNICO DI TORINO
Master of Science in
Mathematical Engineering

Exceptional Model Mining: a logistic regression model on cancer registry data

Master Thesis

Andrea Benevenuta

Supervisors:
Paulo Serra
Wouter Duivesteijn
Harm Buisman
Tania Cerquitelli

Eindhoven, September 2020

A nonna Ada

Abstract

Finding interesting patterns in a cancer registry data can provide oncologists and medical experts with a new perspective to improve healthcare for cancer patients. In this paper, we apply a supervised local pattern mining called Exceptional Model Mining. Its aim is to find subgroups in the data that somehow behave differently from the norm. This behaviour is captured by a model and the interestingness of a subgroup is assessed according to a quality measure. In particular, we develop a logistic regression model and propose a few different quality measures based on statistical tests and probability distributions. Additionally, we provide a statistical test, the permutation test, to assess the interestingness of a subgroup from a statistical point of view.
The results of the experiments show that the proposed model can retrieve some subgroups that may be of interest for doctors. Moreover, the results of the permutation test show that the most interesting subgroups are also statistically significant.

Contents

Chapter 1

Introduction

In 2018, according to the World Health Organization (WHO), 9.6 million people worldwide are estimated to have died from cancer [1], which made it the second leading cause of deaths globally, just after cardiovascular diseases. In the Netherlands, in the same year, nearly 120000 new cases of cancer arose in the population, with breast cancer being the most frequent type (16209 women were affected) and lung cancer being the deadliest type (11000 deaths) [2].
Therefore, it is particularly important to conduct research to better understand this class of diseases and a new perspective could be given by adopting a data-driven approach in a medical domain.

In this regard, work is carried out by IKNL (Integral Kankercentrum Nederland): it is an independent knowledge institute that collaborates with healthcare professionals for the improvement of oncological and palliative care.
IKNL is responsible for retrieving and maintaining the Netherlands Cancer Registry (NCR), a population-wide registry containing diagnosis and treatment data of nearly all cancer cases in the Netherlands since 1989. Many researchers perform epidemiological research using the NCR data, driven by a hypothesis or research question. However, there may be patterns and information hidden in the data: a data-driven approach could be useful to generate hypotheses to feed to actual medical experts to improve healthcare for cancer patients.

In this report, we focus on the application of a specific data mining technique, called Exceptional Model Mining (EMM), to the cancer registry data. The goal of this method is to identify subgroups of the dataset where multiple target attributes interact in an unusual way. This interaction is captured by a model and the more this model behaves differently from the norm (the whole dataset or the complement of the subgroup) the more interesting the subgroup is.

1.1 Lung Cancer Dataset from IKNL

In this report, we want to apply EMM to a real clinical dataset provided by IKNL. There are two main goals that we wish to achieve: finding surprising patterns that could be useful for the scientific community and quantifying the concept of interestingness of the retrieved subgroups through the use of statistics.
In this thesis, the focus is on lung cancer: it is one of the deadliest types of tumor with one of the highest death rates. Thus, it could be interesting to find surprising insights regarding people that suffered from it. Nevertheless, the method that we provide in this project is not strictly connected to lung cancer and can be generalized to any type of cancer.

The dataset contains information regarding patients who have been diagnosed with lung cancer in the Netherlands between the years 2005 and 2018. The provided attributes are both numeric and nominal, with the majority of them being binary.
Some attributes contain information about the patient, such as age (*leeft*), gender (*gesl*), socio-economic status (*ses)*, whether they are dead or alive (*vit_stat, is_dead*). Others relate to the cancer itself: the stage of the tumor (*stadium, ct, cn, cm, ...*), the behaviour (*gedrag*) or the type of cancer (*morf, topog, ...*). There are also geographical attributes to identify where the patient is living (*COROP, PROVINCIE, PC2*) and another important category is represented by those attributes that indicate if the patient received any treatment (*treatment*) and, if so, which specific treatments they were subjected to (*indchorg, indimmuno, indchemo, ...*). The dataset is described more in detail in Chapter 5.

1.2 Research questions

The procedure that has been followed in this project is: select specific target attributes, fit a proper model on them and apply EMM to find the most interesting subgroups in the dataset. Then, validate the significance of the result through statistical analysis. Everything regarding Exceptional Model Mining is described in detail in Section 2.3.

To the best of our knowledge there has been only one study in which EMM was applied to clinical data. In [3], the author applied this data analysis technique to a dataset that consisted of patients who had been diagnosed with breast cancer. The model used in the EMM framework was rather simple: it made use of absolute frequencies and aimed at finding subgroups for which the distribution of a single target attribute (or multiple targets) was remarkably different when compared to the distribution of the same in the whole dataset. In the end, the author found results that were supported by the general knowledge of the scientific community, but were not very surprising from a clinical point of view.

In this report, we have built and analyzed a different model (logistic regression) that makes use of statistical theory and we want to observe if that can lead to the retrieval of surprising insights. In other words, the goal of this project is to answer the following research questions:

How to use Exceptional Model Mining combined with statistical theory to extract noteworthy patterns from cancer registry data?

How to quantify the interestingness of a subgroup in the Exceptional Model Mining framework in an objective way using statistics?

1.3 Summary

This thesis is structured as follows. Chapter 2 describes some data mining techniques with a particular focus on Exceptional Model Mining. It provides an overview of some possible models to apply to the data and the reason why logistic regression is adopted. Chapter 3 exhibites some key concepts, such as what is logistic regression and how to quantify the similarity between two probability distributions. These concepts are then exploited in Chapter 4 to define a proper logistic regression model in the EMM framework. In Chapter 5, the dataset is described as well as the preprocessing phase. Chapter 6 focuses on the preparation phase before conducting the experiments, described in detail in Chapter 7. Chapter 8 explains how to assess the interestingness of a subgroup using the permutation test and Chapter 9 outlines some general considerations on the project as well as some directions that could be considered for a future work.

Chapter 2

Literature overview: Exceptional Model Mining

This study focuses on the application of Exceptional Model Mining. Its purpose is to find subsets of the dataset that somehow behave differently and thus could be interesting.
Before analyzing the details of this technique, it is important to understand where EMM is located within the data mining framework. Therefore, this chapter provides a brief overview of some other existing approaches in the data mining framework and explains why EMM has been chosen. Secondly, it describes some possible models that could be adopted for EMM in our specific case and illustrates why logistic regression has been selected.

2.1 Data mining

The process of "identifying valid, novel, potentially useful, and ultimately understandable patterns in data" is defined as Knowledge Discovery (KD) [4]. Data Mining is the most important step of this process and it consists in the application of data analysis techniques and algorithms to retrieve interesting patterns from the data. Over the years, many techniques have been developed; nevertheless, two main classes of approaches (that may sometimes overlap) can be identified:

- **Predictive approaches** carry out the induction over the current and past data so that predictions can be made. The goal is predicting a target value using supervised learning functions. This category encompasses methods such as classification, regression and time-series analysis.

- **Descriptive approaches** aim at detecting relations and patterns in the current data. It focuses on the conversion of the data into meaningful information for reporting and monitoring. Clustering and association rules belong, for example, to this category.

The distinction between these two classes is not always clear and there are methods that often cross this feeble boundary. One example is with respect to the category of *supervised descriptive rule discovery*, strictly associated to Exceptional Model Mining.
The methods belonging to this category aim at discovering interesting patterns, in the form of rules, by taking into account labeled data. In other words, they make use of supervised learning to solve descriptive tasks: that is why the features of both approaches are somehow encompassed. Three main methods have been identified by Novak Kralj et al.[5] in this category: Contrast Set Mining (CSM), Emerging Pattern Mining (EPM) and Subgroup Discovery (SD).

2.2 Supervised Descriptive Rule Discovery

Contrast Set Mining

CSM [6] has the goal of detecting contrast sets, i.e. conjunctions of attributes and values, that differ meaningfully in their distributions across groups.
More formally, given a set of attributes $A_1, ..., A_k$ and for each A_i a set of values $V_{1i}, ..., V_{mi}$, a contrast set is a conjunction of attributes and values defined on groups $G_1, ..., G_n$, where no attribute is repeated more than once.
Let's consider an example in a medical domain in which we have a dataset of patients diagnosed with cancer in the Netherlands in 2018. In this case, a contrast set could be:

$$smoker = True \quad \wedge \quad gender = male$$

where *smoker* and *gender* are the attributes and $True$ and *male* their values.

Concerning a group G, the support of a contrast set is defined as the percentage of examples in the group for which the contrast set is true. With respect to our example, a group could be identified as the region of residence of a Dutch person. Then, the support of the contrast set defined above would be, out of all people having cancer in that region, the percentage of patients that smoke and are male.

The goal of CSM is to find contrast sets for which:

$$\exists \text{ i,j: P(contrast set = true} \mid G_i) \neq \text{P(contrast set = true} \mid G_j) \tag{2.1}$$

$$\max_{i,j} \mid \text{support(contrast set, } G_i \text{) - support(contrast set, } G_j \text{)} \mid \geq \delta \tag{2.2}$$

where δ is a threshold defined as the minimal support difference.
Equation (2.1) is a statistical significance requirement and ensures that, with respect to the contrast set, there is a true difference between the two groups. A statistical test is performed with the null hypothesis that contrast sets have exactly equal probabilities across groups.
Inequality (2.2) takes into account the size factor because the effect must be large enough to be relevant.
If both requirements are satisfied, we have found a so called *deviation*.
To make things clear, in our example, the above mentioned contrast set is considered significant if the distribution of male smoker patients over all patients that have cancer is very different in a particular region with respect to other regions.

Emerging Pattern Mining

EPM, described in [7], is a data mining technique whose goal is to find emerging patterns (EP). These are defined as subgroups whose supports increase significantly from one dataset to another. This approach can be applied when we want to find relevant contrasts between data classes or when we deal with timestamped databases.
More precisely, let D_1 and D_2 be a pair of datasets and let $supp_{D_i}(X)$ denote the support of the itemset X over the dataset D_i (i=1,2). We define:

$$\text{GrowthRate(X)} = \begin{cases} 0 & \text{if } supp_{D_1}(X) = 0 \text{ and } supp_{D_2}(X) = 0 \\ \infty & \text{if } supp_{D_1}(X) = 0 \text{ and } supp_{D_2}(X) \neq 0 \\ \frac{supp_{D_2}(X)}{supp_{D_1}(X)} & \text{otherwise} \end{cases}$$

The aim of EPM is then, given a certain threshold ρ, to find all itemsets X for which:

$$\text{GrowthRate(X)} \geq \rho$$

An application in the medical domain could be to split the dataset into two sub-datasets, one containing the patients who survived cancer and the other with the patients that died from it.
In this scenario, suppose that we find the EP: (S_1, T_1, T_2), with growth rate of 4 from the not-cured to the cured group. This suggests that, among all cancer patients who presented the symptom S_1 and received both treatments (T_1, T_2), the number of cured patients is 4 times the number of patients who were not cured. Hence, it might be recommendable to apply the treatment combination whenever that particular symptom occurs.

Subgroup Discovery

Subgroup discovery is a supervised pattern mining technique whose aim is to extract interesting rules with respect to a target variable [8].
More specifically, given a population of individuals and a property of those individuals we are interested in, we want to find subgroups that have the most unusual statistical distribution over the target attribute. In other words, we aim at finding rules of the form:

$$R : Subgr_{Descr} \rightarrow Target_{Val}$$

where $Target_{Val}$ is a specific value of the target attribute and $Subgr_{Descr}$ is a conjunction of features (attribute-value pairs) that induce the subgroup.
As an example, suppose that we are analysing the data of students applying for university X. Using SD, we could find:

$$gender = female \wedge income = high \longrightarrow admitted = Yes$$

That means that the subgroup represented by girls living in a family with a high income has an unusual distribution with respect to the target attribute *admitted*. In particular, the distribution has a higher rate of admissions when compared to the distribution of the target in the overall population. That implies that people belonging to this subgroup are more likely to get accepted by university X.

All of these methods, belonging to the category of Supervised Descriptive Rule Discovery, share a common task: to *detect significant deviations* within the data. These deviations are not simply outliers, i.e. data points that differ significantly from the rest of the data. They are characterized by sophisticated structures (e.g. contrast sets, emerging patterns, subgroups), more easily interpretable and actionable than single data points.
In the following section we introduce Exceptional Model Mining. This data mining technique shares common features with the methods just described and, in particular, can be seen as a generalization of Subgroup Discovery. SD aims at finding those subgroups for which the target attribute distribution is significantly different from the norm. In EMM, multiple target attributes can be considered and many models can be applied to examine more complicated relationships.

2.3 Exceptional Model Mining

EMM is a supervised local pattern mining framework with the aim of identifying interesting subgroups in a dataset, i.e. subsets of the dataset that somehow behave differently from the norm. First introduced in [9], it has been further discussed and developed in [10], [11]. In order to understand what a subgroup is and when it is deemed interesting, some notations and definitions are needed. In the following subsections, we explain how we can define a subgroup and assess its interestingness and which is the algorithm that can be exploited to find unusual subgroups.

2.3.1 Descriptions and subgroups

We assume that our dataset Ω is a collection of records r^i $(i = 1, ..., N)$, in which:

$$r^i = (a_1^i, ..., a_k^i, t_1^i, ..., t_m^i) \qquad k, m \in N^+$$

$a_1^i, ..., a_k^i$ are defined as the *descriptive attributes*, $t_1^i, ..., t_m^i$ as the *target attributes*.

A description is a function that maps the descriptive attributes of each record into a 0 or a 1. We say that the description D covers a record r^i if and only if $\mathrm{D}(a_1^i, ..., a_k^i) = 1$.
An example of a description, with respect to our study, would be:

$$\textit{diffgr} = 9 \text{ and } \textit{leeft} > 81 \text{ and } \textit{leeft} \leq 103$$

where *diffgr* is the differentiation grade class of the tumor and *leeft* is the age of the patient.

A subgroup is consequently defined as the collection of records that a description D covers. We denote that with $G_D \subseteq \Omega$.

2.3.2 Model class and quality measure

The core of EMM is the choice of a model class over the target attributes and the choice of a quality measure, a function that assigns a numerical value to a subgroup induced by a description, i.e. $\varphi : G_D \longrightarrow \mathcal{R}$.
This measure indicates how exceptional the model fitted on the targets in the subgroup G_D is, compared to either the model fitted on the targets in the whole dataset Ω , or the model fitted on the targets in the complement G_D^c.
As underlined in [10], different choices can lead to very different results. That is why we must be careful what we compare the subgroup to. However, it often occurs that the chosen model over the target attributes gives you a more precise indication on what the term of comparison will be.

Regarding the quality measure, it generally tends to favor the discovery of smaller subgroups, even though larger subgroups are usually more interesting to analyse. This happens because it is easier to have an unusual behaviour when we consider a small number of observations. Hence, when using a quality measure, the subgroup size should always be taken into account.
One way to deal with that, as suggested in [9], is to multiply the quality measure with the entropy function:

$$\varphi_{ef}(D) = -\frac{n}{N} log\Big(\frac{n}{N}\Big) - \frac{n^c}{N} log\Big(\frac{n^c}{N}\Big) \tag{2.3}$$

where N, n, n^c represent the number of records of respectively the whole dataset Ω, the subgroup G_D, the complement G_D^c. Since the entropy function is maximized for $n = 0.5 * N$, more equal splits, hence larger subgroups, will be favored.
Another way to tackle this problem is to use a quality measure based on a statistical test. The advantage is that we can have a clearer indication of what is significantly interesting because of the underlying statistical theory.

2.3.3 Refining the descriptions

After defining a model over the target attributes and a quality measure, we still need to generate candidate subgroups with the aim of finding the most interesting ones. Hence, it is important to define a way to generate the subgroups, given the descriptive attributes.
Let's call D the description and a_i the attribute of interest. The refinement of a description and hence the generation of a subgroup depends on the type of the descriptive attribute a_i:

- Binary: this is the simplest case and there are just two possibilities (encoded as 0 and 1). Therefore, we can refine our description as follows:

$$D \cap (a_i = 0) \qquad D \cap (a_i = 1)$$

- Nominal: let's suppose that a_i has g values $v_1, ..., v_g$. Then, we can refine the description by distinguishing, for each value, when the attribute assumes that value and when the attribute assumes one of the other g-1 values. More formally:

$$D \cap (a_i = v_k) \qquad D \cap (a_i \neq v_k) \qquad \text{for } k = 1, ..., g$$

- Numeric: in this case we discretize the numerical attribute by creating equal-sized bins. Then, for each split point, we refine by considering when the attribute is less and when it is greater (or equal) than the split point.

  In other words, we first order the values that a_i can assume:

$$v_{(1)}, ..., v_{(n)}$$

  We define the b split points as:

$$s_j = v_{(\lfloor j \frac{n}{b} \rfloor)} \qquad \text{for } j = 1, ..., b$$

  In the end, we add the following refinements:

$$D \cap (a_i < s_j) \qquad D \cap (a_i \geq s_j) \qquad \text{for } j = 1, ..., b$$

2.3.4 Beam search algorithm

Considering the many possible combinations to generate different subgroups, the search space becomes exponentially large with respect to the number of attributes. Since it is generally too expensive to explore it by brute force (analyzing all possible subgroups), many researchers rely on heuristic search.

In the EMM framework, even though there exist some alternatives ([12], [13]), the beam search algorithm [Algorithm 1, [11]] is generally performed. In simple words, the algorithm traverses the subgroup description search space by starting with simple descriptions and refining these along the way, going from generic to specific.
In the first step, all subgroups, whose description is given by just one attribute, are generated. Each subgroup is assessed according to the quality measure φ and the ω most exceptional subgroups are kept (ω is the beam width, a parameter defined before running the algorithm). The method is then iterated and, at each step, the descriptions are refined and the related subgroups assessed again.
Since, at each level, instead of considering just one best partial solution, the best ω partial solutions are kept, the beam search algorithm can be regarded as an extension of a greedy algorithm.

2.3.5 EMM motivation

There are two main reasons why EMM is applied in this project with respect to the cancer registry data.
First, it can be regarded as an actionable and explicable method. As a matter of fact, its aim is not simply to detect outliers, single observations that in most cases are difficult to interpret

and could be even caused by some randomness in the data. EMM focuses on subgroups, coherent structures induced by conditions on the descriptive attributes, that deviate from the norm. The output consists of descriptions of the most interesting subgroups, expressed with a language very similar to the human one and hence easy to understand. The doctors (i.e. the domain experts in this scenario) can then check if the retrieved subgroups derive from common knowledge or are actually interesting and bring a new perspective which is worth investigating.

Another advantage is that EMM does not limit itself by considering one single target attribute (as in SD) but takes into account multiple targets, on which many different models can be built. In other words, EMM is a very flexible method that offers several possibilities: from simply finding unusual distributions of the targets to capturing unusual relationships modeled with logistic regression or Bayesian networks for example. This enables us to identify subgroups that behave differently in terms of a kind of modeling that the domain experts use on a daily basis, which brings the data mining technique closer to the domain-specific world.

2.4 Overview of some EMM instances

The core of Exceptional Model Mining is the EMM instance: the combination of a model class chosen over the target attributes and a quality measure. There are many models that have been applied in the past years, depending on what is deemed interesting in specific scenarios.

In the **Correlation model**, the focus is on the relationship between two numeric targets. The goal is to find subgroups for which the relationship between the targets is significantly different from the one present in the rest of the dataset. According to the properties and features of the data, several quality measures can be defined.
The Pearson correlation coefficient quantifies the linear relationship between the two targets. The drawback is that Pearson coefficient has some underlined assumptions such as normality and it is also heavily affected by outliers. Hence, other quality measures, such as Spearman's rank correlation coefficient or Kendall's Tau, can be used to quantify the relationship between the two targets. In this case, the focus is not anymore on the linearity of the relation but on the more general monotonicity and the model is denominated as *Rank Correlation model*.
In [14], this model, with all three quality measures, was applied to six different datasets to find the subgroups in which there were unusual relationships for example between price and size of a house ([14], Section 5.1) or between age of a woman and number of children ([14], Section 5.4).

When dealing with nominal attributes, we can leverage on the **Association model**. The goal is to find subgroups in which the association between two nominal targets is remarkably different from the association present in the rest of the dataset.
This model was used by Duivesteijn et. al. in [15], a study in which the Exceptional Model Mining framework was applied as an alternative of the classic A/B testing. In this study, instead of just retaining the best variant (A or B) of a product, based on the preference of the population, the association model was applied to find subgroups whose preference was significantly different when compared to the rest of the population. This approach is useful when the company has the possibility to produce both variants, therefore, it can offer the most appealing alternative to specific subgroups of the population.

In [16], Duivesteijn et al. applied the **Linear Regression model**. The idea is to fit two linear regression models (one on the subgroup and one on the whole dataset) and find those subgroups for which the coefficients of the two models are remarkably different. That implies that the relationships between the variables are different and it might turn out to be very interesting to understand where the major discrepancies take place.

In the article, the model was applied to several datasets (e.g. Giffen Behavior Data, Wine data, Housing Data). This means that it is a flexible method that can be applied in numerous scenarios, as long as linear regression is a suitable model, e.g. there is a linear relationship between the independent and dependent variables, there is no or little multicollinearity in the data, homoscedasticity is satisfied, etc.

In [3], **Exceptional Incidence Distribution Mining model** was introduced. It makes use of absolute frequencies and aims at finding those subgroups for which the distribution of a target (or a combination of multiple targets) is unusual when compared to the distribution in the whole dataset. This model was used with respect to cancer registry data and the focus was on patients who were diagnosed with breast cancer. It did not lead to any scientific breakthrough, but that does not mean that it is an invalid model. It might still lead to interesting insights, when applied to other datasets and its strength is that it is very straightforward to interpret.

2.5 Classification with Logistic Regression: motivation

In this report we are dealing with a dataset taken from the National Cancer Registry (the dataset will be described in detail in Chapter 5). The question is: *how to decide which EMM instance (model class and quality measure) would be appropriate for this dataset?*

First of all, it is important to underline that there is no single answer. Given the high flexibility of EMM, there are many models that could be applied to our data, depending also on what we are interested in.
A possible model could be the Exceptional Distribution Incidence Mining model, applied in [3]. The problem is that, as underlined in the previous section, it did not lead to very surprising insights. Therefore, we could study if a different EMM instance might be more suitable in this scenario.

An answer can be found in the **Classification model with Logistic Regression** [Section 5.4 [11]]. The idea is at first to identify, among the target attributes, which will be the response and which the predictors. Then, fit the logistic regression on the subgroup and on the whole dataset (or on the complement of the subgroup) and compare the two models.
We notice that the goal is very similar to the one described previously for the Linear Regression Model. The main difference is that the probability of some obtained event is represented as a linear function of a combination of predictor variables and there is no more the assumption of a linear relationship between dependent and independent variables.

There are three main motivations that justify this choice.
First, it takes into account multiple attributes at the same time, thus providing several insights when analysing a subgroup.
Secondly, the provided dataset on which our study is conducted is full of binary nominal attributes, such as the vital status of the patients (dead or alive) and if they received a particular treatment (yes or no). Among different classification methods Logistic regression is particularly suitable when dealing with this kind of data.
Lastly, even though described in [9], [10], [11], to the best of our knowledge, there is no article in which this particular model was adopted. Therefore, this report aims at understanding if this can be a proper and efficient way to retrieve interesting subgroups within a medical domain in the EMM framework.

Chapter 3

Literature overview: statistics

In Chapter 2, we mentioned that the core of Exceptional Model Mining is the choice of a model and a quality measure. This chapter provides the statistical tools to understand how the Classification with Logistic Regression model can be implemented in the EMM framework.
The first section is devoted to the model, i.e. logistic regression. It provides a description of the main concepts such as the logistic function, the probability of presence of the outcome, the odds ratio. Then, it explains how the coefficients that quantify the relationships among the variables are estimated and how they can be interpreted. In the final part of the section, the focus is instead on the statistical tests that could be applied to assess the significance of the coefficients and on a couple of measures that could help us understand if the fitted model is appropriate.
The second section focuses on how to quantify the difference between two probability distributions. This is functional for the definition, in Section 4.2, of two quality measures that can be used to assess the interestingness of a subgroup.

3.1 Logistic Regression

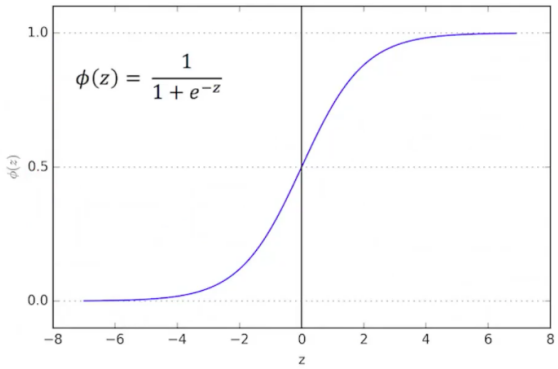
3.1.1 Main concepts

Logistic regression (LR) is a statistical method which is well suited to model the relationship between a categorical response variable and one or more categorical or continuous predictor variables. The main concepts reported in this section have been taken from [17].
There are three types of LR, depending on the nature of the categorical response variable: binomial, nominal, ordinal.
We focus on the binomial logistic regression. In this case, the response (or dependent) variable is dichotomous, i.e. it only contains data coded as 1 or 0; from now on, we denote this as y. All the other features are called predictors or independent variables and can be of any type. We indicate those using the vector $\mathbf{X} = (x_1, x_2, ..., x_p)$.
The predictions of logistic regression are in the form of probabilities of an event occurring, i.e. the probability of $y = 1$, given certain values of input variables $\mathbf{X} = (x_1, x_2, ..., x_p)$. Therefore, one of the objectives of the logistic regression is to model the conditional probability that the outcome is present:

$$\pi(\mathbf{X}) = P_{\mathbf{X}}\Big(y = 1\Big)$$

For ease of notation, we will simply indicate it with π. This probability can only take values in [0,1]. Therefore, instead of fitting a straight line or a hyperplane (as it happens with linear regression), the logistic regression model uses the logistic (or sigmoid) function (Figure 3.1) to map the output of a linear equation to [0,1].

Logistic function: $$logit(z) = \frac{1}{1+e^{-z}}$$



$$domain = [-\infty, \infty] \qquad codomain = [0, 1]$$

Figure 3.1: Plot of the logistic function.

Hence, for multiple logistic regression, the model is the following:

$$\pi = P_{\boldsymbol{X}}\Big(y = 1\Big) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}} = \frac{1}{1 + e^{-\boldsymbol{X'^T}\boldsymbol{\beta}}} \tag{3.1}$$

$$\boldsymbol{X'} = (1, \boldsymbol{X^T}) \qquad \boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$$

Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$logit(\pi) = ln\Big(\frac{\pi}{1-\pi}\Big) = ln\left(\frac{P_{\boldsymbol{X}}\Big(y = 1\Big)}{P_{\boldsymbol{X}}\Big(y = 0\Big)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p = \boldsymbol{X'^T}\boldsymbol{\beta}$$

The probability of the event to happen (π) over the probability that the event does not take place ($1 - \pi$) is defined as the odds:

$$odds = \frac{\pi}{1-\pi} = \frac{\frac{1}{1+e^{-\boldsymbol{X'^T}\boldsymbol{\beta}}}}{\frac{e^{-\boldsymbol{X'^T}\boldsymbol{\beta}}}{1+e^{-\boldsymbol{X'^T}\boldsymbol{\beta}}}} = e^{\boldsymbol{X'^T}\boldsymbol{\beta}} \tag{3.2}$$

and so $logit(\pi)$ is simply the log of the odds.
The odds can be thought as another way to express a probability. To clarify, let's make an example. Suppose that a patient has 80% probabilities of surviving after being diagnosed with colon cancer. Then, the odds are 0.80 / (1 - 0.80) = 4, or 4:1. In other words, for the patient, the probability of surviving is four times higher than the probability of dying.

3.1.2 How to estimate the coefficients

In order to estimate the vector of coefficients $\boldsymbol{\beta}$, the most popular method used in logistic regression is the maximum likelihood estimation.
Let's suppose that we have n observations $y_1, ..., y_n$. So the logistic regression model for each *i-th* observation is:

$$logit(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} = \boldsymbol{X_i'^T \beta} \qquad (3.3)$$
$$\text{with} \qquad \boldsymbol{X_i'} = (1, x_{i1}, ..., x_{ip})$$

Equation (3.3) can be expressed equivalently as:

$$\pi_i = \frac{1}{1 + e^{-\boldsymbol{X_i'^T \beta}}} = \frac{e^{\boldsymbol{X_i'^T \beta}}}{1 + e^{\boldsymbol{X_i'^T \beta}}} \qquad (3.4)$$

To define the likelihood function, we have to consider the fact that $y_1, ..., y_n$ are n independent Bernoulli trials with probability of success equal to $\pi_1, ..., \pi_n$ ($y_i \sim Be(\pi_i)$). Hence, the density function of each single y_i is:

$$f(y_i; \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \qquad i = 1, ..., n$$

The joint distribution of $\boldsymbol{y} = (y_1, ..., y_n)$ is then:

$$f(\boldsymbol{y}; \boldsymbol{\pi}) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_1}$$

Therefore, the **likelihood function** for these n observations is defined as:

$$L(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_1} = \prod_{i=1}^{n} \Big(\frac{e^{\boldsymbol{X_i'^T \beta}}}{1 + e^{\boldsymbol{X_i'^T \beta}}} \Big)^{y_i} \Big(1 - \frac{e^{\boldsymbol{X_i'^T \beta}}}{1 + e^{\boldsymbol{X_i'^T \beta}}} \Big)^{1-y_i}$$

This function expresses the probability of the observed data $(\boldsymbol{X}, \boldsymbol{y})$ as a function of the unknown parameters $(\boldsymbol{\beta})$. The goal is to maximize it so that the resulting estimators are those which agree most closely with the observed data.
In order to estimate the coefficients $\boldsymbol{\beta}$, instead of maximizing the likelihood function itself, it is easier mathematically to maximize the log-likelihood:

$$l(\boldsymbol{\beta}) = log(L(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y})) = \sum_{i=1}^{n} \Big[y_i log\Big(\frac{e^{\boldsymbol{X_i'^T \beta}}}{1 + e^{\boldsymbol{X_i'^T \beta}}} \Big) + (1 - y_i) log\Big(1 - \frac{e^{\boldsymbol{X_i'^T \beta}}}{1 + e^{\boldsymbol{X_i'^T \beta}}} \Big) \Big]$$

Therefore, we want to solve the following maximization problem:

$$\max_{\boldsymbol{\beta}} \quad l(\boldsymbol{\beta})$$

This can be accomplished by solving the system composed of the p+1 equations, obtained by differentiating the log likelihood function with respect to the p+1 coefficients.

$$\begin{cases} \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = 0 \\ \\ ... \\ \\ \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_p} = 0 \end{cases}$$

Apart from some particular cases, the system cannot be solved analytically, hence some numerical algorithms are implemented to get to the solution.

The variances and covariances of the coefficients are then obtained by a process that involves the second partial derivatives of the log likelihood function. First, we compute the following matrix:

$$I(\boldsymbol{\beta}) = \begin{bmatrix} -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0^2} & -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_1} & \cdots \\ \vdots & \ddots & \\ -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_0} & & -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_p^2} \end{bmatrix}$$

$I(\boldsymbol{\beta})$ is called the *observed Fisher information matrix*. The matrix of variances and covariances of the coefficients is then obtained by taking the inverse of the information matrix:

$$VarCov(\boldsymbol{\beta}) = I^{-1}(\boldsymbol{\beta})$$

The estimators of the variances are obtained by evaluating the diagonal elements of the matrix at the estimate of the coefficients $\hat{\boldsymbol{\beta}}$. We can denote those as $\widehat{Var}(\hat{\beta}_j)$. The most important quantity that we are going to use to build statistical tests and confidence intervals is the standard deviation of the estimated coefficient, defined as:

$$\widehat{se}(\hat{\beta}_j) = \left[\widehat{Var}(\hat{\beta}_j)\right]^{\frac{1}{2}}$$

3.1.3 Interpretation

In the linear regression, the interpretation of the coefficients is straightforward, whereas in the logistic regression it is slightly more complicated.
If x_1 is a numeric variable, how do we interpret β_1?
When $x_1 = k$, by Equation (3.3) we have:

$$logit(\pi) = \beta_0 + \beta_1 k + \beta_2 x_2 + ... + \beta_p x_p$$

When $x_1 = k + 1$:

$$logit(\pi) = \beta_0 + \beta_1 k + \beta_2 x_2 + ... + \beta_p x_p + \beta_1$$

So for a unit increase in x_1, with all other predictors held constant, β_1 represents the increase in the log of the odds. The coefficients can also be interpreted in terms of the change in the odds that the response will be positive. For example, by using Equation (3.2), we have that:

$$\frac{odds(x_1 = k+1)}{odds(x_1 = k)} = \frac{e^{\beta_0 + \beta_1(k+1) + \beta_2 x_2 + ... + \beta_p x_p}}{e^{\beta_0 + \beta_1 k + \beta_2 x_2 + ... + \beta_p x_p}} = e^{\beta_1}$$

Therefore, $e(\beta_1)$ represents the odds ratios of a unit increase of x_1, while holding all other predictors at the same values.
Let's consider now the case in which we have a nominal attribute, *nationality*, with three levels: Dutch, Italian and French and the response variable is y = cancer ($y = 1 \rightarrow$ yes, $y = 0 \rightarrow$ no). In general, for nominal attributes, we always need to create a number of predictors equal to the number of levels - 1. In this case:

$$x_1 = \begin{cases} 1 & \text{if nationality is Italian} \\ 0 & \text{otherwise} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if nationality is French} \\ 0 & \text{otherwise} \end{cases}$$

The model, when considering only this nominal attribute, is:

$$logit(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

When considering a Dutch person, the model simply becomes:

$$logit(y) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 = \beta_0$$

That means that "nationality=Dutch" is the reference group (with $x_1 = 0$, $x_2 = 0$) and the intercept β_0 is the log odds in favour of a Dutch person to suffer from cancer. Subsequently, by Equation (3.1), $\frac{1}{1+e^{-\beta_0}}$ is the probability that a Dutch person suffers from cancer.
If we have an Italian patient, the model is:

$$logit(y) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 = \beta_0 + \beta_1$$

The coefficient β_1 is then the change in the log-odds of an Italian getting cancer relative to a Dutch person. e^{β_1} represents instead the odds of having cancer if the patient is Italian over the odds of having cancer if the patient is Dutch (the reference level).
Similarly, when considering a French person, we have that:

$$logit(y) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 1 = \beta_0 + \beta_2$$

The coefficient β_2 is then the change in the log-odds of a French getting cancer relative to a Dutch person.

3.1.4 Hypothesis testing for logistic regression

Hypothesis testing can be defined as the formal procedures used by statisticians to accept or reject statistical hypothesis. A statistical hypothesis is an assumption about a population parameter that may be true or not. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. When performing a statistical test, there are two different types of hypothesis: the null (indicated with H_0) and the alternative hypothesis (indicated with H_a). The outcome of the test can be either to reject the null hypothesis or to fail to reject it.
In this section we describe two popular statistical tests when using logistic regression: the Wald test and the Likelihood ratio test. They can be used to assess the significance of one or multiple logistic regression coefficients.
The concepts reported in this section have been taken from [18].

Wald test (single logistic regression coefficient)

If the aim of our analysis is to understand whether a single logistic regression coefficient $\hat{\beta}$ is statistically significant or not, the Wald test is one of the most popular statistical tests that can be performed. The test is the following:

$$H_0 : \hat{\beta} = 0$$
$$H_a : \hat{\beta} \neq 0$$

Then, we can use the Wald statistic, defined as:

$$W = \frac{\hat{\beta}}{\widehat{se}(\hat{\beta})}$$

i.e. the estimate of the coefficient divided by its standard error.
Under the null hypothesis, $W \overset{H_0}{\sim} N(0,1)$, i.e. the statistic asymptotically follows a standard

normal distribution.[1]
Therefore, if W is large enough we can reject the null hypothesis and claim that the coefficient is statistically significant.
An equivalent way to perform the Wald test is to consider:

$$W^2 = \frac{\hat{\beta}^2}{\hat{se}(\hat{\beta})^2}$$

Under the null hypothesis, $W^2 \overset{H_0}{\sim} \chi^2_1$, i.e. it asymptotically follows a chi-squared distribution with one degree of freedom.
From the Wald statistic W, it is possible to define confidence intervals.
We start from:

$$P\Big(Z_{1-\frac{\alpha}{2}} \leq W \leq Z_{\frac{\alpha}{2}}\Big) = 1 - \alpha$$

where $Z_{1-\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$ are respectively the $1-\frac{\alpha}{2}$ and the $\frac{\alpha}{2}$ 100%-quantiles of the standard normal distribution.
We have that:

$$P\Big(Z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}}{\widehat{se}(\hat{\beta})} \leq Z_{\frac{\alpha}{2}}\Big) = 1 - \alpha$$
$$P\Big(Z_{1-\frac{\alpha}{2}}\,\widehat{se}(\hat{\beta}) \leq \hat{\beta} \leq Z_{\frac{\alpha}{2}}\,\widehat{se}(\hat{\beta})\Big) = 1 - \alpha$$

We can then build our $(1-\alpha)\times 100\%$ confidence interval for $\hat{\beta}$:

$$\Big[\hat{\beta} - Z_{1-\frac{\alpha}{2}}\,\widehat{se}(\hat{\beta}) \quad , \quad \hat{\beta} + Z_{1-\frac{\alpha}{2}}\,\widehat{se}(\hat{\beta})\Big]$$

α represents the probability of rejecting the null hypothesis (in this case that the coefficient $\hat{\beta}$ is not significant) when the null hypothesis is true (type I error) and is called the *significance level* of the test.

$1-\alpha$ is called the *confidence level* and represents the probability that the constructed confidence interval will cover the true unknown parameter β. As underlined in [19], we must be careful with the interpretation of this quantity. $1-\alpha$ is not the probability for the unknown parameter to be within that interval but it is the probability of selecting a sample such that the constructed interval (that depends on the sample) contains the unknown parameter. A usual value to set α is 0.05 . This means that we have 5% chance to commit a type I error and the confidence level of our test is 95%.

The p-value is the smallest significance level at which the null hypothesis is rejected. In other words, it is a measure of the evidence against the null hypothesis H_0: the smaller the p-value, the stronger the evidence against H_0. It is defined as the probability (under H_0) of observing a value of the test statistic the same as or more extreme than what was actually observed. Therefore, the smaller the p-value is the more extreme the observed value will be under the null hypothesis and the stronger evidence we have in favor of the alternative hypothesis.

[1] In this chapter, the symbol $\overset{H_0}{\sim}$ implies that, under the null hypothesis, the statistic on the left asymptotically follows the distribution on the right.

Likelihood ratio test

The Likelihood ratio test (LRT) can be used to assess the significance of both a single and a group of coefficients. The hypothesis is the following:

$$\begin{aligned} H_0 &: \hat{\beta}_g = 0 \qquad \forall g \in G \\ H_a &: \hat{\beta}_g \neq 0 \qquad \text{for at least one } g \in G \end{aligned}$$

where G is the set of indices of the coefficients we are interested to test. The test is rejected if there is at least one coefficient that is significantly different from 0.
The LRT statistic can be defined as the ratio between the log-likelihood of the reduced model (the one without the coefficients $\hat{\beta}_g$) to the current model (the one with the coefficients $\hat{\beta}_g$), multiplied to -2:

$$LR = -2log\Big(\frac{L \text{ without coefficients}}{L \text{ with coefficients}}\Big) = -2log(L \text{ at } H_0) + 2log(L \text{ at } H_a)$$

where L is the Likelihood function.
For a large n (a large number of observations), we have that

$$LR \overset{H_0}{\sim} \chi^2_{p-r}$$

i.e. LR asymptotically follows a chi-squared distribution with p-r degrees of freedom. p is the number of coefficients in the current model, and r is the number of coeffcients in the reduced model.
The Likelihood ratio test can be applied to the logistic regression model to test the significance of a single coefficient. Suppose that we have the following logistic regression model:

$$logit(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

We want to test:

$$\begin{aligned} H_0 &: \hat{\beta}_2 = 0 \\ H_a &: \hat{\beta}_2 \neq 0 \end{aligned}$$

The model, under the null hypothesis H_0 is:

$$logit(\pi_i) = \beta_0 + \beta_1 x_{i1} \qquad \text{with likelihood function: } L_0$$

The model, under the alternative hypothesis H_a is:

$$logit(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \qquad \text{with likelihood function: } L_a$$

Then the LRT statistic is simply:

$$LR = -2log\Big(L_0\Big) + 2log\Big(L_a\Big)$$

and:

$$LR \overset{H_0}{\sim} \chi^2_1$$

Now, suppose that we want to test if a group of coefficients is significant:

$$H_0 : \hat{\beta}_1 = \hat{\beta}_2 = 0$$
$$H_a : \text{ otherwise}$$

The model, under the null hypothesis H_0 is:

$$logit(\pi_i) = \beta_0 \qquad \text{with likelihood function: } L_0$$

The model, under the alternative hypothesis H_a is:

$$logit(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \qquad \text{with likelihood function: } L_a$$

Then the LRT statistic is simply:

$$LR = -2log\Big(L_0\Big) + 2log\Big(L_a\Big)$$

and:

$$LR \overset{H_0}{\sim} \chi^2_2$$

3.1.5 Pseudo R-squared and AUC

In logistic regression, there are many measures that can be used to assess if the model is appropriate or not. One of the most commonly reported as an output of many programming languages, e.g. Python or R, is the pseudo R-squared. The term "pseudo" is used to distinguish that from the R-squared, a statistic present in linear regression that expresses the proportion of variance explained by the model and that can be used to quantify the goodness-of-fit of the model.
Throughout the years, many different pseudo R-squareds have been proposed to quantify how well the logistic regression model fits the data. Though they have some common underlying properties, e.g. they all range from 0 to 1 and the closer to 1 the better the model, there are also some important differences. Unlike the R-squared in linear regression, each of the pseudo R-squareds is defined in a different way and, even on the same model, different pseudo R-squareds can yield very different results.
One of the most popular, usually provided by many packages in Python or R, is the McFadden's pseudo R-squared [20]. It is defined as:

$$R^2 = 1 - \frac{log\big(\hat{L}(M_c)\big)}{log\big(\hat{L}(M_{null})\big)}$$

where M_c represents the current logistic regression model with the predictors included, M_{null} is the model with only the intercept and $\hat{L}$ is the likelihood computed using the estimates of the coefficients plugged into the model.
The likelihood function, in case of logistic regression, is computed from the joint of Bernoulli distributions, hence it ranges from 0 to 1. Therefore, the log of the likelihood is less than or equal to zero and the closer to zero the more likely the model captures the information given by the data. A small ratio of log likelihoods indicates that the full model is a far better fit than the intercept model. Hence, McFadden's pseudo R-squared can be interpreted as the level of improvement over the intercept model offered by the current model.
The closer R^2 is to 1 the better but is there a clear threshold to state that the model is a poor fit? There is no simple answer to that but in [20] McFadden himself claims that "values from .2

to .4 for R^2 represent an excellent fit".

The Area Under the ROC Curve (AUC) can be regarded as another useful measure to assess the performance of our model [17]. ROC curve is a tool to assess the model's ability to discriminate between those subjects who experience the outcome of interest versus those who do not. ROC curve does that by plotting sensitivity, the probability of predicting that a real positive case (1) is a positive, against 1-specificity, the probability of predicting that a real negative case (0) is a positive. In other words, it shows the tradeoff between the true positive rate and the false positive rate of a classifier for an entire range of possible cutpoints.
AUC ranges from 0 to 1 and the closer to 1 the better the performance of our model. An indicative value is 0.5, stating that the fitted model predicts no better than by chance, hence if AUC is less than 0.5 there is something wrong with our model and it would be even more accurate to flip a coin to predict the outcome of an observation.

It is important to underline that there are many other measures that can quantify how well a logistic regression model does at fitting the data or is performing in terms of predictions, but the aim of this research is not to search for the best measure to assess the quality of a model. Nevertheless McFadden's R-squared and AUC are used in our experiments to get a general idea on whether the model is a very poor fit or not because we want to avoid the scenario in which we draw conclusions from a model that does not really reflect the reality of the data.

3.2 Comparing two probability distributions

In this section we tackle the problem of quantifying how similar or different two probability distributions are. This will be useful in the EMM framework in terms of comparing two different groups: the measures outlined in this part will in facts be used to define the quality measures to assess how interesting a subgroup is (Section 4.2).
In statistics, several measures exist to quantify how similar (or different) two probability distributions are [21]. Many of them rely on the concept of entropy.
In thermodynamics, entropy is intended as the degree of disorder of a physical system. In statistics, it can be used as a measure of goodness of fit to quantify discrepancy between two probability distributions or two statistical hypotheses. In information theory, entropy can measure the degree of uncertainty before a statistical experiment takes place. A decrease in uncertainty can be translated into a gain of information [22]. Hence entropy can be interpreted as the gain or the lack of information when using a probability distribution instead of another one.
In the literature, many measures have been proposed: Rényi's divergence, Kullback-Leibler divergence, Jeffreys' divergence, Hellinger distance, Bhattacharyya divergence, Jensen-Shannon divergence, etc .
In [23], a study has been conducted to analyze if there were significant differences when using these measures as goodness-of-fit measures and no remarkable difference was discovered. To the aim of our study, we decided to focus on KL-divergence and Hellinger distance. Both of them are quite popular in statistics and information theory and differ between each other in terms of properties and interpretation.

3.2.1 Kullback-Leibler divergence

Kullback-Leibler divergence [24], or more simply KL-divergence, also known as cross entropy, is a non-symmetric measure of the difference between two probability distributions P and Q over the same random variable X. When P and Q are discrete probability distributions, which is the

relevant case in this project (see Section 4.2), the formula is:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \cdot log\left(\frac{P(x)}{Q(x)}\right) \tag{3.5}$$

The quantity $D_{KL}(P||Q)$ can be interpreted as the amount of information gained when replacing a priori distribution P with a posteriori distribution Q. Alternatively, it can also be seen as the loss of information resulting from the adoption of an empirical distribution Q when the underlying theoretical distribution is P (P is called the reference distribution).
KL-divergence is always non-negative and it is equal to 0 if and only if $P = Q$. However, it is not a metric because it does not satisfy the property of symmetry and the triangle inequality. Furthermore, there are specific cases that must be handled with care: when there exists an $\hat{x} \in X$ for which $P(\hat{x}) = 0$ or $Q(\hat{x}) = 0$. In such cases the logarithm in (3.5) is not well defined and some conventions have been established. If $P(\hat{x}) = 0$ then that specific term of the sum is set to 0. If instead $Q(\hat{x}) = 0$ then that specific term is set to ∞. That means that the KL-divergence is unbounded.

3.2.2 Hellinger distance

An alternative to KL-divergence can be found in Hellinger distance [25]. It is a measure to quantify the similarity between two probability distributions and can be seen as the probabilistic analog of the Euclidean distance.
When dealing with discrete distributions, which is what we focus on in this project (see Section 4.2), Hellinger distance is defined as:

$$H(P,Q) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{x \in X} \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2} \tag{3.6}$$

The main difference with respect to KL-divergence is that Hellinger distance is a proper distance. Let P, Q and T be three probability distributions; Hellinger distance satisfies:

1. $H(P,Q) \geq 0$ and $H(P,Q) = 0 \iff P = Q$ (non-negativity)
2. $H(P,Q) = H(Q,P)$ (symmetry)
3. $H(P,Q) + H(Q,T) \leq H(P,T)$ (triangle inequality)

The symmetry property implies that, unlike KL-divergence, there is no reference probability distribution so that P and Q are treated as "peer distributions".
Another remarkable difference is that Hellinger distance is always well defined. Moreover, it is bounded:

$$0 \leq H(P,Q) \leq 1$$

In the following chapter we apply these concepts to define both the logistic regression model in the Exceptional Model Mining Framework and the quality measures used to assess the interestingness of a subgroup.

Chapter 4

Methodology: implemented quality measures

In Chapter 3, we have seen the main concepts of logistic regression and how to quantify the difference between two probability distributions. In this chapter, we exploit these notions and describe two different approaches to define proper quality measures to assess the interestingness of the subgroups.
In the EMM framework, when using the model of classification with logistic regression, the goal is to find those subgroups for which the relationship between the predictors and the response is significantly different when compared to the relationship present in the dataset or in the complement.
From now on, we suppose that we have identified our p predictors and response among the attributes of the dataset under analysis. Given n observations $y_1, ..., y_n$, the model defined for multiple predictors is:

$$logit\Big(P(y_i = 1)\Big) = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} \qquad i = 1, ..., n$$

where the coefficients β_1,... $,\beta_p$ quantify the relationship between the predictors x_{1i},... $,x_{pi}$ and the log odds of the response y_i.

4.1 Statistical testing approach

In this approach, the idea is to compare the logistic regression model fitted on the subgroup with the logistic regression model fitted on the complement. Let D be a binomial variable, the description, that informs us if a record belongs to the subgroup or not, i.e.:

$$D_i = \begin{cases} 1 & \text{if the record } i \text{ is in the subgroup} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Here D_i is a short notation for $D(a_1^i, ..., a_k^i)$ and $a_1^i, ..., a_k^i$ represent the descriptive attributes that are used to define the subgroup.
We then fit the following model on the dataset:

$$\begin{aligned} logit\Big(P(y_i = 1)\Big) = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \\ + \beta_{p+1} D_i + \beta_{p+2}(D_i \cdot x_{1i}) + ... + \beta_{2p+1}(D_i \cdot x_{pi}) \end{aligned}$$

It is worth noting that we did not simply add the variable D_i, but also all the interaction terms between D_i and the other predictors. This is done in order to allow the intercept and the coefficients associated to all predictors to be different in the subgroup and in the complement. More explicitly, we have:

$$logit\Big(P(y_i = 1)\Big) = \begin{cases} (\beta_0 + \beta_{p+1}) + (\beta_1 + \beta_{p+2})x_{1i} + ... + (\beta_p + \beta_{2p+1})x_{pi} & \text{(subgroup)} \\ \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} & \text{(complement)} \end{cases}$$

What we are interested in is if the coefficients of the two models are significantly different, i.e. if the regression coefficients $\beta_{p+1}, ..., \beta_{2p+1}$ are significant or not. Therefore, we can perform p statistical tests:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0 \end{cases} \qquad j = p+1, ..., 2p+1$$

As described in Chapter 3, there are a couple of statistical tests that can be used to assess the significance of the regression coefficients. Since in this case we are interested in the significance of a single regression coefficient, we can use the Wald statistic:

$$W = \frac{\beta_j}{se(\beta_j)} \overset{H_0}{\sim} N(0,1)$$

Under the null hypothesis, the coefficient divided by its standard error asymptotically follows a standard normal distribution. Alternatively, we can use W^2 which (under the null hypothesis) asymptotically follows a chi-square distribution with 1 degree of freedom.
Each test j provides a p-value (notation: p-val^j). The lower the p-value the more significant the coefficient β_j will be. Our goal is to identify those subgroups for which at least one of these coefficients is really significant. Subsequently, we can define the following quality measure:

$$\varphi_{log}(S) = 1 - \min_{j=p+1,...,2p+1} \text{p-val}^j \tag{4.2}$$

In other words, when we apply this method in the EMM framework, we are interested in finding those subgroups for which at least one of the regression coefficients is significantly different when comparing the subgroup and the complement. The more significant this coefficient is, the lower the p-value and the higher the quality measure in Equation (4.2) will be.

This approach generalizes the one adopted in [11] in the case of logistic regression. In the article, Duivesteijn et al. considered the simpler case of a single predictor:

$$logit\Big(P(y_i = 1)\Big) = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_{1i} & \text{(subgroup)} \\ \beta_0 + \beta_1 x_{1i} & \text{(complement)} \end{cases}$$

and proposed as a quality measure 1 minus the p-value associated to the regression coefficient β_3. With this novel approach we generalize to the case of multiple predictors and that allows to define more complicated logistic regression models.

4.2 A probabilistic approach

The quality measure defined in (4.2) focuses on finding a significant difference of a *single* logistic regression coefficient between the model fitted on the subgroup and the one fitted on the complement. That implies that it is enough for a subgroup to have a very significant difference for one coefficient to be characterized by a high quality measure no matter what all the other regression coefficients are.
The question is: can we define a quality measure that quantifies a more global effect, i.e. takes into account all the coefficients?
Let's fit the same logistic regression model on the subgroup and on the complement:

$$logit\Big(P(y_j = 1)\Big) = \beta_0^S + \beta_1^S x_{1j} + ... + \beta_p^S x_{pj} \quad \text{(subgroup)}$$

$$logit\Big(P(y_k = 1)\Big) = \beta_0^C + \beta_1^C x_{1k} + ... + \beta_p^C x_{pk} \quad \text{(complement)}$$

with $j \in S$ and $k \in S^c$.[1]

If we want to quantify how different the coefficients of the two models are, a trivial idea would be to adopt some standard metrics such as the 1-norm or the Euclidean distance:

$$\varphi_{log1} = \sum_{j=1}^{p} |\beta_j^D - \beta_j^S| \qquad \varphi_{log2} = \sqrt{\sum_{j=1}^{p} (\beta_j^D - \beta_j^S)^2}$$

Coefficients related to different predictors could be affected by a different scale and a difference variance and we are not taking that into account when we simply sum over all the differences, so this is not a good approach.
The idea is then to define a quality measure which is not specifically based on the coefficients, but on the response y_i. We start by fitting the same logistic regression model on the whole dataset and on the subgroup separately. In general, we will get two vectors of coefficients that are different from each other: we can indicate them as $\boldsymbol{\beta^D}$ (dataset) and $\boldsymbol{\beta^S}$ (subgroup). From these vectors, using Equation (3.4), we can then compute the predicted probabilities:

$$\pi_i^D = \frac{e^{\boldsymbol{X_i'^T \beta^D}}}{1 + e^{\boldsymbol{X_i'^T \beta^D}}} \qquad \pi_j^S = \frac{e^{\boldsymbol{X_j'^T \beta^S}}}{1 + e^{\boldsymbol{X_j'^T \beta^S}}} \tag{4.3}$$

Here we use i and j to indicate the observations belonging respectively to the dataset and the subgroup.
One of the assumptions of logistic regression is that the response has a Bernoulli distribution with probability equal to the probability of the presence of the outcome π:

$$y_i^D \sim Ber(\pi_i^D) \qquad y_j^S \sim Ber(\pi_j^S) \tag{4.4}$$

It is clear that the more different the probabilities are, the more dissimilar the responses will be, but how can we compare those?
First of all, it makes sense to compare two responses, one from the dataset and one from the subgroup, relative to the same observation, i.e. where $i = j$. Indeed, we can see from Equation (4.4) that the difference in the responses is reflected by the difference between the probabilities of success π_i^D, π_j^S. Subsequently, by Equation (4.3), since $\boldsymbol{X_i'^T} = \boldsymbol{X_j'^T}$ if $i = j$, the only difference is

[1] This is a simplified notation to indicate that the j-th observation belongs to the subgroup and the k-th observation belongs to the complement.

determined by the vector of coefficients $\boldsymbol{\beta^D}$, $\boldsymbol{\beta^S}$.
Therefore, we are going to consider only the observations belonging to the subgroup and for each observation compare the two responses y_j^D and y_j^S.
Let ϕ be a measure that is able to capture the difference between two Bernoulli distributions. We compute

$$\phi\big(Ber(\pi_j^S), Ber(\pi_j^D)\big)$$

for every observation belonging both to the subgroup and the whole dataset and then take the arithmetic mean

$$\text{quality-measure} = \frac{1}{|S|}\sum_{j\in S}\phi\big(Ber(\pi_j^S), Ber(\pi_j^D)\big) \tag{4.5}$$

where $|S|$ is the size of the subgroup. Hence, Equation (4.5) gives an idea on how, on average, the responses are different between the dataset and the subgroup.

We must now explicitly define the function ϕ in Equation (4.5), knowing that the responses follow a Bernoulli distribution. We have described in section 3.2 two measures that could be suitable for quantifying the difference of two probability distributions: Kullback-Leibler divergence and Hellinger distance. We have seen that, even tough the goal is the same, they differ in terms of properties, e.g. the former is not a distance, whereas the latter is.
Regarding KL-divergence, in the particular case in which we want to compare two Bernoulli distributions, the formula descibed in Equation (3.5) becomes:

$$D_{KL}\big(Ber(\pi_j^S), Ber(\pi_j^D)\big) = \pi_j^S \cdot log\left(\frac{\pi_j^S}{\pi_j^D}\right) + (1-\pi_j^S)\cdot log\left(\frac{1-\pi_j^S}{1-\pi_j^D}\right) \tag{4.6}$$

Since the first element in Equation (4.6) is $Ber(\pi_j^S)$, we are implicitly considering the distribution of the response $y_j^S \sim Ber(\pi_j^S)$ as the reference distribution. If we swapped the order, since KL-divergence is not symmetric, we would get a different value, in general. The quality measure is then:

$$\hat{\varphi}_{D_{KL}} = \frac{1}{|S|}\sum_{j\in S} D_{KL}\big(Ber(\hat{\pi}_j^S), Ber(\hat{\pi}_j^D)\big) \tag{4.7}$$

where $\hat{\pi}_j^D$ and $\hat{\pi}_j^S$ are the estimates of the probabilities π_j^D, π_j^S of the j-th observation.

If instead we consider Hellinger distance, in the case of two Bernoulli distributions, Equation (3.6) can be expressed as:

$$\begin{aligned} H\big(Ber(\pi_j^S), Ber(\pi_j^D)\big) &= \frac{1}{\sqrt{2}}\cdot\sqrt{\left(\sqrt{\pi_j^S}-\sqrt{\pi_j^D}\right)^2+\left(\sqrt{1-\pi_j^S}-\sqrt{1-\pi_j^D}\right)^2} = \\ &= \sqrt{1-\sqrt{\pi_j^S\pi_j^D}-\sqrt{1-\pi_j^S}\sqrt{1-\pi_j^D}} \end{aligned}$$

In the end, the quality measure that will be used in the EMM framework is:

$$\hat{\varphi}_H = \frac{\sum\limits_{j\in S} H\big(Ber(\hat{\pi}_j^S), Ber(\hat{\pi}_j^D)\big)}{|S|} \tag{4.8}$$

where $\hat{\pi}_j^D$ and $\hat{\pi}_j^S$ are the estimates of the probabilities π_j^D, π_j^S of the j-th observation.

Chapter 5

Dataset description and preprocessing phase

This chapter describes more in detail the dataset under analysis. First, it provides a general background for lung cancer: what it is, the symptoms, the risk factors and how it can be treated. Then, it shows an overview of the dataset itself with the description of the main attributes that have been provided. In the end, it describes how the preprocessing of the data was carried out in order to have cleaned data on which to apply the logistic regression model in the EMM framework. As we mentioned in the introduction, the dataset was provided by IKNL and contains information regarding patients who were diagnosed with lung cancer in the period 2005-2018.

5.1 Lung Cancer

Lung cancer (also known as lung carcinoma) is the leading cause of cancer deaths worldwide, the most common type of cancer for men and the third most common for women [1]. It is characterized by an out-of-control growth of cells in the lungs tissues. The lungs are two organs located in the chest whose main task is devoted to respiration. When we breath in with our mouth and/or nose, air travels through the trachea and reaches the lungs via two main branches called bronchi. Within each lung, the bronchi then divides into smaller branches denominated bronchioles which end up in the alveoli. These are tiny air sacs, responsible for the exchange of oxygen and carbon dioxide with the blood: it is at this level that the actual respiration takes place. The right lung is composed of 3 sections, called lobes. The left lung is slightly smaller, with 2 lobes, due to the presence of the heart in that part of the body.
There are two main types of lung cancer that can be identified: non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). This distinction is important because the behaviour of the tumor is rather different and this is reflected by the treatments that can be provided and the survival rates.
NSCLC is the most common type, ranging from 80 to 85 % of total cases. In this category we can identify three main subtypes according to the kind of cells from which the cancer originates: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. The treatments that can be applied in this case depend on how much the cancer has spread and also on other factors, for example the patient's health condition. Some possibilities are surgery (to remove the cancer if it still limited), radiofrequency ablation, immunotherapy, radiotherapy, chemotherapy.
SCLS is less frequent (10-15 %) and in general is considered more dangerous because the tumor tends to grow faster than the NSCLC type: in most cases, when diagnosed, the cancer has already spread. Therefore, the possible treatments are more limited and in many cases doctors make use of chemo or radiotherapy to limit the tumor rather than cure it.
The most common symptoms of lung cancer are: shortness of breath, chest pain, cough, hoarse-

ness, feeling tired or weak, etc.
The most important risk factor is undoubtedly smoking, but other factors can also play an important part, such as air pollution, exposure to substances like radon and asbestos and personal or family history of lung cancer.

5.2 Dataset description

The dataset contains information regarding patients that were diagnosed with lung cancer in the period 2005-2018. In total we have 170329 rows and each of them does not refer to a single patient but to a primary tumor. Therefore, there can be multiple rows referring to the same patient with several primary tumors. The reader can find the meaning of every attribute of the dataset in Appendix A, as well as some tables relative to specific attributes. In this section, we focus on some of them.
What we notice is that we can split the attributes into certain categories according to the area of interest. For example, we have information closely related to the patient: the gender (**gesl**), the age (**leeft**), the vital status (**vitstat**) and the socio-economic status (**ses**). The latter is not the direct income of the individual but is derived from the the wealth of the area of residence (Table A3).
The dataset also contains temporal attributes such as the incidence year (**incjr**), i.e. the year in which the patient was diagnosed with cancer, the follow up in days (**vitfup**), which represents how many days the patient has survived since the diagnosis, the date of vital status (**vitdat**). With respect to the latter, there are checks once a year and, if the person is still alive, the day of the last check becomes the date of vital status; otherwise the date of death becomes the date of vital status.
Similarly, there are several geographical attributes: the region of residence of the patient (**COROP**), the province (**PROVINCIE**), both according to the Nomenclature of Territorial Units for Statistics (NUTS [26]), and even the last two digits of the post code of residence (**PC2**).
There are many attributes strictly related to the cancer itself. The topography (**topog**) is composed of one letter followed by 3 digits, with the first three characters representing the type of tumor (in this case lung cancer is identified with the code C34) and the last digit indicating the specific location (Table A1).
The morphology (**morf**) focuses on the histology of the tumor. There are several morphology codes present in the dataset and each of them records the type of cell that has become neoplastic (abnormal growth) and its biologic activity; in other words, it records the kind of tumor that has developed and how it behaves.
The differentiation grade of the cells (**diffgr**) is a reflection of how abnormal the cells look under the microscope. In cancer, cells become deregulated and proliferate abnormally (dysplasia). As dysplasia develops, the cancer cells lose features of their tissue of origin and become less and less differentiated. This translates into a higher grade of differentiation (Table A2).
One of the most interesting features, when examining a tumor, is certainly its stage (**stage_main**), an indication of the degree of spread of the cancer. Understanding the degree of spread of a cancer brings advantages to oncologists and physicians because it gives them vital information to choose the best treatment options. For example, the treatment for an early-stage cancer may be surgery or radiation, while a more advanced-stage cancer may need to be treated with chemotherapy. There are essentially three different ways a tumor can spread:

- **Direct**: the tumor grows and invade an adjacent structure (local invasion). In the case of lung cancer, the tumor can spread from the lung directly into the chest wall, airways, etc.
- **Lymphatic**: cancer cells enter the lymphatic vessels and lymphnodes. In the case of lung cancer, the tumor could spread to the hilar, the mediastinal and the supercurricular lymphnodes.
- **Hematogeneous**: cancer cells go into the blood circulation and reach distant location with respect to the origin of the tumor. In the case of lung cancer, the tumor could spread to organs such as the brain, the liver, the bones, the adrenal glands.

Stage is commonly defined using the TNM classification system. T stands for *tumor* and focuses on the extent of the local and primary tumor growth (size). N stands for *nodal status* and represents the degree of lymphnodes involvement. M stands for *metastasis* and indicates whether there was a spread of cancer cells from the place where they first formed to another part of the body. T, N, M are therefore a reflection of the three possible ways a tumor can spread: direct, lymphatic or hematogeneous.
Furthermore, there are two main types of staging. Clinical staging is an estimate of the extent of the cancer based on results of imaging tests (x-rays, CT scans, etc.), physical exams and tumor biopsies. Pathological staging is instead conducted when part of the tumor has been surgically removed (generally, nearby lymphnodes are also sampled in such a setting). The pathological stage can then be different from the clinical one and gives the doctors more precise information about the cancer.
In the dataset, we have both attributes related to the clinical stage (**ct, cn, cm**) and the pathological one (**pt, pn, pm**). Since it is not always possible to perform the latter type of staging when examining a cancer, there are many more missing values for **pt, pn, pm** (approximately 143000 for each of them) compared to **ct, cn, cm** (approximately 8000 for each of them).
There is also a specific attribute indicating the presence or not of metastasis (**metastasized**) and two attributes indicating the number of examined lymphnodes (**lyond**) and the number of positive lymphnodes (**lypos**).
In the end, another important category of attributes is composed of the ones related to treatments. The attribute **treatment** indicates whether or not a patient received any treatment. Then there are many Boolean attributes that refer to the specific treatment, e.g. local surgery (**indchlok**), immunotherapy (**indimmuno**), etc . **radio_chemo** is specific for people that received both radio and chemotherapy and takes 4 values, indicating if the two therapies were concurrent, sequential, distinct or it was not possible to determine that. We also have information regarding the first type of hospital the patient came in contact with (**first_hospital_type**), the first treatment they received (**first_hospital_treatment**) and the first hospital ward (**first_specialism**).

5.3 Preprocessing

In every data mining task, having consistent, clean data is of the utmost importance. Data affected by inconsistencies (e.g. outliers, wrong entries) will influence the result of any technique that will be applied on the dataset. It does not exist any perfect method that is suitable for every occasion, indeed the preprocessing depends on the particular dataset itself.
In this section, we explain step by step what has been done in regards of the dataset that IKNL provided.

5.3.1 Dropped attributes

First of all we focus on which attributes are kept for analysis and which ones instead are discarded because not informative (barely no variance, not useful for our analysis). The following is a list of attributes that were dropped from our dataset:

- **episode**: DIA (diagnosis) is the only value, i.e. the tumor was detected after a diagnosis of the patient.
- **topo**: C34 is the only value (the code is indicative for lung cancer);
- **gedrag**: 3 is the only value (it stands for malignant tumor);

- **Age 31-12 yoi, Age 1-1 yoi**: age of the patient computed respectively on December 31st and January 1st to conduct statistical studies with respect to the year. For our research it is not important because a very similar piece of information is included in *leeft*;
- **vitfup**: it is given by *follow_up_years**365.25 so it is redundant. *vitfup* is the followup in days of the patient given by date of vital status - incidence date; *follow_up_years* will be useful for some experiments regarding survival analysis (Section 7.1).
- **vitdat**: date of vital status. It is not interesting for our analysis.

5.3.2 Mappings

There are some attributes for which recoding is required before continuing with the analysis.
In the case of *stage_main*, in the NCR, there are different ways for classifying the stage of the cancer and, in general, TNM is the preferred classification system. However, if there is not sufficient information, the EoD (Extent of Disease) classification is instead adopted. The EoD system is characterized by 6 different values, from 1 to 6, which indicate the degree of spread of the tumor. Our goal is to have for *stage_main* a unique system of classification. Therefore we are going to convert the EoD stage to a TNM stage, according to Table 5.1.

EoD stage	Corresponding TNM stage
1	1
2	1
3	2
4	3
5	3
6	4

Table 5.1: Recoding to convert a stage classified via the EoD system to a stage classified using TNM.

There are also some categorical attributes that have many different levels. This can cause mainly two problems depending on how the attribute is used. If the attribute is set to be a descriptor, the risk is that, when we impose a condition on that attribute to generate a subgroup (*attribute*=value), the subgroup will be too small to be informative. If instead the attribute is used as a predictor in the logistic regression model, the risk is that the model is unstable and convergence may not be reached. This is discussed more in detail in Chapter 6.
Regarding *first_hospital_treatment*, i.e. the hospital of first treatment, the main 3 levels are ALGZK(general hospital), STZ(top clinical), UNIVC(university hospital). There are also some special values such as 968 (general practitioner), 998 (abroad), 999 (unknown) and these will be coded as "OTHER". Then we have many other levels coded in the range 900-1000 that simply refer to a specific radiotherapy institution/department: we are going to code those simply as "RADIO".
Regarding *first_specialism*, the first hospital ward that was visited by the patient after the diagnosis of cancer (e.g. pneumology, cardiology, ...), there are many different levels and some of them with very few observations (in some cases even just one). Therefore, we are going to keep the top 3 most frequent levels ("0400 - Longziekten", "0200 - interne", "2100 - Neurologie") and map all the others as "other".

5.3.3 Correlation analysis

Another important issue to be tackled down is correlation. Indeed, highly correlated attributes hold similar information, hence can be redundant for the analysis. Furthermore, as it will be described in Chapter 6, when applying logistic regression, we want the predictors to be as less

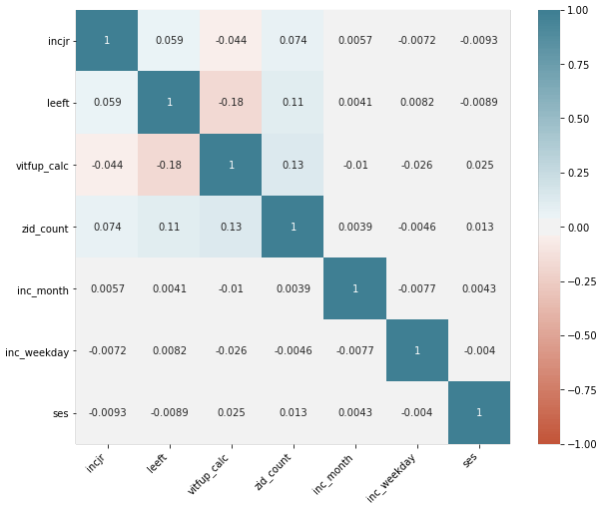


Figure 5.1: Pearson correlation among numeric attributes

correlated as possible between each other, otherwise we might have some problems. We can split the attributes into two categories: numeric and categorical.
Regarding the numeric attributes, we can make use of the classic Pearson correlation coefficient. It is a statistic that measures linear correlation between two variables x and y. Its values can range from -1 to +1, with 0 suggesting no correlation at all and -1, +1 indicating respectively completely negative and positive correlation between the attributes. Formally, it is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x^i - \bar{x}) \cdot (y^i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x^i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y^i - \bar{y})^2}}$$

where x^i, y^i are the i-th observations and $\bar{x}, \bar{y}$ are the sample means.
As we can see from figure 5.1, there is a rather low correlation between the attributes, hence they will not be removed.
Regarding the categorical attributes, we cannot apply Pearson but we have to look for another quantity to measure correlation. Cramér's V [27] can be suitable for our goal: it is one of the most popular measures of association between two nominal variables that can also have more than two levels. It is computed after defining the contingency tables with the two nominal variables. It is defined as:

$$V = \sqrt{\frac{\chi^2/n}{min(k-1, r-1)}} = \sqrt{\frac{\phi^2}{min(k-1, r-1)}} \tag{5.1}$$

where n is the total number of observations, k is the number of columns and r is the number of rows. χ^2 represents Pearson's chi-squared statistic and tells us whether there is a significant relationship between the two variables or not. Given this definition, we could wonder why there is the need to use Cramér's V when we could simply consider χ^2. The problem is that χ^2 by itself does not inform us on how important the relationship between two variables is, it just indicates whether it is significant or not. Cramér's V is then an adjustment of χ^2 and takes values between 0 and 1. The former is indicative of no relationship at all whereas the latter suggests a strong association between the two nominal variables. ϕ is called the mean square contingency coefficient and is equivalent to Cramér's V when we are dealing with a 2×2 contingency table, i.e. when we have two binary variables.

Cramér's V defined by Equation (5.1) is actually a biased estimator and Bartlett[28] proved that, under independence, the expected value of ϕ^2 is:

$$\mathbf{E}\Big(\phi^2\Big) = \frac{1}{n-1}(k-1)(r-1)$$

In this regard, Bergma[29] proposes to use instead:

$$\tilde{\phi}^2 = \phi^2 - \frac{1}{n-1}(k-1)(r-1) \qquad \tilde{k} = k - \frac{1}{n-1}(k-1)^2 \qquad \tilde{r} = r - \frac{1}{n-1}(r-1)^2$$

and, since now $\tilde{\phi}^2$ could be negative and that would not make sense, the suggestion is to correct with:

$$\tilde{\phi}^2_+ = max(0, \tilde{\phi}^2)$$

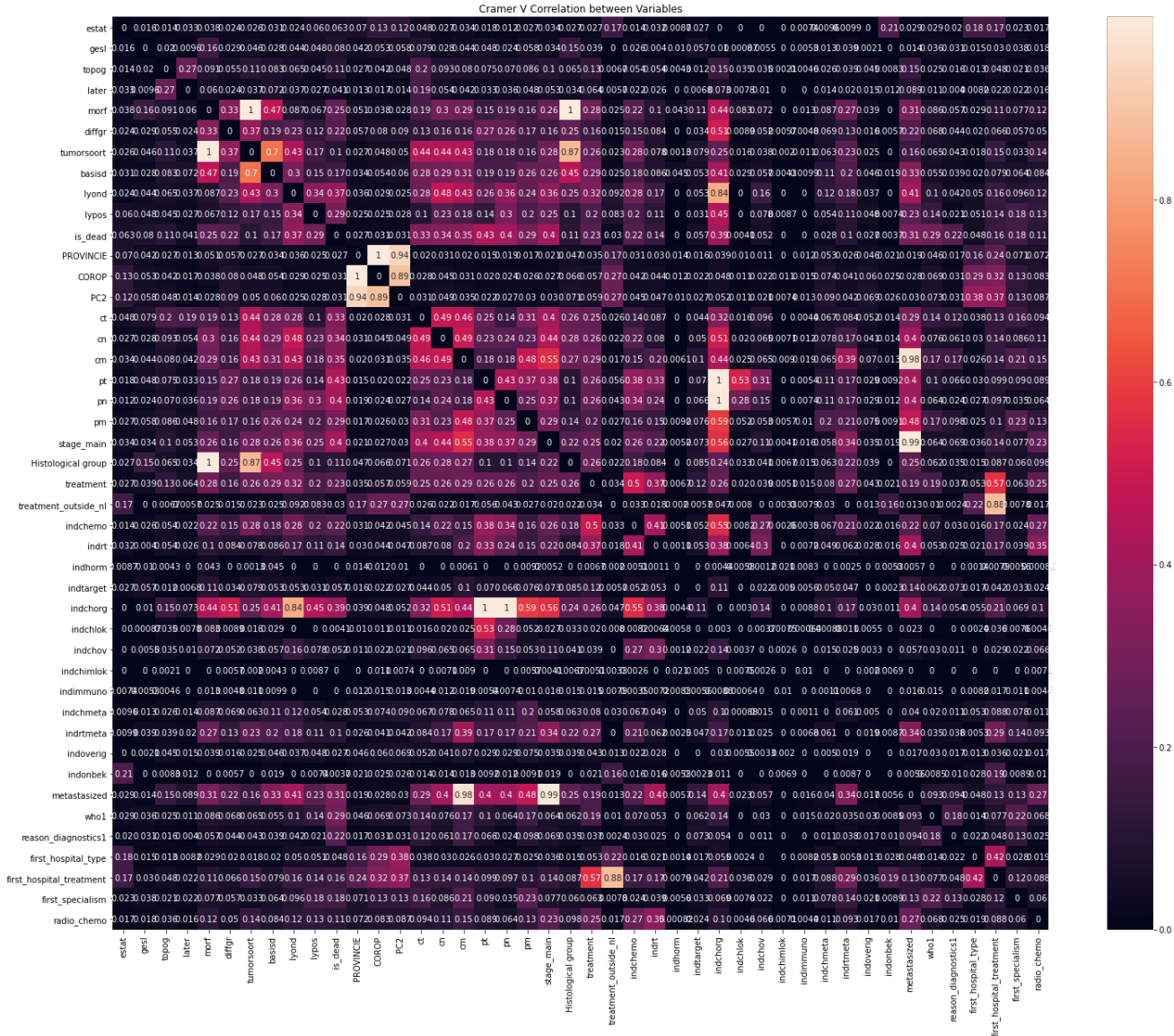
The adjusted Cramér's V statistic is:

$$\tilde{V} = \sqrt{\frac{\tilde{\phi}^2_+}{min(\tilde{k}-1, \tilde{r}-1)}} \tag{5.2}$$

The correlation matrix, for the categorical attributes, calculated using Equation (5.2) is shown in Figure 5.2.

So far, we have described the first steps to analyze and prepare the dataset for the experiments that will be conducted in the EMM framework. The following chapter shows which are the parameters of our experiments, what are the issues that we could encounter when applying the logistic regression model and how to overcome those.

Figure 5.2: Cramer's V correlation among categorical attributes



Chapter 6

Experimental Setup

In the previous chapter we analyzed the dataset in detail, providing a general overview of the attributes that we have, with a focus on which ones to consider later in our analysis and how they are correlated to each other.
This chapter is about the preparation phase before conducting the experiments. In the first section, we describe meticulously the parameters which can be used to define the different experiments. The second section addresses some limitations, caused by the fact that we are using logistic regression, and some ideas on how to overcome these constraints, based also on the preprocessing steps of Chapter 5. Lastly, the third section introduces the concept of marginal effect, an important measure which can be used to interpret the differences between two logistic regression models.

6.1 Beam search parameters

As described in Section 2.3.4, in order to explore many different subgroups and find interesting ones in the EMM framework, the most common choice is to rely on a heuristic: the beam search algorithm. Heuristic is preferred over a brute force approach because in general there would be too many subgroups to analyse.
This section describes the parameters involved in the beam search algorithm: they determine which subgroups can be generated, which logistic regression model we are interested in and how the search is conducted. The algorithm that has been used is the one described in pseudo-code in [11] (Algorithm 1). The code for this project is based on the one implemented for Exceptional Incidence Distribution Mining [3]. Several changes have been made in order to accommodate for new aspects, such as the logistic regression model, the new quality measures, the specific dataset and the output (summaries, plots, ...). Nevertheless, the general structure is fundamentally the same (definition of a subgroup, search, evaluation, ...).
At the beginning of the algorithm, in the first level of the beam, we generate all the possible subgroups by imposing a single condition on one of the descriptive attributes and then we assess them according to the chosen quality measure. At a first glance, it looks like an exhaustive search, because all the possibilities are explored. However, only the ω (with ω width of the beam) subgroups with the highest scores of the quality measure are kept. The real difference occurs in the second level of the beam: at this point only those ω subgroups are further processed. Again we consider all the possible ways to refine those specific subgroups and then we keep the ω subgroups with the highest quality measure. These steps are repeated until we reach the maximal depth of the beam. The output of the algorithm is represented by the best ω subgroups for each single level of the beam, ordered by the score of the quality measure.
The parameters of the beam search algorithm are:

- **descriptors**: attributes that will be used to generate the subgroups. As we described earlier in Section 2.3.3, every time we want to create a subgroup, we impose one or multiple conditions on the descriptive attributes (e.g. *metastasized*=True, *diffgr*=4).

- **width**: number of subgroups that are kept at each level of the beam search. We indicate that with ω.
- **depth**: number of levels to explore. If depth=1, we are only exploring the subgroups defined by a single condition on the descriptors. If depth=2, we are also exploring the subgroups defined by two conditions and so on.
- **score_metric**: here we indicate the quality measure that we are going to use. With respect to our project, we have three options. The first is defined in Section 4.1, in which we take into account the p-values and compare the subgroup to the complement. The second and third options are respectively Kullback-Leibler divergence, defined by Equation (4.7), and Hellinger distance, defined by Equation (4.8). We are also going to multiply these two measures to the entropy function defined by Equation (2.3) to take the subgroup size into account. It is worth noting that in these 2 cases, we are comparing the subgroup to the whole dataset.
- **target_attribute**: list of attributes that are used in the logistic regression model. The first element of the list indicates the response, whereas all the others are the predictors.

Among all these parameters, the most important to define is probably **target_attribute** because it outlines the objective of the experiment itself. We should recall that our goal is to retrieve those subgroups for which the relationships between the predictors and the response are remarkably different from the same relationships present in the dataset or in the complement (according to the chosen quality measure).
The amount of experiments that we could conduct using this particular model is really enormous. Indeed, the dataset is full of binary attributes and each one of them could theoretically be chosen as the response. Furthermore, after the choice of the response, we have plenty of possibilities to define the predictors because any type of attribute (numeric, categorical, boolean) is accepted.
There are some limitations though, due to the fact that we cannot randomly select the targets and hope that the logistic regression model makes sense. There are some prior analysis that are necessary to define a suitable model. In the following section, we explain what are some limitations of this model and how we could overcome them.

6.2 Limitations and possible solutions

In the EMM framework, we are using the (binary) logistic regression model to analyze the relationship between some independent variables (covariates or predictors) and a dichotomous dependent variable (response). Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is linear. However there are still some limitations that must be tackled down.

First of all the problem of **multicollinearity**, i.e. a statistical phenomenon in which two or more predictors are highly correlated with each other [30]. More formally, when referring to a set of variables, we say that the set is collinear if there exists one or more linear relationships among the variables. The presence of multicollinearity among the predictors causes the variances of the parameter estimates to be inflated: this leads to large standard errors and the confidence intervals of the coefficients tend to become very wide. Therefore, it can become more difficult to reject the null hypothesis that the coefficients are not significant and that can lead to incorrect conclusions about the relationships between the predictors and the response. It is very important to avoid that because our focus when using Exceptional Model Mining is on the relationships, hence the coefficients, and we do not want to draw the wrong conclusions. Furthermore, if there is a strong collinearity, a small perturbation of the data can cause a large and unpredictable perturbation on the estimates of the coefficients, making them not very stable and subsequently unreliable.
In order to tackle this problem, the first thing we need to do is to avoid including redundant

variables in the predictors. That is why, during the preprocessing step, we got rid of all the non-informative and superfluous attributes. However, in most cases, it is more common to have attributes that are not redundant but are still highly correlated to each other. That is the reason why it is very helpful to take a look at the correlation matrices computed in Chapter 5 and avoid to select as predictors those attributes that show a high correlation.

Another limitation relates to convergence of the logistic regression model. In some cases, the algorithm used to compute the regression coefficients does not converge and there can be several reasons. We must keep in mind that the method that we use to find the coefficients is the maximization of the likelihood function (or equivalently of the log-likelihood function). Since we are trying to maximize a function, we could encounter the problem of local maxima; in this scenario the algorithm used to find the solution could keep on iterating around the same local maximum, never reaching the global one, i.e. the desired solution. Fortunately this does not happen with the log-likelihood function because it is globally concave, hence there is at most one maximum.
Unfortunately, it is possible that the function has an infinite maximum, thus convergence is never reached, as pointed out in [31]. This can occur in the case of *complete separation*. It happens when there exists a linear function of the predictors that yield prefect predictions of the response. Formally:

$$\exists \boldsymbol{\beta} \in \mathrm{R}^p : \begin{cases} y_i = 0 & \text{if} \quad \boldsymbol{X_i'^T}\boldsymbol{\beta} < 0 \\ y_i = 1 & \text{if} \quad \boldsymbol{X_i'^T}\boldsymbol{\beta} > 0 \end{cases} \quad \forall i = 1, ..., N \tag{6.1}$$

where $\boldsymbol{\beta}$ is the vector of coefficients and $\boldsymbol{X_i}$, y_i are respectively the vector of predictors and the response associated to the i-th observation.
A very similar problem, the *quasi-complete separation*, occurs when:

$$\exists \boldsymbol{\beta} \in \mathrm{R}^p : \begin{cases} y_i = 0 & \text{if} \quad \boldsymbol{X_i'^T}\boldsymbol{\beta} \leq 0 \\ y_i = 1 & \text{if} \quad \boldsymbol{X_i'^T}\boldsymbol{\beta} \geq 0 \end{cases} \quad \forall i = 1, ..., N \tag{6.2}$$

with the only difference that the equalities are also allowed.
For both types of separations, a maximum likelihood estimate does not exist, hence the algorithm stops after a maximum amount of iterations has been reached. Often, separation occurs when the dataset is too small to observe events with low probabilities. Another common cause is when there is a category or range of a predictor with only one value of the response.
While collinearity can be kept more easily under control, by avoiding to include highly correlated variables in the predictors, separation is more of an issue in the EMM framework. In the beam search algorithm, for every experiment, we define a model with a response and some predictors. A logistic regression is fitted on the whole dataset to get an idea of the relationships between the variables. The same logistic regression model, i.e. with the same predictors and response, is fitted on every subgroup explored by the beam search. The issue is that a subgroup could be pretty small, especially when exploring deeper levels of the beam. That can greatly increase the chances of having a categorical predictor with very few observations for a specific level (or a numeric one with a limited range). Hence, we are more likely to encounter a separation problem and subsequently not being able to compute the coefficients.
A possible solution to overcome this limitation would be to run an analysis on every single subgroup and select the best possible model, if there is any. That would imply to select those predictors that can best explain the response variable, that are not too correlated to each other and that do not lead to a complete separation issue. However, there are two main reasons why that would not be ideal. The first one is merely a computational time issue: for every single subgroup we would have to run a separate analysis and this could take a long time. The second one is the core of Exceptional Model Mining itself because it would not make sense to compare the subgroup to the dataset (or the complement) using two very different models, i.e. models in which the predictors are not the same.
Another possible solution is to use the same response and predictors for all the subgroups so that a comparison is possible. Additionally, we should avoid including categorical variables with too

many levels as predictors. In any case, for each experiment, we are going to keep track of the number of times that convergence is reached for the subgroups to see if the attributes that we included as predictors are appropriate or not.

6.3 Interpretation of the results: marginal effects

The aim of Exceptional Model Mining is to find subgroups that behave differently with respect to the norm (i.e. the whole dataset or the complement). After running the beam search algorithm, we end up finding the most interesting subgroups according to the quality measures that we previously decided to adopt. The question is then how to interpret the results that we get. How can we compare the logistic regression model fitted on the subgroup to the logistic regression model fitted on the dataset (or the complement)?
To illustrate better the problem, let's consider two groups on which we fit logistic regression using the same response and predictors:

$$GroupA : logit\Big(P(y=1)\Big) = \beta_0^A + \beta_1^A x_1 + ... + \beta_p^A x_p$$
$$GroupB : logit\Big(P(y=1)\Big) = \beta_0^B + \beta_1^B x_1 + ... + \beta_p^B x_p$$

Group A could represent the subgroup and group B the whole dataset or the complement, for example. The coefficients are indicative of the relationships between the predictors and the response; this relationship is expressed in terms of the logit of the probability that the response is equal to 1.
The idea of making comparisons by using the coefficients, at a first glance, looks very appealing. We could for example compare β_1^A and β_1^B and based on their values, conclude that the relationship between x_1 and y is the same or is different across the two groups. Though this approach is straightforward and rather easy to implement, it has a non-negligible drawback. As Allison points out in [32]: "there is a potential pitfall in cross-group comparisons of logit or probit coefficients that has largely gone unnoticed. Unlike linear regression coefficients, coefficients in these binary regression models are confounded with residual variation (unobserved heterogeneity)". Unobserved heterogeneity is the variation in the dependent variable that is caused by variables that are not observed (e.g. omitted variables or variables that have not been taken into account). This heterogeneity affects the scale of the coefficients making it unreliable to make comparisons across groups unless the residual variation is the same for all the groups.
Mood [33] showed that the issue of unobserved heterogeneity not only represents a problem in the comparison of coefficients across samples, but it is also problematic when comparing the odds or the odds ratio across different groups.
That is why Mustillo and Long have proposed a different method to make such comparisons [34]. Instead of focusing on the coefficients, they show how to compare groups based on predicted probabilities and marginal effects. Though it may not be as immediate as comparing the coefficients right away, probabilities and marginal effects are not affected by residual variation. But what is a marginal effect and how can it be computed with respect to logistic regression?
The marginal effect of a predictor x_j is equal to the change in the probability of the outcome when a change of x_j occurs, while keeping all the other predictors at a fixed value. The marginal effect is defined as marginal change if the change of x_j is infinitesimal . It is instead denominated discrete or finite change if the change of x_j is discrete . Mathematically, if we have y as the response and $\boldsymbol{X} = (x_1, ..., x_p)$ as the vector of predictors, the discrete change for the variable x_j $(1 \leq j \leq p)$ is defined as:

$$DC_j = P\big(y = 1|x_1 = x_1^*,\ \dots, x_j = end,\ \dots, x_p = x_p^*\big) - P\big(y = 1|x_1 = x_1^*,\ \dots, x_j = start,\ \dots, x_p = x_p^*\big)$$

i.e. the difference of the probabilities of $y = 1$ when x_j changes from the value *start* to the value *end*, holding all other predictors at specific values[1].
There are different types of marginal effects based on which values we set the other predictors at. A very common approach is to adopt the average discrete change (ADC). It is the arithmetic mean of the discrete change of x_j computed for each observation within the group, using the observed values of the predictors. Therefore, for every observation we compute the following discrete change:

$$DC_{ij} = P\big(y = 1|x_1 = x_{1i},\ \dots, x_j = end,\ \dots, x_p = x_{pi}\big) - P\big(y = 1|x_1 = x_{1i},\ \dots, x_j = start,\ \dots, x_p = x_{pi}\big) \tag{6.3}$$

where $x_{1i},\ \dots,\ x_{pi}$ are the values of the predictors for the ith observation.
The values of *end* and *start* in Equation (6.3) depend on the type of variable that we are considering. If x_j is numeric, then $start = x_{ij}$, i.e. the actual value of x_j for the ith observation, and $end = x_{ij} + \delta$ where δ is a constant (a typical value is 1). If x_j is binary with two levels coded as 0 and 1, $start = 0$ and $end = 1$. If x_j is categorical with more than 2 levels, it is slightly more complicated. Suppose that x_j has levels $l_1, l_2, ..., l_m$, with l_1 being the reference level in the logistic regression. Then, for every level $\bar{l} \in \{l_2, ..., l_m\}$, we compute a different ADC with $start = l_1$ and $end = \bar{l}$.
After computing the discrete change for every observation, we can compute the average discrete change via the formula:

$$ADC_j = \frac{1}{|G|} \sum_{i \in G} DC_{ij}$$

where $|G|$ is the size of the group.

Another popular marginal effect is the so called discrete change at the mean (DCM). It is defined as the differences of the probabilities of success of the response when x_j goes from *start* to *end*, keeping the other predictors at their mean values $(\bar{x_1},\ \dots, \bar{x_p})$.

$$DCM_j = P\big(y = 1|x_1 = \bar{x_1},\ \dots, x_j = end,\ \dots, x_p = \bar{x_p}\big) - P\big(y = 1|x_1 = \bar{x_1},\ \dots, x_j = start,\ \dots, x_p = \bar{x_p}\big)$$

If x_j is numeric, *start* is equal to the mean value of x_j and $end = start + \sigma(x_j)$, where $\sigma(x_j)$ is the standard deviation of the predictor x_j. If x_j is categorical, *start* and *end* are defined in the same way as with ADC.
When interpreting the results from our experiments, we are going to focus on ADC rather than DCM. The reason is that ADC uses the true values of the observations within the groups thus reflecting the true population of that group, whereas DCM uses the mean values that could be indicative of an individual that is not even present in the group.

[1] $x_1^*, ..., x_p^*$ represent the values at which we set the predictors.

Chapter 7

Experimental Results

After the preprocessing phase, some considerations over the limitations of the logistic regression model and the description of marginal effects to interpret the model, we are ready to conduct some experiments with the aim of finding interesting subgroups. It is important to remind that by interesting we mean those subgroups that show relationships between the response and the predictors that somehow deviate from the norm (the full dataset or the complement). For every experiment we describe which parameters have been set for the beam search algorithm. Regarding the output, we provide a summary of the logistic regression model, a plot of the average discrete changes and, when suitable, a plot of the coefficients of the single predictors. Furthermore, we outline extra information, such as the pseudo R squared and the AUC to get a better understanding of the reliability of the model. We also consider the number of times in which convergence was not reached, trying to fit the logistic regression model on the subgroups.
In the first section, the experiments focus on survivability and all of the three quality measures previously defined are applied: Hellinger distance, KL-divergence and the one defined with the p-values. In the second section, we describe an experiment that has treatment as the key factor to analyze.

7.1 Logistic regression analysis on survivability

In a clinical setting, survivability is often regarded as one of the most relevant factors to analyze. As we mentioned before, lung cancer is among the deadliest types of cancer and one of the main causes is that the cancer is usually diagnosed when it is already spreading, making it more difficult for doctors to contain it and remove it.
First of all, we need to choose which attribute will represent the response in our logistic regression model. To this end, we create a new binary attribute in the dataset, named ***survival_more_1year***, which is equal to 1 if the patient survived for more than 365 days after the first diagnosis, 0 otherwise. In order to generate that, we make use of the attribute *follow_up_years*, i.e. how long the patient survived after the first diagnosis in terms of years, so that:

$$survival\_more\_1year = \begin{cases} 0 & \text{if} \quad follow\_up\_years < 1 \\ 1 & \text{if} \quad follow\_up\_years \geq 1 \end{cases}$$

For the future experiments on survivability, we are going to consider non-small-cell lung cancer (NSCLC), since it is the most common type of lung cancer and has more observations than SCLC (146000 vs 24000).
Let's start by having a look at the distribution of *survival_more_1year*, showed in Figure 7.1. We notice that the response variable is rather balanced with respect to the two classes with a slightly larger number of people that did not survive for more than one year.
In the dataset, we also have information about the current vital status of the patient. If we have a look at the contingency table with the attributes *is_dead* and *survival_more_1year*, displayed

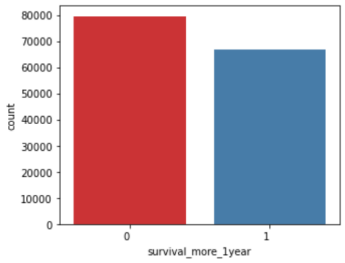


Figure 7.1: Distribution of *survival_more_1year* over the patients that were diagnosed with NSCLC.

Figure 7.2: Contingency table *is_dead* and *survival_more_1year*.

survival_more_1year	0	1
is_dead		
False	95	24270
True	79442	42490

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.1108      0.057     19.527      0.000       0.999       1.222
C(gesl)[T.2]   0.2669      0.014     19.602      0.000       0.240       0.294
C(treatment)[T.True]  2.0353      0.019    107.385      0.000       1.998       2.072
C(stage_main)[T.2]   -0.7242      0.029    -24.977      0.000      -0.781      -0.667
C(stage_main)[T.3]   -1.5892      0.021    -74.885      0.000      -1.631      -1.548
C(stage_main)[T.4]   -2.9714      0.021   -143.792      0.000      -3.012      -2.931
C(stage_main)[T.X]   -1.2721      0.042    -30.492      0.000      -1.354      -1.190
leeft          -0.0170      0.001    -25.408      0.000      -0.018      -0.016
==============================================================================
AUC: 0.844
Average probability: 0.457
Pseudo R-squ: 0.295
```

Figure 7.3: Logistic regression model for survival analysis on the whole dataset.

in Figure 7.2, we observe that there are 95 patients that are still alive (*is_dead*=False) but were classified as if they did not survive for more than one year (*survival_more_1year*=0). If we analyse these patients more in detail, we notice that they all have a follow-up date that has not been updated. In other words, we do not have information regarding the cancer anymore (for example because they left the Netherlands). Since there are only 95 records of this kind, we can simply remove them from the dataset.

After having selected the response, the most important choice to make is which attributes will be the predictors in our analysis. The idea is to receive indications from a domain expert on which could be the most interesting attributes to analyze, select those as predictors and check retrospectively if the model is a suitable one or not with the aid of some statistical measures (p-values, pseudo R-squared, AUC). Therefore, after having received some suggestions from a lung cancer expert, I decided to select the following predictors: *leeft, gesl, treatment, stage_main*. Before proceeding, we want to be sure that these attributes are not highly correlated to each other because that could lead to the issue of multicollinearity. If we have a look at Figure 5.2, we notice that there is no high correlation among these attributes.

Now we are ready to fit the logistic regression model on the whole dataset with, as response, $y = survival\_more\_1year$ and, as vector of predictors, $\boldsymbol{X} = [leeft, gesl, treatment, stage\_main]$. If we have a look at Figure 7.3, we notice that all the coefficients are statistically significant (they all have p-values smaller than 0.05), the AUC is rather high (0.844) and the pseudo R-squared is in the range 0.2-0.4 which is considered good by McFadden [20].

The average probability is obtained by computing the arithmetic mean of the predicted probabilities of all the observations in the dataset. It is equal to 0.457, meaning that, according to the model, a randomly selected person in the dataset has 45.7% probability of surviving more than one year. This is also reflected by the histogram in Figure 7.1, in which the amount of people that did not survive for more than one year is slightly higher than the rest of the population.

Now we can shift our attention to the coefficients estimates to get a clearer view on the relation-

ships between the predictors and the response. Regarding *gesl*, gender, the reference level is *gesl=1*, i.e. gender=Male. The coefficient associated to gender=Female is equal to 0.2669, meaning that being a woman has a positive effect on the logit of the probability of *survival_more_year=1*. We can also interpret it with respect of the odds. The fitted model suggests that, holding all the other predictors at a fixed value, the odds of surviving more than one year for females (*gesl=2*) over the odds of surviving more than one year for males (*gesl=1*) is exp(0.2669) = 1.306. In terms of percent change, we can say that the odds for females are 30.6% higher than the odds for males.
A similar interpretation can be given to *treatment*. In this case, receiving a treatment has a positive effect on survival, with an increase of more than 6 times in the odds of surviving (exp(2.0353)=7.655) for a patient that received a treatment versus a patient that did not receive any.
So far we have considered two dichotomous variables; *stage_main*, instead, is a categorical variable with multiple levels. The reference level is *stage_main=1* and, as expected, the higher the stage the more negative the effect of the coefficient on the logit of the probability of surviving more than one year. According to the model, an unknown stage (*stage_main=X*) has a negative effect on survivability worse than being diagnosed with a stage 2 cancer (-1.2721 vs -0.7242) and slightly better than a stage 3 cancer (-1.2721 vs -1.5892).
In the end, let's consider age (*leeft*), the only numeric attribute in the model. Holding all the other predictors at a fixed value, we have a 1.7% decrease in the odds of surviving more than one year for a one-unit increase in age, since exp(-0.017) = 0.983. For a numeric attribute we can also compute the effect of a wider increase of the variable, by simply multiplying the increase to the coefficient and then use that as the power of e. For example, for a ten years increase in age, we have a 15.6% decrease in the odds of surviving more than one year, since exp(-0.017×10) = 0.844.
The results that we get from this model are not very surprising. It is common knowledge that if you are older, you have not received a treatment or have a higher stage (meaning that the cancer has already started spreading) it is more likely that you will survive for a shorter period of time. What we seek to find, when using the Exceptional Model Mining framework, are subgroups in which the relationships between the response and the predictors are remarkably different compared to the ones observed here.
Now we are ready to conduct some experiment in order to find interesting subgroups.

7.1.1 Experiment 1: quality measure via Hellinger distance

We briefly list the parameters of the beam search algorithm that have been set to conduct the first experiment on survivability:

- **Width**: 5
- **Depth**: 3
- **Descriptive attributes**: [topog, later, diffgr, basisd, PROVINCIE, COROP, PC2, Histological group, metastasized, first_hospital_type, first_hospital_treatment, first_specialism, radio_chemo, reason_diagnostics1, incjr, ses]
- **Target_attributes**:
  y = survival_more_1year $\boldsymbol{X}$ = [leeft, gesl, treatment, stage_main]
- **Quality_measure**: Hellinger distance

After running the beam search algorithm, we find the five most interesting subgroups, shown in Figure 7.4, ordered by the score of the quality measure. We observe that, within all the descriptions that define the five subgroups, there is always at least one attribute set to be equal to "Nan", i.e. a missing value. Furthermore we notice that subgroups 1, 3, 4 are very similar to each other (they all share the double condition `first_hospital_treatment=RADIO, reason_diagnostics1= Nan`). To avoid this redundancy and to find subgroups which are not defined by conditions on

	Subgroup Description	Score	Dimension
1	first_hospital_treatment=RADIO, reason_diagnostics1=Nan, radio_chemo=Nan	0.026549	10929
2	Histological group=Unspecified carcinomas (NOS), radio_chemo=Nan	0.022276	12665
3	first_hospital_treatment=RADIO, reason_diagnostics1=Nan, first_specialism=Nan	0.021656	8350
4	first_hospital_treatment=RADIO, reason_diagnostics1=Nan, diffgr=9	0.019802	9776
5	topog=C340, radio_chemo=Nan	0.019665	8082

Figure 7.4: Top 5 most interesting subgroups retrieved by the beam search.

	Subgroup Description	Score	Dimension
1	first_hospital_treatment=RADIO	0.023586	13027
2	first_specialism=0400 - Longziekten, Histological group=Adenocarcinomas	0.019632	15826
3	Histological group=Unspecified carcinomas (NOS)	0.018596	15048
4	diffgr=2, basisd=7.0	0.017787	12540
5	topog=C340	0.017157	9675

Figure 7.5: Top 5 most interesting subgroups retrieved by the beam search without missing values (Nans) in the description.

```
Start
Level: 1
Total: 218      Fails: 21      Convergence rate: 90.37%

Level: 2
Total: 1231     Fails: 636     Convergence rate: 48.33%

Level: 3
Total: 2231     Fails: 1322    Convergence rate: 40.74%
```

Figure 7.6: Convergence rate with categorical *stage_main*.

```
Start
Level: 1
Total: 218      Fails: 15      Convergence rate: 93.12%

Level: 2
Total: 1237     Fails: 309     Convergence rate: 75.02%

Level: 3
Total: 2247     Fails: 540     Convergence rate: 75.97%
```

Figure 7.7: Convergence rate with numeric *stage_main*.

missing values, we are going to run the same experiment but without considering those subgroups in which at least one descriptor is set to be equal to "Nan".
After running the beam search algorithm again, the five most interesting subgroups are shown in Figure 7.5. The first thing that we notice is that, even though the depth parameter has been set to 3, we do not find subgroups defined by three conditions among the five most interesting ones. This implies that the subgroups defined by three attributes are either not very interesting or very difficult to find because the logistic regression model could not converge, given the smaller size.
To investigate that, we can have a look at the convergence rate shown in Figure 7.6. In the first level of the beam, convergence is reached for 90% of the generated subgroups. On the other hand, when considering the second and the third level, in less than half of the generated subgroups the logistic regression model reaches convergence.
As pointed out in Section 6.2, we can run into the problem of *complete separation*, hence failed convergence, when using logistic regression. One of the most common causes is the presence of a categorical variable with levels characterized by a limited number of observations. In this case the predictor *stage_main* is likely to be the one that causes most of the problems because it has multiple levels, hence fewer observations for each single level. To check the truth of this claim, we can run the beam search again, but in this case we convert *stage_main* into a numeric attribute (we remove the observations with *stage_main*=X). Figure 7.7 shows the convergence rate in this scenario. As we can observe, at a deeper level, the convergence rates are remarkably higher than before (75% vs 40%). However, considering stage as numeric is not correct because the difference between a stage 2 cancer and a stage 1 cancer is not the same as the difference between a stage 3 cancer and a stage 2 cancer. Therefore, we focus on the results found using *stage_main* as a categorical attribute.
Let's have a look at the first subgroup to get some insights and see if it is characterized by interesting features. The subgroup is defined by '`first_hospital_treatment = RADIO`', i.e. the hospital of first treatment is a radiotherapy institution.
Figure 7.8 outlines the summary of the logistic regression model fitted on the subgroup. The AUC

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.4315	0.573	4.243	0.000	1.308	3.555
C(gesl)[T.2]	0.2069	0.046	4.513	0.000	0.117	0.297
C(treatment)[T.True]	0.9742	0.545	1.787	0.074	-0.094	2.043
C(stage_main)[T.2]	-0.9280	0.091	-10.214	0.000	-1.106	-0.750
C(stage_main)[T.3]	-1.7536	0.066	-26.772	0.000	-1.882	-1.625
C(stage_main)[T.4]	-3.4073	0.061	-55.475	0.000	-3.528	-3.287
C(stage_main)[T.X]	-1.0538	0.116	-9.121	0.000	-1.280	-0.827
leeft	-0.0244	0.002	-10.731	0.000	-0.029	-0.020

AUC: 0.825
Average probability subgroup: 0.444 vs Average probability dataset: 0.457
Pseudo R-squ: 0.266

Figure 7.8: Logistic regression model for the subgroup `first_hospital_treatment = RADIO`

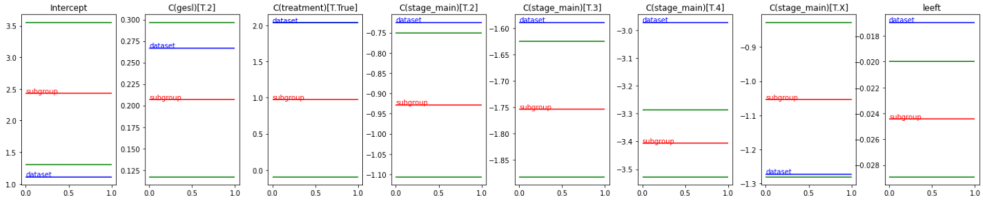


Figure 7.9: Comparison of the dataset and the subgroup coefficients.

(0.825) and the McFadden's R-squared (0.266) are rather high and the average probability of an individual to survive more than one year is very similar to the average probability in the whole dataset (0.444 in the subgroup against 0.457 in the dataset).
Figure 7.9 shows instead a plot of the coefficients. In the plot, for each single predictor, four horizontal lines have been provided: the red one represents the value of the subgroup coefficient, the blue one is instead the value of the dataset coefficient and the green ones are the values of the extremes of the 95% level confidence interval for the subgroup coefficient. A we mentioned in Section 6.3, comparing coefficients across different samples could be misleading because the coefficients might be affected by a different scale. That is why we are going to rely more on the comparison of the average discrete changes, shown in Figure 7.10. For stage 2, we have an average discrete change of -0.1156 for the dataset: this means that in the dataset, on average, a person that goes from stage 1 (the reference level) to stage 2 lowers their probability of surviving more than one year by 11.56% (holding the other predictors at the same value), whereas for the subgroup the decrease of the probability is equal to 15.21 % (ADC = -0.1521). The decrease in both samples is, as expected, larger with a higher stage: on average, a person in the dataset with stage 4 cancer is 47.42% less likely to survive than a patient with exactly the same features but a stage 1 tumor. For the subgroup the decrease is even more emphasized (55.85%).
In general, for `stage_main=2, stage_main=3, stage_main=4` the subgroup ADCs are lower than the corresponding dataset ADCs. The situation is exactly the opposite when we are instead considering `stage_main=X` (-0.2030 in the dataset, -0.1727 in the subgroup). This is also reflected by the histogram in Figure 7.11 that shows the distribution of the response, for each single stage, for both the dataset and the subgroup. The distribution of *survival_more_1year* for the first stage is very similar for both the dataset and the subgroup (85%-15% vs 86%-14%). The proportion of people that survived for more than one year then decreases more rapidly in the subgroup than in the dataset, whereas the distribution is completely swapped when the stage is unknown. These results might suggest that, when the stage of the tumor is known, the stage has a more negative impact on the probability of surviving more than one year for the subgroup, whereas, if it is unknown, the stage has a more negative influence on the probability of surviving more than one year for the dataset. This could be an interesting fact to show to doctors and cancer experts.

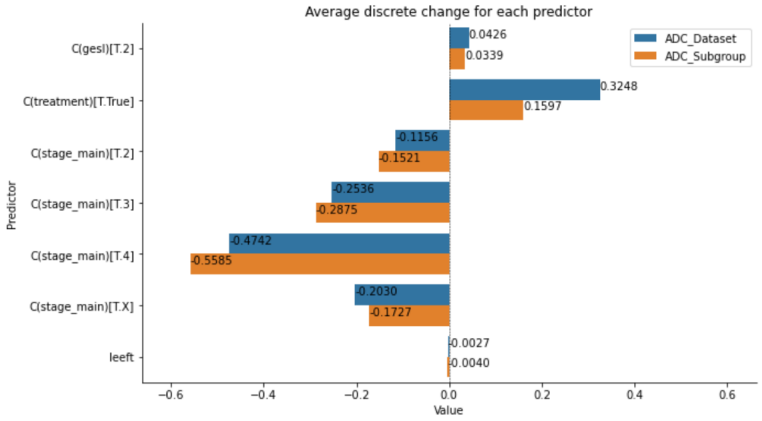


Figure 7.10: Average discrete change for the dataset and the subgroup.

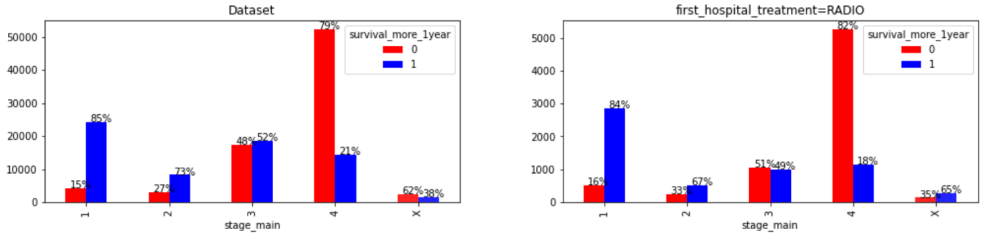


Figure 7.11: Histograms: stage distribution against survival for the dataset and subgroup `first_hospital_treatment = RADIO`.

In the case of age (*leeft*) the ADC in the dataset is -0.0027, indicating that, for a unit increase of age, a patient is on average less likely to survive more than 365 days by 0.27%. In the subgroup, instead, the decrease is more remarkable (0.4%). In other words, we observe a negative impact on survival when there is an increase of age in both situations, but in the subgroup the negative effect seems to be even more accentuated. In Figure 7.12, we can have a look at the distribution of the response with respect to age. It is a specific plot, called violin plot, and it shows the distribution of age, separated according to the response, for both the dataset and the subgroup. The three dashed lines, from bottom to top, represent respectively the first quartile, the median and the third quartile of the distribution of age. In the violin plot relative to the dataset, there is no surprise: we observe that the patients that survived for more than one year are characterized by a lower distribution of age. Regarding the subgroup, both the coefficient and the ADC relative to age are even more negative than the ones of the dataset. Hence, we would expect to have a more emphasized difference in the distribution of age between the people who survived for more than one year and the ones that did not. Surprisingly, this is not the case: the distributions for the two outcomes are almost identical, even with a slightly older population for people with a positive outcome (*survival_more_1year=1*).
This illustrates the power of logistic regression. A simple plot like a histogram, a boxplot or a violin plot can be misleading because it does not take other factors into account. With logistic regression the interpretation is instead the following: in the subgroup, if we consider two patients with the same stage of cancer, the same gender and that both received a treatment (or did not), the older patient is predicted to have a lower chance of surviving more than one year. Therefore, simple plots add extra information but can also be misleading. This highlights a limitation of a model like Exceptional Incidence Distribution Mining, defined in [3], which was based on histograms.
In the end, let's have a look at the treatment coefficient. Compared to the others, it is the only coefficient with a p-value slightly higher then 0.05. If we set the critical value $\alpha = 0.05$, the

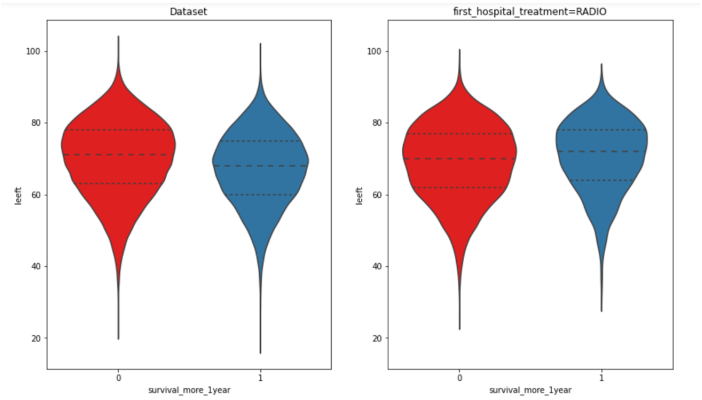


Figure 7.12: Violin plots: age distribution against survival for the dataset and subgroup `first_hospital_treatment = RADIO`.

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.4077      0.179     19.065      0.000       3.057       3.758
C(gesl)[T.2]   0.2059      0.046      4.492      0.000       0.116       0.296
C(stage_main)[T.2]  -0.9303      0.091    -10.242      0.000      -1.108      -0.752
C(stage_main)[T.3]  -1.7545      0.065    -26.792      0.000      -1.883      -1.626
C(stage_main)[T.4]  -3.4067      0.061    -55.473      0.000      -3.527      -3.286
C(stage_main)[T.X]  -1.0520      0.116     -9.107      0.000      -1.278      -0.826
leeft          -0.0245      0.002    -10.752      0.000      -0.029      -0.020
==============================================================================
AUC: 0.825
Average probability: 0.444
Pseudo R-squ: 0.265
```

Figure 7.13: Logistic regression summary for subgroup `first_hospital_treatment = RADIO`, excluding treatment as a predictor.

null hypothesis of the Wald test is not rejected and the coefficient can be regarded as not really significant. This is reflected by the fact that in the subgroup only 23 patients did not receive a treatment against 13004 that did receive a treatment so that there are very few observations for the reference level `treatment=False`. This makes sense if we reflect about the definition of this subgroup, i.e. the hospital of *first treatment* was a radiotherapy institute. The reasons for these 23 patients belonging to the subgroup is that the NCR (National Cancer Registry) guidelines state that a patient can be assigned a first_hospital_treatment even if they have not received any treatment, though it is rather rare.
Since treatment is not significant according to the Wald test, we can drop it and fit the logistic regression model again. By looking at Figure 7.13, we notice the coefficients related to the predictors are basically unaffected by this change. The only difference is in the intercept that shifts from 2.43 to 3.41. Therefore, the considerations on age and stage remain true even when not taking treatment into account.
Let's proceed now with the analysis of the second best subgroup found in this experiment: `first_specialism=0400 - Longziekten, Histological group =Adenocarcinomas`. The first attribute refers to the first specialist visit of the patient after the diagnosis, in this case pneumology. Adenocarcinomas is instead the most frequent subcategory of NSCLC: its origin is in the cells that would normally secrete substances such as mucus. If we look at the fitted logistic regression model in Figure 7.14, we notice that all the coefficients have a p-value lower than 0.05, the AUC is rather high (0.843), as well as the pseudo R-squared (0.303). One difference, with respect to the previous subgroup, is the average probability of a positive response, higher than the one in the dataset (0.534 in the subgroup vs 0.457 in the dataset). If we have a look at Figure

```
==============================================================================================
                            coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                0.8865      0.181      4.888      0.000       0.531       1.242
C(gesl)[T.2]             0.3562      0.040      8.809      0.000       0.277       0.435
C(treatment)[T.True]     2.3033      0.069     33.349      0.000       2.168       2.439
C(stage_main)[T.2]      -0.7188      0.111     -6.485      0.000      -0.936      -0.502
C(stage_main)[T.3]      -1.6013      0.082    -19.415      0.000      -1.763      -1.440
C(stage_main)[T.4]      -3.0096      0.076    -39.426      0.000      -3.159      -2.860
C(stage_main)[T.X]      -1.7527      0.337     -5.197      0.000      -2.414      -1.092
leeft                   -0.0110      0.002     -5.391      0.000      -0.015      -0.007
==============================================================================================
AUC: 0.843
Average probability subgroup: 0.534 vs Average probability dataset: 0.457
Pseudo R-squ: 0.303
```

Figure 7.14: Logistic regression summary for subgroup `first_specialism=0400 - Longziekten, Histological group=Adenocarcinomas`.

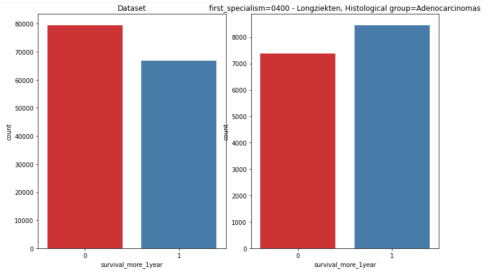


Figure 7.15: Distribution of survival_more_1year over the subgroup `first_specialism=0400 - Longziekten, Histological group=Adenocarcinomas`.

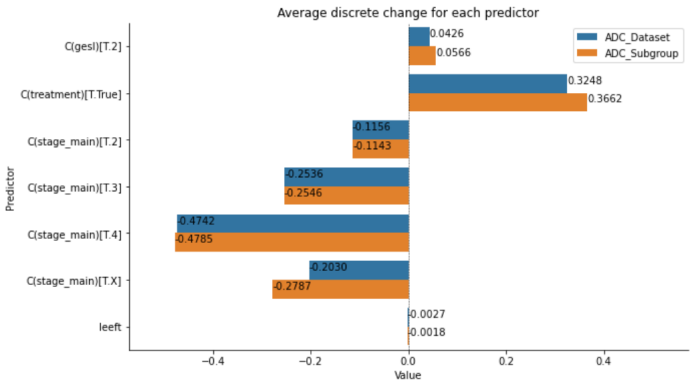


Figure 7.16: Average discrete change for the dataset and the subgroup.

7.15 we notice that the proportion of patients that survived for more than one year is higher than the ones who did not. A possible explanation could be that pneumology is the most accurate specialism for dealing with lung cancer and people that visited other specialists (e.g. cardiology) maybe had other diseases beside lung cancer.
If we look at the average discrete changes in Figure 7.16, we discover that there are other interesting elements regarding this subgroup. The predictors *gesl* and *treatment* are characterized by a more positive ADC in the subgroup than in the dataset, whereas the ADC relative to *leeft* is less negative for the subgroup. By seeing these results, a medical expert could investigate why there is a different distribution in terms of survival in this subgroup and why treatment and gender influence even more positively the probability of surviving for more than one year, whereas age seems to have a more dampened negative influence.

7.1.2 Experiment 2: quality measure via KL-divergence

The second experiment makes use of exactly the same parameters set for the first experiment (the missing values are again not considered part of the description of the subgroup). The only difference is with respect to the quality measure: in this case we are employing KL-divergence multiplied to the entropy function. In Figure 7.17, we can see which are the five most interesting subgroups retrieved by the algorithm. If we compare these findings with the ones discovered using Hellinger distance, we definitely notice some similarities. The most interesting subgroup is still

	Subgroup Description	Score	Dimension
1	first_hospital_treatment=RADIO	0.008237	13027
2	diffgr=2, Histological group=Adenocarcinomas, basisd=7.0	0.005680	5197
3	incjr>2017.0, incjr<=2018.0, Histological group=Adenocarcinomas	0.005661	5662
4	topog=C340	0.005464	9675
5	diffgr=2, basisd=7.0, first_hospital_treatment=STZ	0.004680	6739

Figure 7.17: Top 5 interesting subgroups retrieved using KL-divergence.

```
Start
Level: 1
Total: 218     Fails: 21      Convergence rate: 90.37%

Level: 2
Total: 1228    Fails: 704     Convergence rate: 42.67%

Level: 3
Total: 2219    Fails: 1488    Convergence rate: 32.94%
```

Figure 7.18: Convergence rate, second experiment.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.2817	0.437	2.932	0.003	0.425	2.139
C(gesl)[T.2]	0.4046	0.094	4.295	0.000	0.220	0.589
C(treatment)[T.True]	2.6248	0.234	11.232	0.000	2.167	3.083
C(stage_main)[T.2]	-0.7941	0.141	-5.633	0.000	-1.070	-0.518
C(stage_main)[T.3]	-1.5802	0.126	-12.548	0.000	-1.827	-1.333
C(stage_main)[T.4]	-2.9709	0.123	-24.083	0.000	-3.213	-2.729
C(stage_main)[T.X]	-2.6713	1.010	-2.644	0.008	-4.651	-0.691
leeft	-0.0165	0.005	-3.310	0.001	-0.026	-0.007

AUC: 0.832
Average probability subgroup: 0.845 vs Average probability dataset: 0.457
Pseudo R-squ: 0.271

Figure 7.19: Logistic regression model for the subgroup `diffgr=2, Histological group=Adenocarcinomas, basisd=7.`

`first_hospital_treatment=RADIO` and there is also another identical subgroup: `topog=C340`. Regarding the others, we observe something peculiar: in this experiment we were able to retrieve subgroups defined by three conditions on the descriptors.
The convergence rates shown in Figure 7.18 are also very similar to the ones of the first experiment. They show that the logistic regression reaches less often convergence when exploring smaller subgroups, defined by more conditions on the descriptors.
Since the analysis of the first subgroup would be exactly the same, let's consider the second most interesting one. In this case we are including the patients whose tumor is of the type Adenocarcinoma, where the cells have a moderately low differentiation grade (diffgr=2) and where there has been a histological confirmation of the primary tumor (basisd=7).
The fitted logistic regression model showed in Figure 7.19 has acceptable values for AUC (0.832) and pseudo R-squared (0.271) and all the predictors included are significant according to the Wald test. An element that really stands out is the average probability of a positive response in the subgroup: 0.845. This is even more clear if we have a look at the histogram in Figure 7.20. The proportion of people that survived for more than one year in this subgroup is really high.
We can then analyze more in depth the relationship between the predictors and the response by having a look at the coefficients. Figure 7.21 shows the comparison of the dataset with the subgroup coefficients. Let's remember that the dataset has exactly the same coefficients that were

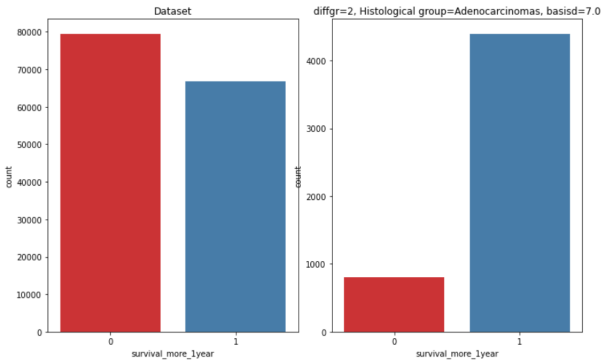


Figure 7.20: Distribution of survival_more_1year over the subgroup `diffgr=2, Histological group=Adenocarcinomas, basisd=7`.

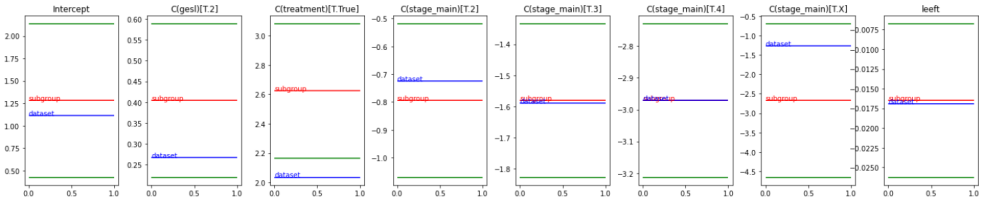


Figure 7.21: Comparison of the dataset and the subgroup coefficients.

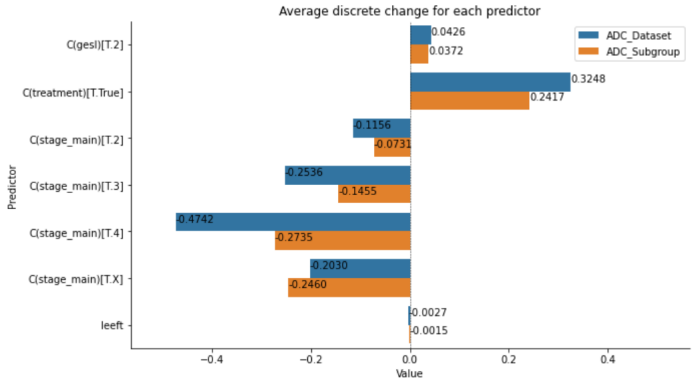


Figure 7.22: Average discrete change for the dataset and the subgroup.

showed in the first experiment.
In this case, treatment looks to be the interesting predictor with a higher value in the subgroup (2.6248), rather than in the dataset (2.0353). That could mean that when treatment=True there is even a higher positive effect on survival, but this could also be due to a different scale of the coefficients in the subgroup and the dataset, therefore it is better to have a look at the average discrete changes.
Figure 7.22 shows the average discrete changes of each predictor for both the dataset and the subgroup. Regarding treatment, in the subgroup we have that on average a person that receives at least a treatment increases their probability of surviving more than one year by 24.17% (ADC=0.2417). This increase is even more accentuated in the dataset (32.48% with ADC=0.3248) and does not reflect what we noticed before by having a look at the coefficients, i.e. that the treatment seemed to have a more positive effect in the subgroup. This fact shows the limitation of the comparison of the coefficient estimates (there could be a different residual variance in the two

	Subgroup Description	Score	Dimension
1	basisd=7.0	1.0	74266
2	first_hospital_treatment=ALGZK, first_hospital_type=ALGZK	1.0	29866
3	Histological group=Unspecified types of cancer, basisd=2.0, diffgr=9	1.0	20537
4	Histological group=Unspecified types of cancer, diffgr=9, first_hospital_treatment=STZ	1.0	6116
5	basisd=2.0, diffgr=9, first_hospital_treatment=STZ	1.0	603

Figure 7.23: Top 5 interesting subgroups retrieved using the quality measure defined in Equation (4.2).

```
Start
Level: 1
Total: 218      Fails: 18       Convergence rate: 91.74%

Level: 2
Total: 1259     Fails: 279      Convergence rate: 77.84%

Level: 3
Total: 2280     Fails: 477      Convergence rate: 79.08%
```

Figure 7.24: Convergence rate, third experiment.

samples and this could affect the scale of the coefficients themselves).
The same phenomenon can be observed if we analyze the predictor *stage_main*. In Figure 7.21 we notice that the coefficients relative to the levels of the tumor stage are more or less the same for the dataset and the subgroup. On the other hand, Figure 7.22 shows that the ADCs are very different: the ADCs of the dataset are almost double the ADCs of the subgroup for stage 2, 3, 4, whereas the ADC for stage=X is slightly less negative for the dataset with respect to the subgroup (-0.2030 against -0.2460).
To summarize, this subgroup can be deemed interesting because of the different effects, between the subgroup and the dataset, that age and stage have on the probability of surviving more than one year.

7.1.3 Experiment 3: quality measure via statistical tests

In the third experiment we are still using the same target attributes, depth and width but the quality measure is defined via the p-values by the formula (4.2). Another remarkable difference is that, with this quality measure, we compare the subgroup to the complement instead of the whole dataset. Figure 7.23 shows the five most interesting subgroups retrieved by the algorithm. The subgroups are different from the ones found with the other quality measures but with almost the same descriptive attributes. We have again `first_hospital_treatment`, `basisd`, `diffgr`, `Histological group`) as the main attributes used to define the most interesting subgroups. Additionally, all five subgroups reach the maximum possible score on the quality measure: 1. In other words, there is at least one predictor whose coefficient is significantly different in the subgroup and in the complement, according to the Wald test.
Figure 7.24 shows the convergence rates. In this case they are quite high also at a deeper level of the beam. This can be explained by the fact that we always fit the model on the whole dataset instead of on the single subgroup as in the two previous experiments. Since we have more observations, the risk of complete separation is reduced, hence convergence is reached more often.
Let's analyze the most interesting subgroup according to the algorithm: `basisd = 7.0`, i.e. there has been a histological confirmation of the primary tumor. Figure 7.25 shows the logistic regression fitted on the subgroup and the complement. The attribute *is_in_subgroup* is the variable defined in (4.1): it is a binary attribute that indicates whether we are in the subgroup (*is_in_subgroup*=1) or not (*is_in_subgroup*=0).
The coefficients defined by the interaction term with *is_in_subgroup*, in Figure 7.25, are the most interesting ones because they inform us, for every predictor, if there is a statistically significant difference between the coefficient of the complement and the one of the subgroup. For example, *C(treatment)[T.True]*= 1.8886 indicates that in the complement there is an increase of 1.8886 for the logit of the probability of *survival_more_1year=1* when a patient receives a treatment. The coefficient *is_in_subgroup:C(treatment)[T.True]* is equal to 0.3409 and has a p-value < 0.05, indicating a significant difference in terms of treatment for the two samples. The coefficient relative to treatment in the subgroup is given by the sum of these 2 coefficients, i.e. 1.8886+0.3409=2.2295. This difference is also reflected by the average discrete changes, shown in Figure 7.26. The ADC for treatment relative to the complement is 0.2978, whereas the corresponding ADC in the subgroup is 0.3578. Both these findings seem to suggest that if a patient receives a treatment in the subgroup, their chance of surviving for more than one year is increased more than in the rest of the dataset.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.8281	0.082	10.098	0.000	0.667	0.989
C(gesl)[T.2]	0.2887	0.019	14.967	0.000	0.251	0.326
C(treatment)[T.True]	1.8886	0.024	78.176	0.000	1.841	1.936
C(stage_main)[T.2]	-0.9212	0.055	-16.794	0.000	-1.029	-0.814
C(stage_main)[T.3]	-1.5071	0.033	-45.123	0.000	-1.573	-1.442
C(stage_main)[T.4]	-2.9409	0.032	-91.458	0.000	-3.004	-2.878
C(stage_main)[T.X]	-1.2293	0.049	-25.269	0.000	-1.325	-1.134
is_in_subgroup	0.3925	0.115	3.401	0.001	0.166	0.619
is_in_subgroup:C(gesl)[T.2]	-0.0388	0.027	-1.420	0.156	-0.092	0.015
is_in_subgroup:C(treatment)[T.True]	0.3409	0.040	8.473	0.000	0.262	0.420
is_in_subgroup:C(stage_main)[T.2]	0.2313	0.065	3.550	0.000	0.104	0.359
is_in_subgroup:C(stage_main)[T.3]	-0.1085	0.043	-2.501	0.012	-0.194	-0.023
is_in_subgroup:C(stage_main)[T.4]	0.0456	0.042	1.074	0.283	-0.038	0.129
is_in_subgroup:C(stage_main)[T.X]	-0.0842	0.114	-0.736	0.462	-0.309	0.140
leeft	-0.0128	0.001	-13.638	0.000	-0.015	-0.011
is_in_subgroup:leeft	-0.0076	0.001	-5.638	0.000	-0.010	-0.005

AUC: 0.845
Average probability: 0.457
Pseudo R-squ: 0.297

Figure 7.25: Logistic regression summary for subgroup `basisd=7.0`.

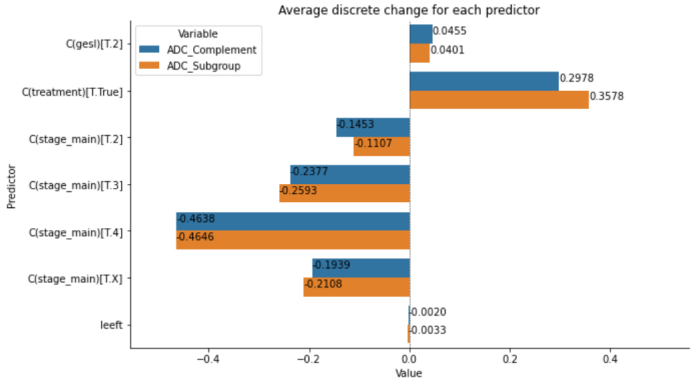


Figure 7.26: Average discrete change for the dataset and the subgroup.

This might turn out to be an interesting insight for the doctors.
Another interesting predictor is age. The coefficient *is_in_subgroup:leeft* is significant and, since it is negative (-0.0076), it shows that in the subgroup the effect of an increase of age has a more negative effect on the logit of the probability of surviving for more than one year. This is also suggested by the ADCs: the ADC in the subgroup is -0.0033, meaning that for a unitary increase of age, a person on average decreases their possibility of survival by 0.33%. In the rest of the dataset the decrease of the probability is smaller (0.2%), since the ADC relative to age is equal to -0.020. These findings suggest that an increase of age has a more negative impact on a 1-year survival in the subgroup than in the rest of the dataset. For the doctors it could be worth investigating on why there is a different effect of age on survival in the subgroup with respect to the rest of the dataset.

```
================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept                8.0419      0.060    134.954      0.000       7.925       8.159
C(gesl)[T.2]            -0.0576      0.014     -4.225      0.000      -0.084      -0.031
C(stage_main)[T.2]      -0.3227      0.036     -9.015      0.000      -0.393      -0.253
C(stage_main)[T.3]      -1.2153      0.024    -49.898      0.000      -1.263      -1.168
C(stage_main)[T.4]      -1.9307      0.022    -86.015      0.000      -1.975      -1.887
C(stage_main)[T.X]      -2.5443      0.041    -62.449      0.000      -2.624      -2.464
leeft                   -0.0822      0.001   -113.467      0.000      -0.084      -0.081
ses                      0.0292      0.002     12.786      0.000       0.025       0.034
================================================================================
AUC: 0.769
Average probability: 0.730
Pseudo R-squ: 0.163
```

Figure 7.27: Logistic regression model for treatment analysis on the whole dataset.

```
Start
Level: 1
Total: 168      Fails: 7        Convergence rate: 95.83%

Level: 2
Total: 844      Fails: 368      Convergence rate: 56.40%

Level: 3
Total: 1520     Fails: 729      Convergence rate: 52.04%
```

Figure 7.28: Convergence rates for treatment analysis.

	Subgroup Description	Score	Dimension
5	incjr>2009.0, incjr<=2011.0	0.014037	20600
1	PROVINCIE=Zuid-Holland	0.010731	30385
4	first_specialism=0400 - Longziekten	0.009515	37149
3	first_hospital_type=UNIVC	0.009097	13763
2	first_specialism=0200 - interne	0.008913	3173

Figure 7.29: Top 5 most interesting subgroups retrieved by the beam search for treatment analysis.

7.2 Logistic regression analysis on treatment

In this section, we propose another experiment using the attribute *treatment* as a response. It is equal to 1 if a patient received at least a treatment after being diagnosed with cancer, 0 otherwise. Following the recommendations of a lung cancer expert, the selected predictors are: *gesl* (gender), *stage_main* (tumor stage), *ses* (socio-economic status), *leeft* (age).
Figure 7.27 shows the summary of the logistic regression model fitted on the whole dataset. The coefficients suggest that being a woman, being older or having a higher stage of tumor have a negative impact on the logit of the probability of receiving a treatment. Additionally, the coefficient relative to an unknown stage of cancer is even more negative than the coefficient relative to a stage 4 tumor (-2.5443 vs -1.9307), meaning that an unknown tumor stage leads to an even lower chance of receiving a treatment, according to the model. It is worth noting that socio-economic status is instead characterized by a positive coefficient (0.0292). This means that if a patient has a higher socio-economic status they have a higher probability of receiving a treatment.
Now we want to find subgroups that behave somehow differently from the norm. In this case we run the beam search algorithm with width=5, depth=3, descriptive attributes=[*PROVINCIE, COROP, PC2, first_specialism, first_hospital_type, reason_diagnostics1, incjr*] and using as quality measure the one defined with Hellinger distance. The response and the predictors are the same that were defined for the whole dataset.
As we can see from Figure 7.28, the convergence rate is much higher at the first level of the beam and is not reached at the second and third level for almost half of the generated subgroups.
Figure 7.29 shows instead the top five interesting subgroups retrieved by the algorithm. Looking at these results, we notice that all of the subgroups are defined by a single condition on the descriptive attributes. We are going to focus our attention on the second most interesting subgroup, i.e. `PROVINCIE=Zuid-Holland`. This is because having a difference in a particular geographical area can be clinically relevant.
In Figure 7.30, we can observe the logistic regression model fitted on the subgroup. All the coefficients are significant according to the Wald test, the average probability for a patient receiving a treatment (0.704) is comparable to the one relative to the whole dataset (0.730) and the AUC (0.766) and the pseudo R-squared (0.160) have reasonable values. To see if there are differences

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	7.8253	0.124	62.892	0.000	7.581	8.069
C(gesl)[T.2]	-0.0729	0.029	-2.533	0.011	-0.129	-0.016
C(stage_main)[T.2]	-0.4175	0.073	-5.753	0.000	-0.560	-0.275
C(stage_main)[T.3]	-1.2911	0.050	-25.871	0.000	-1.389	-1.193
C(stage_main)[T.4]	-1.8162	0.046	-39.413	0.000	-1.907	-1.726
C(stage_main)[T.X]	-2.4295	0.088	-27.622	0.000	-2.602	-2.257
leeft	-0.0827	0.002	-54.242	0.000	-0.086	-0.080
ses	0.0416	0.005	9.087	0.000	0.033	0.051

AUC: 0.766
Average probability subgroup: 0.704 vs Average probability dataset: 0.730
Pseudo R-squ: 0.160

Figure 7.30: Logistic regression model for the subgroup `PROVINCIE=Zuid-Holland`.

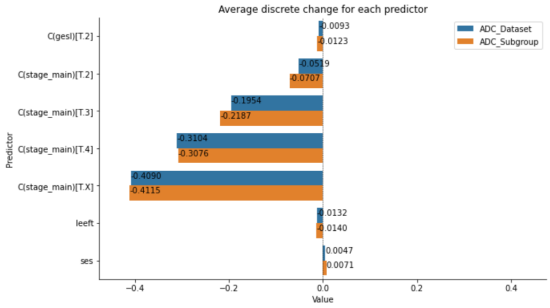


Figure 7.31: Average discrete changes for the dataset and the subgroup.

between the model fitted on the subgroup and the one fitted on the dataset, we can have a look at Figure 7.31 that shows the average discrete changes. The most relevant difference is with respect to the attribute *ses*. For the dataset model, an increase of one unit of *ses* increases on average by 0.47% the probability of receiving a treatment (ADC=0.0047). For the subgroup model instead, a unit increase of *ses* on average leads to an increase of 0.71% of the probability that a patient receives at least a treatment (ADC=0.0071). This fact may be deemed interesting by doctors.

7.3 Considerations on the results of the experiments

In light of the results of these experiments, we can notice what are the advantages and the disadvantages of adopting a logistic regression model in the EMM framework. As we mentioned before, logistic regression can be used to find the relationships between a binary variable and one or more independent variables of any type and this allows to take more factors into account in the same experiment. Simple plots, such as violinplot or histograms, though easy to interpret, could be misleading since they show the relationship between the response and a specific predictor without considering other factors, whereas logistic regression reflects more the complexity of real data.
One could argue that a drawback of using such model is the time needed to analyse the results. Indeed there are several different elements to check for every subgroup to make sure that the model is a good fit: the p-values for the single coefficients, the AUC and the pseudo R-squared. However, the analysis allows to find multiple insights regarding different variables because for every subgroup we have the average distribution of the response, the plot of the coefficients, the plot of the average discrete changes and all these elements contribute to the interpretation of the relationships present in the model.
Regarding survivability, we have found that for some subgroups there was a different relationship between one or more predictors and the response. For example, a more negative effect of the

increase of age on the probability of survival in the subgroup where patients had their first treatment in a radiotherapy institute.
Regarding the quality measures, the one implemented with Hellinger distance and the other with KL-divergence yielded similar results in terms of most interesting subgroups and convergence rates. The main difference was instead with respect to the third quality measure based on the p-values. In this case, convergence was reached more often and the subgroups retrieved were different compared to the previous two experiments. Additionally, this quality measure is focused on a single coefficient rather then all the predictors but this can be seen as a different way to find interesting results rather than a limitation.
It is worth noting that, even though we made use of three different quality measures, the descriptors that defined the most interesting subgroups were approximately the same. Not a single subgroup among the most interesting ones was defined by a condition on the geographical attributes (*PROVINCIE, COROP, PC2*). This fact might suggest that, with respect to survival, there are no significant differences in different areas of the Netherlands, suggesting that the healthcare system is working homogeneously across the country.
Regarding the experiment in the treatment analysis, the retrieved subgroup might be deemed interesting by doctors because the model shows that a higher socio-economic status of a patient has a more positive impact on the probability of receiving a treatment in the province of South Holland, with respect to the whole Netherlands.

Chapter 8

Permutation test

So far we have discussed how to find interesting subgroups in the EMM framework with the aid of logistic regression. In order to explore the numerous possibilities and generate the subgroups, we have implemented the beam search algorithm. However there is a potential pitfall of this algorithm: it will always find something which is deemed 'interesting' because it will always return, as output, the ω most interesting subgroups (ω = width of the beam). Therefore, when looking for exceptional subsets, some of them could be deemed interesting even though they are not (*false discoveries*) and are caused by random artifacts in the data. In our case, we want to understand if the score of the subgroup on the quality measure is statistically different from 0, hence the subgroup is actually interesting. In other words, every time we find a subgroup with a certain score $\bar{\varphi}$ we implicitly perform a statistical test with the null hypothesis $H_0 : \bar{\varphi} = 0$: the further the score is from 0 the more interesting the subgroup is according to the algorithm. However, with the beam search algorithm, we explore a lot of different subgroups and there is the potential risk of rejecting the null hypothesis and considering the subgroup interesting even if it is not. This is the effect of a well known problem called the **Multiple Comparison Problem** [35], which states that, when considering a large number of candidates for a statistical hypothesis, for some of them we will inevitably reject the null hypothesis even when it is true (type I error).
More formally our problem is: given a dataset Ω, a quality measure φ and a possibly interesting subgroup S found through EMM, determine the statistical significance of the subgroup S.
In [36], Gionis et al. propose a technique, called swap randomization, which can be used to assess the results of a data mining technique. In [10], Duivesteijn describes how to apply the swap randomization technique more specifically for the Exceptional Model Mining framework. This technique is part of a broader class of methods called the permutation methods. The rationale behind these methods is that, if we change the labels of the observations (in our case the target attributes) and we consider every possible permutation of the labels, we get the exact distribution under the null hypothesis of the test statistic. Hence we can assess whether the subgroup of interest is statistically interesting or not.
Let's see how this can be applied to our study. Let's suppose that, after running the beam search algorithm, we find that $\bar{S}$ is the most interesting subgroup, i.e. it is characterized by the highest score on the quality measure φ. The subgroup $\bar{S}$ is defined by one or more conditions on the descriptors and these generate a partition in the dataset, i.e. the records that belong to the subgroup and the ones that do not. In order to generate a random permutation, we swap the target attributes across the partition, maintaining both the distribution within the single target column and the dependencies between different targets intact. Let's call this random permutation $\bar{P}$. We have that the target attributes of some observations that belonged to the subgroup are now part of the complement and the opposite is also true. Then, we can assess how interesting the randomly generated subgroup is according to the quality measure φ. We can repeat this process for all the possible permutations. At the end we get the exact distribution of the scores of the quality measure under the null hypothesis. If the score of the original subgroup $\bar{S}$ is extreme enough in this distribution, i.e. the p-value is small enough, we can reject the null hypothesis and regard it

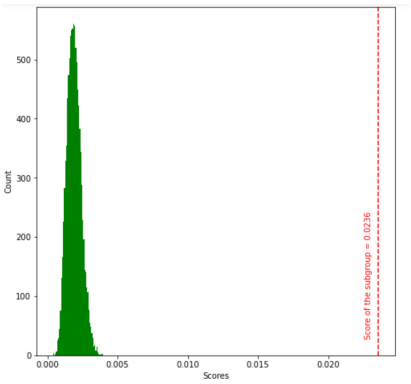


Figure 8.1: Distribution of the scores of the permutations and the subgroup `first_hospital_treatment=RADIO` (quality measure defined with Hellinger distance).

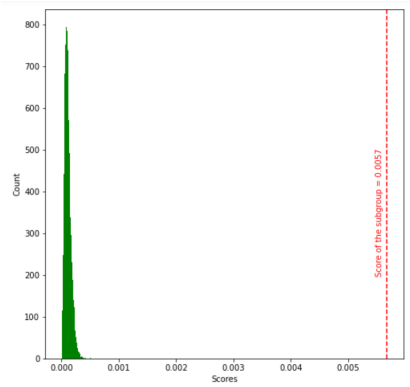


Figure 8.2: Distribution of the scores of the permutations and the subgroup `diffgr=2, Histological group=Adenocarcinomas, basisd=7` (quality measure defined with KL-divergence).

as interesting also from a statistical point of view.
The great advantage of permutation tests is that they do not make any assumption on the distribution of the data (other than the observations being exchangeable under the null hypothesis) and the p-value is the result of simulations rather than formulas of parametric distributions. The disadvantage is mainly computational because there may be too many possible orderings of the data to conveniently allow complete enumeration. Indeed, if N is the number of rows in the dataset and n is the number of observations in the subgroup, the number of possible permutations is equal to $\binom{N}{n} = \frac{N!}{n!\cdot(N-n)!}$. Therefore, the number of possible random permutations can become very large as the size of the dataset increases. A solution to that is to generate the distribution of false discoveries by Monte Carlo sampling, which takes a relatively small random sample of the possible replicates.
The next section exhibits the results of the permutation test applied on some of the subgroups retrieved in the experiments conducted in Chapter 7.

8.1 Experiments

In this section we apply the permutation test to some of the retrieved subgroups in Chapter 7 to see if they are statistically significant.
We start from the subgroup `first_hospital_treatment=RADIO`. In this case we generate 10000 random permutations and assess them using as quality measure the same adopted in the experiment, i.e. the Hellinger distance adjusted with the entropy function. Figure 8.1 shows the distribution of the scores of the random permutations (in green) and the score of the actual subgroup (in red). We can see that the value is very extreme with respect to the others. In this case the simulated p-value would be smaller than $\frac{1}{10000} = 0.0001$, thus indicating that the retrieved subgroup is statistically significant.
Let's consider the subgroup retrieved in the second experiment, using KL-divergence as a quality measure: `diffgr=2`, `Histological group=Adenocarcinomas`, `basisd=7`. Figure 8.2 shows the distribution of the scores of the random permutations of the subgroup, assessed with KL-divergence adjusted with the entropy function, and the score of the actual subgroup. We notice that again the value is very extreme with respect to the others with a p-value smaller than $\frac{1}{10000} = 0.0001$, thus indicating that the retrieved subgroup is statistically significant.
A similar situation occurs with the subgroup `basisd=7.0` in the third experiment as well. Figure 8.3 shows the scores of the permuted subgroups and the one relative to the subgroup. In this case,

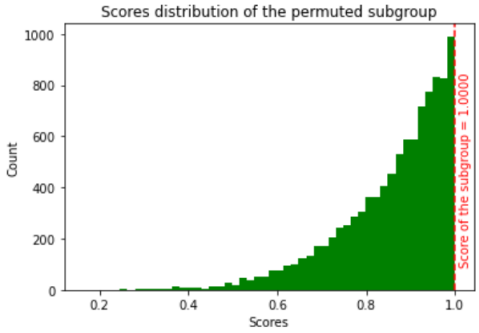


Figure 8.3: Distribution of the scores of the permutations and the subgroup `basisd=7.0` (quality measure defined with statistical approach using p-values).

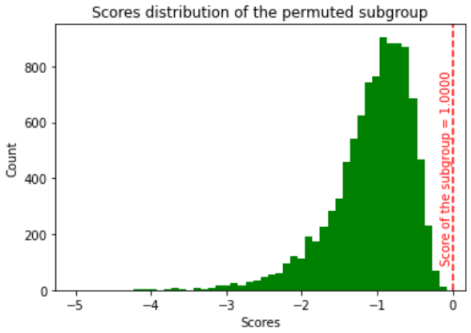


Figure 8.4: Distribution of the scores of the permutations and the subgroup `basisd=7.0` after applying the transformation $log_{10}(1 - score)$.

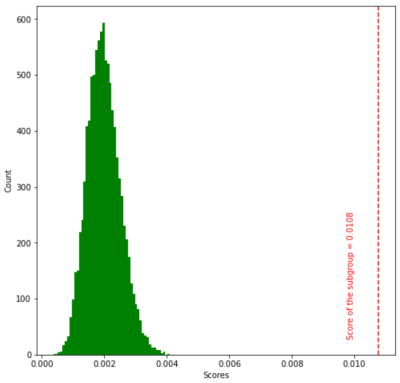


Figure 8.5: Distribution of the scores of the permutations and the subgroup `PROVINCIE=Zuid-Holland` (quality measure defined with Hellinger distance).

we must keep in mind that the highest possible score for the adopted quality measure is 1. To have a clearer view of the subgroup score, Figure 8.4 shows the same scores presented in Figure 8.3 after applying the following transformation: $log_{10}(1 - score)$. It is even more clear from this picture how extreme the quality measure of the subgroup is compared to the rest of the distribution.
Regarding the experiment relative to treatment, Figure 8.5 shows shows the distribution of the scores of the random permutations and the score of the actual subgroup assessed using Hellinger distance. Again 10000 permutations were generated and the quality measure of the subgroup is statistically significant.
We notice that the score of all the subgroups are very extreme compared to the scores of the permutations. A reason could be that the total number of permutations is much larger than the sample that we considered through Monte Carlo simulation. For the four subgroups that we have assessed, there are respectively $\binom{146202}{13027}$, $\binom{146202}{5197}$, $\binom{146202}{74266}$ and $\binom{146202}{30385}$ possible permutations and these numbers are much larger than 10000. The reason why we did not compute all the possible permutations is due to computational time and that is why we relied on Monte Carlo simulation. Despite that, the score of the actual subgroup undoubtedly stands out with respect to the others and that can have another explanation. As Leeuwen et al. point out in [37], one should "expect the unexpected". The authors shows that, when using pattern mining techniques with the aim of finding interesting subgroups (e.g. Subgroup Discovery and Exceptional Model Mining), it is very likely to find subgroups with a high quality measure and so the test for statistical significance could be too lenient.

Chapter 9

Conclusions and future research

The goal of this project was primarily to extract interesting patterns from the cancer registry data handled by IKNL, in order to provide doctors and cancer experts with a new perspective on the Netherlands Cancer Registry. In particular, the focus was on a specific data mining technique called Exceptional Model Mining which aims at finding coherent subsets of the dataset, i.e. subgroups, that behave somehow differently from the norm. This behaviour is captured by a model and the interestingness of the subgroup is assessed according to a quality measure. The great advantage of using EMM is its flexibility given by the possibility of choosing many different models that can offer different insights on the data.
The second goal of this project was to understand whether the retrieved subgroups were interesting not only from a data mining perspective but also from a statistical point of view.
In other words, in this thesis we set out to answer two specific research questions, as outlined in Section 1.2. Here, we answer them in turn.

How to use Exceptional Model Mining combined with statistical theory to extract noteworthy patterns from cancer registry data?

To answer the first research question, we first located EMM in the data mining framework. EMM is closely related to a category of data-driven approaches called supervised descriptive rule discovery [5] which aim at finding surprising patterns that deviate from the norm. The main methods belonging to this category are Contrast Set Mining [6], Emerging Pattern Mining [7] and Subgroup Discovery [8] and EMM [9] can be regarded as a generalization of SD.
In Chapter 2, we provided an overview of some models that could be adopted in the EMM framework and found out that, with respect to cancer registry data, only a model based on absolute frequencies, named Exceptional Distribution Incidence Mining [3], was applied.
The novelty of this Master's thesis project is the implementation of the logistic regression model in the EMM framework: it has been briefly presented in [9], [10], [11] but, to the best of our knowledge, never implemented in a clinical context. Logistic regression can be used to explain the relationships between a binary dependent variable and multiple independent variables of any type. This model exploits the full potential of Exceptional Model Mining because it considers multiple targets at the same time.
The novelty of this thesis can also be found in the implementation of a few quality measures that could be used to assess whether a subgroup is interesting or not, exploiting both statistical tests (Section 4.1) and probability distributions (Section 4.2).
After a preprocessing phase on the dataset (Chapter 5) and an analysis on the beam search algorithm parameters (Chapter 6), we conducted some experiments. Regarding the experiments, it is important to underline that we made use of the average discrete changes to interpret the difference between the logistic regression model fitted on the subgroup and the model fitted on the whole dataset (or on the complement in case we used the quality measure based on the p-values).

We could have also interpreted the results by observing the differences of the coefficients in the two models. However, as Allison points out in [32], the coefficients could be affected by a different scale in two different groups due to unobserved residual variance and that is why average discrete changes have been preferred.
The results in Chapter 7 showed that the logistic regression model is able to find subgroups that behave in a different way in terms of one or multiple predictors with respect to the norm. For example, the subgroup represented by patients that went to a radiotherapy institution as the first hospital of treatment (cf Section 7.1.1). In this case we found out that an older age has on average a more negative impact on the probability of surviving for more than one year compared to the impact it has on a random patient in the whole dataset (Figure 7.10). By just observing the violinplot relative to age and survival_more_1year (Figure 7.12), we would have not been able to extract this piece of information. This showed that the logistic regression model can give a more precise idea about the relationships present in the data by considering multiple factors at the same time. Plots such as violinplots or histograms provide extra information but might also be misleading because they do not take many variables into account. Another noteworthy result was found in the experiment relative to treatment (Section 7.2). In this case, we discovered that there is a more positive effect of a higher socio-economic status of a patient on the probability of receiving a treatment in the province of South-Holland compared to the rest of the dataset. These results might be deemed interesting by doctors and oncologists.

How to quantify the interestingness of a subgroup in the Exceptional Model Mining framework in an objective way using statistics?

To answer the second research question, we decided to implement the permutation test. It is a non-parametric technique that allows to understand if the generated subgroup is statistically interesting or it is instead caused by random artifacts present in the data. The idea is to generate all the possible permutations of a subgroup by swapping the target attributes and then assess them according to the same quality measure used to evaluate the original subgroup. The results presented in Chapter 8 showed that the subgroups analyzed in the experimental phase were also interesting from a statistical point of view. In many cases the quality measure associated to the subgroup was characterized by a remarkably higher score compared to the score of the other permutations. A possible explanation is given in [37], in which the authors claim that, when using pattern mining techniques with the aim of finding interesting subgroups, the probability of finding a completely random subgroup with a large score tends to be very small.
It is important to underline that the permutation test is not restricted to the logistic regression model but could be applied to other models present in the Exceptional Model Mining framework.

9.1 Limitations and future work

The experiments conducted in Chapter 7 showed that there are also some limitations that comes with the adoption of the logistic regression model in the EMM framework: for example, the problem of *complete separation* of the data [31] that makes convergence impossible for certain subgroups. Indeed, the convergence rates outlined in Figures 7.6, 7.18, 7.28 were very high with respect to the first level of the beam (around 90%) and remarkably lower in the second and third level (around 40-50%). In other words, convergence was more difficult to reach in case of more specific and subsequently smaller subgroups. Furthermore, not every model is a good fit for logistic regression: there must be a careful study in the selection of the response and the predictors with a good trade-off between clinical and statistical importance of the selected target attributes.

Regarding the interpretation of the results, a new direction that could be explored is to implement a statistical test which indicates when the average discrete changes from the two samples (subgroup and dataset or subgroup and complement), relative to the same predictor, are significantly different from each other. Mize et al. discuss this topic [38] and propose a solution with the implementation of the method SUEST from the programming language Stata. The idea is to estimate the variances of the two average discrete changes that we want to compare and the covariance between them and then perform a Wald statistical test to verify if the difference is significant or not. A possible future work would be to adapt this approach to the EMM framework. Having a statistical basis that indicates if the difference of the ADCs is significant may help with the interpretation of the results.

Regarding the permutation test, the main limitation is that it is computationally expensive to generate all the permutations and assess them according to the chosen model and quality measure. That is why in Chapter 8 a Monte Carlo simulation has been implemented. The results that we found from applying the permutation test on the subgroups retrieved from the experiments confirm that the subgroups are interesting from a statistical point of view. However, the very extreme scores of the subgroups with respect to the randomly generated permutations seem to suggest that one should "expect the unexpected" when using pattern mining techniques, as underlined in [37]. A possible future direction could be to define alternative statistical tests that are stricter than the permutation test in assessing the statistical significance of the retrieved subgroups.

Bibliography

[1] World Health Organization, "Cancer." Available at https://www.who.int/health-topics/cancer#tab=tab_1. 1, 24

[2] World Health Organization, "Fact sheet - Cancer Netherlands 2018." Available at https://gco.iarc.fr/today/data/factsheets/populations/528-the-netherlands-fact-sheets.pdf. 1

[3] C. Attanasio, "Exceptional Incidence Distribution Mining on a Nationwide Cancer Registry: a Descriptive Approach," Master's thesis, Eindhoven University of Technology, 2019. 2, 9, 31, 41, 54

[4] U. Fayyad, G. Piatseky-Shapiro, P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," 1996. 3

[5] P. Kralj Novak, N. Lavrac, G. I. Webb, "Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining," *Journal of Machine Learning Research*, vol. 10, pp. 377–403, 2009. 3, 54

[6] S.D. Bay, M.J. Pazzani, "Detecting Group Differences: Mining Contrast Sets," *Data Mining and Knowledge Discovery*, vol. 5, pp. 213–246, 2001. 4, 54

[7] G. Dong, J. Li, "Efficient mining of emerging patterns: discovering trends and differences," *fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 43–52, 1999. 4, 54

[8] F. Herrera, C.J. Carmona, P. González, M.J. Del Jesus, "An overview on subgroup discovery: foundations and applications," *Knowl Inf Syst*, vol. 29, pp. 495–525, 2010. 5, 54

[9] W. Duivesteijn, A. Feelders, D. Leman, "Exceptional Model Mining," *Machine Learning and Knowledge Discovery in Databases*, pp. 1–16, 2008. 5, 6, 9, 54

[10] W. Duivesteijn, *Exceptional model mining.* PhD thesis, Leiden University, http://hdl.handle.net/1887/21760, 09 2013. 5, 6, 9, 51, 54

[11] W. Duivesteijn, A. Feelders, A. Knobbe, "Exceptional Model Mining: Supervised descriptive local pattern mining with complex target concepts ," *Data Mining Knowledge Discovery*, vol. 30, pp. 47–98, 2015. 5, 7, 9, 21, 31, 54

[12] S. Moens, M. Boley, "Instant Exceptional Model Mining Using Weighted Controlled Pattern Sampling," *Advances in Intelligent Data Analysis*, vol. xiii, pp. 203–214, 2014. 7

[13] T. E. Krak, A. Feelders, "Exceptional Model Mining with Tree-Constrained Gradient Ascent," *2015 SIAM International Conference on Data Mining*, 2015. 7

[14] L. Downar, W. Duivesteijn, "Exceptionally monotone models—the rank correlation model class for Exceptional Model Mining," *Knowledge and Information System*, vol. 51, pp. 369–394, 2016. 8

[15] W. Duivesteijn, T. Farzami, T. Putman, E. Peer, H.J.P. Weerts, J.N. Adegeest, G. Foks, M. Pechenizkiy, "Have It Both Ways—From A/B Testing to A&B Testing with Exceptional Model Mining," *RECML PKDD 2017*, pp. 114–126, 2017. 8

[16] W. Duivesteijn, A. Feelders, A. Knobbe, "Different Slopes for Different Folks," 2012. 8

[17] D.W.Hosmer, S.Lemeshow, *Applied Logistic Regression.* Wiley Series in Probability and Statistics, John Wiley & Sons, 2000. 10, 18

[18] L. Wasserman, *All of Statistics. A Concise Course in Statistical Inference.* Springer Texts in Statistics, Springer, New York, NY, 2004. 14

[19] F. Abramovich, Y. Ritov, *Statistical Theory A Concise Introduction.* Texts in statistical science, Taylor & Francis Group, 2013. 15

[20] D. McFadden, "Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments," *D. Hensher and P. Stopher (eds.), Behavioural Travel Modelling, Croom Helm*, pp. 279–318, 1977. 17, 37

[21] F. Liese, I. Vajda, "On Divergences and Informations in Statistics and Information Theory," *IEEE Transactions on Information Theory* , vol. 52, pp. 4394–4412, 2006. 18

[22] Jensen G, Ward RD, Balsam PD, "Information: theory, brain, and behavior," *Exp Anal Behav.* , vol. 100, pp. 408–431, 2013. 18

[23] A. Evren, E. Tuna, "On some properties of goodness of fit measures based on statistical entropy," *IJRRAS*, vol. 13, 2012. 18

[24] S. Kullback, R. Leibler, "On information and sufficiency," *Ann. Math. Statist*, vol. 22, pp. 79–86, 1951. 18

[25] P. Harsha, "Lecture notes on communication complexity," September 2011. 19

[26] European Commission, "Nomenclature of territorial units for statistics." Available at `https://ec.europa.eu/eurostat/web/nuts/background`. 25

[27] H. Cramér, *Mathematical methods of statistics.* Princeton mathematical series, Almqvist & Wiksells, 1946. 28

[28] M. S. Bartlett, "Properties of sufficiency and statistical tests.," *Society of London. Series A, Mathematical and Physical Sciences*, vol. 160, pp. 268–282, 1937. 29

[29] W. Bergsma, "A bias-correction for Cramér's V and Tschuprow's T," *Journal of the Korean Statistical Society*, vol. 42, pp. 323–328, 2013. 29

[30] H. Midi, S. Sarkar, and S. Rana, "Collinearity diagnostics of binary logistic regression model," *Journal of Interdisciplinary Mathematics*, vol. 13, pp. 253–267, 2013. 32

[31] P.Allison, "Convergence Failures in Logistic Regression," *SAS Global Forum 2008*, 2008. 33, 55

[32] P. Allison, "Comparing Logit and Probit Coefficients Across Groups," *Sociological Methods and Research*, vol. 28, pp. 186–208, 1999. 34, 55

[33] C. Mood, "Logistic regression: Why we cannot do what we think we can do, and what we can do about it," *European Sociological Review*, vol. 26, pp. 67–82, 2010. 34

[34] J. S. Long, S. Mustillo, "Using predictions and marginal effects to compare groups in regression models for binary outcomes," *Sociological Methods & Research*, 2018. 34

[35] J. Ludbrook, "Multiple comparison procedures updated," *Clinical and Experimental Pharmacology & Physiology*, vol. 25, pp. 1032–1037, 1998. 51

[36] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas, "Assessing data mining results via swap randomization," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, 2007. 51

[37] M. van Leeuwen, A. Ukkonen, "Expect the unexpected - on the significance of subgroups," vol. 9956, pp. 51–66, 2016. 53, 55, 56

[38] T. D. Mize, L. Doan, J. S. Long, "A general framework for comparing predictions and marginal effects across models," *Sociological Methodology*, vol. 49, pp. 152–189, 2019. 56

Appendix A

Dataset attributes

Table A1: Lung cancer topography values (**topog**)

Code	Meaning
C340	Main bronchus
C341	Upper lobe, lung
C342	Middle lobe, lung (right lung only)
C343	Lower lobe, lung
C348	Overlapping lesion of lung
C349	Lung, NOS

Table A2: Differentiation grade classes (**diffgr**)

Code	Meaning
1	Well differentiated
2	Moderately differentiated
3	Poorly differentiated
4	Undifferent, anaplastic
9	Grade or differentiation unknown, not applicable or not determined

Table A3: Social-economic statues values (**ses**)

Range of values	Meaning
1-2-3	Poor
4-5-6-7	Middle class
8-9-10	Rich

Table A4: Basis of diagnosis classes (**basisd**)

Code	Meaning
1	Clinical examination only (case history and physical examination).
2	Clinical diagnostic examinations, exploratory surgery or autopsy (without microscopic confirmation).
4	Specific biochemical and / or immunological laboratory tests.
5	Hematological or cytological confirmation of the primary tumor or metastases, or there is microscopic confirmation but it is unclear whether this is cytology or histology.
6	Histological confirmation only of metastasis, including confirmation in case of autopsy.
7	Histological confirmation of the primary tumor, or unclear or histological confirmation of the primary tumor or a metastasis. And/or autopsy (with histological confirmation).
8	Histological confirmation due to obduction.

Table A5: Tumorsoort classes (**tumorsoort**), with corresponding morphologies (**morf**)

Tumorsoort Code	Meaning	Morphologies
302310	Non-small-cell lung carcinoma	8010-8020, 8022-8035, 8046-8230, 8243-8246, 8250-8576, 8972, 8980-8982, 9110
302320	Small-cell lung carcinoma	8002, 8021, 8041-8045
302330	Carcinoid of the lung	8240-8242, 8248-8249
302340	Other / unspecified lung cancer	8000-8001, 8003-8005, 9990, 8720-8790
302350	Pleuropulmonal blastoma	8973

Table A6: Lateralization classes (**later**)

Code	Meaning
1	Left
2	Right
3	Medial
4	Double sided
X	Unknown

Table A7: Dataset attributes

Attribute	Meaning	Attribute	Meaning
incjr	Incidence year	stage_main	Stage of the cancer
estat	Registration status of the episode	Histological group	Type of tumor cells
gesl	Gender	treatment	Whether the patient received at least a treatment
leeft	Age at incidence	treatment_outside_nl	Whether the patient received at least a treatment outside the Netherlands
topog	Topography	who1	Performance status
later	Lateralisation	first_hospital_type	First hospital the patient came in contact with
morf	Morphology	first_hospital_treatment	Hospital of first treatment
diffgr	Differentiation grade	first_specialism	First hospital ward
tumorsoort	Tumor type	reason_diagnostics1	Diagnosis reasons
basisd	Basis of diagnoses	ses	Social economic status
ct	Clinical tumor stage	radio_chemo	Radiotherapy and chemotherapy concurrent, sequential or separate
cn	Clinical lymphnodes stage	indchemo	Chemotherapy
cm	clinical metastasis stage	indrt	Radiotherapy
pt	patological tumor stage	indhorm	Hormonal therapy
pn	patological lymphnodes stage	indtarget	Targeted therapy
pm	patological metastasis stage	indchorg	Organic Chirurgy
lyond	Number of examined lymphnodes	indchlok	Local Surgery
lypos	Number of positive lymphnodes	indchov	Other Surgery
zid_count	Number of records related to one patient	indchimlok	Local chemo or immunotherapy
inc_month	Incidence month	indimmuno	Immunotherapy
inc_weekday	Incidence day of the week	indchmeta	Surgery aimed at metastasis
follow_up_years	how many years that the patient survived after the first daignoses	indrtmeta	Radiotherapy aimed at metastasis
is_dead	Whether the patient is still alive or not	indoverig	Other treatments
PROVINCIE	Province	indonbek	Unknown treatment
COROP	Region	metastasized	Metastasis
PC2	Last two digits of the postal code		