UNIVERSITAT POLITÈCNICA DE CATALUNYA

FACULTAT D'INFORMÀTICA DE BARCELONA

Master degree course in Innovation and Research in Informatics

Master Degree Thesis

Reducing waiting times and crowding in hospital emergency departments using Machine Learning

Supervisor:
prof. Ricard Gavaldà

Candidate
Silvia Casola

27-06-2018

Summary

Emergency department physicians receive patients in a wide range of conditions and must be able to take sensitive decisions in a small amount of time.

Lack of resources, in the form of medical personnel, diagnostic tools, and beds, as well as improper emergency room access, often translate to high emergency room crowding.

In addition to the lower perceived quality of service, emergency room crowding and, in particular, excessive waiting times, are linked to major risks for the patient health (complications, readmissions, leaves without being seen, greater hospital length of stay etc) and higher mortality rate.

Analyzing the case of the *Consorci Sanitari de Terrassa*, in agreement with which this work has been developed, we propose a methodology to predict hospital admission from the emergency department using machine learning, right after triage.

Being able to anticipate hospitalizations at a so early stage could theoretically allow eliminating waiting times between the physician final decision and the moment the patient is actually admitted in a hospital room.

Starting from a baseline which only uses data produced in the emergency room, this thesis shows that integrating data from the patient's medical history considerably improves the model's predictive power. A first methodology to remove trivial cases is also discussed.

The obtained results show that a prediction could be obtained with a nontrivial accuracy even when using highly interpretable models.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the broader health system, emergency rooms are arguably the single facility which needs to be able to address the major number of diverse events.

Because of their nature, in fact, emergency rooms must deal with patients which arrive without a prior appointment because they experience an acute condition for which they need immediate treatment.

Moreover, emergency rooms are often used by subjects which have no access to other forms of medical assistance and require a consultant in response to unexpected and worrying symptoms.

In this scenario, emergency rooms are often the primary access point to hospitals: it is in the emergency room that the decision whether to visit and treat the patient without a hospitalization or to admit him/her to the hospital is taken.

In the latter case, the patient can directly be admitted to the general hospital connected to the emergency room. However, it might also be the case that the hospital is unable to host the patient, either because it is not featured with the necessary diagnostic and therapeutic tools, or since the patient's condition requires a specific intervention; it might also be the case that, while the hospital could theoretically meet the patient's needs, it is congested and can not receive new incoming patients. In case the general hospital connected to the emergency room cannot admit the patient, he or she needs to be derived to another health centre among the ones available in the area. The patient derivation can be a long procedure, in which several actors must be coordinated.

In all the above cases, once the target hospital has been defined, a set of bureaucratic procedures must be started to register the patient and officially admit him or her.

Several studies have shown that longer waiting times result in greater risk of readmission in less than 30 days, complications and, ultimately, higher mortality.

[20] is a systematic review showing how emergency room crowding positively correlates with mortality both among patients admitted to the hospital and discharged home; an association is also reported with higher rates of patients leaving the emergency room without being seen. The link between waiting time, hospital length of stay and high rate of left without being seen is also reported in [15], where accelerated care at triage is shown to lead in a decrease of the above metrics.

In [28], waiting room time is found to be a strong predictor of perceived compromised care.

In light of these findings, a great effort has been spent by the scientific community in finding ways to reduce waiting time in the emergency room.

Different works have proposed ways of minimizing emergency department crowding and reducing waiting times, for example by anticipating the triage phase, often with the help of computer application, or by serving in parallel different subpopulations of patients.

Moving from these considerations, this work proposes a methodology to minminimize the waiting times between the specialist's decision to admit the patient to the hospital and the actual moment the person is admitted to a hospital room. As anticipated, the time between the two events can be considerable due to managing issues. In order to shorten admitted waiting times, we propose a method to obtain an estimation of the likelihood of admission at a very early stage of the emergency room process, to anticipate any action needed for proper hospitalization.

The project described in this thesis aims to build a model enabling emergencies rooms to estimate the likelihood of admission just after the triage phase, in which the level of severity of the patient's condition is assessed and his symptoms are firstly verified. This phase precedes the specialist consultant, which usually determines, with the help of other diagnostic analysis, whether the patient's status require admission to the hospital.

We are aware that human intervention, in the form of an actual examination by a physician, is far from being unnecessary, especially for critical decisions as the ones described above.

However, knowing with a high level of confidence that a person will later need to be admitted to the hospital can make the procedure to prepare a placing for the patient (as well as any bureaucratic procedure needed) start earlier, highly reducing the waiting time between the doctor decision and the actual admission, while having likely no impact on the patient's well-being and a fixed and likely low cost for the hospital in case of false positives.

In this chapter, we briefly describe our case study and define the goal of the developed project.

1.1 Context

The *Universitat Politècnica de Catalunya* (UPC) has a collaboration agreement with *Consorci Sanitari de Terrassa* (CST) to work on medical domain problems. As part of this agreement, the University receives previously pseudo-anonymized patients' data for the structures in the *Consortium.* In particular, in this project, electronic health records (CMBDs) from five different health facilities will be used. The nature and structure of the data is described in Chapter 4.

The pseudo-anonymization, performed by the *Consorci Sanitari de Terrassa* consists in removing all fields with identifying personal information (e.g., name, identification documents, address etc) and replacing the unique identifier within the Catalan information system with a hashed alphanumeric code. This hashed code allows us to find the same patient across files but does not identify him or her in real life, while theoretically allowing the *Consorci Sanitari de Terrassa* to recover the identity of any particular patient that is deemed 'interesting' by an analysis.

The *Consorci Sanitari de Terrassa* setting is particularly interesting for many reasons. First of all, the data to which we get access come from a number of different medical facilities, including emergency rooms, hospitals, primary care centres, sociosanitary centres and psychological facilities. The joint analysis of this data allows building a broader picture on the patients' situation and previous clinical history. Secondly, all data is managed by the same organization, so that their joint access could practically be possible in practice, while in non-related organizations the problem of integrated access to distributed data is hard to solve.

Thirdly, the *Consorci Sanitari de Terrassa* is a relatively 'closed' system: being the city less subject to touristic flows than bigger cities (e.g, Barcelona), the majority of the people registered in the emergency room data are resident in the area. This means a) it is more likely that the emergency room is accessed for actual urgent situations, and not as an improper way to receive sanitary assistance for people who can not benefit from other services and b) that is more likely that people using the emergency room also have used other medical facilities (e.g., primary care, mental health centres, hospitals, etc.) in the past and that data is available.

1.2 Goal of the Project

The primary goal of this project is to propose a number of models to predict the relative probability of admission to hospital for emergency room patients at triage time, before a proper evaluation of the patient condition, leading to a diagnosis, is carried out by a specialist.

This would allow for waiting times reduction and ultimately lead to a better patient experience and recovery on one hand while making the hospital able to better use its resources on the other hand.

We chose to construct models by excluding diagnostic data (e.g., diagnostic images) and using less expressive but easily available electronic health records. Since the high majority of electronic facilities produce electronic health records, our base model could be used, with minor modifications only, in many real world emergency room facilities; in settings in which data can easily be shared among different health centres, as in the case of the *Consorci Sanitari de Terrassa*, a highly accurate model, integrating other forms of health data could also be deployed.

We require the model at hand to be accurate in its results and interpretable in its internal logic: it is important for healthcare professionals who are not data scientists to understand the basis on which decisions are taken in complex situations.

Starting from a baseline which only uses the emergency rooms' own electronic records, we aim at showing that integrating data from other types of emergency facilities highly improves the quality of the prediction and propose a way to integrate such data.

Finally, we show how the model performs once trivial situations are excluded.

1.3 Structure of the Report

This thesis is structured in the following way: after this brief introduction, in Chapter 2 we present the background of this work, including a brief presentation of the Catalan healthcare system in general and of the emergency room organization and general protocol in particular. We then describe how machine learning can help in health-related problems, and briefly describe the main machine learning algorithms and concepts used in the making of this work.

In Chapter 3 we present a brief review of the literature of machine learning and heath applications, focusing specifically on techniques used for emergency rooms' crowding prevention and managing techniques.

In Chapter 4 we describe the datasets used to build the model, including the main variables and descriptive statistics for such data.

The following chapters describe the models developed, with a focus on the technical side.

We will first present a base model using emergency room data only (Chapter 5), which will later be integrated with data generated by other health facilities (Chapter 6). In Chapter 7 we propose a way to exclude trivial cases from the classification.

Finally, in Chapter 8, we draw our conclusions and present some possible extension of the presented work.

Chapter 2

Background

In this chapter, we briefly present the context for which this model has been developed.

We first focus on the health facilities, briefly describing their management; then, we focus on the emergency room, describing their general organization and protocols. Secondly, we describe some of the areas of health care in which machine learning has contributed. In the last section, we describe the basic machine learning concept that we are going to use in the following chapters.

2.1 Emergency Room Context

2.1.1 Catalan Health Care System Organization

The Catalan health system is centered on a publicly funded system, accessible to all residents of Catalonia.

Generally, the main services offered are:

- Primary attention: the primary attention centres (CAP) are the first health facility a person is supposed to contact in case he or she is suffering a health problem. All other service, in fact, should be accessed via the primary attention centres.

  A special type of primary attention centre, the CUAP (*Centres d'atenció primària, continuada i urgent*), are specifically designed to function continuously through the day and deal with emergency situations.

- Specialized attention and general hospitals (*hospitals d'aguts*): they offer ambulatory specialized medical consultations, hospitalizations for severe illness, surgery and similar services. Urgent situations and emergencies are addressed in the hospitals' Emergency Department.

- Sociosanitary attention: they are centres designed for chronically ill people, or people with disabilities or little autonomy, which need a wide range of continuities assistance, or palliative care.
- Mental health: such centres include ambulatory centres and psychiatric hospitals, among the others.
- Emergency centres: emergencies are dealt with in a number of ways, depending on their nature. Emergency centres include CUAPs and hospitals' emergency departments. Inspected situations are also dealt by phone (ie., the health professionals advise the patient on what to do autonomously or which centre to contact, a doctor is sent to the patient's place or an ambulance is called).

In the following section, we will focus on how emergency rooms are organized and what the main steps are in addressing an incoming patient.

2.1.2 Emergency Room Protocol

The general organization of an emergency room can highly vary across different structures and hospitals.

However, there are some steps which are generally followed.

- The patient arrives at the hospital either by his own mean or by ambulance or helicopter.
- The non-critical patient proceeds to the registration desk where she identifies herself, provides the required documents (including insurances, when necessary) and follows a brief series of bureaucratic procedures.

  This step is skipped in case the patient is in a visibly life-threatening condition, as patients in need of immediate care are sent directly to a treatment area.
- The patient interacts with a health professional (typically a nurse) explaining what symptoms or conditions led to her arrival to the centre.

  The health professional briefly visits the patient and assigns a numeric code defining how urgently the situation must be addressed by a specialist. This operation is called 'triage' and it is of high importance since it determines most of the patient's waiting time. The assigned code ranges from 1 (for highly urgent situations, with an immediate risk for the patient's life) to 5 (for nonurgent situations who can wait). When using an electronic health record, the symptoms described by the patient are encoded in the variable *Motiu* using a numeric code. In the case of our interest, the used code is ICD-9-CM (International Classification of Diseases, 9^{th} revision - Clinical Modification, *CIM-9-MC* in Catalan). [1]

Depending on the situation, some complementary diagnostic proofs can be advised at this stage (for example, X-rays).

- At this point, the patient must wait for a specialist to visit him or her and formulate a final diagnosis or require other diagnostic procedures.

  Patients are often divided into different groups which are treated in parallel and for which the assigned amount of resources depends on the triage level.

  Once a diagnosis is finally formulated, the specialist decides if the patient can be directly discharged, treated in the emergency department and then sent home or if a hospitalization is required.

- In case the patient needs to be admitted to the hospital, if the emergency has been dealt with in an emergency department linked to a general hospital and the hospital provides the technologies needed to address the patient's condition, the patient is often hospitalized there directly. Otherwise, in case the hospital is full or is unable to address the patient's needs, other structures must be contacted to find the proper arrangement.

In this work, we aim at providing a way of anticipating such procedures, in order to reduce the patient's waiting time. Our work shows that it is possible to predict whether a patient will need to be hospitalized with nontrivial confidence just after the triage has been conducted and his or her symptoms have been noted by a health professional.

Having such a prediction, the hospital could start the bureaucratic procedures needed to hospitalize the patient, or contact other health centres if it were not possible to hospitalize the patient in the same structure. This could drastically diminish the waiting time after the specialist's visit, in case the prediction is confirmed.

2.2 Machine Learning in Healtchare

Machine learning defines a set of techniques that use statistics to enable computers to improve their performances on a task with data (ie., 'learn') without being explicitly programmed.

Recently, the improvements in such techniques, as well as the unprecedented amount of publicly available data, have exponentially increased the number of machine-learning-based application in the most diverse fields.

Healthcare is one of such fields.

Machine learning has been applied to solve a wide range of health-related problems. Examples include but are not restricted to:

- Diagnosis, including medical images processing for diseases identification (examples are [8, 22]).
- Robotics surgery [25, 31].
- Drug discovery [30] and drug-drug interaction identification [33], including personalized treatments.
- Epidemic prediction and prevention [9].
- Predict patients' no show [17], hospital readmissions [18], hospital time of stay [16] and similar issues.

Here, we focus specifically on hospitalization management in the emergency room.

Some of the applications listed, such as the ones dealing with the diagnosis of a disease, often rely on data (for example, diagnostic images) that are difficult to obtain.

In this work, we decided to only use electronic health records to develop our model. Though the information contained in these data is generally limited, these data are already produced in a great quantity from hospitals, so that our model could immediately be used, with minor modifications, by any emergency room producing electronic health records.

2.3 Machine Learning Concepts

In this section, we will briefly describe the machine learning techniques used in our work.

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. [26]

Historically, a wide range of algorithms has been developed to enable computers to 'learn'. In this work, we focus on classification algorithms.

A classification algorithm is a supervised learning algorithm designed to assign each input vector to one of a finite number of discrete categories (classes). [11]

Among the many well-known classification algorithms, we mainly focused on decision trees and random forests, as both provide results that can be interpreted to various degrees, unlike for example neural networks.

We say that a classification algorithm is ’interpretable’ when at least part of the process that led to the final classification can be understood by a nontechnical observer.

Following this definition, a classification tree is a very interpretable model, as a nontechnical reader could look at a visual representation of the tree and easily follow the rules applied to split each note from the tree root to the ultimate leaf. As a result, trees can be described as a ’white box’ model: the explanation for any observable situation in the model is easily explained by boolean logic.

Random forests often produce better accuracy than decision trees but are less interpretable, as no intuitive visual representation can be easily obtained. However, it is possible to rank features by the ’importance’ the model assigned to each one, and this is often satisfying for the non-technical user.

In the following, we briefly present those algorithms.

2.3.1 Classification Tree

Tree-based methods partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one. [21]

Different implementations have been proposed. In our work, we use an optimized CART [13] implementation provided by the scikit-learn python library [27], a free, open-source library, widely used for many classical machine learning tasks.

Given a vector of features $x_i \in \mathbb{R}^n$, $i = 1, ..., l$ and a vector of labels $y \in \mathbb{R}^l$, a decision tree recursively partitions the space so that the data points with the same labels are split in partitions in which impurity is minimized.

If one calls Q the data instances at each node m, each split $\theta(j, t_m)$ consists of a feature j and threshold t_m such that Q is split into

$$Q_{left}(\theta) = (x, y) : x_j \leq t_m$$

and

$$Q_{right}(\theta) = (x, y) : x_j > t_m$$

where $Q_{left}(\theta)$ and $Q_{right}(\theta)$ are the data instances in the left and right child node, respectively.

The impurity at node m is computed using some function $G()$, whose choice might vary.

The global cost function for a split is θ

$$H(Q, \theta) = \frac{n_{left}}{N_m} G(Q_{left}(\theta)) + \frac{n_{right}}{N_m} G(Q_{right}(\theta))$$

The best split θ^* is then chosen as the one minimizing the cost function $H(Q, \theta)$.

This procedure is repeated recursively in Q_{left} and Q_{right} until some stopping criterion is met.

The measure of impurity used in this work is the Gini impurity index, defined as

$$G(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

where p_{mk} is the proportion of observations belonging to class k in each region R_m

$$p_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

being N_m the number of observations in R_m and $K = \{0, 1\}$ for a binary classification task.

As a stopping criterion, we define a threshold on the minimum number of data instances Q to split node m. For nodes containing fewer instances than the threshold, the recursive process is stopped.

2.3.2 Random Forest

A random forest [12] is classification algorithm which combines the predictions of several classification trees to improve robustness and avoid overfitting.

Two are the main modification to the algorithm made for each tree in the ensemble: first, each tree is built from bootstrap sample (ie., a sample drawn with replacement) from the training set data; second, the split chosen at each node is the best split among a random subset of all the original features.

Given the set of trees, in the used scikit-learn implementation their probabilistic prediction is then averaged.

The main parameters to define in a random forest are the number of estimators (the larger the better in terms of accuracy, at the cost of a greater training time) and the maximum number of features to be considered at each node.

In this work, we set the number of estimators to 50 and the maximum number of features to $\sqrt{p}$, where p is the original number of features. For each individual tree in the ensemble, a threshold on the minimum number of samples needed to further split the node is set (see 2.3.1); the Gini index is used to compute the impurity measure.

While the nature of a random forest does not make it possible to understand the complete sequence of decisions which lead to the final classification, it is possible to consider the relative importance of each feature in the classification.

The rank is computed by considering at which depth the feature is used in each estimator. Features used near the tree root contribute more to the classification (ie., are more discriminative) than features used near the leaves, as they impact a larger number of samples. Thus, the relative importance of a feature can be computed by considering the fraction of the samples it contributes to. The relative importance of each feature is then averaged among all trees to compute the final rank. See the original description for details [12].

Chapter 3

State of the Art

In this chapter, we describe some of the main scientific production in healthcare as approached under a statistical and computer-science point of view, with a specific attention to machine learning inspired solutions.

As described in section 2.2, the recent scientific contribution has been huge and has ranged in a number of different areas, including disease prevention, diagnostics, therapy and service management.

In the latter area of study, the scientific production focuses on identifying the critical issues impacting the quality of service and their source or in proposing methods to eliminate such issues.

Some of this production focuses on admission prediction, which is directly related to the topic analyzed in our work.

In [24], for example, the author proposes a method to evaluate an individual's risk of being admitted to hospital as an emergency inpatient within a year. The aim of the study is to be able to identify patients at higher risk, to allow the health system to explicitly target such patients with preventive and anticipatory care. The method, called SPARRA (Scottish Patients At Risk of Readmission and Admission), uses a huge quantity of data (4.2 million patients are analyzed) from the Scottish National Health System. A logistic regression-based algorithm is trained using patients' medical history (hospitalizations, pharmaceutic prescriptions, emergency department records, outpatients and psychiatric admissions) during a given period; this information is then linked to whether or not the patient experienced an emergency admission in the following year to compute his or her degree of risk.

While the problem addressed by the paper is closely related to the one analyzed in our work, SPARRA predicts admission in a wide period; our work, on the other hand, tries to use timely data to predict the outcome of each single emergency room contact.

Note, however, how the method used in the paper is a 'white box' which lets the authors identify risk factors in addition to obtaining the prediction itself.

Many of the studies targeting quality of service in the emergency departments have identified crowding as a great reason of concern.

In a systematic review of the English-language literature for the years 1989–2007 [10], emergency room crowding is associated with an increased risk of in-hospital mortality, longer times to treatment for patients suffering pneumonia or acute pain, and a higher probability of leaving the Emergency Department against medical advice or without being seen.

[29] identifies among the commonly proposed solutions of crowding additional personnel, observation units, hospital bed access, nonurgent referrals, ambulance diversion, destination control, crowding measures, and queuing theory.

Note how most of these solutions imply a nontrivial additional cost for the hospital.

Recently, machine learning has increasingly being used to address a number of problems strictly related to crowding and the emergency department in general.

An electronic triage [23] which outperforms the Emergency Severity Index (ESI) in accurately differentiating patients using machine learning was recently proposed by Levin et al.

The study uses a random forest trained using several patient's triage data to predict the need for critical care, emergency procedures, and inpatient hospitalization. The predicted outcomes are then translated to a triage level.

The used data were age, sex, arrival mode, vital signs (temperature, pulse rate, respiratory rate, systolic blood pressure, and oxygen saturation), reason of admission and active conditions documented in the electronic health record.

Even if the central problem the study aims to address is different from the one under consideration in this work, the model developed to predict hospitalizations is conceptually very similar to the one developed in this work; the used data, however, are somehow more specific and are not always easily available from electronic records.

The problem specifically addressed by this work has also recently been studied in [19].

Here, data from two hospitals in Northern Ireland are used to predict the admissions from the emergency department.

Three algorithms were used to build the predictive models: logistic regression, decision trees, and gradient boosted machines (GBM).

The models predicted whether the patient is admitted to hospital based on the hospital site, the date and time of attendance, the age and gender, the arrival model, the care group (a category indicating the pathway a patient should take), the triage and whether the patient had a previous admission to the hospital within the last week, month, or year. No previous medical history data is used.

Although GBM performed best, the paper proposes logistic regression as a candidate for implementation when interpretability is important.

With respect to the reviewed literature, our approach presents many points of contact. First of all, many of the considered variables (e.g., demographic, arrival model, triage, motivation, previous admissions) are consistently used in the literature for such previsions. Many of the reviewed articles also share a special attention for interpretable models.

The main novelty of our approach, on the other hand, consists of a systematic use of the patient past medical history, from different types of health facilities, to perform the prediction.

Chapter 4

Dataset

This chapter describes the datasets used to develop the models.

Five different datasets were available, each referring to a different type of *Conjunt Mínim Bàsic de Dades* (CMBD).

In the following, we first explain what a CMBD is and its main purposes; then, we briefly describe every single dataset, including its structure, the main variables we used in the model and some descriptive statistical information.

4.1 Conjunt Mínim Bàsic de Dades (CMBD)

The *Conjunt Mínim Bàsic de Dades* or CMBD - Catalan for 'Minimal Basic Data Set' is a set of standards for encoding patient and attention information defined in detail by the Catalan Department of Health. The standard comprises six different file formats to cover different types of attention:

- Emergencies (*'Urgències'*)
- Acute Hospitalizations (*'Hospital d'aguts'*)
- Primary Care (*'Atenció primària'*)
- Ambulatory Mental Health
- Socio-sanitary (Long-term, low-intensity hospitals, e.g., for recovery or people who can't make independent life)
- Mental Health Specialized Hospitals

The Acute Hospital level is standardized over all Spain and is in fact an extension of a European standard [32] that sets the minimum that should be gathered in all EU countries. The other file types have only partially implanted in the rest of Spain, and Catalonia is unique in that it has defined all six.

The primary use of the CMBD is to establish a common standard of encoding for the information that the central agency in each region or country needs to collect from all health-care organizations. It is not intended to contain all clinical information collected about a patient or contact, which would be really complicated because of the diversity of software in use. Instead, it defines a number of variables that are easy to extract and encode from any system (say, with a SQL query). The main uses that it currently have are 1) statistical studies, for example for planning the deployment of new infrastructures and 2) billing, namely, a hospital uses CMBD to report all the actions it has performed on patients and gets billed accordingly by the central funding agency (in the public or publicly-funded systems).

4.2 Emergency Centre Dataset

The central dataset used in the making of this work is a database of electronic records produced in the emergency rooms, as defined in the CMBD-UR standard. [4]

Each row of the dataset represents a contact, defined as any type of assistance given to a patient by any of the emergency resources. These include hospitals' emergency services, specialized primary attention centres (*dispositius d'atenció primària d'alta resolució*) and the System of Medical Emergencies (SEM), which deals with patients either by phone or outside of conventional medical centres (public spaces, the patient's house etc).

In the following, we briefly describe the variables we took into account to develop our model. For a more comprehensive description of variables, please refer to the official CMBD-UR manual.

We uniquely identify patients by their *Historia* (History), which is a patient id previously pseudo-anonymized by the hospital.

Some of the patient's personal data are recorded: in particular, the patient Date of birth (*D_naix*) and Sex (*Sexe*), encoded as 0 for males, and 1 for females.

Information on the patient's residence includes his or her City (*Muni*), District (*Distr*), and Country (*Pais*), all encoded using numeric codes. The variable *Pais* represents the country of residence, while *Pais_orig* is the country of birth.

The variable *T_act* (Type of activity) refers to the place of intervention and, in our dataset, only assumes values corresponding to the ER centre or to the patient

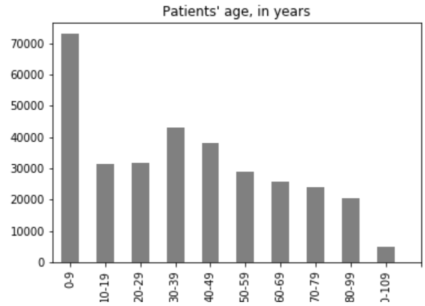

Figure 4.1: Age distribution in the CMBD-UR

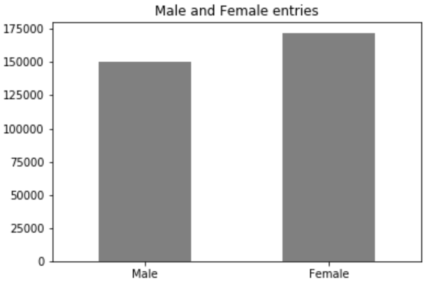

Figure 4.2: Male and female patients in the CMBD-UR

own place. The first value is much more common than the second one, as figure 4.3 shows.

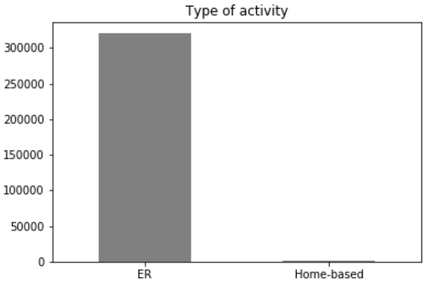

Figure 4.3: Type of activity in the CMBD-UR

The variable *Iniciativa* (Initiative) encodes the entity that took the initiative to bring the person to the emergency centre. It is encoded using five categories: the patient's own initiative (1), the legal guardian initiative (2), medical initiative (3, if the patients arrive from another medical centre), authority initiative (4, e.g. the police) or a juridical order (5).

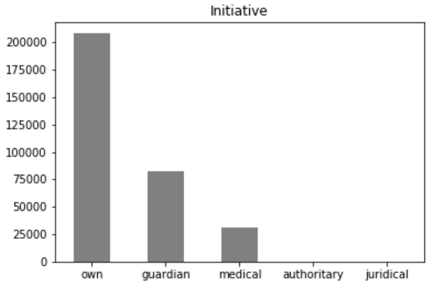

Figure 4.4: Initiative in the CMBD-UR

'Mitja_a' (Mean of transport) refers to the means of arrival to the emergency room; it can refer to the patient's own means (1, car, public transportation etc), the ambulance (2), the helicopter (3) or other means.

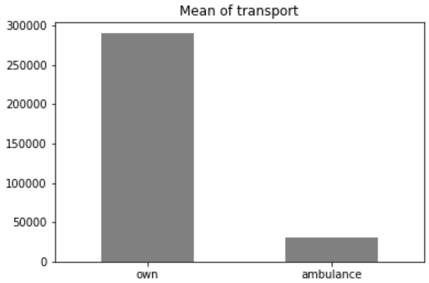

Figure 4.5: Mean of transport in the CMBD-UR

Where the patient was before coming to ER is recorded in the variable *Precedencia* (Precedence). The value 1 represents the patient's house or any other private or public place which is not a general or psychiatric hospital (2), a socio-sanitary health centre (3), a mental health centre (4), a public or private primary attention centre (5); the value 6 is used if the patient was in a home-hospitalization regime, while 7 means that the patient arrived following an external consult of the same hospital.

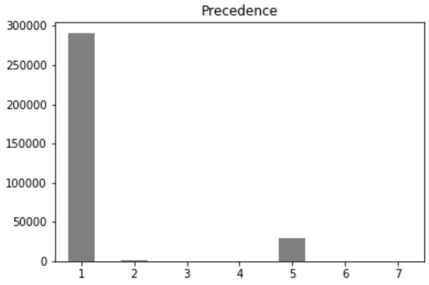

Figure 4.6: Precedence in the CMBD-UR

The variable *Pr_ungencia* identifies in which emergency centre (if any) the patient was prior to his or her arrival to the hospital (the value is 0 if the patient does not come from another emergency centre).

The urgency code which was activated is registered in the variable *Code_urg*, which defines the protocol (where 0 indicates that no particular protocol was activated) used to face a given situation. Example include ictus, heart attack, sepsis etc.

The day and time of the start of the contact and of its finish are recorded in Starting Date, Starting hour and End Date, End Time, respectively.
Our dataset includes contacts from 31-12-2014 to 03-12-2017.

In addition, the dataset includes important diagnostic and procedural variables.

The variable *Triatge* (Triage) refers to the triage level of the patients, which represent his or her degree of priority. The triage code is represented by a number from 1 to 5 (where 1 means 'resuscitation needed, with life-threatening conditions (immediate attention needed)' and 5 means 'not urgent situation, which can be programmed, with no risk for the patient'.
The variable is NULL in the 0.6% of cases.
Date and time of the triage are recorded in the variables *Date of triage* and *Hour of triage.*

The variable *Motiu* (Reason) refers to what motivated the patients to go to the centre, as described in the previous chapter.
The variable is encoded using ICD-9 codes. In this representation, the first three digits represent the more general pathology or condition, which could further

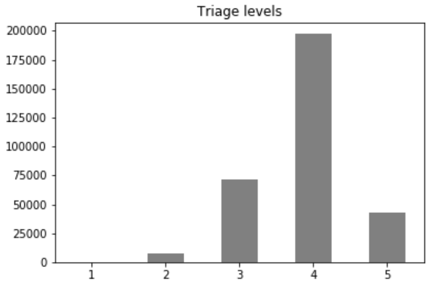

Figure 4.7: Triage levels in the CMBD-UR

be specified by using up to two other digits. For example, the code 300 indicates ’Anxiety, dissociative and somatoform disorders’, while 300.2 refers to the subclass of ’phobic disorders’, further specifying that 300.21 is ’agoraphobia with panic disorders’.

The variable assumes 269 different values in our dataset, once preprocessed keeping the first three digits only, and is NULL in the 0.5% of the cases.

The main diagnosis, as well as the secondary diagnostics, are also recorded.

Other information includes the external causes leading to the condition and all principal and secondary proceedings.

Finally, the variable *S_alta* (Exit situation) includes information on the patient’s situation when leaving the hospital.

It can assume the following values: exit with post-triage derivation (0); exit (1); admission to the same centre (2); exit with post-assistance derivation (3); voluntary exit (4); evasion or administrative discharge (5); death (6); home hospitalization (7); arrival with cardiorespiratory arrest (8).

Values 1 and 3 refers to situations in which the patient is sent to another medical centre for various reasons, either immediately after triage or after a visit from a health professional.

Value 8 identifies people who arrive in the emergency room with an irreversible cardiorespiratory arrest which led to death before arriving at the centre. We removed these patients from our analysis.

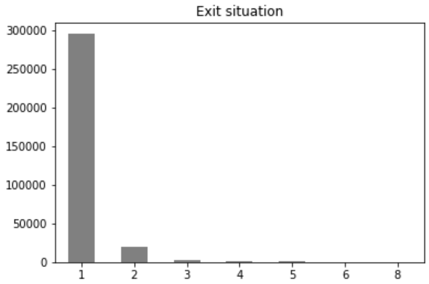

Figure 4.8: ER exit situation in the CMBD-UR

4.3 Hospital Dataset

The second dataset collects entries from the CMBD-HA (*hospital d'agut*), which contains information about conventional hospitalizations, major and minor surgeries, day-hospitals (*hospital de día*) and home hospitalizations.

In the following, we briefly describe the main variables. A more comprehensive description of variables can be found in the official CMBD-HA manual [3].

In addition to the unique patient identifier *Historia* and the patient's personal data, the dataset includes the hospitalization start date and time and the end date and time.

The variable *C_ingres* (circumstance of entry) defines if the visit was programmed or not, while *C_alta* describes the exit circumstance.

The type of activity is encoded in the variable *T_act* which assumes three values: conventional hospitalization, major ambulatory surgery and day-hospital (among the possible ones).

The variable *T_visita* describes the type of visits (initial visit or a follow-up).

The diagnosis, the procedures, and similar information are also included, with a similar interpretation than the one described in section 4.2.

Moreover, some specific diagnostic variables are included: *UCI* defines if the patient needed intensive care, with other variables recording its starting and ending date and time.

The variable *F_Apache* (Acute Physiology and Chronic Health Evaluation II) describes the grade obtained by the patient in the Apache II test, which evaluates a set of physiological parameters of the patient in the first 24 hours of intensive care.

A_funcional, *A_cognitiu* and *A_social* describe the result of functional, cognitive and social tests, respectively, as performed when the patient leaves the hospital.

4.4 Primary Care Dataset

This dataset contained entries from the CMBD-AP [6].

The basic unit of the register is the visit, which is considered as any existential activity performed for a patient by any professional in a CAP or a continued attention centre and which generates an entry in the patient's clinical history.

In our analysis, we consider data gathered between 2015 and 2017. Since the format of the transmitted data slightly changed in 2016, we actually obtained two different datasets.

Each entry contains the *Historia* and the personal data of the patient, as well as the starting and ending dates and times. The variable *C_contacte* describes if the visit was previously programmed or not.

The variable *T_act* describes the type of activity involved, namely if the activity was carried out in a health centre, at the patient's house (or socio-sanitary centre), by phone or electronically.

The variable *Programa* is either missing or has value *ATDOM* (domiciliary attention).

The field *Derivació* encodes if the patient was sent to another health centre or specialist, and its type (specialist; CUAP or emergency room; socio-sanitary unit; mental health centre; drug addiction centre; sex-health centre; work-health centre).

An interesting variable is *Tprof*, which encodes the main professional figure which conducted the visit: family doctor (1), pediatrician (2), dentist (3), nurse (4), social worker (5).

Other variables include the principal diagnostic and when it was first diagnosed, secondary diagnostics, performed procedures etc.

4.5 Mental Heath Dataset

Another dataset taken into account was the CMBD-CSM (*Conjunt Mínim Bàsic de Dades dels Centres de salut mental ambulatòria*) [2]

In addition to the identifier (*Historia*) and personal data, the dataset contains information on the health centre, if any, the patient was prior to admission (*Pr_ingres*), the circumstance of the admission (*C_ingres*, namely if the visit was programmed or not), the exit state *C_alta*, and the classical diagnostical variables recording principal and secondary diagnosis, external causes and procedures.

Moreover, the CMBD contains some variables which are specific for the ambulatory health centres.

The variable PV describe if the entry corresponds to a first visit of the patient with a psychologist or psychiatrist or to a follow-up; in that case, the variable SV contains the number of visits the patient had during the notification period, while the variable PC contains the number of complementary proofs for a specific evaluation in the same period.

The notification period corresponds to a natural trimester (January-March, April-June, July-September, October-December). As a result, information has a lower granularity than the one contained in the previously CMBDs, where each entry corresponded to a single contact.

The number individual (*PI*), group (*PG*) and family (*PF*) psychotherapy session in the notification period is included.

In addition, the dataset contains number of times a patient required the attention of a nurse (*AI*), of a social worker (*TS*) or needed home-based attention (*VD*); the number of non-programmed visits is contained in *VNP*.

The variable *TMS* (Severe mental disorder) indicates if the patient carries a pathology associated with a severe mental disorder (eg. schizophrenia, major depression, severe paranoia, autism etc), while *Prg_TMS* indicates if the patient follows a specific treatment program for his or her severe mental disorder.

4.6 Sociosanitary Dataset

The last dataset taken into account corresponds to the entries of a CMBD-RSS (*Conjunt Mínim Bàsic de Dades dels recursos sociosanitaris*) [5].

The basic unit of the register is the episode. This includes the hospitalization, ambulatory visits, and home care. Data is registered at the entrance, periodically, if the patient's condition drastically changes and at exit (practices vary depending on the type of service the patient needs).

Many different subcategories of assistance (*Codi_ambit*) are theoretically possible. In our case, the only two possible values are *SLE* (*Sociosanitaris llarga estada*), for long-stay and *SPA* (*Sociosanitaris pal · liatives i altres*), for palliative care.

The type of activity refers to hospitalization (21*), home-based care (22*) or ambulatory care (23*).

The date in which the entry was inserted is in the variable *D_valor* in case of SLE records; in case this variable is not present, we refer to the date the data was transmitted to estimate it.

In addition to the common diagnostic variables, the CMBD also contains specific variables.

They are related to the cognitive state (e.g. *Coma*, *Memòria recent*), to the communication capability, the mood and the behavioural patterns, the physical autonomy, the capability of continence, some commonly diagnosed diseases (e.g. *Diabetis mellitus*) or other health problems (e.g. *Febre*); other variables are related to the nutrition, the skin, the state of day-time wake, the pharmacy administration, treatments and special procedures (e.g. *Quimioteràpia*), or other therapies.

In our analysis, we only considered a subset of these variable, sometimes aggregating similar variables.

Chapter 5

Basic Predictive Model

In this chapter, we propose a model to predict a patient's probability of admission to hospital following an emergency contact, developed by using the previously described CMBD-UR dataset only.

5.1 Model Definition

For each contact in the CMBD-UR dataset, given some of the variables available right after the triage assessment, we want to predict the probability that the contact will result in a hospital admission, using machine learning.

We suppose the following variables are theoretically available: all the patient's personal variable (sex, date of birth, geographical information); who took the initiative to bring the patient to the emergency room (*Iniciativa*); the mean of transport used (*Mitja_arr*); if the patient comes from another health centre (and its type) or not (*Precedencia*); if the patient comes from another emergency centre and its type (*Pr_urg*); the triage level; the symptoms described in the patient's (*Motiu*). We also take into account if the assistance has been carried out directly in the emergency room or elsewhere (type of activity).

We model the underlying problem as a supervised classification problem, in which the previously described variables play the role of features and the exit status, *C_alta*, plays the role of label. The label will be binarized, to reduce the classification problem to an admitted (1) versus discharged (0) binary classification. The specific exit status (e.g, derivation to another hospital), depends on several factors (for example the number of free beds in the hospital) and is not interesting for our purposes.

The main aim of the model is to output a probability of admission.

To do so, instead of simply considering the class in which the entry has been classified, we retrieve the probability that the entry belongs to the positive class (admitted patients).

Since the final decision on the admission depends on numerous, various and not easily predictable factors (for example, hospitals tend to be more congested in given periods of the year), in fact, we want to obtain not only a binary label indicating admission or discharge but also the probability of such admission.

Moreover, having a rank of probabilities could theoretically allow the hospital to estimate the global cost of admission and discharges depending on how many patients are admitted, and find the optimal threshold which maximizes the patient's well-being while minimizing costs.

Among the requirements for the model, interpretability played an important role, as anticipated.

Knowing the main factors or - even better - the whole sequence of rules by which a decision is made can lead the user to a better confidence about model reliability and helps to highlight bias. Moreover, it could help in gaining an insight into the main 'factors of risk' which determine admission from the emergency department.

5.2 Data Preprocessing

In order to obtain the required model, a set of proprocessing step was needed.

Data with unknown target were removed, as well as cases for which the exit state indicated that they arrived at the emergency with a cardiorespiratory arrest (so that no other intervention was possible).

After this action, the dataset contained 321691 rows.

Moreover, the dates of birth were transformed into a more intuitive age representation.

We decided to remove any subclass of the variable *Motiu* and keep only a 3-digit representation. This decision was made in order not to further increase the number of distinct categories and avoid sparsity.

Moreover, three new variables were manually created.

The first one, (*#ER_urg*), is a counter of the number of the previous contact with the emergency room the patient had in the past. In practice, given a patient code, his or her records are put in chronological order and the counter is incremented for each new record (having the first record in time *#ER_urg* $= 1$).

The second one, *#H_urg* corresponds to the number of times a patient went to the ER and was later admitted to the hospital, as suggested by his exit status, in the past. In this case, the first time a patient gets admitted to the hospital, the counter is not increased to 1, to avoid to implicitly suggest the label (otherwise, all cases with *#ER* $= 0$ could be automatically classified as non-hospitalized).

The third new variable, $\#H\_norm\_urg$ is a normalized version of the second one, obtained as $\#H\_norm\_urg = \frac{\#H\_urg}{\#ER\_urg}$.

In order to train the model on binary classified data, the variable indicating the exit state was binarized as follows: patients with label 1, 4 and 5 were considered non-entry (meaning they were not admitted to hospital), and mapped to the value 0, while all the other categories were mapped to 1, meaning they were admitted to the hospital.

This lead to a very unbalanced situation, as shown in figure 5.1. Only 7.4% of the contacts, in fact, resulted in a hospital admission.

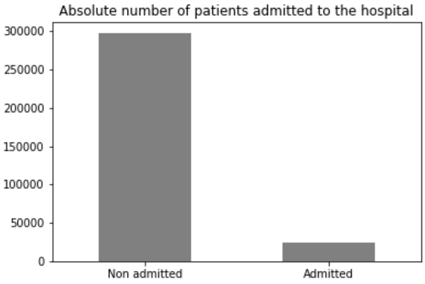

Figure 5.1: Absolute number of patient admitted to the hospital

5.3 Model Evaluation

The main method to evaluate the model was the following: we first obtained the probability that each contact belongs to the positive class of admitted patients and sorted such probabilities in descending order.

We put on the x-axis a counter which increments with the descending probabilities $(1, 2, ...N$ where N is the total number of contacts under consideration).

We retrieved the true label associated with the data and put on the y-axis the cumulative sum of such labels; in this way, the y-axis represents the number of truly admitted patients.

For example, imagine ten patients

$$x = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}]$$

(where each x_i is described by a number of features) with true labels

$$y = [0, 1, 1, 0, 1, 0, 0, 0, 0, 1]$$

.

Say that the probability predicted by our model for each patient is

$$p = [0.1, 0.9, 1, 0.5, 0.7, 0.3, 0.2, 0, 0.8, 0.6]$$

The patients and their relative label will be sorted in order of decreasing probability of admission

$$x_{sorted} = [x_3, x_2, x_9, x_5, x_{10}, x_4, x_6, x_7, x_1, x_8]$$
$$y_{sorted} = [1, 1, 0, 1, 1, 0, 0, 0, 0, 0]$$

A curve will be drawn putting on the x-axis the values 1 to 10, while the y-axis will contain the cumulative sum of true entry

$$y_{cumsum} = [1, 2, 2, 3, 4, 4, 4, 4, 4, 4]$$

The area under this curve (computed numerically) will be the evaluation metric for our model.

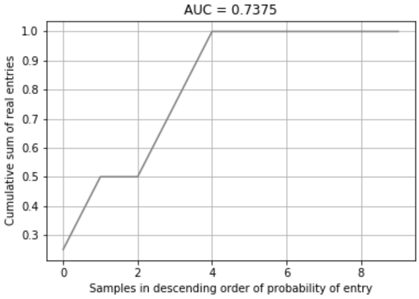

Figure 5.2: AUC example

The area under the Receiver operating characteristic (ROC) [21], using the true positive rate against the false positive rate, is also computed.

The area under the curve (obtained with our method or with a ROC) is an estimation of the model performances: good models will have areas close to 1, while bad models (close to random) will have area under curve close to 0.5.

We chose these methods of evaluation and not, for example, accuracy, for a number of reasons.

First, the aim of the classifier is not to simply predict the class of each contact, but to produce the probability of the a record to belong to one class (the positive class in our case), for reasons already explained.

Secondly, both our primary evaluation method and the Receiver operating characteristic are well suitable for situations in which there's a great class imbalance, as they evaluate if the produced probabilities are in consistent ordering.

Using again an example to illustrate class imbalance, say we have

$$x = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, ...x_{20}]$$

with true labels

$$y = [0, 1, 1, 0, 0, 0, 0, 0, 0, 0, ..., 0]$$

. in which only 2 out of 20 true labels belong to class 1.

In case our classifier is able to output probabilities which distinguish between the two classes, for example

$$p = [0, 0.99, 0.98.0.1, 0.11, 0.12, 0{,}13, 0.14, 0.15, 0.16, ..., 0.26]$$

one would have a very high area under curve, both for our evaluation method (AUC = 0.9375) and for a ROC curve (AUC = 1).

In case, on the contrary, the system is unable to correctly predict probabilities and generates, say,

$$p = [0, 0.1, 0.11.0.12, 0.13, 0.14, 0{,}15, 0.16, 0.17, 0.18, ..., 0.28]$$

the obtained area under the curve would be 0.1 for our evaluation metric and 0.05 for a ROC.

In case the majority class (in our case 0) is well predicted as having a low probability, while the minority class is predicted randomly (say, One record is predicted as having very high probability and the other as having very low probability), as, for example in

$$p = [0.1, 1, 0, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, ..., 0.27]$$

one obtains an area under curve of 0.4875 and 0.5 for our evaluation method and the ROC curve respectively, reflecting the behaviour in the positive class (despite being highly underrepresented in the data).

Concluding, our evaluation metrics evaluate to which extent the model is able to discriminate the classes involved by analyzing if the produced order of probability is consistent with the true labels. This process is insensitive to class imbalance.

Model evaluation is performed using hold-out, with 80% of the data used for training and 20% used for test.

5.4 Learning Models

As previously explained, interpretability was one of the most important requirements for the model.

As a result, we mainly focused on two learning models: random forest and simple classification tree.

We will describe the two approaches, including the results obtained, separately.

For each model, we performed some preprocessing steps on the data, transforming the categorical variables into their one-hot representation and the ordinal variables to their ordinal representation.

Tree-based methods do not support missing values (which are very few in the dataset) so that we imputed using the mean.

For each learning model, we performed two experiments.

In the first one, we only considered the data which are directed available when the patient arrives to the Emergency Department, excluding any past history. The considered variables are sex, age, the type of activity (in one-hot representation), the precedence (*Precedencia*, in one-hot), the means of arrival to the emergency room (one-hot), the triage level (ordinalized), the entity which took the initiative (*Iniciativa*, in one-hot), the symptoms which motivated the patient to go to the ER (*Motiu*, one-hot), the type of health centre the patient was in (*Precedencia*, in one-hot), the previous ER, if any (*Pr_urgencia*, in one-hot) and the emergency code activated (*Code_urg*, one-hot).

In the second one, we also included the past Emergency Department data as encoded in the variables (*#ER_urg*, *#H_urg*, and *#H_norm_urg*).

5.4.1 Random Forest

We used 50 estimators and decided to tune the minimum number of samples required to split an internal node; for this value, we trained the model in the range $2^6, 2^7, ...2^{11}$ and took the value $2^8 = 256$ $2^9 = 512$ for the two experiments, respectively.

5.4.2 Classification Tree

We selected $2^{11} = 2018$ as the optimal minimal number of samples to split, in both cases.

5.5 Results

5.5.1 Random Forest

The results obtained using random forest are reported in the following.

Table 5.1 reports the parameters used to train the model, for the two different settings.

When considering only timely data available when the patient arrives to the Emergency Department, without considering his past contact (figures 5.3 and 5.4), we obtain an area under the curve of 0.877 and 0.9072 for our our evaluation method and the Receiver Operating Characteristic, respectively.

When also using ER historical data (figure 5.5 and 5.6), the obtained area under curve using our validation method is 0.8917; using a validation based on the ROC curve, we obtain an area of 0.9231.

Note that using historical data (even when limited to the electronic records produced in the Emergency Department), we obtain an improvement of 1.5% in the area under curve.

Table 5.2 reports a list of the most important features used by the model, when also using historical data.

As shown in table, the new handcrafted variables play a very important role in the classification, as the two first most important variables refer to the patient's past history..

Random forest Parameters

	Without past data	With past data
Number of estimators	50	50
Number of samples for split	256	512
Number of features for best split	$\sqrt{p}$	$\sqrt{p}$

Table 5.1: Random forest: parameters for the emergency room data

5.5.2 Classification Tree

The results obtained using the classification tree are reported in the following. Table 5.3 reports the parameters used to train the model in the two experiments.

The area under the curve obtained was of 0.8829 (AUC for our evaluation method, figure 5.7) and 0.9139 (for ROC curve, figure 5.8) when using only timely data; it slightly increased to 0.8875 (figure 5.9) and 0.918 (figure 5.6) for the two evaluation methods when using the patient's past ER data only.

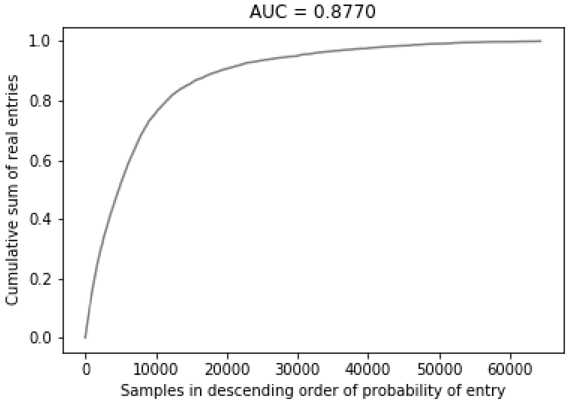

Figure 5.3: AUC for the ER data using a Random Forest, without historical data

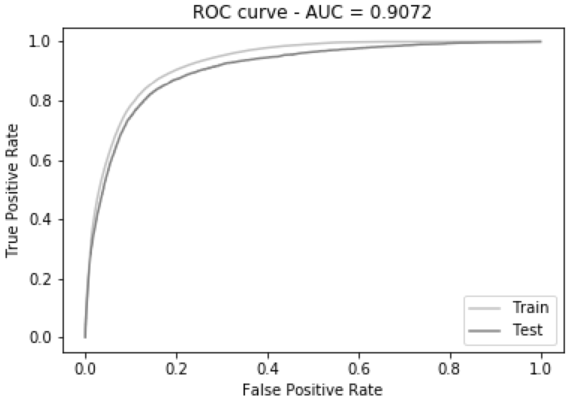

Figure 5.4: ROC AUC for the ER data using a Random Forest, without historical data

Note that in this experiment, all used past data come from the CMBD-URs, so that we use record coming from the emergency room only and not linked to other types of health services.

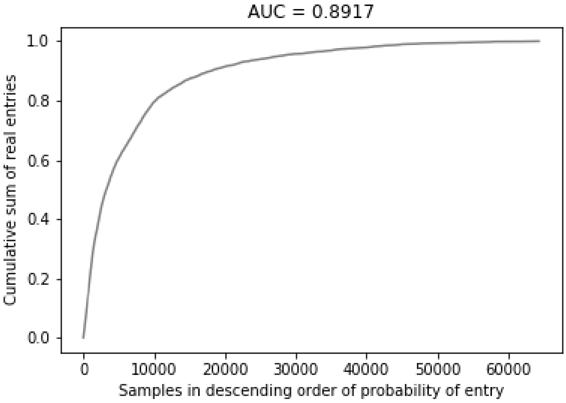

Figure 5.5: AUC for the ER data using a Random Forest, including ER historical data

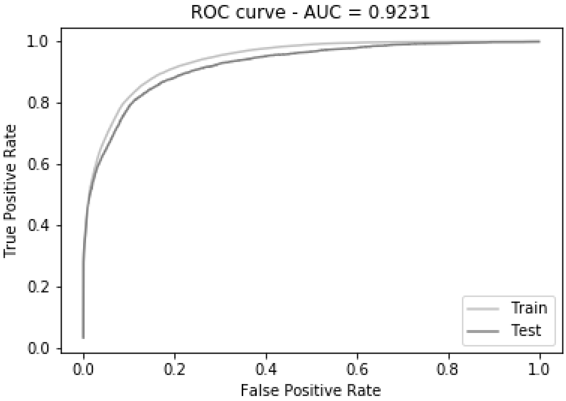

Figure 5.6: ROC AUC for the ER data using a Random Forest, including ER historical data

Feature	Importance
#H_norm_urg	0.240
#H_urg	0.207
Motiu: full-term uncomplicated delivery	0.140
Mitja_arr: 1 (own)	0.076
Mitja_arr: 2 (ambulance)	0.059
Age	0.056
Motiu: Heart failure	0.015
#ER	0.014
Motiu: Chronic bronchitis	0.011
Motiu: Pneumonia, organism unspecified	0.009
Motiu: Dyspnea and respiratory abnormalities	0.009
Motiu: Transcervical fracture	0.008
Precedencia: 2 (psychiatric hospital)	0.007
Code_urg: 0 (none)	0.007
Motiu: Other nonorganic psychoses	0.006
Iniciativa: 2 (guardian)	0.006
Motiu: Other disorders of eye	0.006
Motiu: Acute, but ill-defined, cerebrovascular disease	0.005

Table 5.2: Random forest: feature importance for emergency data only

Classification Tree Parameters

	Without past data	With past data
Number of samples for split	2048	2048

Table 5.3: Classification tree: parameters for the emergency room data

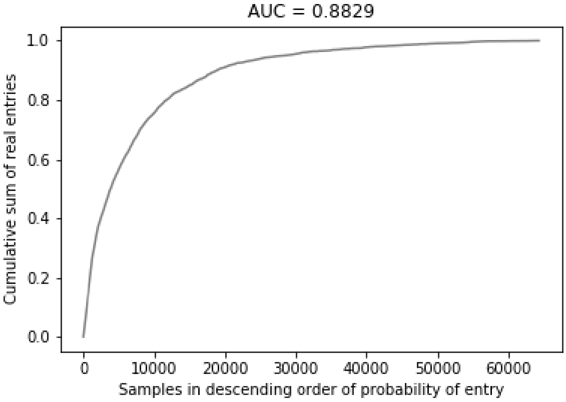

Figure 5.7: AUC for the ER data using a Classification Tree, without historical data

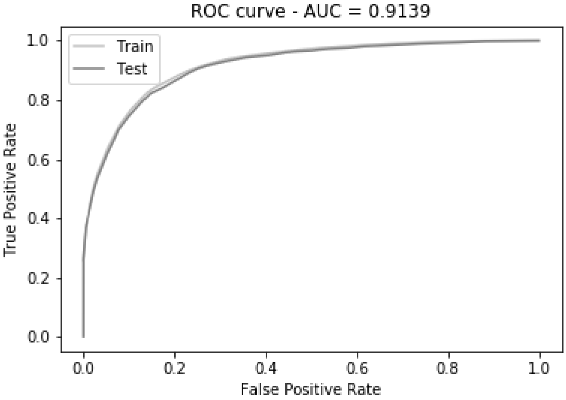

Figure 5.8: ROC AUC for the ER data using a Classification Tree, without historical data

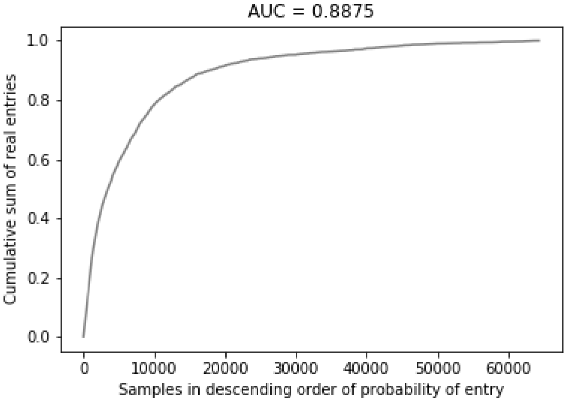

Figure 5.9: AUC for the ER data using a Classification Tree, including ER historical data

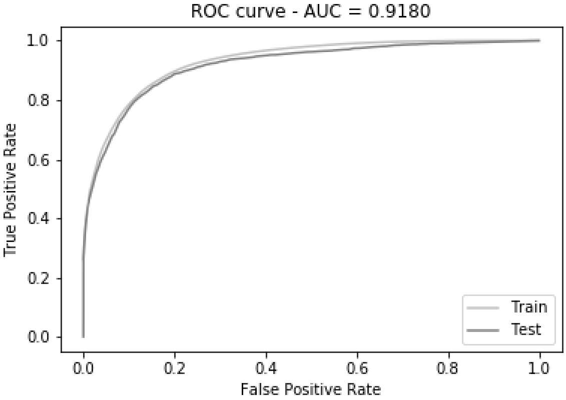

Figure 5.10: ROC AUC for the ER data using a Classification Tree, including ER historical data

Chapter 6

Predictive Model: Clinical History Integration

In Chapter 5, we presented a basic model obtained by only using the data contained in the CMBD-UR dataset.

In this chapter, we want to investigate whether integrating data from the other CMBDs would improve the model predictive power.

We will use historical information that could be available at triage time only. While these data are not necessarily accessible at the moment for an emergency-room professional, this could be made possible, especially in a context of a *Consorci Sanitari.*

In the following, we will present a model which uses data related to the patient's clinical history to predict his/her hospitalization at triage time.

6.1 Data Preprocessing

Excluding data cleaning, the aim of this step is to 'summarize' the data available in each CMBD for a given patient at a given point in time.

Conceptually, when one queries a dataset looking for entries of a specific patient, the system could return no data if the patient has never used that service, or one or more rows depending on how many times the patient used it.

In the latter case, we need to represent the returned data using variables which summarize multiple records into a single value which could later be used in the integrated model.

To do so, we will manually create a number of feature for each dataset.

We explain the data cleaning and the feature creation process for each CMBD separately.

6.1.1 Hospital Data

After a first data cleaning, we sorted the data using the time and date of the contact as index. The following variables were created:

- *#Visits_ha*: a counter of records for each patient, starting from the first available record in time and ending at the day of the contact under consideration.

  For example, for patient A, one would have:

Id	Date_hospitalization	#Visits_ha
A	01-01-2015	1
A	05-07-2015	2
A	13-11-2016	3

- *#Visits_non-progr_ah*: a counter which is incremented only considering visits which where not previously programmed, as reported in the variable *C_contact* of the CMBD.

For each contact, only the patient id (*Historia*), the date of the hospitalization and the new variables were saved.

6.1.2 Primary Care Dataset

We cleaned, unified (two different primary care available, with a slightly different format, for 2015 and 2016-2017) and sorted the data.

After some experimental work, we created the following variables:

- *#Visits_ap*: a counter of records for each patient, starting from the first available record in time and ending at the day of the contact under consideration.

- *#Visits_derivation_ap*: a counter incremented only considering visits after which the patient was sent to another, specialized, centre (as reported in the *Derivaciò* until 2015 and in the *DER1_D1 ... DER3_D6* variable afterwards).

  For example, for patient A, one would have:

Id	Date_visit	#Visits_ap	#Visits_derivation_ap
A	01-01-2015	1	1
A	05-07-2015	2	1
A	13-11-2016	3	2

where after the first and third visit the patient was sent to a specialist.

- *#Visits_non-progr_ap*: a counter which is incremented only considering visits which were not previously programmed, as reported in the variable *C_contact* of the CMBD.
- *#Visits_no-dentist_ap*: a counter which is incremented only considering visits which where not addressed by a dentist. The type of specialist who took care of the patient is originally recorded in the variable *Tprof*.
- *#Visits_doctor_ap*: a counter which is incremented only considering visits which where addressed by a doctor (for adult people) or a pediatrician.

For each contact, only the patient id, the date of the contact and the new variables were saved.

6.1.3 Mental Heath Dataset

After some experimental work, we created the following variables:

- *#Visits_mh*: a counter of visits for each patient; in this case, we did not simply count the number of records, but we sum the number of first and secondary visits for each notification period.
- *#Visits_non-progr_mh*: a counter of not previously programmed visits.
- *#Th_individual_mh*: the cumulative sum of the number of individual therapy sessions (originally in the *PI* variable).
- *#Th_group_mh*: the cumulative sum of the number of group therapy sessions (originally in the *PG* variable).
- *#Th_famil_mh*: the cumulative sum of the number of familiar therapy sessions (originally in the *PF* variable).
- *#Th_mh*: the cumulative sum of the number of all the therapy sessions. For each record, $\#Th\_mh = \#Th\_individual\_mh + \#Th\_group\_mh + \#Th\_famil\_mh$. We created this variable as a summary of the previous ones.

- *#Nurse_mh*: the cumulative sum of the times the patient needed the attention of a nurse.
- *#Social-w_mh*: the cumulative sum of the times the patient needed the attention of a social worker.
- *#Home_mh*: the cumulative sum of the times the patient needed a domestic visit.

For each contact, only the patient id (*Historia*), the date and the new variables were saved.

6.1.4 Sociosanitary Dataset

For the sociosanitary data, we selected the reference date as follows: we used the date the measurement was taken if that was not NULL; otherwise, we used the date the data were transmitted.

After some experimental work, we created the following variables:

- *#Visits_ss*: a counter of records for each patient, starting from the first available record in time and ending at the day of the contact under consideration. This measure is less accurate than the similar ones reported for the other dataset since a record can be inserted for various reasons (start of the hospitalization, change of condition, periodical measurements etc), but it still provides an idea of how much attention the patient needed.
- *#SLE_ss*: a counter of records for each patient, considering the ones of long-term care only.
- *#SPA_ss*: a counter of records for each patient, considering the ones of palliative care only.

  This distinction was necessary since there are patients which are initially hospitalized under a long-term care regime but then (possibly due to the worsening of their condition) start to have SPA records only.

In addition to these variables, we also grouped the diagnostic ones.

- *Neuro_ss*: groups the value of the diagnostic variables which are associated with a brain illness or condition (for example brain paralysis, hemiplegia etc).
- *Wound_ss*: groups the value of the diagnostic variables which are associated with wounds.
- *Tumor_ss*: groups the value of the diagnostic variables which are associated with tumors and related therapies.

- *Breath_ss*: groups the value of the diagnostic variables which are associated with breathing problems.
- *Therapies_ss*: groups the value of variables which are associated with therapies.
- *Diagn_ss*: groups all the diagnostic variables.

Even though these variables are a very coarse representation of the information, they can still give an idea of the seriousness of the patient's condition.

In addition to the id, the date and the new variables, also the original variables indicating the patient condition were saved.

6.2 Data Integration

Session 6.1 describes the variables which were created from each available CMBD to enrich the information with the other patient's electronic records.

Given a CMBD type, the data we obtained consist of the patient id, the date in which they were taken (or the date in which the data were transmitted, in case such information was unavailable), and a number of variables, most of which are counters or zero-one indicators.

Id	Date	Var1	Var2	...	VarN
A	01-01-2015	1	0	...	0
B	05-07-2015	2	1	...	1
B	15-09-2015	3	1	...	0
A	13-11-2016	3	1	...	1
C	13-11-2016	1	0	...	0

Given a row from the preprocessed CMBD-URG, we conceptually extracted the patient id and the date of the contact; given these values, for each other preprocessed CMBD, we extracted, among the rows corresponding to the same patient (if any) the one having the closer previous date.

Having found the correct row, we unified the two CMBDs by adding the new variables to the CMBD-URG.

For example, given the following row from the CMBD-URG dataset:

CMBD-URG				
Id	Date	Var1_urg	...	VarN_urg
A	05-07-2016	1	...	4

and the following data for the CMBD_AP:

CMBD-AP				
Id	Date	Var1_ap	...	VarM_ap
A	02-04-2014	0	...	1
A	05-04-2015	2	...	9
C	05-07-2016	3	...	7
B	11-07-2016	1	...	4
A	05-09-2016	1	...	4
A	05-11-2016	1	...	4

We selected the second row and obtained:

CMBD-URG + CMBD-AP							
Id	Date	Var1_urg	...	VarN_urg	Va1_ap	...	VarM_ap
A	05-07-2016	1	...	4	2	...	9

We used the same procedure to integrate all CMBDs.

Table 6.1 reports the percentages of rows in the CMBD-URG for which it was possible to find a correspondence in each CMBD.

For entries which did not have a correspondence, the value zero or NaN (NULL) was used depending on the variable.

	CMDBD-HA	CMDBD-AP	CMDBD-CSM	CMDBD-RSS
CMBD-URG	23.1%	59.1%	3.5%	3.6%

Table 6.1: Percentage of ER data for which corresponding information was found in other CMBDs

6.3 Model Evaluation

The model was evaluated following the methodology explained in section 5.3.

6.4 Learning Models

A random forest and a classification tree were trained using all variables.

Results are reported in the following.

6.5 Results

6.5.1 Random Forest

The results obtained using random forest are reported in the following.

Table 6.2 reports the parameters used to train the model. Figure 6.1 and 6.2 report the area under curve obtained (0.9284 and 0.9629 for our evaluation method and for the ROC curve respectively).

It can be noted that the area under curve increased of more than 4% with respect to the model obtained using the CMBD-URG data only.

Moreover, the table of features clearly shows that the new features are very important for the decisions taken by the model: the most important variable in the classification is, in fact, the number of past non-programmed hospitalization. Other important features include sociosanotary records and primary care visits performed by a doctor.

Random forest Parameters	
Number of estimators	50
Number of samples for split	128
Number of features for best split	$\sqrt{p}$

Table 6.2: Random forest: parameters using data for all CMBDs

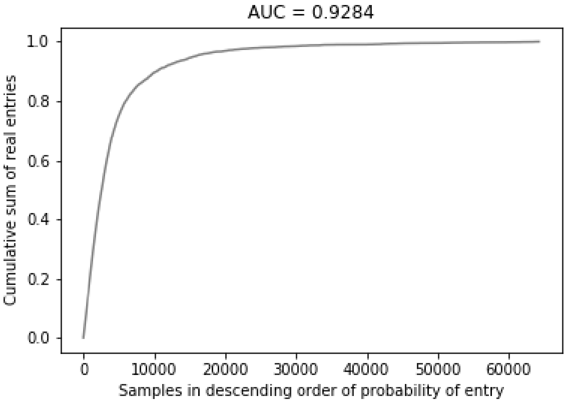

Figure 6.1: AUC for the all data using a Random Forest

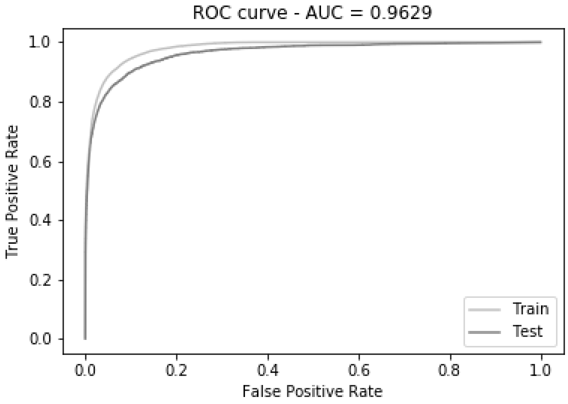

Figure 6.2: ROC AUC for the all data using a Random Forest

Feature	Importance
#Visits_non-progr_ah	0.174
#H_norm_urg	0.121
#H_urg	0.119
#HA_all	0.105
Triatge > 4	0.070
#SPA_ss	0.041
Motiu: full-term uncomplicated delivery	0.040
#ER	0.037
Triatge > 3	0.035
#Visits_ss	0.033
Age	0.024
Mitja_arr: 1 (own)	0.019
Mitja_arr: 2 (ambulance)	0.013
#Visits_doctor_ap	0.012
#Visits_no-dentist_ap	0.012
#Visits_ap	0.009
#Visits_non-progr_ap	0.007
Motiu: Heart failure	0.005
Code_urg: 0 (none)	0.005

Table 6.3: Random forest: feature importance for all CMBDs

6.5.2 Classification Tree

Table 6.4 shows the parameters used in the training of the classification tree; figures 6.3 and 6.4 show the areas under the curve (0.8875 and 0.918 when using our evaluation method and the ROC curve respectively).

Again, note that the result is 4% higher than the one obtained using emergency room data only.

Classification Tree Parameters	
Number of samples for split	2048

Table 6.4: Classification tree: parameters using data for all CMBDs

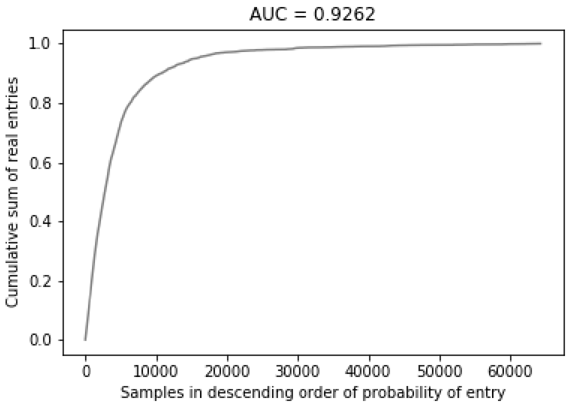

Figure 6.3: AUC for the all data using a Classification Tree

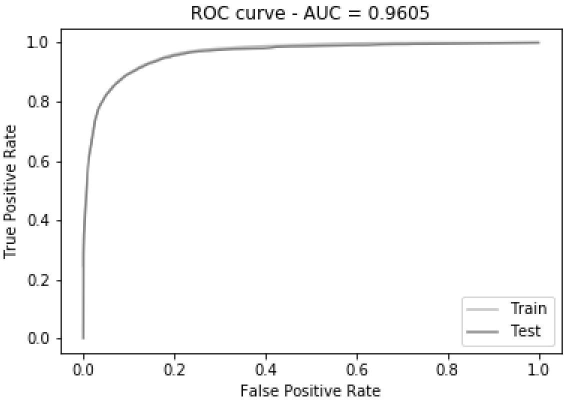

Figure 6.4: ROC AUC for the all data using a Classification Tree

Chapter 7

Removing Trivial *Motius*

In the previous chapters, we described a model using data available when a patient arrives at the emergency room, as well as his clinical history, to predict whether or not he or she will need to be admitted to the hospital, in order to reduce waiting times.

The previously described model is a classifier built by using all the available records. One could object that some of the cases into consideration are actually rather easy to be classified by a human health professional.

A person which arrives at the emergency room with a superficial injury, for example, is very unlikely to be later hospitalized regardless of the other parameters. On the contrary, someone with a major heart attack will for sure require hospitalization.

In this chapter, we describe a model obtained by removing such cases from the classification. The goal is double: to force the classifier building algorithm to focus on the cases which are harder to predict by humans, and to make a more interpretable model in which at least some cases have a trivial interpretation.

The model works in two different steps: first, it identifies the main *Motius* (reasons, as registered when the patient first arrives at the emergency room) which almost always imply hospitalization or discharge and directly estimate the probability of entry for the record using the variable *Motiu* only; then, after having removed these records, it performs the classification on the remaining data, and obtains the probabilities of entry from the classifier, using all variables. The results of the two steps are then combined to obtain a final result.

In the following, we describe how these steps are performed and report the results.

7.1 *Motius* Identification

The aim of this step is to identify records which could be predicted with a very high confidence using the variable *Motiu* only.

The variable *Motiu* is the one which better represents a condition. Some of the described conditions are often associated whith risky conditions (for which an hospitalization is often required) or with trivial ones, easily solvable in the emergency room directly.

Moving from this consideration, we try to extract the values of the variable *Motiu* for which the classification is trivial.

Given the training data, for each *Motiu* m, we define the probability of admission simply as the number of records in which the variable *Motiu* has value m for which the exit status is 1 (hospitalized), normalized by the total number of records for m

$$p(admission)_{m \in M} = \frac{\#Admitted_m}{\#Total_m}$$

In order to obtain *Motius* which lead to very high or very low probability of admission, we define the following thresholds:

- *T_cases*: this value defines the minimum number of cases with the considered *Motiu* which must be present in the training data to extract a 'unary rule' (ie. automatically classify records with the given *motiu*). The idea behind this threshold is that data must be numerous enough to generalize.
- *T_Admission*: this is the minimum value $p(admission)_{m \in M}$ can have in order to create the rule of the type $m \implies admission$.
- *T_Discharge*: this is the maximum value $p(admission)_{m \in M}$ can have in order to create the rule of the type $m \implies discharge$.

Given the prior distribution of classes, while *T_Discharge* must be very low (ie. almost all records with the given *Motiu* must be hospitalized), while *T_Admission* could, in theory, be less strict.

Given a *Motiu* m with a total of N cases, the model computes the probability of admission.

Then, if $N > T\_cases$ and $p(admission)_{m \in M} > T\_Admission$, for all records with the given *Motiu*, the probability of admission will directly be $p(admission)_{m \in M}$ (these *Motius* are considered as trivially implying admission).

The same appens if $N > T\_cases$ and $p(admission)_{m \in M} < T\_Discharge$ (these *Motius* are considered as trivially implying discharge).

If none of these conditions holds, no decision is taken at this step.

7.2 Learning Model

Trivial records as identified during the previous procedure are removed from the training set.

A learning algorithm is then run on the remaining records, following the same procedure illustrated in the previous chapters.

In particular, a Random Forest was used as learning model; however, similar results could be obtained by using a classification tree or other algorithms. The complete set of features form all CMBDs were used in the classification.

7.3 Model Evaluation

In order to evaluate the model, we assign $p(admission)_{m \in M}$ computed on the training set as probability to each test record whose *Motiu* has been identified as leading to admission or discharge in the training phase.

After removing these elements from the test set, we predict the remaining test records using the learned model and retrieve the probability associated to each entry.

In order to obtain a final evaluation figure, we *calibrate* the probabilities obtained from the learning model and later use results from both steps to compute the area under curve.

Calibration is the process by which a raw score provided by a classifier is turned into an actual probability of belonging to one class. For example, an instance may have a score of 2.3 for class 1 (which cannot even be interpreted as a probability). After calibration, the calibrated score could be 0.7. This means that the instances with raw score about 2.3 have probability about 70% of belonging to class 1.

Some predictors (e.g., logistic regression) produce calibrated scores by design, while for others (e.g., Naive Bayes, Support Vector Machines and Random Forests) the raw scores must be calibrated to correspond to probabilities.

Looking at the probabilities predicted by a Random Forest, one sees that the ones close to 0 or 1 are very rare. This is due to the fact that the model averages a set of basic predictions obtained from trees.

From [7]: "Because predictions are restricted to the interval [0,1], errors caused by variance tend to be one-sided near zero and one. For example, if a model should predict p = 0 for a case, the only way bagging can achieve this is if all bagged trees predict zero. If we add noise to the trees that bagging is averaging over, this noise will cause some trees to predict values larger than 0 for this case, thus moving the average prediction of the bagged ensemble away from 0. We observe this effect most strongly with random forests because the base-level trees trained with random forests have relatively high variance due to feature subsetting.”

There are several calibration algorithms. We have used the one based on on Platt's sigmoid model [14].

While calibration was not important for evaluation using the previously described methods, because the area under curve only depends on the relative order of probabilities, it has to be used now because we must merge the instances from the two steps; thus, we need to merge using the same type pf values i.e. probabilities computed in step 1 and predicted probabilities (from the classifier), not raw scores.

The results obtained varying the values of the thresholds are reported in the following section.

7.4 Results

In this section, we report the results obtained when modifying the values of the previously specified thresholds.

First of all, we establish two baselines:

- Predict the probability of entry based on *Motiu* only (for non-null *Motius*). For each *Motiu*, the associated probability of entry $p(admission)_{m \in M}$ was computed. The classifier was run for entries with null *Motiu* only. Probabilities obtained through the classifier were calibrated and the join area under the curve, for all entries, resulted in 0.8195.

- Predict the probability always based on all features (as done in the previous chapter). The obtained area under curve is 0.9284.

Note that using a proper classifier makes the result almost 11% better than the one obtained by considering the symptoms of the patients as noted at triage (*Motiu*) only.

Given this baseline, we try to investigate results using more realistic values for the thresholds.

For all the following experiments, we will set to $T\_cases = 10$. Note that this value has only been chosen in order to exclude extremely rare conditions. A more realistic value should be chosen in accordance with a domain expert.

Following a 'conservative' approach, we set $T\_Admission = 1$ and $T\_Discharge = 0$ for the first classification. In practice, only *Motius* which always imply discharge or admission (in 100% of the cases in training case), will have their probability of entry estimated using *Motiu* only.

Doing so, we only remove 1% of records in the first step.

In this setting, no *Motiu* implies admission, while 29 imply discharge. Table 7.1 reports a summary of the *Motius* which were surely associated with a discharge.

Motiu	Number of cases
Fracture of one or more phalanges of foot	191
Disorders of iris and ciliary body	168
Fracture of carpal bone(s)	27
Disorders of external ear	108
Pruritus and related conditions	150
Burn of wrist(s) and hand(s)	71
Open wound of genital organs (external), including traumatic amputation	17
Sexual and gender identity disorders	31
Effects of heat and light	31
Foreign body in trachea, bronchus, and lung	57
Superficial injury of hand(s) except finger(s) alone	35
Atopic dermatitis and related conditions	63
Disorders of lacrimal system	303
Open Wound Of Upper Limb	14
Other disorders of female genital organs	45
Strabismus and other disorders of binocular eye movements	18
Other diseases due to viruses and Chlamydiae	62
Superficial injury of trunk	11
Dislocation of foot	13
Multiple fractures of hand bones	16
Erythematous conditions	22
Sprains and strains of sacroiliac region	76
Burn of trunk	13
Infectious mononucleosis	11
Effects of air pressure	11
Other venereal diseases	22
Gingival and periodontal diseases	23
Superficial injury of other, multiple, and unspecified sites	17
Foreign body in anus and rectum	10

Table 7.1: Motius having zero probability of entry

Using a Random forest with 50 estimators (requiring at least 64 samples to split an internal node), a joint area under curve of 0.9229 was obtained (while the classifier considering non-trivial *Motiu* only scored AUC=0.9230).

Note that using a less strict threshold, as for example $p(admission)_{m \in M} < 0.005$, meaning considering trivial all *Motius* which imply admission in less than 0.5% of cases, 14.43% of cases are classified according to their *Motiu* only.

The joint area under curve obtained with this setting is 0.9149, while the one of the non-trivial classifier is 0.9229 (both slightly decrease).

Further increasing $p(admission)_{m \in M}$ to 1% would already consider 64% of cases as trivial.

No *Motiu* leads to admission in more than 95% of the cases.

In general, one could say that a balance should be found among the advanced deriving from considering some *Motius* as trivial and the inevitable error made by using such a simple classification algorithm as the one discussed for the 'first step' is: note, in fact, that the area under curve decreases as a greater portion of data is considered as trivially classifiable.

The optimal value for the thresholds could possibly be chosen in accordance with a domain expert.

An alternative to this approach could be to obtain a list of trivial *Motius* from a domain expert and exclude those from the classification. In this way, the model could be developed paying attention to the 'grey area' only.

Chapter 8

Conclusions and Future Work

Following a widely active area of research, this work proposes a methodology to predict the probability of hospital admission from the medical department.

Firstly, it considers a model built using data directly collected in the emergency room only and shows that the probability of admission can be predicted with a nontrivial accuracy.

In this setting, the features which better predict the admissions are the number of past emergency room visits and hospitalizations, the means of arrival, the age and some specific conditions (e.g, pregnancy).

Secondly, it proposes a new way to integrate the patient's medical data from other health services and shows that the quality of the prediction improves.

These data show that some features obtained from the past medical history (e.g., the number of non programmed hospital admissions, the total number of hospital record and some sociosanitary features) are highly important in the prediction.

The results obtained for each setting are reported in table 8.2

As the table shows, including past data helps both when using Emergency Department data only (in the best setting, we obtain an improvement of 1.5%), and when using data from other health centres. Including both ER past data and information from the other CMBDs, we obtain an improvement of about 4.5% for the Classification Tree and of 5.15% for the Random Forest. Note that all values are above 87%.

Regarding the classification obtained using a two-steps approach, the area under curve slightly decreases both when totally removing trivial *Motius*, and when trying

Results

Setting	AUC	ROC AUC
CMBD-UR data (without historical ER variables)		
Classification Tree	0.8829	0.9139
Random Forest	0.877	0.9072
CMBD-UR data (with historical ER variables)		
Classification Tree	0.8875	0.918
Random Forest	0.8917	0.9231
All CMBDs		
Classification Tree	0.9262	0.9605
Random Forest	0.9284	0.9629

Table 8.1: Results obtained without removing trivial cases

Results

Setting				
T_cases	*T_Admission*	*T_Discharge*	AUC without trivial cases	Join AUC
0	0	1	0.8207	0.8195
(Predict all record based on *Motiu* only)				
10	1	0	0.9229	0.9230
10	0.98	0.005	0.9229	0.9149
10	0.98	0.01	0.9174	0.9062

Table 8.2: Results obtained by removing trivial cases

to join the prediction obtained in the two classification steps.

However, a classifier using this kind of approach might be preferred for two reasons: first, it makes the classification straight forward for trivial yet common cases, for which the medical professionals are surely well trained; secondly, it would allow future work to focus on complex cases only, studying *ad hoc* models to better discriminate among intricate situations.

When evaluating which model to deploy, interpretability should be central. Even if classification trees are often regarded as intuitive both in their representation and in their functioning, their practical interpretative power depends on the dimensions of the tree (very big trees become difficult to interpret); moreover, if has been proven that a small modification in the data can produce very different splits. Random forests, on the other hand, produce a clear rank of feature, even if they can't be intuitively represented.

As a practical recommendation, the obvious need to make this work really relevant is to implement this method as part of the software that the triage personnel use for their work.

Right now, the developed models are separated from the software used to register patients' data.

In a stressful situation, as the Emergency Department is, the professional cannot be expected to be using two piece of software simultaneously, so that our work should be integrated in the existing one and emit its predictions, for example, as a non-intrusive label that is computed in the background and appears automatically on the patient's record once all relevant information has been introduced. This integration work should probably be performed by the hospital's IT department.

Future works should consider other ways to engineer features from the available data (for example, trying to better use the diagnostic data, which have been dismissed in most of this project) and possibly, test the model on a real setting to obtain a forward prediction study.

In a multidisciplinary team setting, it could also be worth excluding some trivial cases from the model training data and focus on cases for which a prediction is difficult for a domain expert only. This possibility is briefly investigated in this work but should be better evaluated.

Moreover, the actual reduction in waiting time that can be achieved should be considered both from a theoretical point of view and once the system is studied in real setting.

Considering the false positive rates and the false negative rates of the classifiers, the number of patients that need admission, their current waiting time, and the extent to which the system can anticipate the admission process, a first evaluation of the improvements introduced by the system could evaluated, but a pilot study in a real system is surely need before a formal adoption.

In the long run, being able to predict admissions with a nontrivial confidence (which is theoretically allowed by the system developed in this work) could impact emergency room crowding and reduce waiting times, reducing the index of premature death, the rate of complications and even costs for the system.

The developed project only represents the first step in this direction, but the obtained results show that it is a direction worth to be taken.

Bibliography

[1] International classification of diseases - 9 - cm. https://wonder.cdc.gov/wonder/sci_data/codes/icd9/type_txt/icd9cm.asp, 1979. [Online; accessed June-2018].

[2] Manuals de notificació dels requeriments específics del CMBD, Centres de salut mental ambulatòria. http://catsalut.gencat.cat/web/.content/minisite/catsalut/proveidors_professionals/registres_catalegs/registres/cmbd/manuals_notificacio/cmbd_smental.pdf, 2014. [Online; accessed May-2018].

[3] Manuals de notificació dels requeriments específics del CMBD, Hospitals generals d'aguts. http://catsalut.gencat.cat/web/.content/minisite/catsalut/proveidors_professionals/registres_catalegs/registres/cmbd/manuals_notificacio/cmbd_aguts.pdf, 2016. [Online; accessed May-2018].

[4] Manuals de notificació dels requeriments específics del CMBD, Urgències. http://catsalut.gencat.cat/web/.content/minisite/catsalut/proveidors_professionals/registres_catalegs/documents/manual_cmbd_ur.pdf, 2016. [Online; accessed May-2018].

[5] Manuals de notificació dels requeriments específics del CMBD, Recursos sociosanitaris. http://catsalut.gencat.cat/web/.content/minisite/catsalut/proveidors_professionals/registres_catalegs/registres/cmbd/manuals_notificacio/cmbd_socio.pdf, 2017. [Online; accessed May-2018].

[6] Manuals de notificació dels requeriments específics del CMBD, Atenció primària. http://catsalut.gencat.cat/web/.content/minisite/catsalut/proveidors_professionals/registres_catalegs/documents/manual_cmbd_ap.pdf, 2018. [Online; accessed May-2018].

[7] Niculescu-Mizil A. and Caruana R. Predicting good probabilities with supervised learning. 2005.

[8] Marco Alban and Tanner Gilligan. Automated detection of diabetic retinopathy using fluorescein angiography photographs. 2016.

[9] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011.

[10] Steven L Bernstein, Dominik Aronsky, Reena Duseja, Stephen Epstein, Dan Handel, Ula Hwang, Melissa McCarthy, K John McConnell, Jesse M Pines, Niels Rathlev, et al. The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine*, 2009.

[11] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[12] L. Breiman. Random forests. *Machine Learning*, 2001.

[13] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[14] Platt J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. MIT Press, 1999.

[15] T. C. Chan, J. P Killeen, Kelly D., and Guss D. A. Impact of rapid entry and accelerated care at triage on reducing emergency department patient wait times, lengths of stay, and rate of left without being seen. *Annuals of Emergency Medicine*, 2005.

[16] Mao-Te Chuang, Ya-han Hu, and Chia-Lun Lo. Predicting the prolonged length of stay of general surgery patients: a supervised learning approach. *International Transactions in Operational Research*, 2018.

[17] Carlos Elvira, Alberto Ochoa, Juan Carlos Gonzalvez, and Francisco Mochón. Machine-learning-based no show prediction in outpatient visits. *International Journal of Interactive Multimedia and Artificial Intelligence*, 2018.

[18] Andres Garcia-Arce, Florentino Rico, and José L Zayas-Castro. Comparison of machine learning algorithms for the prediction of preventable hospital readmissions. *Journal for Healthcare Quality*, 2018.

[19] Byron Graham, Raymond Bond, Michael Quinn, and Maurice Mulvenna. Using data mining to predict hospital admissions from the emergency department. 2018.

[20] A. Guttmann, M. J. Schull, M. J. Vermeulen, and T. A. Stukel. The relationship between emergency department crowding and patient outcomes: A systematic review. *Journal of Nursing Scholarship*, 2013.

[21] T. Hastie, Tibshirani, and J R Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[22] Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. 2017.

[23] Scott Levin, Matthew Toerper, Eric Hamrock, Jeremiah Hinson, Sean Barnes, Heather Gardner, Andrea Dugas, Bob Linton, Tom Kirsch, and Gabor Kelen. Machine-learning-based electronic triage more accurately differentiates

patients with respect to clinical outcomes compared with the emergency severity index. 2017.

[24] Ahmed Mahmoud. Scottish patients at risk of readmission and admission (sparra). *International Journal of Integrated Care*, 2016.

[25] H. Mayer, F. Gomez, D. Wierstra, I. Nagy, A. Knoll, and J. Schmidhuber. A system for robotic heart surgery that learns to tie knots using recurrent neural networks. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[26] T. Mitchell. *Machine Learning.* McGraw Hill., 1997.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[28] J. M. Pines, C. Garson, W. G. Baxt, Rhodes K. V., Shofer F. S., and Hollander J. E. Ed crowding is associated with variable perceptions of care compromise. *Academic Emergency Medicine*, 2008.

[29] Nathan R Hoot and Dominik Aronsky. Systematic review of emergency department crowding: Causes, effects, and solutions. 2008.

[30] Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

[31] Carol E. Reiley, Erion Plaku, and Gregory D. Hager. Motion generation of robotic surgical tasks: Learning from expert demonstrations. *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010.

[32] F. H. Roger. The minimum basic data set for hospital statistics in the ec. 1981.

[33] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *CoRR*.