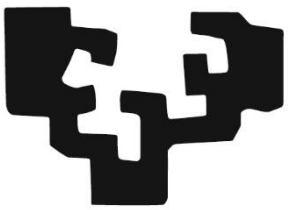


Computational Intelligence Contributions to Readmission Risk Prediction in Healthcare Systems

by
Arkaitz Artetxe

Submitted to the Department of Computer Science and Artificial Intelligence, in partial fulfilment of the requirements for the degree of Doctor of Philosophy

PhD Advisors:

Prof. Manuel Graña At The University of the Basque Country

and

Dr. Andoni Beristain At Vicomtech-IK4

Universidad del País Vasco
Euskal Herriko Unibertsitatea
Donostia-San Sebastián

2017

Acknowledgements

Special thanks to my advisor, Prof. Manuel Graña for his experience, support and guidance. Without him this Thesis would not have been possible.

I am also grateful to Dr. Andoni Beristain for his technical collaboration and direction at Vicomtech-IK4.

I would also like to thank to all my workmates and colleagues at Vicomtech-IK4 Research Center who helped me in this task. Eskerrik asko.

Arkaitz Artetxe

Computational Intelligence Contributions to Readmission Risk Prediction in Healthcare Systems

by

Arkaitz Artetxe

Submitted to the Department of Computer Science and Artificial Intelligence, in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Abstract

The Thesis tackles the problem of readmission risk prediction in healthcare systems from a machine learning and computational intelligence point of view. Readmission has been recognized as an indicator of healthcare quality with primary economic importance. We examine two specific instances of the problem, the emergency department (ED) admission and heart failure (HF) patient care using anonymized datasets from three institutions to carry real-life computational experiments validating the proposed approaches. The main difficulties posed by this kind of datasets is their high class imbalance ratio, and the lack of informative value of the recorded variables. This thesis reports the results of innovative class balancing approaches and new classification architectures.

Keywords: Readmission, class imbalance, classification, heart failure, machine learning, artificial intelligence

“Nola aldatzen diren gauzak, kamarada.”

Hertzainak

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides a general introduction to the Thesis, providing a brief presentation of its contents, motivation, supporting publications and structure. It is structured as follows: Section 1.1 presents the main motivations of the Thesis. Section 1.2 summarizes the main methodological and technical contributions. Section 1.3 enumerates the publications obtained during the research associated with this Thesis studies. Finally, Section 1.4 describes the structure of the Thesis.

1.1. Motivation

The application of predictive analytics techniques in the medical and clinical practice is gaining momentum because they can improve healthcare in several ways [Mortazavi2016]. Specifically, risk prediction models are widely used to predict the level of risk of individual patients or patient groups for different types of diseases and populations. Those models facilitate the identification of patients potentially at high risk so that resources can be used more efficiently in terms of cost-benefit.

In hospitals inside public and private healthcare service networks, there is a growing concern on the quality and sustainability of the service. Readmission events, defined as returning admissions to a hospital after a short time (below some specified threshold) after discharge from hospital, are widely recognized as healthcare quality indicators. Readmission threshold is a matter of political choice. Readmissions are costly events that impose tremendous burden on patients and on healthcare systems [Wallmann2013, Dharmarajan2013]. Preventable readmissions are related to suboptimal care during hospitalization and poor management of the discharge process [Swain2015, Balla2008]. Thus, hospital readmissions are becoming a strong concern of hospitals and policy makers as a measure of the quality of given care and have been adopted by many organizations as quality indicators [Baillie2013]. Centres of Medicare and Medicaid Services

(CMS) in the USA [CMS2011] and policy makers in UK [Kmietowicz2010] have introduced financial penalties to hospitals with high readmission rates by reducing the payment of patients readmitted within 30-day of discharge. This is a widely-used readmission threshold, but there are some studies where they use 28 days [Betihavas2015, Tsui2015], and we have even dealt with a short 3 days' threshold in one of the studies reported in this Thesis.

Readmission risk prediction models have become effective tools that help medical decision making and provide several benefits to both healthcare providers and patients [[Zheng2015]. Predictive models facilitate identification of patients at high risk for hospital readmissions and potentially enable direct specific interventions toward those who might benefit most [Walraven2010]. Interventions involving issues such as medication reconciliation, patient education, telephone follow-ups among others, have shown to effectively reduce readmission rates for patients after hospital discharge [Kripalani2014, Urma2017, Leppin2014].

However, some studies agree in concluding that predictive models based on administrative and clinical data discriminate poorly on readmissions [Ross2008, Kansagara2011, Dharmarajan2013, Mortazavi2016, [Krumholz2016]. The inherent difficulty of the problem and the limited discriminant power of the variables recorded in the dataset (i.e. the problem may be far from being linearly separable) may be the cause to the modest performance of the risk prediction models.

Most of the models in the literature are based on traditional statistics, mainly logistic regression and survival analysis [Ross2008**,** [Zheng2015]. Increasingly, authors propose machine learning as one of the best ways in which data can be used to extract knowledge [Kadi2017].

Machine learning techniques can improve both discrimination and range of prediction over traditional statistical techniques, with the ability to leverage all available data and their complex relationships [Mortazavi2016].

This Thesis aims to contribute to the field of readmission risk prediction modelling by providing comparative studies on the application of state-of-the-art and some innovative machine learning techniques for model building. We have tackled the problem in two different medical areas which are commented below, namely *emergency department readmissions* and those related with *heart failure* patients.

1.1.1. Emergency Department Readmissions

The aging of global population is a recognized fact. The number of people aged over 65 is projected to grow from an estimated 524 million in 2010 to nearly 1.5 billion in 2050 worldwide [WHO2011]. This trend has a direct impact on the sustainability of health systems, in maintaining both public policies and the required budgets.

This growing population group represents an unprecedented challenge for healthcare systems. In developed countries, older adults already account for 12 to 21% of all Emergency Department (ED) visits and it is estimated that this will increase by around 34% by 2030 [Carpenter2011]. Older patients have increasingly complex medical conditions in terms of their number of morbidities and other conditions, such as the number of medications they use, existence of geriatric syndromes, their degree of physical or mental disability, and the interplay of social factors influencing their condition [Kansagara2011]. Recent studies have shown that adults above 75 years of age have the highest rates of ED readmission, and the longest stays, demanding around 50% more ancillary tests [Lopez2011]. Notwithstanding the intense use of resources, these patients often leave the ED unsatisfied, with poorer clinical outcomes, and higher rates of misdiagnosis and medication errors [Han2009] compared to younger patients. Additionally, once they are discharged from the hospital, they have a high risk of adverse outcomes, such as functional worsening, ED readmission, hospitalization, death and institutionalization [Guidelines2014].

1.1.2. Heart Failure

Heart failure (HF) is a clinical syndrome characterized by typical symptoms (e.g. breathlessness, ankle swelling and fatigue) caused by a structural and/or functional cardiac abnormality, resulting in a reduced cardiac output and/or elevated intra-cardiac pressures at rest or during stress. Demonstration of an underlying cardiac cause is central to the diagnosis of HF. This is usually a myocardial abnormality causing ventricular dysfunction or abnormalities of the valves, pericardium, endocardium, heart rhythm and conduction [Ponikowski2016]. The prevalence of HF is approximately 1–2% of the adult population in developed countries, rising to ≥10% among people >70 years of age [Mosterd2007]. Cardiovascular diseases and pathological processes such as HF have the highest 30-day readmission rates [Jencks2009]. In USA, it is estimated that almost half of the Medicare beneficiaries are readmitted within 6 months after a hospitalization for congestive HF [Krumholz1997].

1.2. Thesis Contributions

The following are the technical and methodological contributions to the field of predictive models for readmission in this Thesis:

- We carry out a systematic literature review, through a thorough analysis of the most significant and recent literature on readmission risk prediction modelling

- We contribute an innovative ensemble method that combines data resampling with bootstrap aggregating (bagging) and an ensemble of Extreme Learning Machine (ELM) and Decision Tree (DT) pairs for modelling heavily imbalanced datasets
- We carry out a detailed analysis of the state-of-the-art approaches addressing the issue of class imbalance. Different methods for alleviating the majority class bias are evaluated using real life medical datasets
- We present a real-life application of the recently published Anticipative Hybrid Extreme Rotation Forest (AHERF), which is a heterogeneous ensemble classifier that anticipates which classifier architecture is better suited for the problem domain at hand
- We provide an overview and evaluation of common approaches for feature selection in readmission risk prediction. We evaluate the performance of some of the most relevant techniques in the field using real use-case data
- We design and implement a software toolbox for predictive modelling in readmission, distributed as open source software[1]. In addition, a synthetic dataset is included for testing purposes, which has been generated in accordance with the statistics of the real datasets used in this Thesis

1.2.1. On the curation of the experimental datasets

Although it's been extensively reported, we believe that it is necessary to stress the importance of data preparation in general, and data cleansing in particular. Often, when working with most of publicly available datasets, most data preprocessing is already done and thus, it is transparent to the data scientist, whose effort can be focused on carrying out *machine learning* experiments.
However, the real-life datasets that we have been provided with are in a quite different state. Most of times, data was delivered to us in several ASCII or spreadsheet files, often containing incongruences and erroneous data. Let this example help illustrate the problem: One of the datasets came from a study regarding a telemonitoring program for specific kind of patients. The user interface presented to the patients in the telemonitoring program (via a PDA) provided a free text field for reporting data such as weight or systolic and diastolic blood pressure, instead of using proper float or integer constrained set of form fields (even with input coherence check). Therefore, I had to spend long working hours figuring out the way the patient wrote down the data, and implementing appropriate scripts to filter millions of data entries. This minor design flaw means that data is unreliable even after such cleaning since patients will always be able to

[1] https://github.com/aartetxe/par-toolbox

invent innovative ways of annotating the required measurement values if no guidance or control is provided. The consequence is that most of the researchers in the data science field spend more time data sanitizing than in model building. As a corollary, we learnt that data preparation is a time and resource consuming task intrinsic to data mining that shouldn't be underestimated, under penalty of jeopardizing a data analysis project (in terms of both tasks and budget).

1.3. Publications

The following publications are the direct result of the works reported in this Thesis.

1. Arkaitz Artetxe, Manuel Graña, Andoni Beristain, Sebastián Ríos. Balanced training of a hybrid ensemble method for imbalanced datasets: A case of emergency department readmission prediction. Neural Computing and Applications (2017) (Accepted). **[JCR (2016): 2.505, 5-year: 2.012, Q2]**

2. Arkaitz Artetxe, Manuel Graña, Andoni Beristain, Sebastián Ríos. Emergency Department Readmission Risk Prediction: A Case Study in Chile. In International Work-Conference on the Interplay Between Natural and Artificial Computation (pp. 11-20). Springer, Cham (2017).

3. Arkaitz Artetxe, Nekane Larburu, Nekane Murga, Vanessa Escolar, Manuel Graña. Heart Failure Readmission or Early Death Risk Factor Analysis: A Case Study in a Telemonitoring Program. In International Conference on Innovation in Medicine and Healthcare (pp. 244-253). Springer, Cham (2017).

4. Arkaitz Artetxe, Borja Ayerdi, Manuel Graña, Sebastian Rios, Using Anticipative Hybrid Extreme Rotation Forest to predict emergency service readmission risk. Journal of Computational Science, vol. 20, p. 154-161 (2017). **[JCR (2016): 1.748, 5-year: 2.009, Q2]**

5. Arkaitz Artetxe, Andoni Beristain, Manuel Graña, Ariadna Besga. Predicting 30-Day Emergency Readmission Risk. In Proceedings of the International Joint Conference SOCO'16-CISIS'16-ICEUTE'16. ICEUTE 2016. Advances in Intelligent Systems and Computing, vol. 527, pp.3-12. Springer, Cham (2016).

Other publications by the PhD student not directly related to the topics presented in this Thesis:

6. Arkaitz Artetxe, Gorka Epelde, Andoni Beristain, Ane Murua, Roberto Álvarez. Gaining Insight from Physical Activity Data using a Similarity-based Interactive Visualization. In

Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications ISBN 978-989-758-175-5, pages 115-122. DOI: 10.5220/0005675701150122 (2016)

7. M. Alberich, A. Artetxe, E. Santamaría-Navarro, A. Nonell-Canals, G. Maclair, GENESIS - Cloud-Based System for Next Generation Sequencing Analysis: A Proof of Concept, Innovation in Medicine and Healthcare 2016, Smart Innovation, Systems and Technologies, vol 60, pp 291-300. Springer, Cham (2016).

8. Álvarez, R., Murua, A., Artetxe A., Epelde G. & Beristain A. A platform for user empowerment through Self Ecological Momentary Assessment / Intervention. Proceedings of 5th EAI International Conference on Wireless Mobile Communication and Healthcare (2015)

9. Carrasco, E., Sánchez, E., Artetxe, A., Toro, C., Graña, M., Guijarro, F., Susperregui J.M., Aguirre, A. Hygehos Home: an innovative remote follow-up system for chronic patients. Innovation in Medicine and Healthcare 2014, 207, 261 (2015).

10. Iker Mesa, Eider Sanchez , Carlos Toro , Javier Diaz , Arkaitz Artetxe , Manuel Graña , Frank Guijarro , Cesar Martinez , Jose Manuel Jimenez , Shabs Rajasekharan , Jose Antonio Alarcon & Alessandro De Mauro: Design and Development of a Mobile Cardiac Rehabilitation System. Cybernetics and Systems: An International Journal, 45:2, 92-108 (2014) **[JCR (2014): 0.84, 5-year: 0.968, Q3]**

11. Arkaitz Artetxe, Andoni Beristain, Luis Kabongo. Activity Classification Using Mobile Phone based Motion Sensing and Distributed Computing. Studies in health technology and informatics, 207, 1-10 (2013)

12. Arkaitz Artetxe, Eider Sanchez, Carlos Toro, Cesar Sanín, Edward Szczerbicki, Manuel Graña, Jorge Posada: Impact of Reflexive Ontologies in Semantic Clinical Decision Support Systems. Cybernetics and Systems: An International Journal 44(2-3): 187-203 (2013) **[JCR (2013): 0.507, 5-year: 0.77, Q3]**

13. Eider Sanchez, Carlos Toro, Arkaitz Artetxe, Manuel Graña, Cesar Sanín, Edward Szczerbicki, Eduardo Carrasco, Frank Guijarro: Bridging challenges of clinical decision support systems with a semantic approach. A case study on breast cancer. Pattern Recognition Letters 34(14): 1758-1768 (2013) **[JCR (2013): 1.062, 5-year: 1.466, Q3]**

14. Iker Mesa, Eider Sanchez, Javier Diaz, Carlos Toro, Arkaitz Artetxe. GoCardio: A novel approach for mobility in cardiac monitoring. InImpact: The Journal of Innovation Impact, vol. 6(1), p. 110 (2016)

15. Arkaitz Artetxe, Eider Sanchez, Carlos Toro, Cesar Sanín, Edward Szczerbicki, Manuel Graña, Jorge Posada: Speed-up of a Knowledge-Based Clinical Diagnosis System using Reflexive Ontologies. KES 2012: 1480-1489 (2012)

16. Eider Sanchez, Carlos Toro, Arkaitz Artetxe, Manuel Graña, Eduardo Carrasco, Frank Guijarro: A Semantic Clinical Decision Support System: conceptual architecture and implementation guidelines. KES 2012: 1390-1399 (2012)

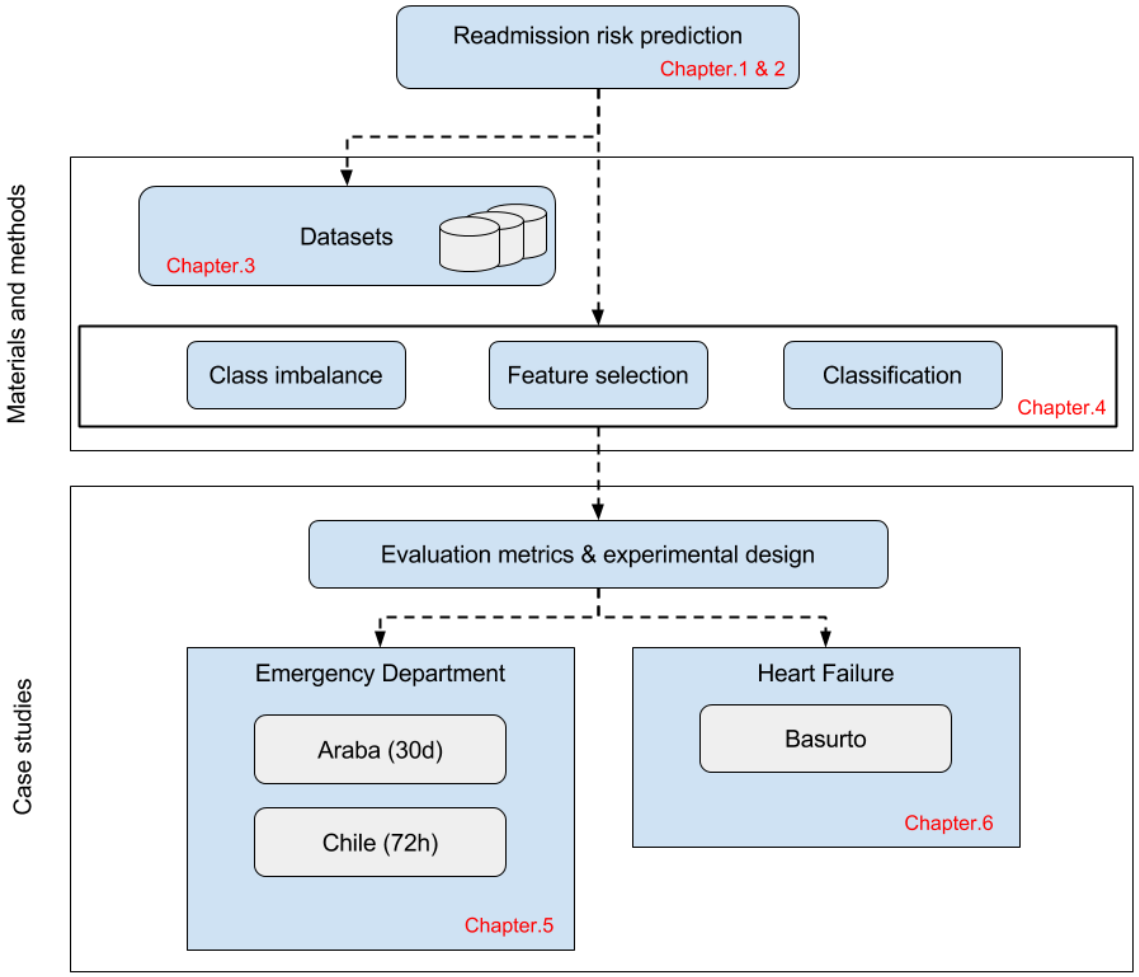


Figure 1.1. Thesis structure

1.4. Structure of the Thesis

The contents of the Thesis (shown diagrammatically in Figure 1.1) are structured as follows:

- Chapter 2 introduces the readmission risk prediction problem and provides background information related with the main contributions of this Thesis. This chapter contains a detailed systematic review of the state-of-the-art techniques and methodologies related with readmission risk prediction models.
- Chapter 3 describes the datasets that were used for the computational experiments carried out in this Thesis.
- Chapter 4 provides the definition of the computational methods used in the experiments, along with a description of the related methodological issues. Several feature selection techniques, class imbalance addressing approaches and classification algorithms used in this Thesis are presented.
- Chapter 5 presents the experimental results of the studies developed in the two areas of healthcare tackled in this Thesis: Emergency Department and Heart Failure readmission risk prediction.
- Chapter 6 provides the conclusions of the Thesis and proposes some future work.

Complementarily, 2 appendices are included in the Thesis.

- Appendix A: Describes the HF patient telemonitoring program, which is part of the working scenarios of this Thesis.
- Appendix B: Presents the results of the systematic review on readmission risk prediction models.

Chapter 2

State of the Art

This chapter provides a description of the Thesis' most relevant concepts by a systematic literature review, in which a thorough analysis of the most significant readmission risk predictive modelling studies is carried out. Next, we briefly describe heart failure from the medical point of view, discussing some predictive modelling studies related to this disease.

2.1. Predictive Models for Readmission Risk: A Systematic Review

Readmission prediction models are not new, and there exists a plethora of studies addressing this problem. A query about readmission prediction in Google Scholar returns about 28,500 hits, which is a clear indicator of the interest of the scientific community in the topic. The high number of published studies covers a wide spectrum of approaches, which justifies the work on a thorough review in order to achieve a map of the relevant procedures and issues regarding the topic.

Some authors have performed bibliographic review studies with the objective of synthesizing the literature on prediction models for the estimation of readmission risk. In 2011 Kansagara et al. [Kansagara2011] presented the most referenced systematic review paper about this topic. It was focused on model description and performance comparison in order to assess model suitability for clinical or administrative use. Authors conclude that most readmission risk prediction models perform poorly so that efforts to improve their performance are still needed. The study also concludes that readmission risk prediction is a complex problem by nature, with many inherent difficulties and inescapable traps, such as the small number of variables which are very noisy and not very much informative.

In 2015 Swain et al. [Swain2015] conducted a semi-systematic review of readmission predictive factors from predictive modeling papers published prior to March 2013. This review was, to some degree, based on [Kansagara2011] since its citations were automatically included within the potentially relevant article's list. Other studies concentrate on a certain subpopulation rather than covering all the published risk prediction models. Ross et al. [Ross2008] conducted a review of statistical models for the readmission of heart failure (HF) patients. This work included the identification of analytic models, apart from identifying patient characteristics associated with readmission. A more recent study from Leppin et al. [Leppin2014] reviewed randomized trials that assessed the effect of interventions intended to prevent 30-day hospital readmissions.

Most of the previous review studies have focused on measuring the discrimination ability of the models and identifying predictive characteristics associated with readmission. In different but related fields, review studies targeting the analysis of data analysis approaches can be found. For instance, [Kadi2017] is a recent systematic literature review on data mining techniques applied in cardiology.

Nevertheless, to our knowledge no review study covering data mining techniques, including feature selection and class imbalance, has been presented in the field of readmission prediction.

2.1.1. Research methodology

A systematic review is a formal method that enables the identification, assessment and interpretation of all available studies relevant to a specific research question, topic area or subject of interest [Brereton2007]. Systematic reviews differ from narrative reviews in that they are based "on a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant primary research, and to extract and analyse data from the studies that are included in the review" [Khan2003].

In this work, we conduct a systematic review following the three stages proposed by Tranfield et al. [Tranfield2003], namely planning, conducting and reporting, as illustrated in Figure 2.1.

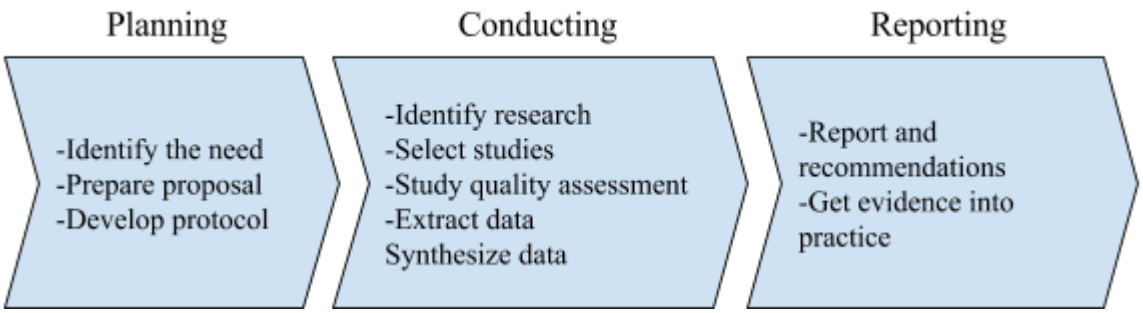


Figure 2.1. Phases of a systematic review according to [Tranfield2003]

According to this methodology, first we define the research questions. Secondly, we define the search strategy by identifying the source databases and the inclusion and exclusion criteria. Next, we present the data extraction procedure and, finally, we present the results.

2.1.2. Research questions

The overall objective of our systematic review is to identify and analyse the most significant research studies carried out on the topic of readmission risk prediction. More precisely, this review examines the data analysis methods utilized in these studies, paying special attention to data mining techniques. Table 2.1shows the research questions that guided this review.

Table 2.1. Research questions

#	**Research Question**	**Rationale**
Q1	Which data analysis methods were used in readmission risk prediction?	To identify the most common procedures that are applied for model construction in readmission risk prediction.
Q2	Which data mining techniques were used in readmission risk prediction?	To identify which data mining techniques are used for readmission risk prediction model construction.
Q3	What is the overall performance of models in readmission risk prediction?	To assess the discrimination ability of the models in readmission risk prediction.

Given that the second research question (Q2) is broad, it was divided into three sub-questions:

- **Q2.1** Classification algorithms,
- **Q2.2** Feature selection techniques, and
- **Q2.3** Techniques addressing class imbalance issues.

2.1.3. Search strategy

Search engines

We chose PubMed and Google Scholar search engines to retrieve the primary literature references. Google Scholar was selected because of its broad coverage of general scientific publications, while PubMed provided access to the more specialized MEDLINE (Medical Literature Analysis and Retrieval System Online) database.

Table 2.2 shows the search strings used with the search engines. The search strings were designed to achieve an appropriate trade-off between coverage and manageable size of the retrieved reference list.

Table 2.2. Search strings

Database	**Search term**
Google Scholar	((readmission) OR (rehospitalization)) AND ((“prediction model”) OR (“predictive model”) OR (“risk model”))
PubMed	((readmission*) OR (rehospitalization*)) AND (("prediction model") OR ("predictive model"))

Additionally, we added the reference lists of main review articles to the references used in the analysis, assuming their quality.

Search limits

The following search limitations were applied:

- **Peer-reviewed journal articles in English**

  We limited the search to indexed journal articles written in English language. Peer-reviewed journal articles are considered to provide a good view of accepted and validated methodologies and knowledge.

- **Search within**

  We performed the search using all fields available, that is, we do not restrict the search to the title and abstract or to a particular subject area. Our main goal was not to disregard high impact papers due to restrictive search conditions.

- **Published between**

  We did not restrict our search to a precise time frame. Citations were collected on February 15, 2017 so that very few studies published in 2017 were included. It’s worth noting that, due to the delays related to journal publishing, some studies accepted for publication in late 2016 may not be included.

Moreover, we excluded studies whose target population were patients that underwent certain surgical procedure for being too specific.

Data extracted from the publications

For each study included in the review, we have extracted and summarized data associated with the research questions defined. Analytic model was extracted in relation to Q1. We collected the AUC metric (Area Under the Roc Curve) if reported as the canonical measure for discrimination

ability (Q3) of the models. The different sub-questions of Q2 lead us to the collection of the following information: Feature selection technique, procedures addressing class-imbalance applied, and readmission rate, which is directly related to the imbalance-ratio (see Section 4.1). Additionally, target population, readmission threshold (in hours, days or months) and dataset size (number of instances of the dataset) were also collected.

2.1.4. Results

In this section, we present the results of our systematic review study. First, we present an overview of the results and following we discuss specific research questions.

Overview of selected studies

As shown in Figure 2.2, we gathered 208 eligible references from the search engines. Duplicated references (32) in the merged list retrieved from both databases were excluded. To this list, we added references extracted from the reference lists of main review articles in the literature (43 additional references). At this step, we had a dataset consisting of 219 potentially relevant references for analysis and review. In the following step, 95 references were excluded based on the review of the title and abstract. Further reviews excluded 58 articles that did not fulfil the predefined inclusion criteria. 32 of them were excluded due to the language and peer-review criteria. 11 citations were excluded for not including a readmission prediction model and 15 were discarded for being out of the review scope.

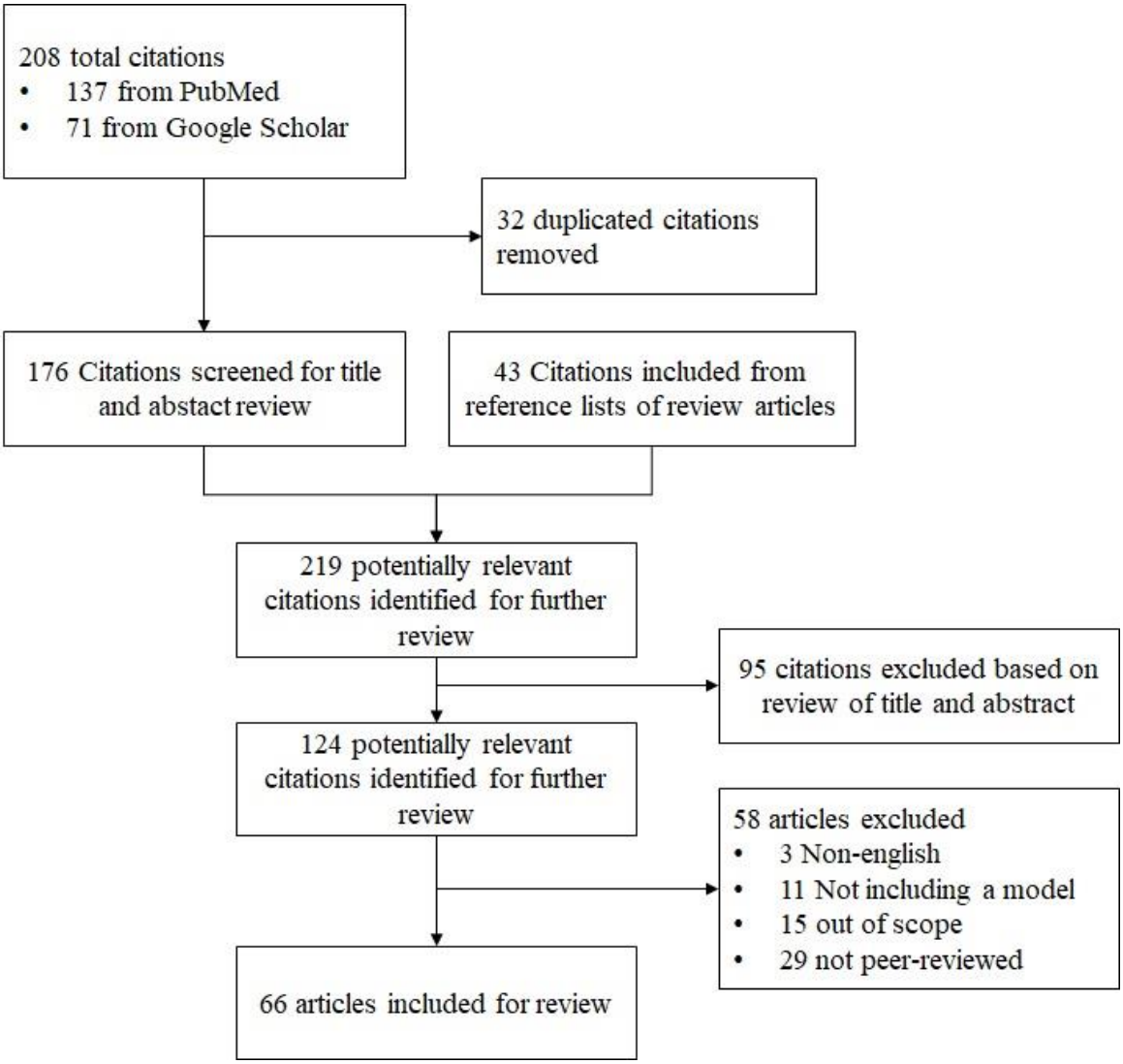


Figure 2.2. Flow diagram of the selection process

Figure 2.3 shows the number of papers (only covers studies included in the review) per year. Can be noticed that the number of papers has increased in recent years, reaching a peak in 2015 and 2016. Nevertheless, it's worth noting that the number of papers corresponding to 2017 is not the final count since the search was performed on 15th February 2017. In addition, the value for 2016 should also be considered with caution, since probably some 2016 papers were not yet indexed on early 2017, when this survey was carried out.

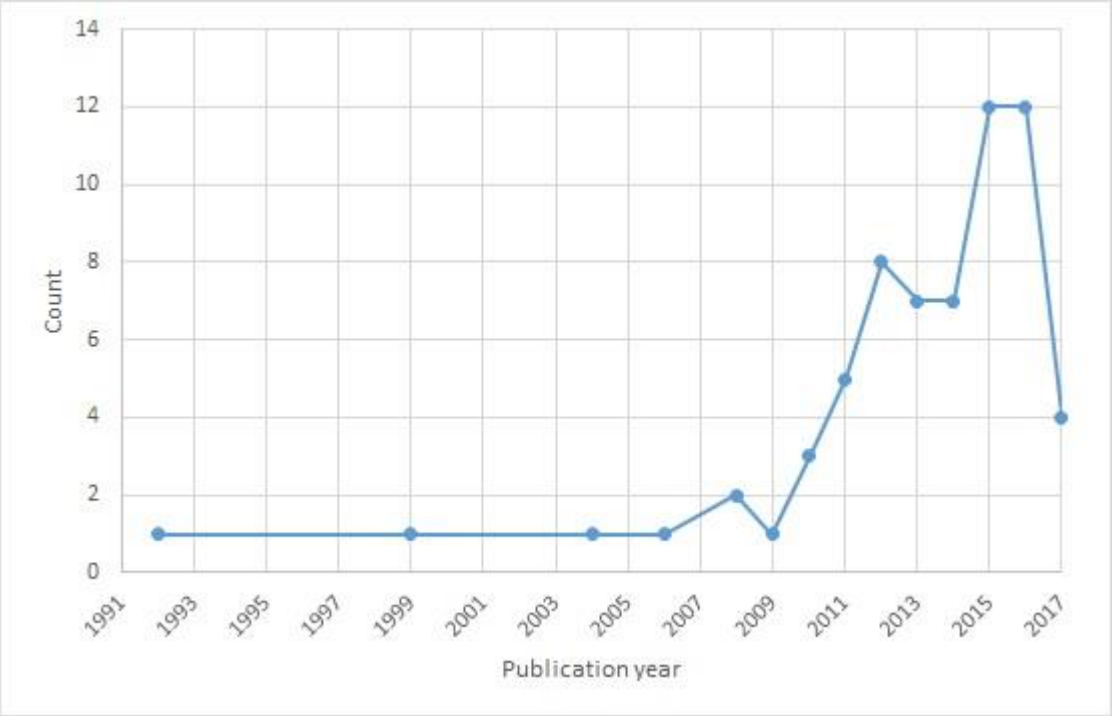


Figure 2.3. Number of publications per year

Data analysis methods

Readmission risk prediction has been addressed from different perspectives. Early studies have been conducted using conventional statistical multivariate modelling, which has been widely used in medical research. Fundamentally two different but related procedures from classical statistical modelling approaches have been used: regression analysis and survival analysis. Both techniques consist basically in defining a binary outcome (readmitted or not), and fitting a multivariate model over a given set of samples (aka instances), each composed of multiple features (aka variables, predictors or covariates) describing the facts related to the event, such as patient demographics, physiological state, etc.

Regression analysis estimates the probability of the target variable from some linear combination of the predictors. Binary logistic regression is a regression model where the target variable is

binary, that is, it can take only two values, 0 or 1. It is the most utilized regression model in readmission prediction, where the output is modelled as readmitted (1) or not readmitted (0). Survival models, on the other hand, use the features to compute an estimate of the time that passes before the event of interest (i.e. readmission) occurs.

In recent years, machine learning and data mining have emerged as approaches with big potentiality to improve the prediction ability of the readmission risk prediction models. Those techniques include classification algorithms widely used in multiple fields for predictive modelling of the most diverse tasks. However, machine learning techniques are not limited to the construction of the classifier, but they also encompass a wider set of techniques such as feature selection, variable discretization and normalization, missing value imputation, and many others.

Figure 2.4 presents a simplified taxonomy of procedures held in the studies included in this review. Figure 2.5 shows the evolution in time of the proportion of modelling techniques regarding the type of approach. A trend can be devised where machine leaning (ML) techniques emerged during the last years are gaining relevance over the classical techniques.

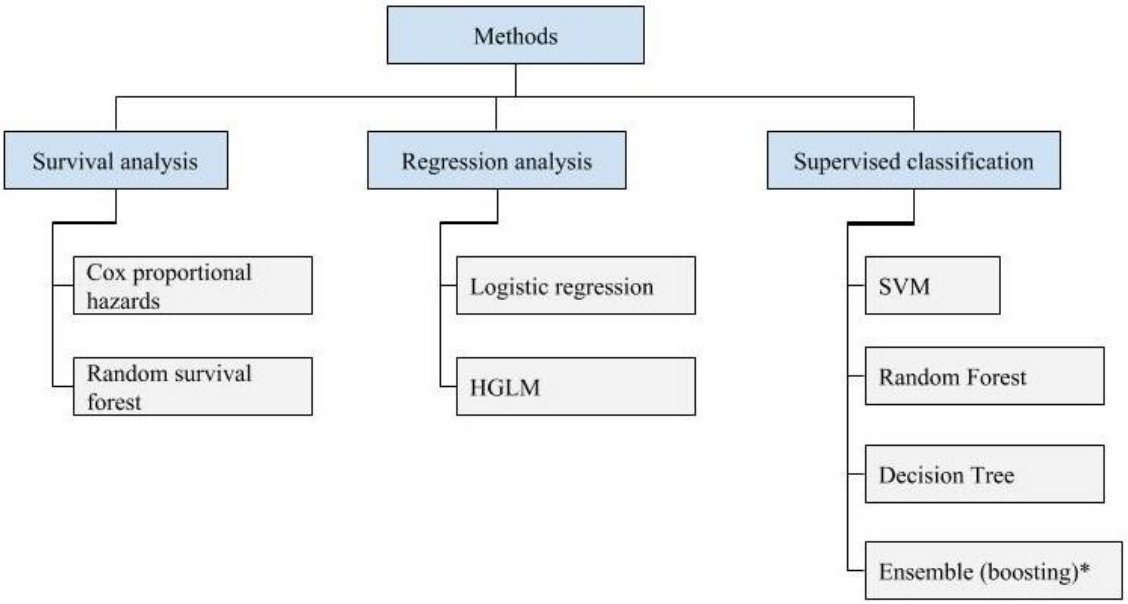


Figure 2.4. Taxonomy of data analysis methods

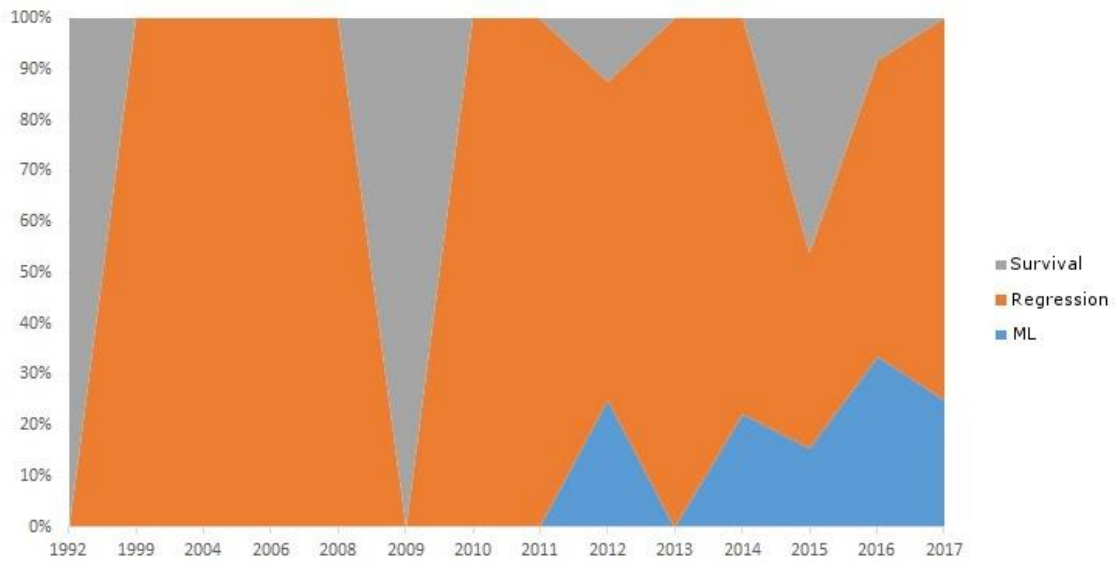


Figure 2.5. Distribution of methods per type and year (note that years without any publication included in the study are not present)

Feature selection techniques

Feature selection, aka feature subset selection (FSS), is a common practice in many data analytic fields aiming to identify the most significant variables of a dataset. In medicine, it is of special importance since it allows identifying the key factors associated to a disease, or a specific risk condition. Moreover, feature selection is of special interest for its ability to reduce the number of features, what simplifies model's complexity and reducing overfitting. This is particularly important in the clinical environment, where data acquisition is often related to costly procedures.

In the context of readmission risk prediction, feature selection is tightly related to the classification model used. Here we can clearly distinguish the classical approach, consisting on a regression analysis procedure preceded by a univariate parametric or model-free method that selects the most significant variables to be included in the model. The most extended feature selection procedure is to carry out a univariate/bivariate analysis by means of statistical tests such as Student's t-test, chi2, Wilcoxon or ANalysis Of VAriance (ANOVA), among others. Significant predictors from the univariate/bivariate analyses are then included in the final model. Variables with p-values lower than a pre-established threshold (typically 0.001 [AbdelRahman2014] though it may change from one study to another) are considered statistically significant features. A more refined hybrid approach that includes a stepwise [Greenland1989] approach is widely utilized with regression-based models.

The (logistic) regression with a multi-step heuristic approach consists in the following steps:

1. Univariate variable selection (optional): For every feature, a univariate logistic regression model is built. Only features with a p-value from a Likelihood Ratio test below a specified threshold are retained.
2. A multivariate logistic regression is built on a stepwise fashion. There are two basic approaches:
   a. Forward selection: initializes the model with an empty set of selected features and iteratively adds features, retaining only those whose addition shows statistically significant improvement of the fit.
   b. Backward elimination, which initializes the model with the whole set of features proceeding by iteratively removing the features that do not improve (or do worsen) the model fit.
3. A final logistic regression model is built using the features selected in previous steps.

There is a wide variety of feature selection techniques that are utilized in the studies following data mining approaches for readmission risk prediction. Abdelrahman et al. [AbdelRahman2014] systematically evaluated different feature selection and ranking methods such as wrapper subset selection, information gain, gain ratio and symmetrical uncertainty. Cai et al. [Cai2016] used a correlation-based feature selection (CBFS) method for selecting the most significant features. Some other authors follow an embedded feature selection approach, which consists in conducting the feature search within the classifier itself, as part of the learning process. Nevertheless, many of the DM papers do not report the use of any specific feature selection approach.

Class imbalance

In readmission prediction, as well as in many other fields (e.g. fraud detection or fault diagnosis), instances of the event of interest are outnumbered by the "other events" instances. In supervised classification, data imbalance occurs when the *a priori* probabilities of the classes are significantly different, i.e. there exists a minority (positive) class that is underrepresented in the dataset in contrast to the majority (negative) class. Often the goal is the detection of the minority class instances, while the majority class is the collection of "other things" in the universe where classification is desired. Most classification algorithms assume equal *a priori* probabilities for all classes, so when the training dataset is imbalanced, the resulting model is biased towards the majority class.

Readmission prediction is an intrinsically imbalanced problem. All-population 30-day readmission rate is estimated in a 20% [Jencks2009], although it varies greatly depending on multiple factors (e.g. readmission threshold, subpopulation characteristics etc.). The level of class

imbalance of a dataset is given by the imbalance ratio (IR), so that a IR of 1:10 expresses that for each sample of the positive class, there are 10 samples of the negative class.

Unlike most classification algorithms used in machine learning (e.g. decision trees or linear discriminant analysis), linear regression is not affected by class imbalance (at least for modestly imbalanced data) [Crone2012]. Thus, while regression-based approaches do not suffer the class imbalance problem, it is a relevant problem that arises when machine learning approaches are implemented.

As shown in Table 2.3, most studies using machine learning algorithms do not report the use of any procedure correcting class imbalance. Among those who actually do something, resampling is the most utilized strategy to overcome class imbalance, either subsampling the majority class or oversampling the minority class.

Table 2.3. Class imbalance addressing methods in readmission risk prediction

Paper	Class imbalance addressing	IR
amalakuhan2012	-	1:2.1
Au2012	-	1:5.3
abdelrahman2014	-	1:5.3
walsh2014	sub-sampling	1:14
Yu2015	-	1:5.3
Zheng2015	random oversampling	1:4.6
cai2016	-	-
fisher2016	-	1:4
turgeman2016	boosting	1:3.6
Mortazavi2016	sub-sampling/oversampling/weighting[1]	1:6.8
bergese2017	-	1:45.5

[1]weighting chosen for final model

Most studies included in this review employ basic random over or sub sampling techniques. There are some cases [[Zheng2015] using more sophisticated resampling methods, such as SMOTE (Synthetic Minority Oversampling Technique) [Chawla2002]. In [Mortazavi2016] different methods are compared, including resampling (oversampling and subsampling) and cost-sensitive learning (weighting). According to the authors, weighting achieved the best results, and, hence, they included it in their final model. By contrast, Turgeman et al. [Turgeman2016] used boosting, which is an ensemble meta-algorithm also known to overcome the bias towards the majority class.

2.1.5. Discussion

Different approaches reported in different studies cannot be directly compared since each study has its own particular characteristics in population, definition of the problem, computational methods and evaluation metrics.

The Area Under Receiver Operating Characteristic (ROC) Curve or c-statistic is the standard *de facto* metric for measuring the discrimination ability of readmission risk prediction models. The main goal of some papers is to identify predictors associated to readmission. Often, this kind of studies do not provide the c-statistic as the overall performance metric. Regarding the discrimination ability of the models, most papers report modest AUC scores, mostly below 0.75,

in agreement with the results presented in [Kansagara2011]. Nevertheless, 16 models reported AUC scores above 0.75 and 21% of the studies did not report AUC metric. However, discrimination of the models is not comparable since it is greatly influenced by the population subject to study as well as by factors such as readmission length threshold.

The most widely used readmission threshold is 30-day. It is used by 75% of the reviewed papers, although we found time spans ranging from 48 hours to 1 year. Consequently, readmission rates also vary depending on this threshold. Longer readmission thresholds are related to higher readmission rates and vice versa. However, other factors such as the population subject to study or the type of clinical study, can greatly influence this indicator.

There exists a discussion about what separates "traditional" applied statistics from machine learning. Even though we consider that there is considerable overlap among them, in this work we separate "traditional" methods from data mining or machine learning techniques. All the same, we are aware that many researchers accept that regression analysis (which we have excluded from machine learning) does actually make part of machine learning. We found very few studies were both approaches are compared under the same conditions. Most salient is the work by Futoma et al. [Futoma2015], where a comparison of logistic regression, stepwise logistic regression, random forest, SVM and deep neural networks is presented. Authors conclude that overall predictive accuracy can be improved moving from standard logistic regression to more complicated non-linear models although resulting models may be difficult to tune and interpret.

2.2. Heart Failure readmission risk

Heart failure (HF) is a clinical syndrome characterized by typical symptoms (e.g. breathlessness, ankle swelling and fatigue) caused by a structural and/or functional cardiac abnormality, resulting in a reduced cardiac output and/or elevated intra-cardiac pressures at rest or during stress. Demonstration of an underlying cardiac cause is central to the diagnosis of HF. This is usually a myocardial abnormality causing ventricular dysfunction or abnormalities of the valves, pericardium, endocardium, heart rhythm and conduction [Ponikowski2016]. The prevalence of HF is approximately 1–2% of the adult population in developed countries, rising to ≥10% among people >70 years [Mosterd2007]. Cardiovascular diseases such as HF have the highest 30-day readmission rates [Jencks2009]. In USA, it is estimated that almost half of the Medicare beneficiaries are readmitted within 6 months after a hospitalization for congestive HF [Krumholz1997].

Over the last 30 years, improvements in treatments and their implementation have increased survival but the outcome often remains unsatisfactory. Most recent European data (ESC-HF pilot

study) demonstrates that 12-month mortality rates for HF patients are between 7% and 17%, and the 12-month hospitalization rates are between 32% and 44% [Ceia2002].

The negative effects of cardiovascular disease (CVD) are not limited only to the individual's health. When CVD causes hospitalizations, short-term expenses tend to be extremely high. Costs include ambulance rides, diagnostic tests, hospital stays, and immediate treatment that may include surgery. Short-term costs aside, CVD remains expensive for the long-term due to the price of drugs, tests to monitor the progress of the disease, and frequent doctor appointments [HSA2011]. The high cost of CVD is compounded by the lack of productivity and income that such patient may have [Anand2006]. Additionally, high rates of readmission after hospitalization for HF impose tremendous burden on patients and on the healthcare system.

In this context, predictive models facilitate the identification of patients at high risk for hospital readmissions and potentially enable direct specific interventions toward those who might benefit most by identifying key risk factors. However, current predictive models using administrative and clinical data discriminate poorly on readmissions [Kansagara2011]. That is the reason why some studies have been developed in order to try to define whether machine learning would enhance prediction [Mortazavi2016].

Nevertheless, it remains unclear whether it is possible to predict and prevent hospital readmission and mortality in patients with HF. Currently, there are several healthcare programs where patient monitoring is carried out, so that clinicians can check patients' progress [Riley2009, Cleland2005, Lusignan2001, U4H2017].

In some cases, clinicians define some simple rules, so that they can get some alerts that may indicate the deterioration of a patient [Mosterd2007]. In other cases, as shown in Mobiguide EU project, the system implements the local clinical guidelines and extend them to guide patients during their daily life [Ceia2002]. However, due to the lack of time of clinicians and the lack of suitable IT solutions, clinicians do not exploit the monitored information.

2.2.1. Related Studies

Most studies follow methodologies based on statistical approaches, where logistic regression and Cox proportional hazard models are the most extended techniques. Among the studies that follow a traditional regression-based approach, a common procedure for dimensionality reduction is to apply wrapper feature selection techniques known as stepwise procedures, namely forward selection, backwards elimination or stepwise regression. These techniques consist in sequentially adding or removing features into/from a feature subset according to the estimated performance of a multivariate regression model. Often, a previous univariate feature selection is performed, where not-significant features (those with a p-value greater than a given threshold) are removed. With

this preliminary step, it is intended that only significant features are passed to the following feature selection step

Some authors [Mortazavi2016] have pointed out that other approaches for risk prediction, such as machine learning, can be utilized to achieve better performance, comparing the predictive performance of traditional statistics methods (logistic regression and Poisson regression) and machine learning methods (Random Forest, Boosting and SVM). In [Kadi2017] a review of studies on the application of data mining techniques in cardiology is presented where Neural Networks, Decision Trees and SVMs are identified as the most frequently used predictive techniques. Some recent studies [Au2012, [Zheng2015, Turgeman2016] make use of machine learning techniques, where Support Vector Machine (SVM) and Random Forest (RF) are the most utilized algorithms. In [Au2012] authors undertook RF analysis for predicting unplanned readmission or death within 30 day of discharge after a HF hospitalization. Prediction ability of the features selected by RF were compared with the variables in the LACE score [Walraven2010]. On the other hand, [Turgeman2016] presented an ensemble algorithm combining boosted decision trees and SVM.

Zheng et al. [[Zheng2015] studied the risk prediction of hospital readmissions in HF patients using metaheuristic and data mining approaches. Authors indicate the need of compensation strategies that address the class imbalance, suggesting over-sampling techniques such as SMOTE.

2.3. Conclusions

Although classical statistical techniques have prevailed and are still popular techniques in medical studies, machine learning approaches have emerged in the last years as a promising set of techniques that can improve the predictive ability of readmission risk prediction models. Still, univariate and stepwise regression are the dominant modelling methods, while additional feature selection methods are infrequent. Within the studies that use data mining techniques, we found that class imbalance is only addressed in a minority of them, though it is a major shortcoming of conventional machine learning.

Regarding feature selection techniques, we observed that conventional univariate approaches are the most extended. Stepwise regression is also an extended feature reduction procedure intended to produce parsimonious models. Recent studies introducing machine learning techniques report promising results and anticipate advantages over classical methods. Nevertheless, further comparative studies are needed to assess the real impact of this techniques in the domain of readmission risk prediction. Moreover, further areas of machine learning such as feature selection, class imbalance or variable discretization remain still largely unexplored.

In readmission prediction like in many other medical fields, data is intrinsically class-imbalanced. General 30-day readmission rate varies from 11% to 25% [Jencks2009, Silverstein2008, Desai2012] depending on the population subject to study. In supervised classification, class imbalance imposes a bias towards the majority class that leads to a higher misclassification rate of the minority class instances (which are usually the most interesting ones from the practical point of view [Lopez2013]). Although this effect is not that significant for regression analysis, most of machine learning techniques assume equal a priori probability for all the classes, so that class imbalance is an issue that must be addressed.

Feature selection is another challenge that must be tackled when applying machine learning techniques in the readmission prediction domain. Feature reduction is of great importance since it reduces noise, avoids collinearity and reduces the cost since in a clinical context, measuring variables may be expensive. Traditional procedures include univariate parametric methods such as t-test, chi2 or regression [Bradford2016] and wrapper methods, mainly stepwise regression. Nevertheless, machine learning methods can be used to improve model's performance thanks to their ability to leverage all available data and their complex relations.

Chapter 3

Dataset

This chapter is devoted to the description of the datasets that support the experimental works of this Thesis. These datasets were directly provided by physicians through their information and communication (ITC) services, such as the one of Osakidetza, the Basque public health service provider, and the Hospital José Joaquín Aguirre of the Universidad de Chile. We received the data after being anonymized, so that all issues of ethics and data privacy were already solved by the providers. In the following sections, we present the three datasets that we used, named according to their place of origin: University Hospital of *Araba*, Hospital José Joaquín Aguirre of the Universidad de *Chile*, and Hospital of *Basurto*.

3.1. University Hospital of Araba Dataset

The original dataset was collected by Dr. Ariadna Besga during June 2014 and was composed of 802 admissions registered at the two hospitals that form the University Hospital Araba, namely Hospital Txagorritxu and Hospital Santiago Apostol. After filtering the Emergency Department (ED) admissions, the dataset was composed of 462 admission samples of 360 unique patients.
The final curated dataset used for the experiments was presented by Besga et al. in [Besga2015]. It encompases data of 360 patients divided into four groups, namely:

1. Case management (CM), which is the most general category of data encompassing all categories not covered by the specific categories
2. Patients with chronic obstructive pulmonary disease (COPD),
3. Heart failure (HF) and
4. Diabetes Mellitus (DM).

For each patient, a set of 97 variables were collected, divided into four main groups: i) Sociodemographic data and baseline status, ii) Personal history, iii) Reasons for consultation/

Diagnoses made at ED and iv) Regular medications and other treatments. The dataset contains missing values. Table 3.1 shows the distribution of the number of variables of each category.

Table 3.1. Distribution of variables by category from the University Hospital of Araba

Variable	**No. (%) of variables n=96**
Sociodemographic and baseline status	4 (4.2)
Personal history	43 (44.8)
Reasons for consultation	16 (16.7)
Regular medications	33 (34.3)

In order to build our model following a binary classification approach, the target variable was set to *readmitted*/*not readmitted*. Those patients returning to ED within 30 days after being discharged are considered readmitted (value=1), otherwise are considered as not readmitted (value=0).

It is noteworthy that one patient returning the first day and another returning the 30th are both considered as *readmitted*. On the other hand, a patient returning the 31th day is considered as *not readmitted*, while in practice underwent a readmission Table 3.2 shows the distribution of readmission rate across different subpopulations.

Table 3.2. Comparative information about the subpopulations of the dataset from the University Hospital of Araba

	Overall no. of patients *n=360*	Readmission within 30 days, no. (%) of patients	
		No n=296 (82.2)	Yes n=64 (17.7)
Case management	94 (26.1)	73 (77.7)	21 (22.3)
Heart failure	70 (19.4)	62 (88.6)	8 (11.4)
Chronic obstructive pulmonary disease	80 (22.2)	64 (80)	16 (20)
Diabetes mellitus	116 (32.2)	97 (83.6)	19 (16.4)

We observe that readmission rate varies greatly depending on the subpopulations, ranging from 11.4% to 22.3% for HF and case management respectively. We also notice that data is not well

balanced in terms of different population stratum. Figure 3.1 shows the distribution of some demographic features across the mentioned subpopulations. It can be appreciated (Figure 3.2) that the class distributions are imbalanced, though the actual imbalance ratios vary greatly between population strata.

Table 3.3 reproduces the most significant variables of each subpopulation according to a t-test for the significant differences among the mean values between readmitted and non-readmitted patients as (Besga et al. in [Besga2015]).

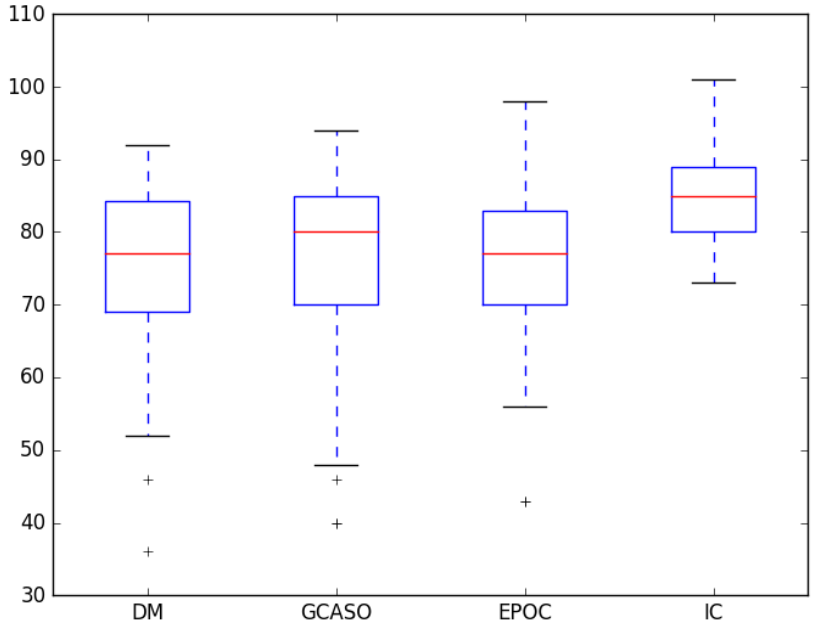


Figure 3.1. Boxplots of age at admission time across different population stratum

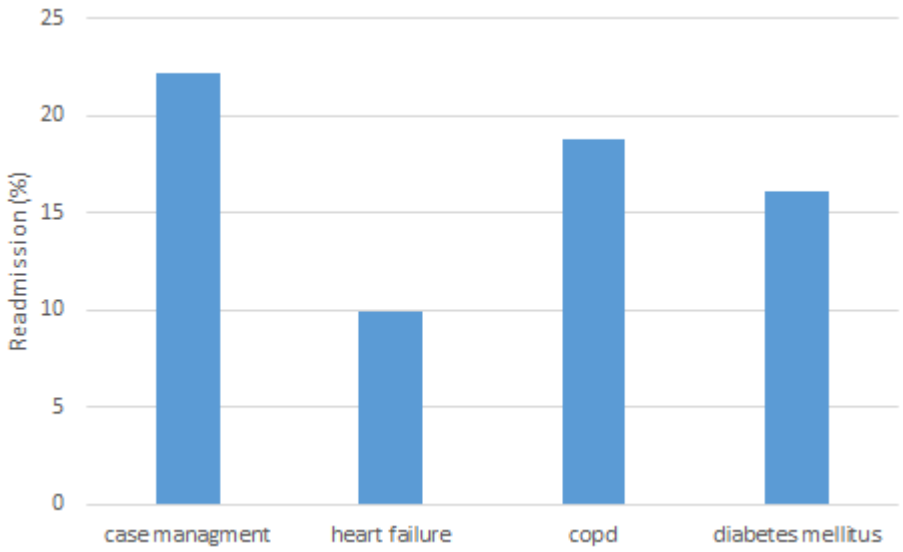


Figure 3.2. Readmission rate across different population stratum

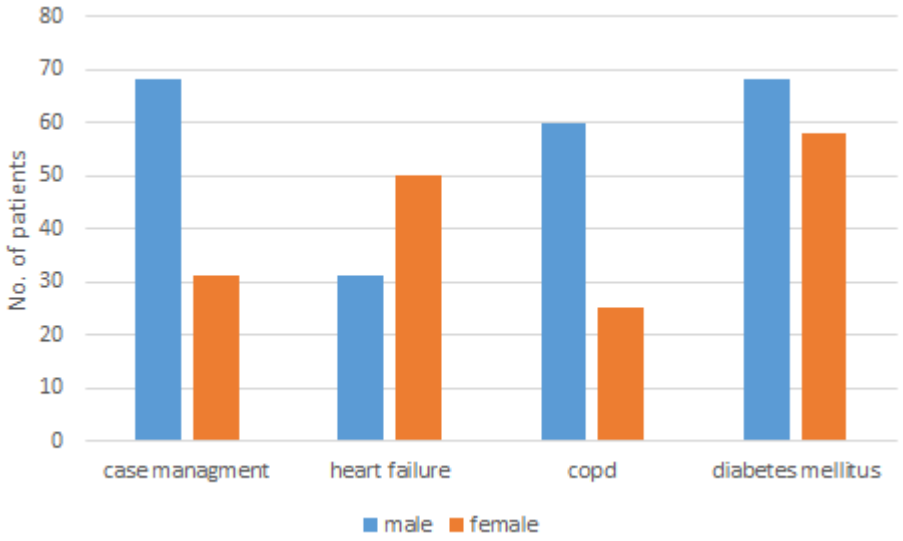


Figure 3.3. Distribution of number of patients per sex in the University Hospital of Araba dataset across different population stratum

Table 3.3. Most significant variables for each population stratum in the University Hospital of Araba dataset according to t-test, extracted from [Besga2015]

Feature	p-value
Case management	
Patient age on admission	0.0054
Considered useful to make a follow-up call	0.0087
Acute myocardial infarction	0.0066
Thyroid disease	0.0013
Use of antipsychotics	0.0039
Use of inhalers	0.0034
Diagnosis of COPD	0.0021
Heart failure	
Acute myocardial infarction	0.0001
Dementia	0.0001
Number of medications prescribed on ED discharge	0.0000
Diagnosis of gastrointestinal illness	0.0020
COPD	
Dementia	0.0071
Depression	0.0038
Use of anticoagulants	0.0071
Genitourinary problems	0.0021
Use of opioids	0.0021
History of falls	0.0071
Diabetes mellitus	
Organic lesions	0.0006

3.2. University of Chile Dataset

This dataset is composed of ED admission events of 102,534 patients divided into 2 groups, namely adults and paediatrics, which amounts to 156,120 admission cases recorded between January 1st, 2013 and August 31, 2015 from the electronic medical records of the Hospital José Joaquín Aguirre de la Universidad de Chile. At admission time a set of 17 variables were collected. The variables or features are categorized into three main groups: i) Sociodemographic data and baseline status, ii) Personal history and iii) Reasons for consultation or diagnoses made at admission. The dataset contains missing values.

3.2.1. Data pre-processing

Data was provided in a large ASCII text file containing 156,120 admission records corresponding to 102,534 different patient identities. After parsing the data, we built a dataset combining admission and patient-related data. Next, we cleaned the data by removing inconsistent and missing samples. Missing values were imputed using the arithmetic mean for continuous variables and the mode for categorical variables.

For each admission of a patient to the ED we calculated the number of days elapsed since his last visit. In order to build our model following a binary classification approach, the target variable meaning was set to readmitted/not readmitted. Those patients returning to the ED within 72 hours after being discharged were considered readmitted, otherwise they were considered not readmitted.

Notice that a patient returning the very first day after discharge and another one returning the third day are both considered as readmitted. On the other hand, a patient returning the 73rd hour from discharge is considered as not readmitted.

After removing inconsistent and missing samples the dataset was composed of 99,858 instances.

3.2.2. Description

Table 3.4 shows the distribution of admissions and readmission records according to gender, broad pathology class (general medicine, traumatology, paediatric, and gyneco-obstetrics), and the assigned triage. Class distribution shown in Table 3.4 indicates an imbalance ratio (IR) of approximately 1:28, which is a strong case of class imbalanced data. Notice that most admissions correspond to general medicine, followed by the paediatric admissions, however if we consider readmissions, the paediatric segment of the population is responsible for more than half (56%) of

the readmissions, with some implications on the causes. Note also that triage III accounts for most admissions and readmissions (75%).

Figure 3.4 shows the distribution of readmission class among some attributes of our dataset. Readmissions (shown as green columns) are much less frequent than normal admissions, i.e. the dataset is heavily imbalanced. Some details, such as the greater frequency of readmission for people in the age range 20-30, can be appreciated. Still, there is not enough evidence that allow to use a single variable for the prediction.

The description of each patient contains a categorical variable encoding the admission motivation, this encoding into more than 500 topics is given by the electronic medical record implementation. Table 3.5 contains the more frequent causes of admission and readmission, those accounting for 1.5% of the cases or more. The non-informative category "OTHERS" is the most frequent, and the most frequent causes for admission appear also as causes of readmission. In our current implementation, this variable has been encoded with a vector of binary valued features, one per each admission motivation category. This approach is equivalent to unfold a subspace of dimension 500 to represent the variable motive. Additional features correspond to the encoding of the triage, demographics variables such as age, sex, adult or paediatric patient, and physiological variables such as blood pressure, temperature, heart rate, respiratory rate, glucose levels, and others. Hence, feature vector dimension is greater than 500, which is an already very high dimension. Table 3.6 contains the descriptive statistics of the main variables of the University of Chile dataset. In some variables, we give the mean and standard deviation *mean (SD)*, for other we give the number of instances and the percentage in the whole population. Figure 3.5 shows the histograms of the causes for admission (a) for the entire population, (b) for the non-readmitted patients, and (c) for the readmitted patients. It can be appreciated that the OTHERS motivation is rather salient in all situations, while the next five most frequent motivations are common to the readmitted and non-readmitted patients, though in different orders of importance, exception made of the fever cause, which is much more prevalent in readmissions.

Table 3.4. Statistics of ED admissions from 2013 to 2016. Age mean and standard deviation. Remaining rows give the number of records and the percentage relative to the total Columns correspond to no readmission, readmission, and total number of records. By rows, we give the total number and percentage of the total population of the occurrence of each kind of gender, class of pathology, and triage assigned upon arrival.

	72h readmission		**Total**
	No (n=148617)	Yes (n=5674)	n=154291

Age	Years (%)	33.3 (24.8)	22.2 (24.6)	32.9 (24.9)
Gender	Male	69106 (46.5)	2832 (49.9)	71983 (46.6)
	Female	79511 (53.5)	2842 (50.1)	82353 (53.4)
Pathology	General Medicine	91566 (61.6)	2375	93941 (60.9)
	Traumatology	16651 (11.2)	325	16976 (11)
	Paediatric	39999 (26.9)	2964	42963 (27.8)
	Gynaeco-obstetrics	401 (0.3)	10	411 (0.3)
Triage	I	649 (0.4)	8	567 (0.4)
	II	17280 (11.6)	501	17781 (11.5)
	III	111310 (74.9)	4309	115619 (74.9)
	IV	19057 (12.8)	848	19905 (12.9)
	V	321 (0.2)	8 (0.1)	329 (0.2)

Table 3.5. Distribution of causes of admission and readmission cases. GAP general abdominal pain, 1/3DF up to three days fever; 24HF 24 hours; fever; HA headache; D diarrhoea; T throwing up; EP epigastric pain; LuP lumbar pain; GD general discomfort; LegP leg pain; AD acute dyspnoea.

Admission		**Readmission**	
Motive	**%**	**Motive**	**%**
OTHER	14.22	OTHER	30.13
GAP	8.21	GAP	8.20
24HF	5.53	1/3DF	5.40
COUGH	5.47	COUGH	4.28
HA	4.93	24HF	4.10
1/3DF	3.65	HA	3.04
GD	2.33	D	2.59
EP	2.22	T	2.43
T	2.21	EP	1.86
D	2.16	LuP	1.51
LegP	2.11		
LuP	2.06		
AD	1.57		
FP	1.55		
NAUSEA	1.44		

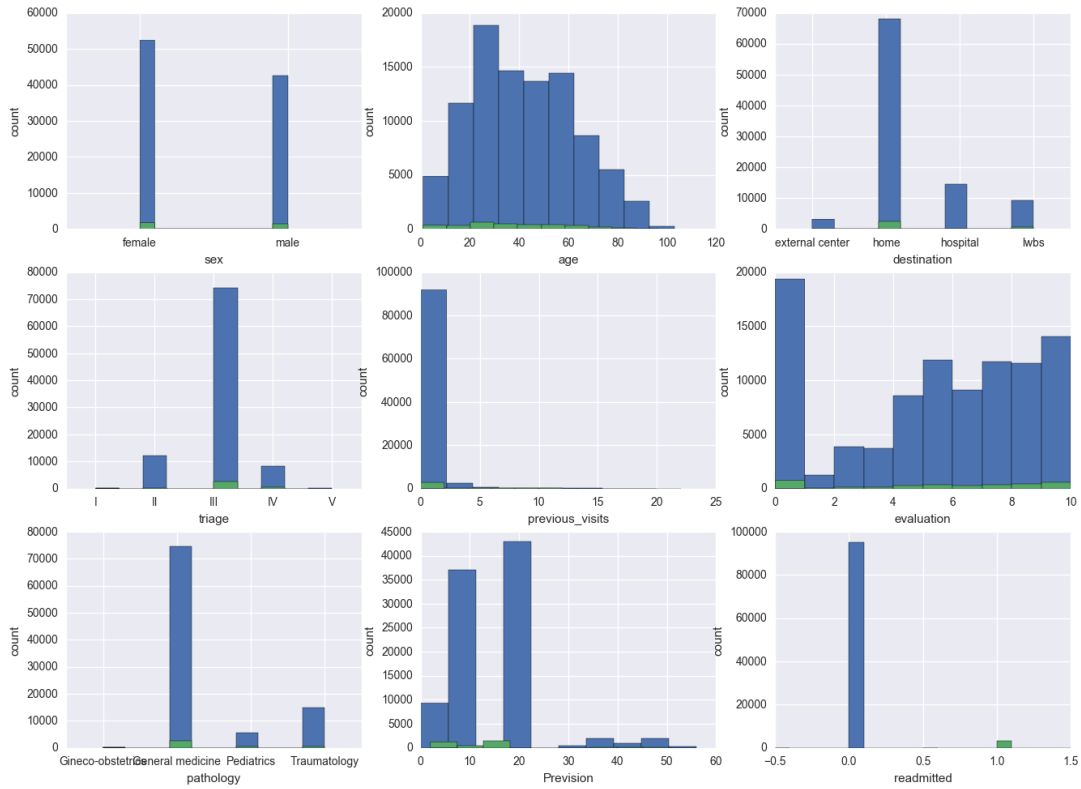
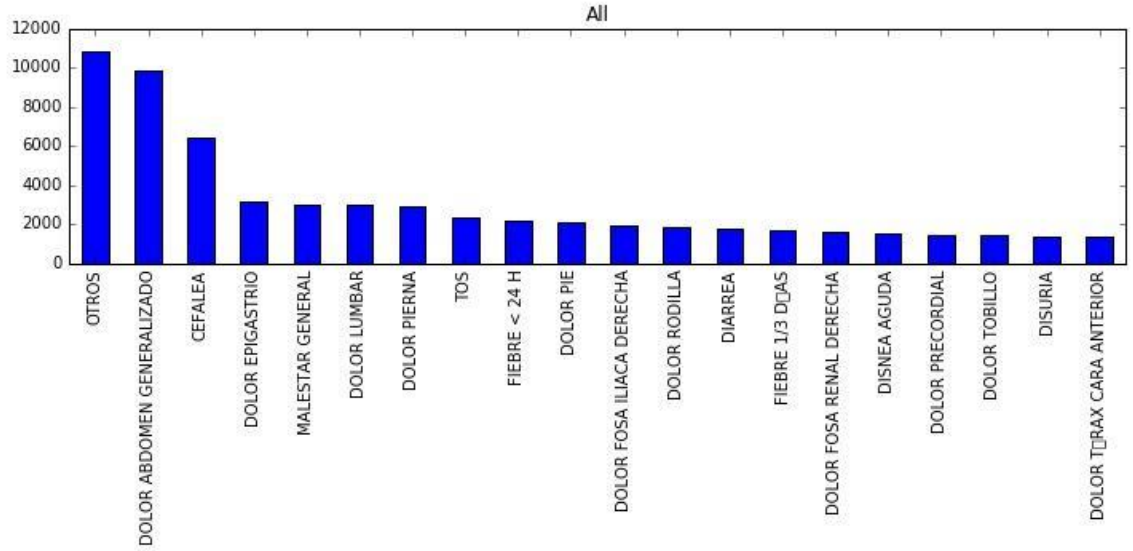


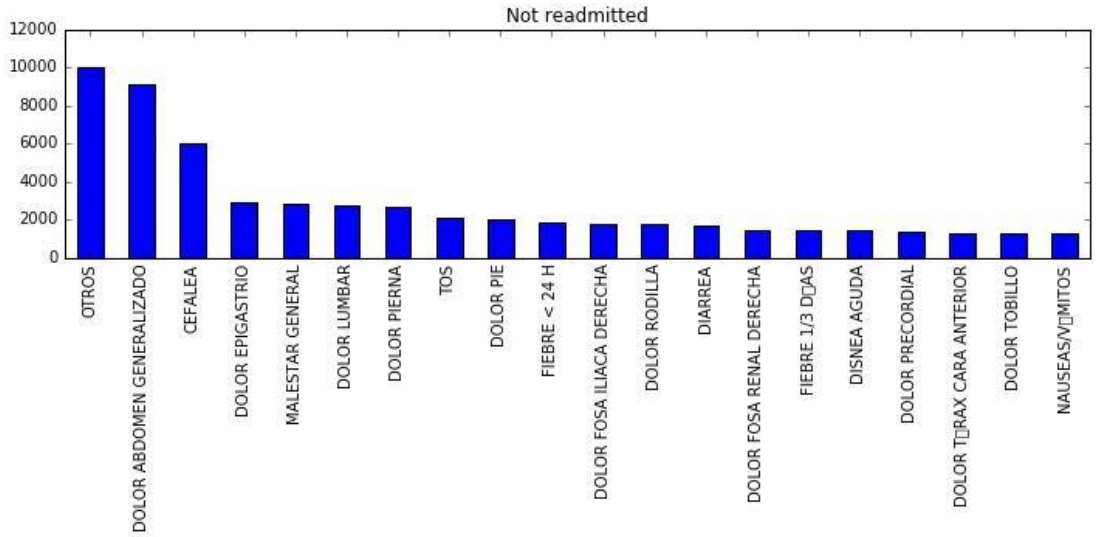
Figure 3.4. Distribution of readmission class among different attributes (readmission in green, regular admissions in blue) in the University of Chile dataset. From left to right, top to bottom: sex, age, destination after discharge, triage, previous visits, evaluation, pathology, prevision, and readmission

Table 3.6. Descriptive statistics of the Hospital of Chile variables

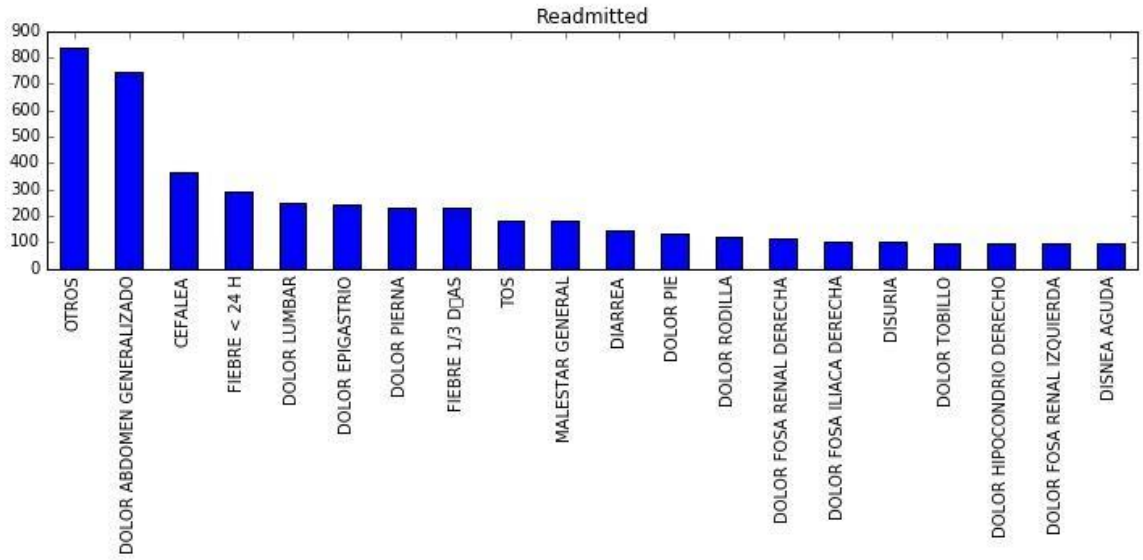
Variable	All patients n=99858	Readmitted n=3425	Not readmitted n=96433
age, mean (SD)	41.0 (22.4)	36.1 (22.9)	41.2 (22.4)
male sex (%)	44956 (45.0)	1624 (1.6)	43332 (43.4)
daytime (%)	69321 (69.4)	2171 (2.2)	67150 (67.2)
evaluation, mean (SD)	5.0 (3.3)	4.8 (3.5)	5.0 (3.3)
fragility idx, mean (SD)	0.0 (2.5)	0.0 (2.3)	0.0 (2.5)
triage (%)			
I	182 (0.2)	2 (0.0)	180 (0.2)
II	12694 (12.7)	317 (0.3)	12377 (12.4)
III	77813 (77.9)	2718 (2.7)	75095 (75.2)
IV	9131 (9.1)	387 (0.4)	8744 (8.8)
V	38 (0.0)	1 (0.0)	37 (0.0)
pathology (%)			
Gineco-obstetrics	236 (0.2)	6 (0.0)	230 (0.2)
General medicine	77192 (77.3)	2458 (2.5)	74734 (74.8)
Pediatrics	7094 (7.1)	563 (0.6)	6531 (6.5)
Traumatology	15336 (15.4)	398 (0.4)	14938 (15.0)
destination (%)			
External center	3372 (3.4)	116 (0.1)	3256 (3.3)
Home	71999 (72.1)	2703 (2.7)	69296 (69.4)
Hospital	14700 (14.7)	61 (0.1)	14639 (14.7)
Left without being seen	9787 (9.8)	545 (0.5)	9242 (9.3)
reason for consultation (%)			
Cephalea	6421 (6.4)	192 (0.2)	6229 (6.2)
Pain - abdomen gen.	9861 (9.9)	404 (0.4)	9457 (9.5)
Pain - epigastrium	3177 (3.2)	143 (0.1)	3034 (3.0)
Pain - lumbar	2964 (3.0)	107 (0.1)	2857 (2.9)
Pain - foot	2909 (2.9)	92 (0.1)	2817 (2.8)
General malaise	3027 (3.0)	78 (0.1)	2949 (3.0)
Other	10867 (10.9)	374 (0.4)	10493 (10.5)
...			
saturation, mean (SD)	96.6 (9.6)	96.2 (12.1)	96.6 (9.5)
tad, mean (SD)	74.1 (22.3)	67.6 (29.4)	74.3 (21.9)
tas, mean (SD)	125.8 (35.9)	114.5 (48.8)	126.2 (35.3)
temperature, mean (SD)	35.9 (4.5)	35.5 (5.9)	35.9 (4.4)
heart rate, mean (SD)	87.2 (22.3)	92.7 (29.1)	87.0 (22.0)
breath rate, mean (SD)	17.0 (5.6)	15.1 (7.6)	17.0 (5.5)1
Prevision (%)			
2	5943 (6.0)	180 (0.2)	5763 (5.8)
5	3641 (3.6)	108 (0.1)	3533 (3.5)
6	27903 (27.9)	1022 (1.0)	26881 (26.9)
9	11060 (11.1)	432 (0.4)	10628 (10.6)
18	44464 (44.5)	1468 (1.5)	42996 (43.1)
35	1011 (1.0)	30 (0.0)	981 (1.0)
37	1103 (1.1)	33 (0.0)	1070 (1.1)
48	2074 (2.1)	70 (0.1)	2004 (2.0)
...			


All
12000
10000
8000
6000
4000
2000
0
OTROS
DOLOR ABDOMEN GENERALIZADO
CEFALEA
DOLOR EPIGASTRIO
MALESTAR GENERAL
DOLOR LUMBAR
DOLOR PIERNA
TOS
FIEBRE < 24 H
DOLOR PIE
DOLOR FOSA ILIACA DERECHA
DOLOR RODILLA
DIARREA
FIEBRE 1/3 D□AS
DOLOR FOSA RENAL DERECHA
DISNEA AGUDA
DOLOR PRECORDIAL
DOLOR TOBILLO
DISURIA
DOLOR T□RAX CARA ANTERIOR

(a)


Not readmitted
12000
10000
8000
6000
4000
2000
0
OTROS
DOLOR ABDOMEN GENERALIZADO
CEFALEA
DOLOR EPIGASTRIO
MALESTAR GENERAL
DOLOR LUMBAR
DOLOR PIERNA
TOS
DOLOR PIE
FIEBRE < 24 H
DOLOR FOSA ILIACA DERECHA
DOLOR RODILLA
DIARREA
DOLOR FOSA RENAL DERECHA
FIEBRE 1/3 D□AS
DISNEA AGUDA
DOLOR PRECORDIAL
DOLOR T□RAX CARA ANTERIOR
DOLOR TOBILLO
NAUSEAS/V□MITOS

(b)



(c)

Figure 3.5. Histograms of the 20 most common reasons for consultation for a) all admissions, b) non-readmissions and c) readmissions

3.3. Hospital of Basurto Dataset

3.3.1. Context

Since 2014 up to March 2017, 193 HF patients were included in a telemonitoring program at the OSI Bilbao-Basurto, Spain. During the program, patients were monitored using validated devices that collected health status data as well as self-reported data from questionnaires.
The prospective study included 193 patients that underwent a hospitalization or emergency visit due to decompensation of heart failure (with need and administration of diuretics) and were diagnosed of HF by a cardiologist. Patients with myocardial infarction or percutaneous coronary intervention in the last 3 months and patients with a coronary artery bypass graft, valve replacement or correction in the last 6 months were excluded (refer to Appendix A for further details) from the study.

3.3.2. Preprocessing

Data was provided spread across several spreadsheet files, containing from few hundred entries up to about a million. Datasets were related using the pseudonymized patient's unique identifier.
Data contained 704 admission records corresponding to 193 different patient identities, along with up to 900,000 monitorization data entries. After parsing the data, we built a dataset combining admission and patient-related data. Next, we cleaned the data by removing inconsistent and missing samples. Missing values were imputed using the arithmetic mean in case of continuous variables and the mode in case of categorical variables.
The primary prediction outcome was readmission or mortality due to heart failure within 30 days after discharge. A committee of physicians studied each potential readmission to determine whether the primary cause was related to HF. A new binary variable named *readmission* was created, which encodes whether the patient was readmitted within the first 30 days from hospital discharge.
We defined admission event as the unit of analysis. Admissions corresponding to the same patient were considered separately if the time between hospitalizations was greater than 30 days. Planned admissions (those hospitalizations scheduled by physicians beforehand) were excluded from the dataset. For each admission instance in the dataset, clinician staff recorded monitorization data consisting of diverse medical parameters (e.g. blood pressure, heart rate, weight) and self-reported information gathered using a questionnaire.

3.3.3. Description

Each instance in the dataset contained variables grouped in i) baseline status data of the patient, ii) monitorization data and iii) other meta-data. A complete list of variables is shown in Table 3.7.

- *Baseline Status*: data that corresponds to the first seen by a physician when entering the study. This information includes patient demographic information, such as year of birth and gender, but also clinical data, such as the hospitalization date, type of heart disease and hemodynamic parameters such as heart rate, systolic and diastolic blood pressure, blood check-up data, pharmaceutical treatment and other non-cardiac comorbidities.
- *Monitored Data*: data that is monitored by the patient remotely every week (with a frequency that varies from 3 to 7 days per week), which contains patient vital signs (such as heart rate, systolic/diastolic blood pressure, weight and oxygen saturation) and a questionnaire about the patient condition (e.g. *During the last 3 days, have you been having your medications as prescribed?*).
- *Meta-data*: Includes data about the admission itself, such as length of stay (LOS), type of admission or season.

Table 3.8 shows a summary of the patient characteristics and their distribution according to the output class. The dataset class imbalance is quite high, as illustrated in the histogram plots for some selected variables in Figure 3.6.

Table 3.7. Description of the variables in the Hospital of Basurto dataset.

Feature	**Description**
Clinical history	
AGE	Age of the patient (years)
SEX	Sex of the patient
SMOKER	Does the patient smoke? (yes/no/former)
WEIGHT	Weight of the patient (kg)
HEIGHT	Height of the patient (cm)
HR	Heart Rate (bpm)
SO2	Oxygen saturation (%)
SBP	Systolic Blood Pressure (mmHg)
DBP	Diastolic Blood Pressure (mmHg)
LVEF	Left Ventricular Ejection Fraction (%)
FIRSTDIAG	Years since first diagnostic
LOS	Length of stay (days)
Implanted device	yes/no
Needs of oxygen	yes/no
Therapies	
THERAPY_1	Furosemide
THERAPY_2	Torasemide
THERAPY_3	Thiazide
THERAPY_4	MRAs (Mineralocorticoid/aldosterone receptor

	antagonists)
THERAPY_5	ACEIs (Angiotensin-converting enzyme inhibitors)
THERAPY_6	ARB (angiotensin receptor blocker)
THERAPY_7	Beta blockers
THERAPY_8	Ivabrandine
THERAPY_9	Digoxin
THERAPY_10	Anticoagulants
THERAPY_11	Antiplatelet therapy
THERAPY_12	Oxygen therapy
THERAPY_13	Antiarrhythmic drugs
THERAPY_14	Lipid lowering therapy
Laboratory	
UREA	Urea (mg/dl)
CREATININE	Creatinine (mg/dl)
SODIUM	Sodium (mEq/L)
POTASSIUM	Potassium (mEq/L)
HEMOGLOBIN	Hemoglobin (g/dl)
TOTAL_CHOLESTEROL	Total cholesterol (mg/dl)
LDL_CHOLESTEROL	LDL cholesterol (mg/dl)
HDL_CHOLESTEROL	HDL cholesterol (mg/dl)
TRYGLICERIDES	Triglycerides (mg/dl)
Comorbidities	
COM_1	Acute coronary syndrome
COM_2	Peripheral vascular disease
COM_3	Stroke
COM_4	Dementia
COM_5	Chronic obstructive pulmonary disease
COM_6	Connective tissue disease
COM_7	Peptic ulcer disease
COM_8	Mild liver disease
COM_9	Diabetes mellitus
COM_10	Hemiplegia
COM_11	Moderate / severe renal disease
COM_12	Complicated Diabetes Mellitus
COM_13	Any tumour
COM_14	Leukemia
COM_15	Lymphoma
COM_16	Moderate/severe liver disease
COM_17	Metastatic solid tumour
COM_18	Anxiety/depression
COM_19	Osteoarthritis/arthrosis/spondylitis
COM_20	Osteoporosis
COM_21	Sinus rhythm
COM_22	Atrial fibrillation
COM_23	Pacemaker rhythm
Questionnaire	

Q1	With respect to previous three days, I feel:
Q2	Does the medication do me good?
Q3	In the last 3 days, have I taken any medication without supervision from my doctor?
Q4	Am I following the diet and exercise recommendations given by my doctor and nurse?
Q5	In the last 3 days, my ankles are:
Q6	Can you take walks like previous days?
Q7	Do I feel breathless or shortness of breath when I lie in bed?
Q8	Do I notice that I have begun to have cough or to expel phlegm?
Q9	Have I noticed fatigue at rest?
Q10	If fatigue – Can I take walks on flat?
Q11	If fatigue – At what level of effort I notice fatigue?

Table 3.8. Summary of characteristics and its distribution. Mean and standard deviation is reported for continuous variables and percentage for categorical ones.

Feature	**All patients (n=193)**	**Readmitted (n=40)**	**Not readmitted (n=153)**
Age, mean (SD)	77.4 (11.2)	77.0 (12.2)	77.5 (11.0)
Male sex (%)	111 (57.5)	26 (13.5)	85 (44.0)
Smoke			
- Yes	97 (50.3)	17 (8.8)	80 (41.5)
- No	43 (22.3)	11 (5.7)	32 (16.6)
- Former	33 (17.1)	9 (4.7)	24 (12.4)
- Unknown	20 (10.4)	3 (1.6)	17 (8.8)
LVEF	41.7 (15.3)	37.4 (14.3)	42.8 (15.4)
First diagnostic	6.6 (7.5)	9.9 (9.9)	5.8 (6.5)
Implanted device	44 (22.8)	12 (6.2)	32 (16.6)
Need oxygen	13 (6.7)	3 (1.6)	10 (5.2)
Urea	72.7 (37.6)	81.7 (42.9)	70.3 (36.0)
Creatinine	1.3 (0.5)	1.4 (0.6)	1.3 (0.5)
Sodium	139.9 (4.2)	138.8 (5.1)	140.2 (3.8)
Potassium	4.3 (0.8)	4.4 (0.7)	4.3 (0.8)
Haemoglobin	13.3 (11.1)	16.2 (23.9)	12.5 (2.7)
Sinus rhythm	73 (37.8)	15 (7.8)	58 (30.1)
Atrial fibrillation	107 (55.4)	21 (10.9)	86 (44.6)
Pacemaker rhythm	25 (13.0)	6 (3.1)	19 (9.8)

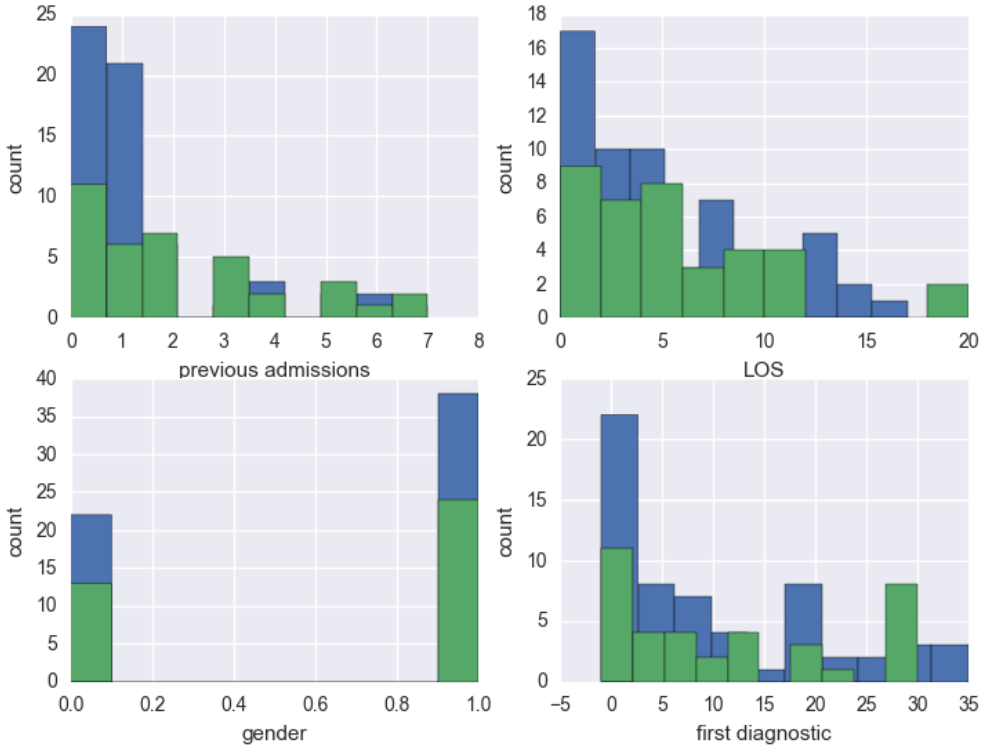


Figure 3.6. Distribution of readmission class for different attributes in the Hospital of Basurto (readmission in green, blue otherwise).

Chapter 4

Methods

In this chapter we gather the computational methods used in the experiments, along with a discussion of methodological issues that have to be taken into account.
First, in Section 4.1 we discuss the problem of class imbalance which strongly affects the classification performance, and that is present in the datasets that we have been dealing with. Next, in Section 4.2 we present feature selection processes that we have applied. Finally, in Section 4.3 we give short descriptions of the classification algorithms that we have used, because most are well known from the literature.

4.1. Class Imbalance

4.1.1. Introduction

In supervised classification, we say that a dataset is imbalanced when the a priori probabilities of the classes are significantly different, i.e. there exists a minority (positive) class that is underrepresented in the dataset in contrast to the majority (negative) class [Haixiang2017, Sun2009, Yang2006]. The minority class can have the meaning of a rare event, such as an alert condition, an intrusion in a security system, or a disease in a population. Such situations appear in healthcare as well as in many other fields, e.g. fraud detection, cybersecurity, communications, fault diagnosis, etc. Often the minority class is the target class to be predicted because it is related to the highest cost/reward events [Lopez2013]. Most classification algorithms assume equal *a priori* probability for all the classes, or equivalently equal cost to errors in classification, so that when this premise is violated the resulting classifier is biased towards the majority class, i.e. it has a higher predictive accuracy over the majority class, but poorer predictive accuracy over the minority class. Although imbalanced data classes have been recognized as one of the key

problems in the field of data mining [Yang2006], it is not usually taken into account in the literature of readmission risk prediction.

A measure of class imbalance is given by the imbalance ratio (IR), defined as the ratio of the number of instances in the majority class and the number of those in the minority class. Computational studies have shown that conventional classifier performance deteriorates even with moderate imbalance ratios [Mazurowski2008]. Figure 4.1 depicts a taxonomy of the methods developed to deal with class imbalance [Lopez2013] where three main techniques are identified, namely *preprocessing*, *cost-sensitive learning*, and *ensemble* techniques. Following we give an overview of the different strategies.

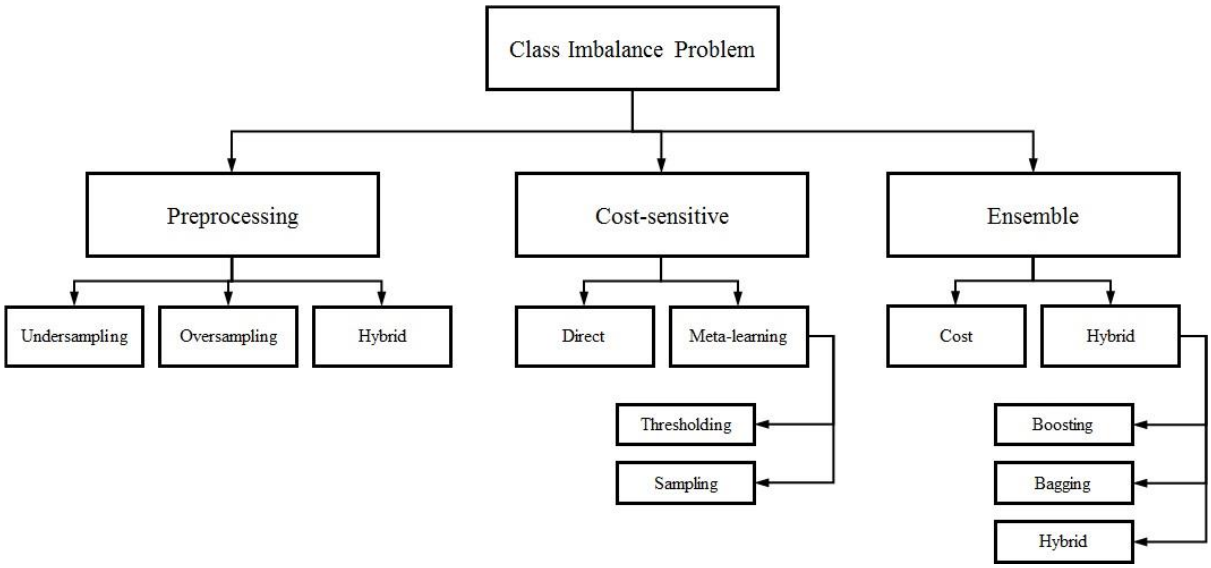


Figure 4.1. Taxonomy of Class imbalance problem addressing techniques extracted from [Lopez2013]

4.1.2. Preprocessing

Methods following this strategy carry out resampling of the original dataset in order to change the class distribution. Sometimes they are referred as data-level methods. Resampling techniques (illustrated in Figure 4.2) can be divided into three groups:

- Undersampling techniques deleting instances of the majority class,
- Oversampling techniques, that replicate or create new instances of the minority class, and
- Hybrid techniques that combine both resampling techniques.

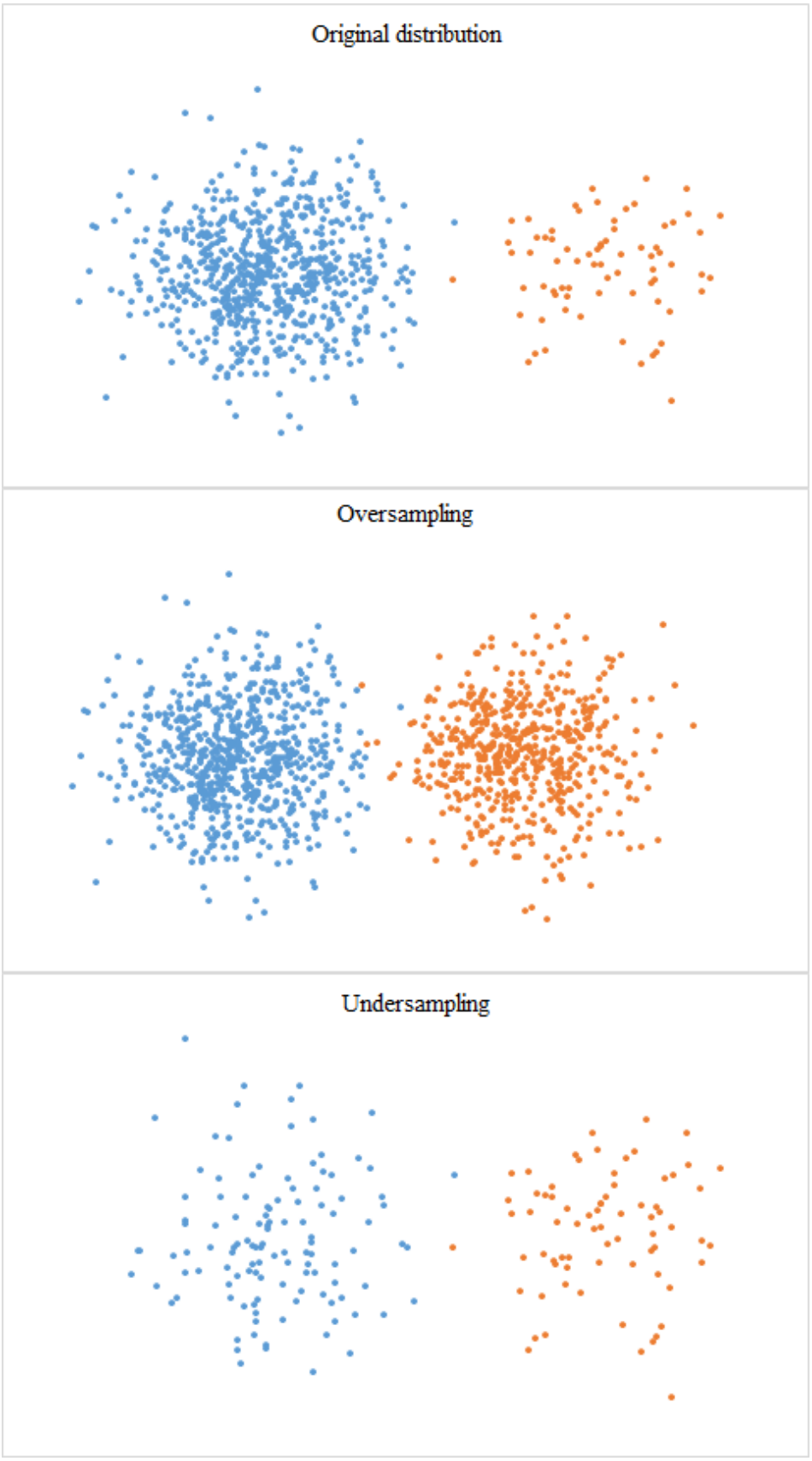


Figure 4.2. Undersampling and oversampling techniques, effect on the sample distribution on a 2D dataset.

Undersampling

The simplest undersampling method is random undersampling, which consists of randomly deleting majority class instances in order to achieve class balance. More sophisticated approaches have been proposed, most of them distance-based methods. Among the most representative methods we find NearMiss [Mani2003], which selects the samples to be removed based on k-NN search. A similar data cleaning method consists of removing Tomek links [Tomek1976] which are defined as instances that are each other's closest neighbours, but belong to different classes. Other approaches introduce the use of clustering techniques, such as SBC (under-sampling based on clustering) [Yen2009] which consists in clustering the samples and selecting the cluster containing the most majority samples. The samples to be removed are randomly selected among the majority class instances of the selected cluster. Undersampling is often criticized because of the information loss that instance deletion may produce. Hence, it is common practice to use this method only when a very high number of possibly redundant majority samples are present in the dataset.

Oversampling

Random oversampling is the simplest oversampling method, which consists of randomly replicating minority class samples. Despite its simplicity, this method leads easily to overfitting, since it generates exact copies of existing instances [Lopez2013]. In order to deal with such problems, more sophisticated techniques have been proposed. Synthetic Minority Oversampling Technique (SMOTE) is probably the most applied oversampling technique. This method over-samples the minority class by creating synthetic instances based on its nearest neighbours [Chawla2002]. Algorithm 4.1 and Figure 4.3 illustrate the SMOTE procedure.

Depending on the percentage of synthetic samples that want to be generated (in respect to the original minority class instances), some -or all- minority samples are selected. Having specified beforehand the number of nearest neighbours k, for each sample, the k nearest neighbours are found using the Euclidean distance. Once the nearest samples are selected, a random value between 0 and 1 is generated and multiplied to the distance of each feature between the actual instance and the neighbour. In other words, the vector of coefficients of a random convex linear combination is generated and applied to the k nearest neighbours in order to create a new sample.

Adaptive Synthetic Sampling Approach (ADASYN) [He2008] is similar to SMOTE but instead of generating an arbitrary number of instances per minority sample, it uses the concept of "difficulty in learning" concept, so that more synthetic data is generated for minority class samples that are harder to learn.

Algorithm 4.1. Synthetic Minority Oversampling Technique [Ditzler1997]

Input: Minority data $\mathfrak{D}^{(t)} = \{\boldsymbol{x}_i \in X\}$ where $i = 1,2,\dots,T$
Number of minority instances (T), SMOTE percentage (N), number of nearest neighbors (k)

for $i = 1,2,\dots,T$ **do**
1. Find the k nearest (minority class) neighbors of $\boldsymbol{x}_i$
2. $\widehat{N} = \lfloor N/100 \rfloor$

while $\widehat{N} \neq 0$ **do**
1. Select one of the k nearest neighbors, call this $\overline{\boldsymbol{x}}$
2. Select a random number $\alpha \in [0,1]$
3. $\widehat{\boldsymbol{x}} = \boldsymbol{x}_i + \alpha(\overline{\boldsymbol{x}} - \boldsymbol{x}_i)$
4. Append $\widehat{\boldsymbol{x}}$ to $\mathcal{S}$
5. $\widehat{N} = \widehat{N} - 1$

end while
end for
Output: Return synthetic data $\mathcal{S}$

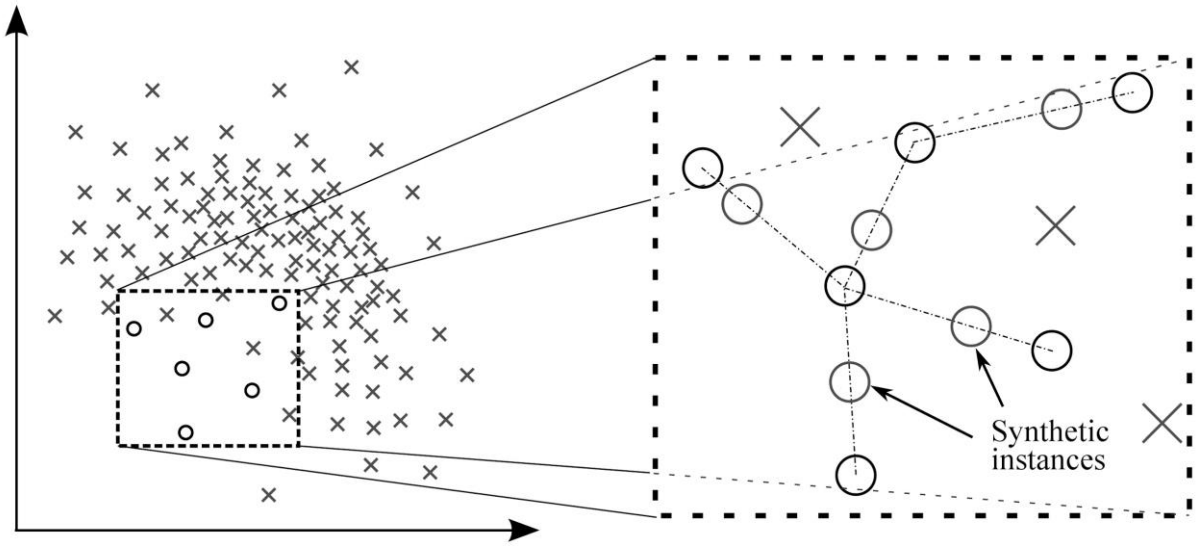


Figure 4.3. Synthetic instance generation with SMOTE [Borovicka2012]

4.1.3. Cost-sensitive learning

Cost-sensitive methods are based on the idea of compensating for the class imbalance of the dataset, without modifying the actual class distribution. Learning methods for classifier building are guided by the minimization of some cost function. The simplest cost formulation attributes cost 1 to a misclassification and 0 to a correct classification. The strategy followed by cost-sensitive learning methods is to assign different cost values to each class misclassifications, so that the bias towards the majority class is balanced by the lower cost of misclassifications. A cost

matrix is built assigning cost values to the entries of the confusion matrix giving (see Table 4.1 for the two-class case). The usual approach is to penalize misclassifications of the minority class. The diagonal elements are usually set to zero, meaning that correct classification has no cost [Kotsiantis2006].

Table 4.1. Cost matrix for binary classification

		Predicted Class	
		Positive	Negative
Actual Class	Positive	C_{TP}	C_{FN}
	Negative	C_{FP}	C_{TN}

Cost sensitive methods are categorized into the following groups:

- **Direct methods**, that introduce the misclassification cost within the classification learning algorithm. For the case of a classification tree, it can be done by minimizing the cost of each node of the tree.
- **Meta-learning**, where the learning algorithm itself is not modified. Instead, a preprocessing (or postprocessing) mechanism is introduced to handle the costs. Meta-learning methodologies can be divided into two categories, namely *thresholding* and *sampling*.

4.1.4. Ensemble classifiers

Ensemble methods rely on the idea that the combination of many "weak" classifiers can improve over the performance of a single monolithic classifier [Galar2012]. They are divided in two groups, namely *cost-sensitive* ensembles and *data and algorithmic* approaches.

- **Cost-sensitive** ensemble techniques, are analogous to cost-sensitive methods mentioned earlier, although in this case, the cost minimization is undertaken by the boosting algorithm. Different variants of AdaBoost such as AdaCost [Fan1999] and other modifications such as AdaC1, AdaC2 and AdaC3 [Sun2007] are some representative examples of this type of techniques.
- **Data and algorithmic** approaches, which embed a data preprocessing technique in an ensemble algorithm. Depending on the ensemble algorithm they use, three groups are identified: i) Boosting, ii) Bagging and iii) Hybrid.

Bagging

Bagging [Breiman1996] consists in creating bootstrapped replicas of the original dataset with replacement (i.e. different copies of the same instance can be found in the same bag), so that different classifiers are trained on each replica (Algorithm 4.2). In the original bagging proposal, each new dataset or bag maintained the size of the original dataset. Nevertheless, UnderBagging and OverBagging strategies embed a resampling process, so that bags are balanced by means of undersampling or oversampling techniques. To classify an unseen instance, the output predictions of the weak classifiers are collected performing a majority vote in order to produce the joint ensemble prediction. In this group we find, among others, algorithms like SMOTEBagging [Wang2009] or UnderBagging which embed undersampling within the ensemble algorithm. We propose RUSBagging which carries out a random undersampling for each bag generated in the ensemble creation. An individual weak classifier is trained from the data in each bag. Figure 4.4 depicts the bagging with resampling procedure.

Algorithm 4.2. Pseudocode of bagging [Du2012]

Input = Training sample S, classifier h, iterations T

Output = $\underset{y \in Y}{\operatorname{argmax}} \sum_{i:L_i(x)=y} 1$

For i=1 to T

  Si = bootstrap sample from S

  hi = train a classifier using S_i

End for

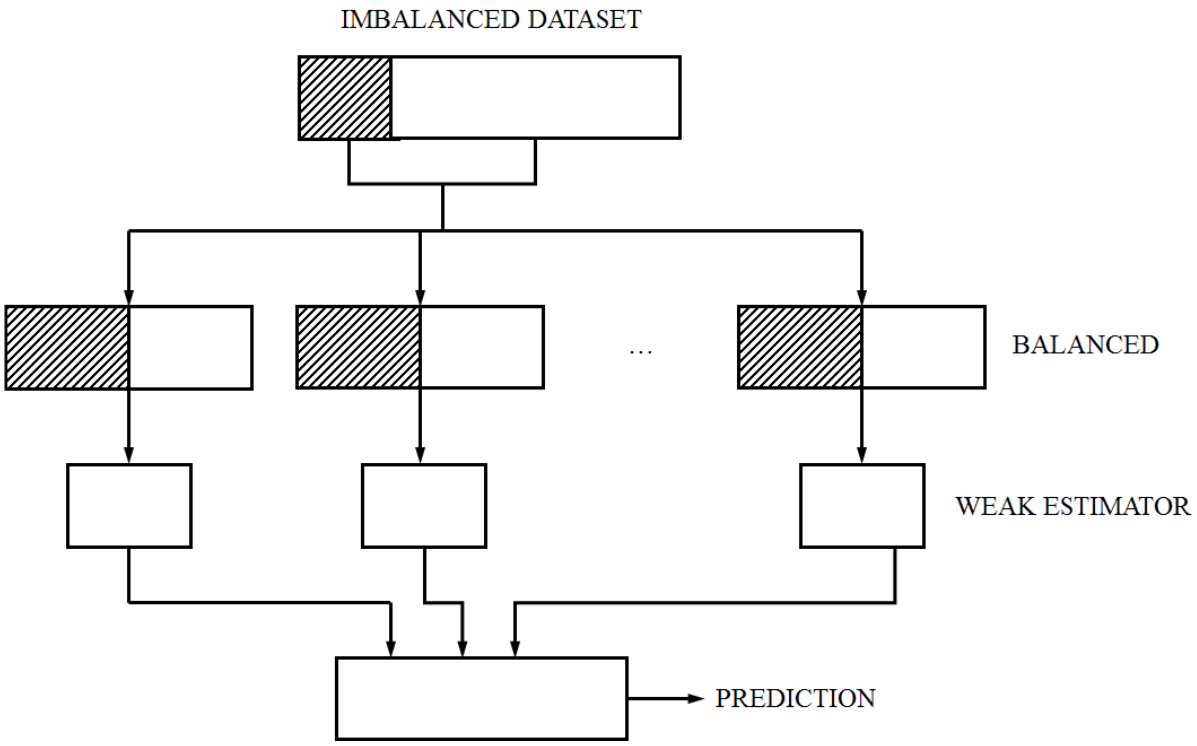


Figure 4.4. Bagging with resampling

Boosting

Most boosting algorithms (there exist multiple variations) consists of iteratively training weak classifiers and combining the outputs to create a strong classifier. In the first iteration the base classifier uses the original dataset, where all the instances are assigned equal weight. In each iteration the weights are updated so that misclassified instances gain weight (i.e. we pay more attention on these observations). The weak classifier is added to the final classifier until the termination criteria is fulfilled (maximum number of iterations reached or a given accuracy threshold achieved).

AdaBoost [Freund1995] is the most representative algorithm of this type of ensemble techniques. This technique has been combined with different resampling strategies, leading to methods such as SMOTEBoost [Chawla2003] (in combination with SMOTE oversampling) or RUSBoost [Seiffert2010] (using random undersampling) among others.

4.2. Feature Selection

The feature set is the set of variables that are input to the classifiers. Features may be produced by transformations of the original variables describing the dataset items, or they can be a subset of the original variables. Feature selection is the process of obtaining a subset of the original variable set containing the relevant features by discarding redundant or irrelevant variables. Dataset instances are described by a vector of variables $X = (x_1, \ldots, x_n)$ and a class label. The goal of feature subset selection is to find an optimal feature subset $X' \subset X$ so that the accuracy of the classifier is maximal.

Feature selection is an important step in model building since it allows model complexity reduction, and makes it more efficient in terms of performance. Often, when dealing with high-dimensional spaces, predictive models tend to overfit as the number of features grows. This phenomenon, known as the curse of dimensionality, causes a degradation of model's performance due to the high number of variables, as shown in Figure 4.5. Moreover, resulting models are easier to interpret from domain expert's perspective. This point is especially important in medicine, where clinicians are reluctant to use complex black-box-type models and demand interpretable solutions.

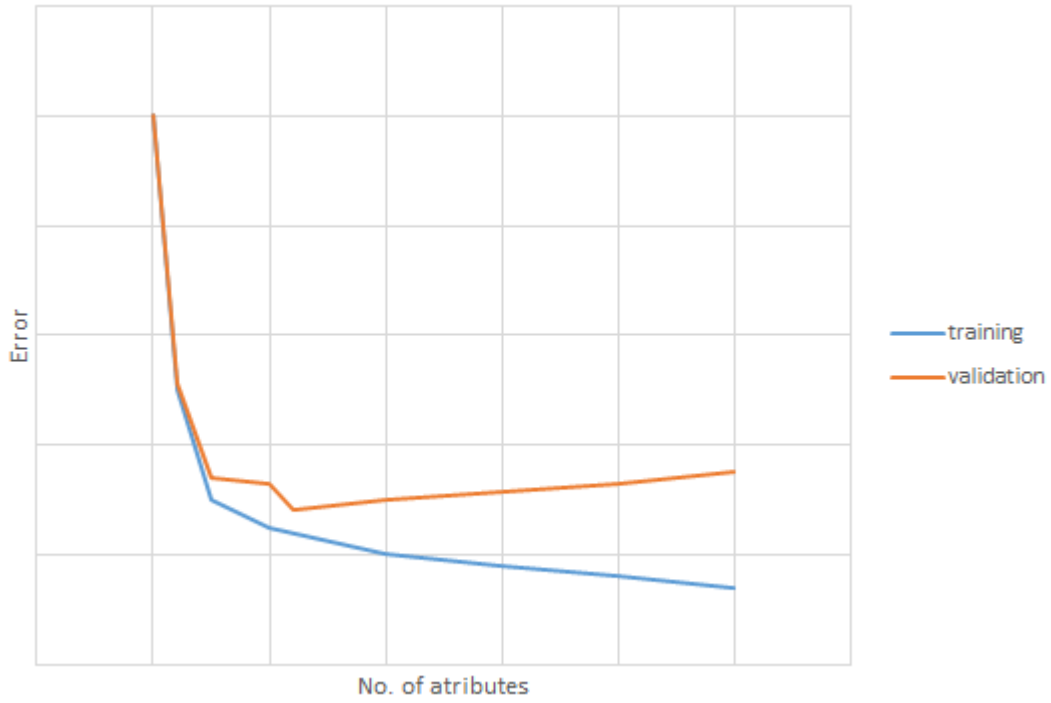


Figure 4.5. Curse of dimensionality

According to the taxonomy of feature selection techniques defined by Kohavi et al. [Kohavi1997] the methods can be grouped as follows (see Figure 4.6):

- Filter Methods
  - Univariate
  - Multivariate
- Wrapper Methods
- Embedded Methods

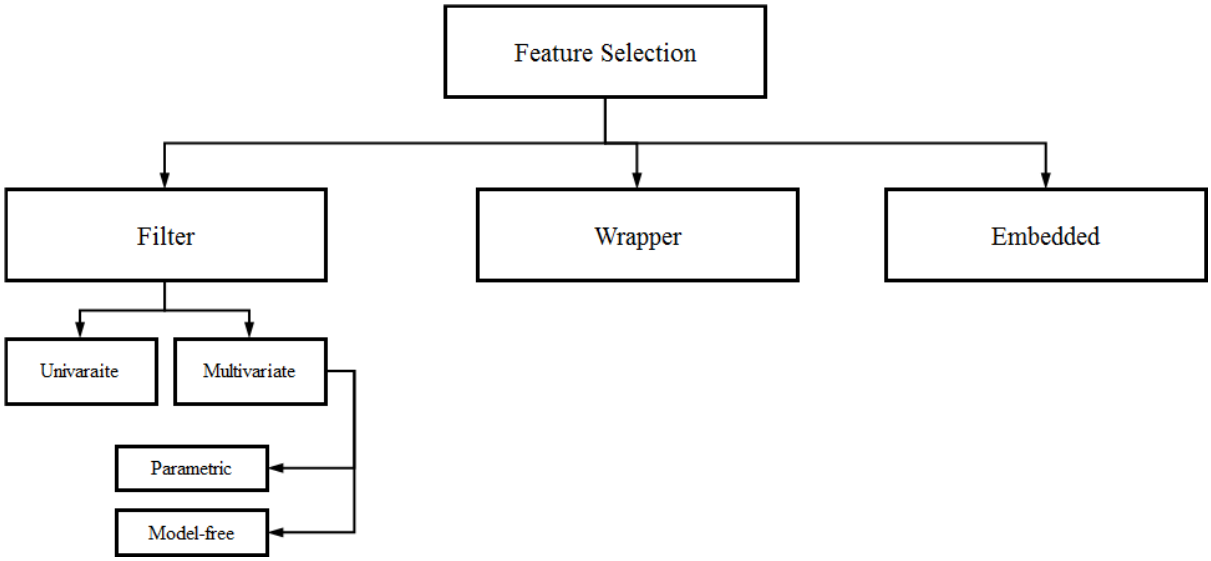


Figure 4.6. Taxonomy of feature selection techniques according to [Kohavi1997]

Following, the different techniques are briefly explained.

4.2.1. Filter Methods

According to [Kohavi1997] filter methods attempt to assess the predictive value of features from the data, without recourse to the classifier learning algorithm. A scoring function **S(i)** is computed for each input variable x_i, (i^{th} component of **X**) according to its corresponding **c** value. Frequently features are ranked according to their relevance, assuming that high scores indicate high relevance and vice-versa. Eventually low-scoring features are removed, so that won't be eligible for further analysis or imputation to the classification algorithm. As shown in Figure 4.7, in the filter approach feature selection is applied as a pre-process of the dataset, regardless of the algorithm to be used in the classification phase.

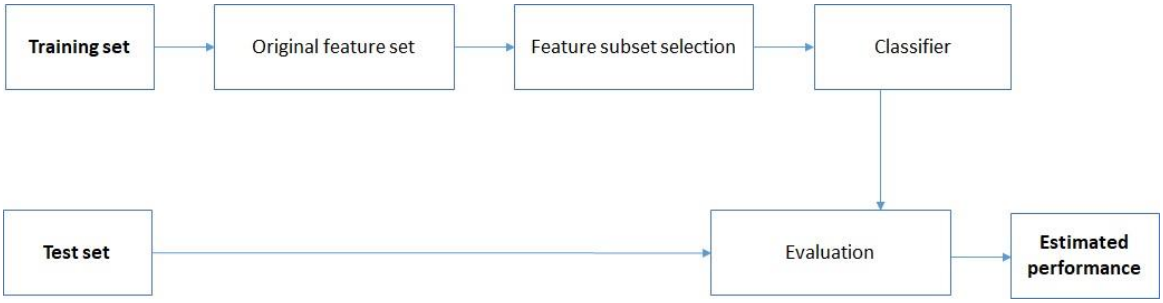


Figure 4.7. Filter approach for feature selection

In terms of computation, filter methods are efficient and scale well since they require only to compute n scores. However, its main advantage, that is, being classification algorithm agnostic, is at the same time one of its biggest disadvantages: It ignores the effects of the selected feature subset on the performance of the classification algorithm. Another disadvantage that is usually pointed is that the proposed techniques are univariate [Saeys2007]. It means that each feature is considered independently, ignoring interactions between features. Not taking into account feature interactions can lead to model's suboptimal performance, since features containing valuable interaction information but with low independent score are not included in the model. In order to overcome the problem of ignoring feature interactions, different multivariate techniques have been proposed. (e.g. correlation-based feature extraction [Hall1999]).

4.2.2. Wrapper methods.

Unlike filter approaches, which ignore the biases of the classification algorithm, the wrapper approach, shown in Figure 4.8, makes use of a classifier for scoring the feature subset's predictive power. As pointed in [Kohavi1997] the classifier is considered a black box, as no knowledge of the algorithm is needed, just the interface. Wrapper methods conduct a search through the feature subset space for a good subset, where subsets are evaluated according to classifier's estimated accuracy. Classification model's accuracy is usually estimated using cross-validation.

Although in cases where the number of features is not too large an exhaustive search may be practicable, the problem is known to be NP-hard, what makes this approach computationally intractable [Guyon2003]. Since an exhaustive search of the space is impractical, often a search procedure guided by a heuristic function is defined. Multiple search strategies have been proposed, including hill-climbing, best-first or genetic algorithms among others.

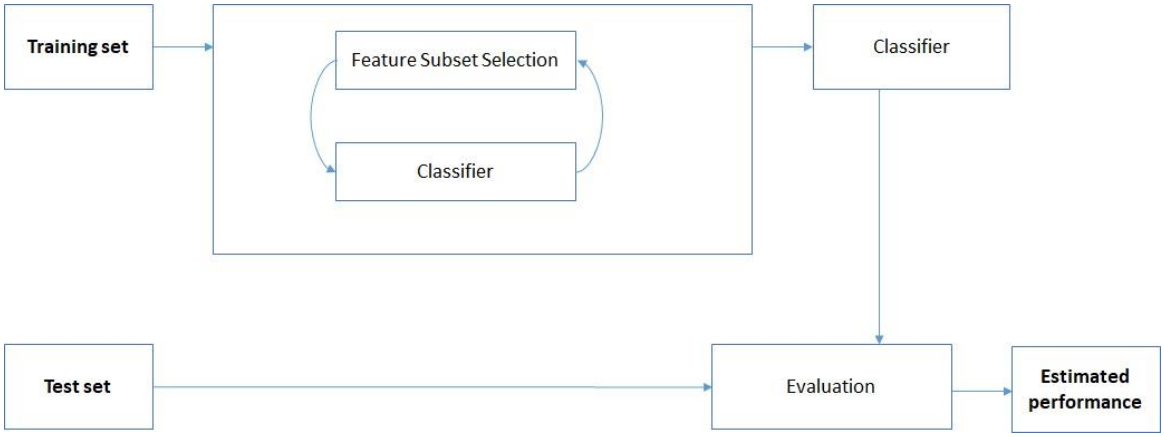


Figure 4.8. Wrapper approach for feature selection

One of the advantages of the wrapper approach is that interactions and dependencies between features are taken into account. Another advantage is that, unlike the filter approach, wrapper methods are linked to the classification model, so that the interactions of the feature set with the prediction model are considered. Nevertheless, a common drawback is that this approach is more prone to overfit to the training data. Wrapper methods are also criticized because their high computational cost, although efficient search strategies can alleviate the problem to a great extent. In the following, we briefly introduce two simple and widely used greedy search strategies, namely Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS):

Sequential Forward Selection (SFS)

Starting from an empty set S', sequentially add the feature x that maximizes the evaluation measure J when is combined with S',

Algorithm 4.3. Pseudocode of SFS

1. Start with empty set S' = {0}
2. **While** no improvement in J in last j steps **or** S'==S
   a. $x' = \underset{x \notin Y_k}{\operatorname{argmax}} J(S' \cup \{x\})$
   b. $S' = S' \cup \{x'\}$
3. **end while**

Sequential Backward Selection (SBS)

Starting from the full feature set, sequentially remove the feature x' that least reduces (or increases) the evaluation measure J when is removed from S',

Algorithm 4.4. Pseudocode of SBS

1. Start with full feature set S’ = S
2. **While** no improvement in J in last j steps **or** S’=={0}
    a. x’ = argmax J(S’ - {x}), x ϵS’
    b. S’ = S’-{x’}
3. **end while**

4.2.3. Embedded methods.

In those methods, the search is conducted within the classifier itself, as part of the learning process. Embedded methods, in the same manner as wrapper methods, are tied to a specific classification algorithm. Nevertheless, the computational cost is significantly lower for embedded methods compared to wrapper methods and are less prone to overfitting than the latter. Common embedded methods include decision tree algorithms including random forest and logistic regression, among many others [Saeys2007].

4.3. Classification

In this Section, we describe the classifier building problem as a supervised learning problem. Moreover, the chapter describes the main classification algorithms that were employed during the different experiments carried out in the context of this Thesis.

4.3.1. Definition of the problem

In supervised classification, a classifier is a prediction model built using a -training- dataset. The dataset is composed of a set of M instances, where each instance is described by a vector of features **X** = (x_1, …, x_n) and the class label C = {c_1, …, c_n}. The classifier can be defined as a function g that returns the c value given a feature vector **X** (i.e. predicts the class of the input instance):

$$g: X \rightarrow C$$

$$g(x) = \underset{c}{argmax} f(x, c)$$

where f defines a scoring function. A well-known principle in machine learning is that we cannot expect a classifier architecture to outperform all others over all problem domains, which has been stated as the no free lunch theorem [Wolpert1996, Wolpert1997]. Thus, it is common practice to compare different classification algorithms and ensembles, in order to find the configuration that provides the best bias-variance trade-off. Following we briefly describe the classification algorithms that we have utilized in the experimental works carried out in this thesis.

4.3.2. Logistic Regression

Logistic regression is a linear classifier that measures the relationship between one or more independent variables and the binary target variable (multinomial logistic regression is used when the target variable can take more than two values). This model estimates the probability of the target variable given some linear combination of the predictors by fitting a logit function, as

$$logit(p_i) = ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im}$$

where p_i is the probability that the target variable is true given some linear combination of the predictors, given by

$$p_i = P(y_i = 1 \mid x_i)$$

$x_i = \{x_{i1}, \ldots, x_{im}\}$ are the predictors (features) of the model, β is the intercept and $\beta = \{\beta_1, \ldots, \beta_m\}$ are the regression coefficients. The probabilities p_i and the regression coefficients are determined

by optimization procedures such as maximum likelihood estimation. The probability of the target variable being true is equal to the logistic function of the linear regression expression, as

$$p_i = \frac{1}{1 + e^{-(\alpha + \boldsymbol{\beta} \cdot \boldsymbol{x_i})}}$$

4.3.3. Gradient Boosting

Gradient boosting is an ensemble-based classifier that produces many weak prediction models iteratively, usually decision trees, gathering them into a single stronger learner. It uses gradient descent optimization algorithm to minimize a cost function (loss function) iteratively fitting a model in the negative gradient direction.

Given a loss function L(y, F(x)), we want to obtain an estimate $\hat{F}(x)$ of the function value F* that minimizes the expected value of the loss function,

$$F^* = \arg\min_F E_{y,\mathbf{x}} L(y, F(\mathbf{x})) = \arg\min_F E_{\mathbf{x}} \left[E_y(L(y, F(\mathbf{x}))) \mid \mathbf{x} \right].$$

Gradient boosting follows an additive expansion approach, so that $\hat{F}(x)$ is formed by a weighted sum of functions h(x;a):

$$F(\mathbf{x}; \{\beta_m, \mathbf{a}_m\}_1^M) = \sum_{m=1}^{M} \beta_m h(\mathbf{x}; \mathbf{a}_m).$$

Where h(x;a) is a function of input variables x characterized by parameters a_m, m= 1, …, M. As fitting h at each step is computationally impractical, gradient descent is used as an optimization algorithm.

Algorithm 4.5. Gradient boost [Friedman2001]

$F_0(\mathbf{x}) = \arg\min_\rho \sum_{i=1}^{N} L(y_i, \rho)$
For $m = 1$ to M do:
  $\tilde{y}_i = -\left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x}) = F_{m-1}(\mathbf{x})}, \; i = 1, N$
  $\mathbf{a}_m = \arg\min_{\mathbf{a}, \beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(\mathbf{x}_i; \mathbf{a})]^2$
  $\rho_m = \arg\min_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i; \mathbf{a}_m))$
  $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}; \mathbf{a}_m)$
endFor
end Algorithm

Gradient tree boosting is a specific adaptation of the more general gradient boosting algorithm which uses decision trees (typically CART trees) as base learners.

4.3.4. Support Vector Machine

Support Vector Machines (SVM) [Burges1998, Vapnik1998] look for the set of support vectors that allow to build the optimal discriminating surface in the sense of providing the greatest margin between the classes. In this way, the decision function can be expressed in terms of the support vectors only:

$$f(\mathbf{x}) = sign\left(\sum \alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + w_0\right)$$

where $K(x_i, x_j) \equiv ø(x_i)^T ø(x_j)$ is a kernel function, α_i is a weight constant derived from the SVM process and the s_i are the support vectors [Vapnik1998]. Nonlinear kernel functions filling some conditions allow to map a nonlinearly separable discrimination problem into a linearly separable equivalent problem in higher dimensional space. For training, the SVM approach solves the dual optimization problem is

$$\min_{\alpha} \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha}$$

subject to $y^T\alpha = 0, 0 < \alpha_i \leq C, i = 1, \ldots, l$ where **e** is the vector of all ones, C > 0 is the upper bound on the error, Q is an l x l positive semi-definite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$. Model selection in SVM involves the selection of the appropriate kernel function as well as tuning of its parameters, which not trivial task [ICS2016]. Often, radial basis function kernel or RBF kernel is used, defined as

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$$

Where x and x' are two samples represented as feature vectors.

4.3.5. Decision Tree

Decision Trees (DT) [Breiman1984, Quinlan1993] are built by recursive partitioning of the data space using a quantitative criterion (e.g., mutual information, gain-ratio, gini index), maybe followed by a pruning process to reduce overfitting. Tree leaves correspond to the probabilistic assignment of data samples to classes. One of the most popular implementations of the algorithm is C4.5 [Quinlan1993] which is an extension of the previous ID3 [Quinlan1986] algorithm. At each node, the algorithm selects the feature that best splits the samples according to the normalized information gain.

Algorithm 4.6. Pseudocode of a decision tree
Preconditions:
Sample set S = {x_i, y_i}, i=1, …, n
F features

```
Tree = {}
For each f in F do:
  Compute normalized information gain if splitting on f
End for
f_max = feature with the highest normalized information gain ratio
S' = subsets generated according to f_max
Tree = Create a decision node that tests f_max in the root
For each S':
  Tree' = C4.5(S')
  Append Tree' to the corresponding branch or Tree
End for
Return Tree
```

4.3.6. Random Forest

Random Forest [Breiman2001] is an ensemble classifier consisting of multiple decision trees trained using randomly selected feature subspaces. This method builds multiple decision trees at training phase. Often, a pruning process is applied to reduce both tree complexity and training data overfitting. In order to predict the class of a new instance, it is put down to each of these trees. Each tree gives a prediction (votes) and the class having most votes over all the trees of the forest will be selected (majority voting). The algorithm uses the bagging method [Breiman1996], where each tree is trained using a random subset (with replacement) of the original dataset. In addition, each split uses a random subset of features.

Algorithm 4.7. Pseudocode of Random Forest

```
Preconditions:
Sample set S = {x_i, y_i}, i=1, …, n
B = Number of trees
F features
for i=1 to B do:
  S' = A bootstrap sample from S by randomly selecting n' samples out of a
set of n samples, with replacement)
  h_i = Train a decision tree on S' using a random subset of F features:
  for each node.
        f = random subset of F
        Split on best feature in f
```

```
  end for
  H = H + h_i
end for
return majority vote of trees in H
```

One of the advantages of random forests is that generally they generalize better than decision trees, which tend to overfit and that naturally perform some feature selection. They can also be run on large datasets and can handle thousands of attributes without attribute deletion.

4.3.7. Extreme Learning Machine

Extreme Learning Machines (ELM) [Huang2006, Huang2011, Huang2015] was proposed as a very fast training algorithm for single-layer feedforward neural networks (SLFN). The ELM avoids gradient descent of the input to hidden layer weights by performing a random sampling, equivalent to a random subspace projection. The training problem reduces to the estimation of the output weights by linear least squares resolution of the network response minimizing the classification error, often solved by the Moore-Penrose generalized pseudo-inverse. Randomization of hidden layer weights introduce training instability which has been tackled in many ways. Ensembles of ELM, such as the Voting ELM [Ayerdi2015, Chyzhyk2015], and the HERF [Ayerdi2014], help improve the training stability. The sought effect is that the individual classifier errors compensate in the limit when the ensemble size grows, assuming that the probability distribution of the individual classifier error is symmetric around zero.

4.3.8. Adaptive Hybrid Extreme Rotation Forest (AHREF)

The Anticipative Hybrid Extreme Rotation Forest (AHERF) algorithm was originally presented in [ICS2016] which is a heterogeneous ensemble classifier that anticipates the correct fraction of instances from each basic classifier architecture to be included in the ensemble.

The training and testing phases of this method are summarized in Algorithm 4.8. We specify the training and test phases of each cross-validation fold. For training, first, a model selection phase is performed, where 30% of the training data is used. This size of model selection data is a balance between an appropriate sampling of the data distribution and allowing data for ensuing ensemble training and testing, because model selection data cannot be reused for ensemble cross-validation. For each classifier type described in the previous section, a 5-fold cross-validation is carried out on the model selection data (line M3). The individual model selection cross-validation average accuracies are ranked, so that r_k is the ranking value of the k-th classifier type (line M4). Then

(line M5), each classifier is assigned a selection probability according to the expression $p_k = \frac{Fib((C+1)-r_k)}{\sum_{t=1}^{C} Fib(i)}$, where Fib(i) is the i-th value of the Fibonacci series.

The ensemble strategy cross-validation is carried out on the remaining 70% data, involving a 10-fold cross-validation process. Notice that the test data size at each fold is reduced to a 7% of the available data, hence larger model selection data cannot be afforded because of the risk of test data misrepresenting the actual data distribution. The following steps are carried out at each fold: for each classifier D_i in the ensemble the first step is the construction of the randomized rotation matrix (line 3) which requires the random partition of the set of features into a K subsets (line 4). For each subset of features $F_{i,j}$, the algorithm extracts the corresponding sample values in a matrix $X_{i,j}$ (line 6), used to build a component $C_{i,j}$ rotation matrix (line 7). The randomized rotation matrix R_i^{α} is built by composing the component rotation matrices reordering the columns in order to match the original variable ordering, as detailed in [ICS2016]. Next (line 9) there is a random decision on the type of the classifier, using the selection probabilities $\{p_k\}$ (built in line M5). Finally, the D_i classifier is trained on the rotated data. In the test phase, a new vector x^{test} is first applied each classifier in the ensemble, obtaining a class hypothesis d_i, (line C2). Majority voting is implemented as follows: the counter c_ω has the number of classifiers that have casted their vote for class ω, (line 3, where $\delta_{i,j}$ is the Kronecker's delta function). Finally, the class with the maximum votes is selected (line C4) and returned as the classification result.

Algorithm 4.8. Anticipative Hybrid Extreme Rotation Forest

Training Phase

Given

X : z-scores of input dataset ($n \times N$ matrix).

Y : the labels of the dataset ($1 \times N$ matrix)

L : the number of classifiers in the ensemble

K : the number of feature subsets

Begin

Anticipative Model selection

M1 Select 30% of the dataset for model selection

M2 For each classifier type $k = 1, \ldots, M$

M3 Perform 5-fold cross-validation, obtain accuracy A_k

M4 Rank A_k, assigning r_k to the k-th classifier

M5 Assign selection probability $p_k = \frac{Fib((C+1)-r_k)}{\sum_{i=1}^{C} Fib(i)}$, $k = 1, \ldots, M$

On the 70% unused data, perform 10-fold cv, at each fold:

Ensemble construction on each training fold

2 For each individual classifier D_i, $i = 1 \ldots L$

3 Computation of rotation matrix R_i^{α}:

4 Partition F into K random subsets: $F_{i,j}; j = 1 \ldots K$

5 For each F_{ij}, $j = 1 \ldots K$

6 - Let $X_{i,j}$ be the subset of X corresponding to features in $F_{i,j}$.

7 - $C_{i,j}$ obtained from PCA on $X_{i,j}$

8 Compose R_i^{α} using matrices $C_{i,j}$.

9 Decide the model of D_i sampling $\{p_k; k = 1, \ldots, M\}$

10 Train classifier D_i on training set $(R_i^{\alpha} X, Y)$ or (X, Y)

End ensemble construction

Test on each testing fold

Let Ω be number of classes

C1 For each unknown $\mathbf{x}^{test}$ z-scores.

C2 $d_i = D_i(R_i^{\alpha} \mathbf{x}^{test}); i = 1, \ldots, L$

C3 $c_{\omega} = \sum_{i=1}^{L} \delta_{d_i,\omega}; i = 1, \ldots, L$

C4 $c^{test} = \arg\max_{\omega} \{c_{\omega}, \omega = 1, \ldots, \Omega\}$

4.3.9. Miscellaneous commonly used classifier learning

k-Nearest Neighbours

k-Nearest Neighbours k-Nearest Neighbours (k-NN) is the simplest formulation of the supervised training, where the training samples are used as class prototypes. The class assigned to

the test input pattern is the result of majority voting on the K closest training patterns according to some defined distance in pattern space, which most often is the Euclidean distance.

Adaboost

Adaboost Adaptive Boosting (AdaBoost) [Schapire1999, Freund1995] is a meta-algorithm for machine learning that can be used in conjunction with many other learning algorithms to improve their performance. AdaBoost trains a weak classifier in a series of rounds $t = 1, \ldots, T$. For each iteration the distribution of sample weights W_t is updated, indicating the importance of examples in the data set for the classification. On each round, the weights of each incorrectly classified example are increased (or alternatively, the weights of each correctly classified example are decreased), so that the new classifier focuses more on those examples.

Chapter 5

Results

This chapter reports the results of various experiments on readmission risk prediction carried out in this Thesis. We have grouped the experiments according to the target subpopulation they are referred to, namely emergency department admissions and heart failure patients.
The chapter is structured as follows: Section 5.1 defines the evaluation metrics that we have used to evaluate and report the results. Section 5.2 briefly presents the methodology followed to perform the experiments in the experimental design subsection. Next, Sections 5.3 and 5.4 present the results of various experiments carried out on Emergency Department and Heart Failure readmission prediction.

5.1. Evaluation Metrics

Whenever we conduct an experiment it is crucial to define beforehand the metric that will be used to measure the performance. Often, there exist multiple metrics that can be applied, each having its own characteristics; its benefits and drawbacks. Hence, it is important to choose the correct metric for our specific scenario, to avoid reporting meaningless results. In supervised classification, the confusion matrix -also known as *error matrix*- is the keystone of every evaluation metric. A confusion matrix has two dimensions, namely the actual and the predicted class, with two classes each -positive and negative-. Table 5.1 shows the confusion matrix of a two-class classifier.

Table 5.1. Confusion matrix for a binary classifier

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

In the following we define the evaluation metrics that were used in our experiments.

Accuracy

In binary classification, accuracy is defined as the proportion of true results among the total population:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

where TN is a true negative, TP a true positive, FN is a false negative and FP a false positive. In heavily imbalanced datasets it is not very meaningful because a simple strategy such as always assigning each test sample to the majority class provides high accuracy.

Sensitivity or Recall

Sensitivity is a classification performance measure defined as the proportion of correctly classified positives:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Sensitivity provides more information about the success on the target class.

Specificity

Specificity is defined as the proportion of negatives that are correctly identified as such:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Precision

The precision is the ability of the classifier not to label as positive a sample that is negative.

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

F-measure

F-measure is defined as the harmonic mean that combines the values of precision and recall, so that:

$$\mathrm{F}_{\mathrm{score}} = 2 \times \frac{precision \times recall}{precision + recall}$$

ROC curve

Receiver Operating Characteristic (ROC) curves are 2-D graphs used to represent the trade-off between the True Positive rate (sensitivity) and False Positive rate (1-specificity). Figure 5.1 shows an example of a ROC curve.

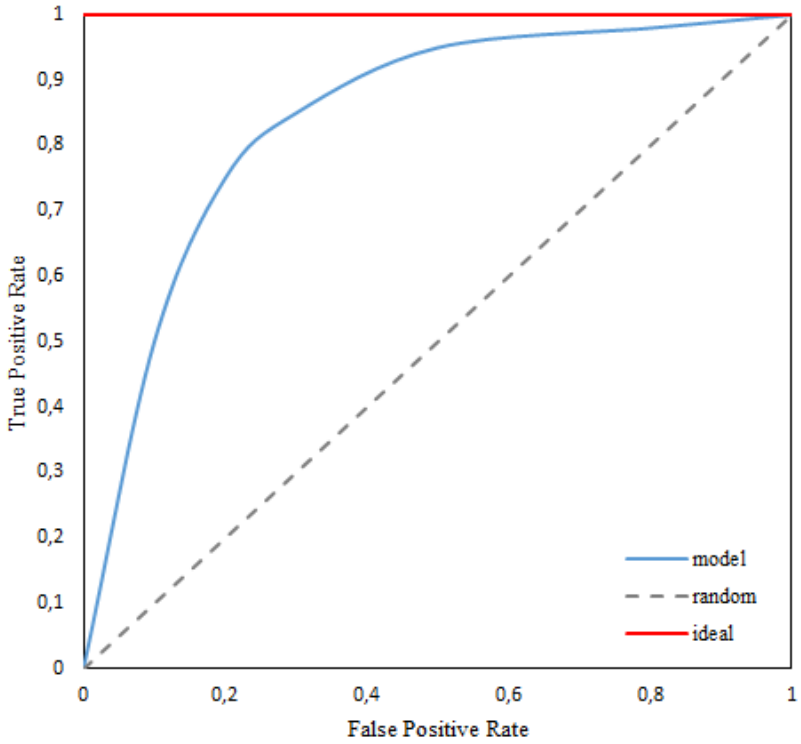


Figure 5.1. Example of a ROC curve.

AUC

The Area Under ROC Curve (AUC) sometimes referred as c-statistic, shows the trade-off between the sensitivity or TP_{rate} and FP_{rate} (1 - specificity):

$$\mathrm{AUC} = \frac{1 + TP_{rate} - FP_{rate}}{2}$$

where the True Positive rate is equal to the Sensitivity and the False Positive rate is defined as

$$\mathrm{FP}_{\mathrm{rate}} = \frac{FP}{FP + TN}$$

This metric is considered the de facto standard evaluation score in the field of readmission prediction.

Precision recall curve

When we talk about AUC we usually refer to the area under the ROC curve, although this is not necessarily like that. Technically speaking AUC can refer to any kind of curve. Unlike receiver operating characteristic curve, precision recall curves (such as Figure 5.2) are not influenced by the large values of TN, so that it is considered more suitable to be used in scenarios where the negative class outnumbers the positive class. Thus, when dealing with class-imbalanced datasets, it would be more meaningful to use the precision-recall curve rather than the ROC curve.

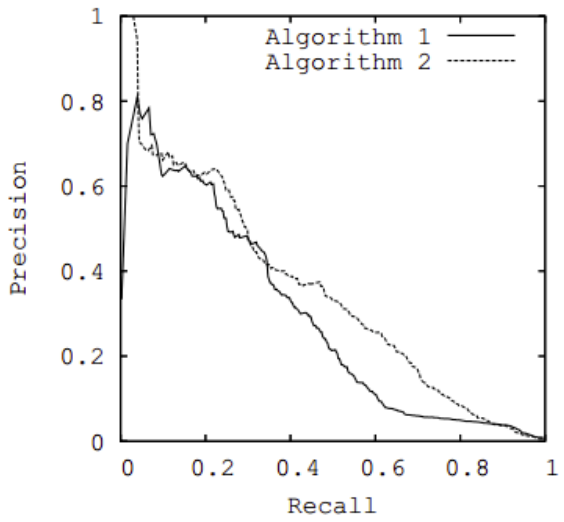


Figure 5.2. Example precision-recall curve.

5.2. Experimental Design

This subsection is intended to describe which is the methodology that we have used in most of our computational experiments.

5.2.1. Defining the outcome

We design our experiments as a two-class classification problem solved by supervised learning. In order to do so, we must define the outcome, which in our experiments is mainly the readmission variable (or death, depending on the dataset). The readmission event is illustrated in Figure 5.3.

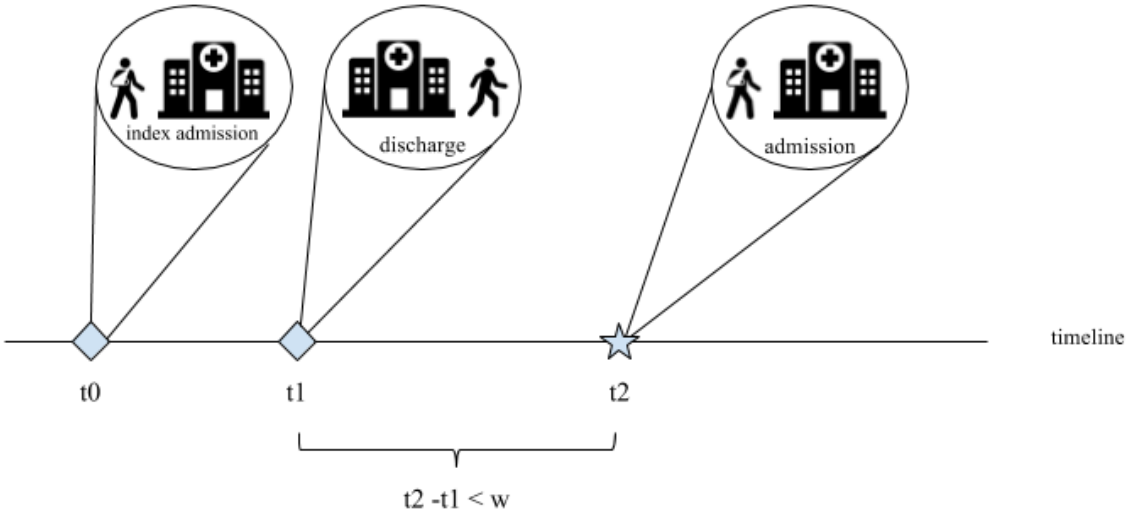


Figure 5.3. Hospital readmission event.

Datasets are curated versions of the raw EHR extracted from the hospital which is usually structured in an event-centred basis. Our events are hospital admission, either regular admissions or ED admissions, although death events can also be present. Those events need to be grouped by patient, so that a timeline-like schema *per* patient is built, as shown in Figure 5.4. Then, we define a window length *w* that will serve to encode the dichotomous outcome as readmitted or not-readmitted. Window length, or readmission threshold, is usually set to 30-days, although virtually any threshold can be applied (e.g. 72 hours, 28 days, etc.). When the time span between a discharge and the subsequent admission is lower than *w*, the index admission is labelled as readmitted (i.e. a positive class).

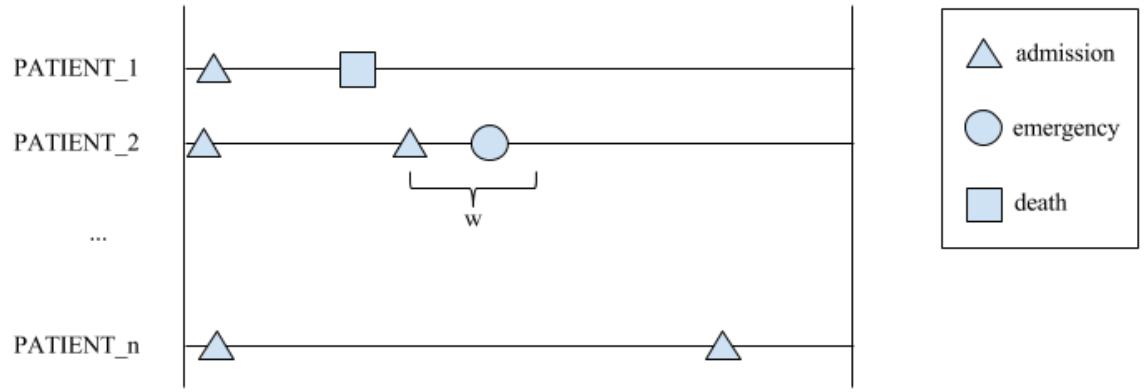


Figure 5.4. Different events among patients

5.2.2. Validating the model

The main premise of any validation schema is to ensure the independence of the training and testing datasets, that is, instances used to train the model can't be used to test the model. Traditionally, clinical studies construct and validate models following a percentage-split approach, by which training and validation sets are randomly split in a certain proportion, e.g. 50/50 or 70/30. Despite its simplicity, this approach produces unstable results, since it is sensible to the split selection (unless dataset is very large). Another widely extended method uses bootstrapping, consisting on random sampling with replacement, which reports more stable results.

Machine learning generally uses a cross-validation methodology when it comes to supervised classification. K-fold cross-validation is probably the most widely used method, since it has the advantage of using all the samples available, while providing balanced reports (not too optimistic nor pessimistic). Nevertheless, in order to avoid any random-related bias, it is common practice to repeat the process n times and to report the average scores. Figure 5.5 depicts the k-fold cross-validation process.

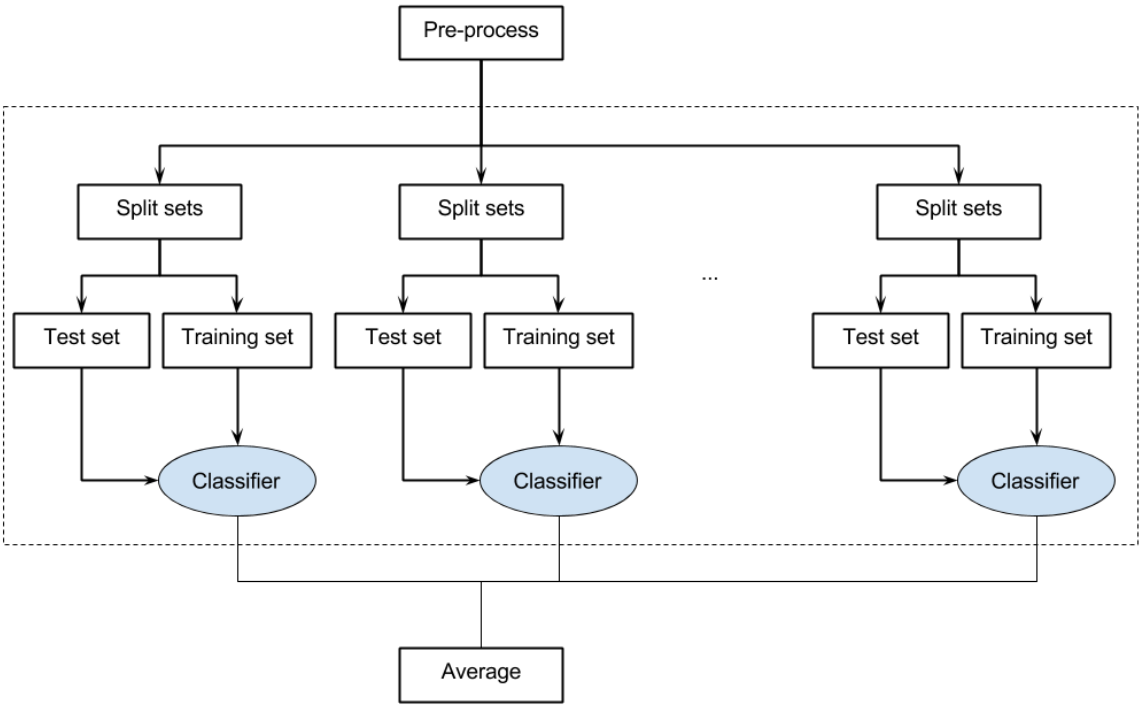


Figure 5.5. Flowchart of k-fold cross-validation

When working on a cross-validation scenario it is crucial to scrupulously preserve the independence of the test and validation sets through the whole pipeline. The very basic rule is that any training task that is performed in a supervised way must be held within the training set of each split. For instance, if we do feature selection or minority class oversampling before splitting our data, our model may suffer some kind of bias due to the use of training information in the validation phase.

Figure 5.6 shows an example process consisting on a data preparation process followed by k-fold cross-validation. Note that class balancing is performed after splitting the data only on the training dataset. Next, different models are built using the classification algorithms subject to comparison.

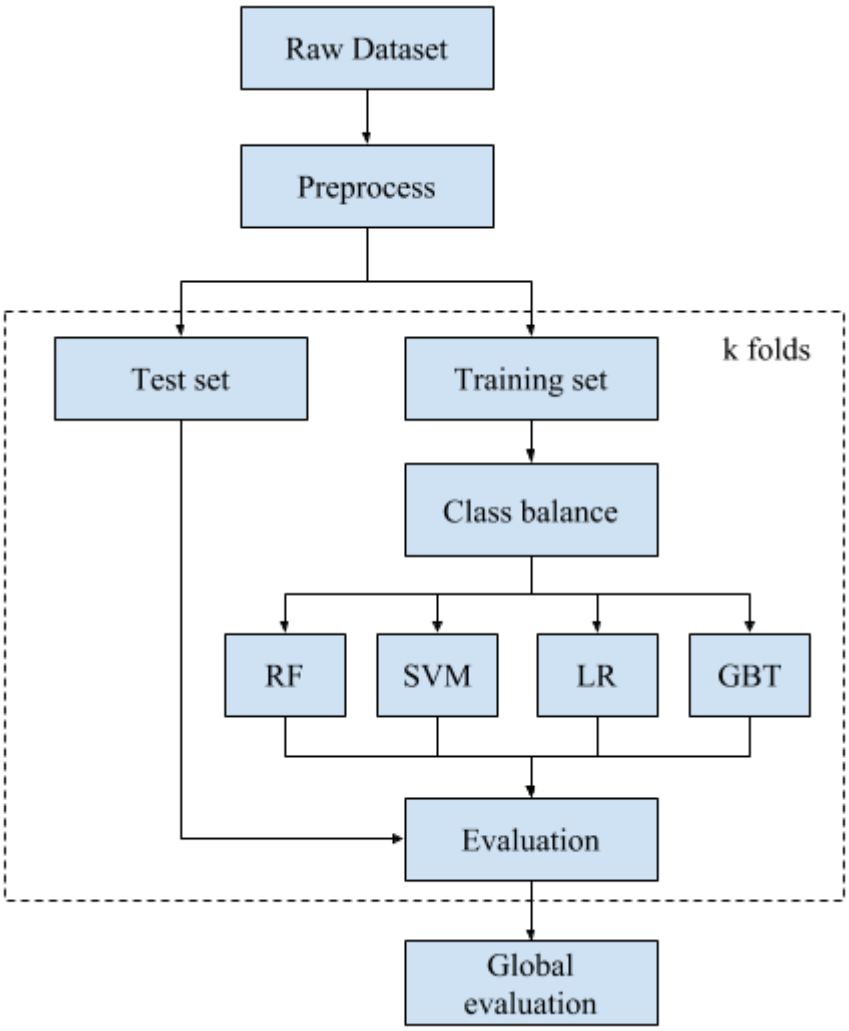


Figure 5.6. Flowchart of an example experiment

5.3. Emergency Department Readmission Prediction

In this section, we present the experiments that we performed in order to tackle the Emergency Department (ED) readmission risk prediction. Our first approach started with "Araba" dataset, targeting the widely-used time span of 30 days. Following we performed a series of experiments using the "Chile" dataset, targeting short-time readmissions (less than 72 hours).

5.3.1. Hospital Universitario Araba dataset

The dataset, presented in Section 3.1, is composed of 360 instances containing 97 features and the dichotomous outcome is set to "readmitted within 30-days from discharge".

Methods

All the experiments were conducted using 10-fold cross-validation. The evaluation metrics that we have used are: sensitivity, specificity and accuracy. In order to avoid any random number generation bias, we have conducted 10 independent executions with different random generating seeds and averaged the results obtained.

Table 5.2. Distribution of variables by category

Variable	**No. (%) of variables n=96**
Sociodemographic and baseline status	4 (4.2)
Personal history	43 (44.8)
Reasons for consultation	16 (16.7)
Regular medications	33 (34.3)

According to the data shown in Table 5.2 our dataset has a high dimensional feature space. In this scenario we have carried out some feature selection techniques. The goal is to find a feature subset that would reduce the complexity of the model, so that it would be easier to interpret by physicians, while improving the prediction performance and reducing overfitting.

We used the following feature selection approaches:

- **InfoGain filter**: It evaluates the worth of a feature by measuring the information gain with respect to the dependent variable. The output of this filter is a list of the attributes ranked by their predictive importance.
- **Wrapper**: Wrapper methods evaluate subsets of variables, that is, unlike filter methods, do not compute the worth of a single feature but the whole subset of features. We have selected SVM

as the classification algorithm and AUC as evaluation measure. Since an exhaustive search is impractical due to space dimensionality, we used heuristics, following a greedy stepwise approach.

Results

Besides the original four subpopulations shown in Table 5.3, we have considered an additional fifth dataset that encompasses all of them.

Table 5.3. Comparative information about the subpopulations of the dataset

	Overall no. of patients *n=360*	Readmission within 30 days, no. (%) of patients	
		No	Yes
		n=296 (82.2)	n=64 (17.7)
Case management	94 (26.1)	73 (77.7)	21 (22.3)
Heart failure	70 (19.4)	62 (88.6)	8 (11.4)
Chronic obstructive pulmonary disease	80 (22.2)	64 (80)	16 (20)
Diabetes mellitus	116 (32.2)	97 (83.6)	19 (16.4)

Class balancing

Table 5.4. Confusion matrix of SVM on the diabetes mellitus dataset

	Readmitted	**Not readmitted**
Readmitted	97	0
Not readmitted	19	0

As shown in Table 5.4, class imbalance is causing an accuracy paradox. If we just look at the accuracy of the model we get an 83.62% although SVM just behaves as using only the greatest *a priori* probability to make the classification decision. There are several methods that can be used in order to tackle the class imbalance problem. Building a more balanced dataset is one of the most intuitive approaches. In our experiment, we have used under-sampling as a preliminary approach and continued with an over-sampling using synthetic samples.

Undersampling with random subsample.

Given that there is a low number of samples for the minority-class, which is also the most relevant for classification, we can anticipate that reducing the amount of samples for the majority-class to be comparable to the minority-class and avoid the class imbalance will lead to a model with poor generalization capability.

Focusing on the diabetes mellitus subpopulation dataset, it is composed of 97 instances belonging to the *not-readmitted* class and only 19 of the *readmitted* class. An experiment consisting of subsampling the dataset to a distribution of 1:1.5 between the minority and majority classes, and then applying a Random Forest classifier shows the following results in Table 5.5.

Table 5.5. Comparison of performance evaluation metrics for RF over original and under-sampled versions of diabetes mellitus dataset

Dataset	Accuracy	Sensitivity	Specificity
Original	84.48	10.52	98.96
Under-sampled	61.7	31.57	82.14

As seen in Table 5.5, although the classification sensitivity has increased, it is still low (31.57%) despite the sacrifice of both accuracy and specificity performance. Taking into account the low number of instances contained in our dataset, we don't consider under-sampling an effective approach.

Oversampling with SMOTE.

We used Synthetic Minority Oversampling Technique (SMOTE) for oversampling the minority class. In order to avoid overfitting, we applied SMOTE at each fold of the 10-fold cross validation. If oversampling is done before 10-fold cross-validation, it is very likely that some of the newly created instances and the original ones are both in the training and testing sets, thus causing performance metrics being optimistic.

Our approach is to test the performance of two classifiers, namely SVM and RF, using the over-sampled dataset, in order to compare it with the results obtained using the original imbalanced dataset. The experiment will be carried out by generating a model for each of the subpopulations on each of the specified scenarios. Table 5.6 shows the results of our experiment.

Table 5.6. Performance comparison using SVM and RF classifiers on original and over-sampled datasets

		original			over-sampled		
		specificity	sensitivity	accuracy	specificity	sensitivity	accuracy
Case management	SVM	1	0.42	0.87	0.98	0.42	0.86
	RF	1	0.42	0.87	1	0.42	0.87
Heart failure	SVM	1	0	0.88	0.90	0.12	0.81
	RF	1	0	0.88	1	0	0.88

COPD	SVM	1	0	0.80	0.81	0.37	0.72
	RF	1	0.37	0.87	1	0.43	0.88
Diabetes mellitus	SVM	1	0	0.83	0.88	0.15	0.76
	RF	1	0.10	0.85	0.96	0.10	0.82
All	SVM	1	0.21	0.86	0.78	0.40	0.71
	RF	1	0.28	0.87	0.99	0.28	0.86

Results show that class-balanced dataset achieved better sensitivity than the original dataset. Nevertheless, both accuracy and specificity achieve worse results. It is worth noting that while performance is similar for both classifiers using the original dataset, SVM performs much better (in terms of sensitivity) when using the over-sampled version. At last, we observe that sensitivity improvement is rather small and it is obtained mainly at the expense of worsening both sensitivity and accuracy.

Feature selection

Our dataset has a high dimensional feature space. With the use of feature selection algorithms, we want to find a feature subset that would reduce the complexity of the model (so that it would also be easier to interpret by the physicians) while improving the prediction performance and reducing overfitting. For that purpose we are using a filter method, with InfoGain as metric and a wrapper method. The experiment consists in training a SVM and a RF classifier using the original feature set and the generated feature subsets. The performance of the classifiers will be compared in terms of sensitivity, specificity and accuracy for each of the subpopulations.

It's worth noting that the feature selection must be done using cross-validation. If full training set is utilized during attribute selection process, the generalization ability of the model can be compromised.

Table 5.7. Performance comparison of both feature selection methods

		infoGain			wrapper		
		specificity	sensitivity	accuracy	specificity	sensitivity	accuracy
Case management	SVM	0.98	0.33	0.84	0.94	0.23	0.78
	RF	0.89	0.38	0.77	0.89	0.38	0.77
Heart failure	SVM	1	0	0.88	0.90	0.12	0.81
	RF	0.96	0.25	0.88	0.98	0	0.87

COPD	SVM	1	0.18	0.83	0.95	0.37	0.83
	RF	0.93	0.43	0.83	0.92	0.37	0.81
Diabetes mellitus	SVM	1	0	0.83	0.98	0.05	0.83
	RF	0.96	0.05	0.81	0.98	0.05	0.83
All	SVM	0.99	0.10	0.83	0.97	0.14	0.83
	RF	0.92	0.25	0.80	0.95	0.18	0.81

In Table 5.7 the results of the experiment are shown. According to these results, although in some cases the sensibility has been increased, overall the results are not as promising as expected. Actually, even though models are much simpler than the original model (i.e. the one using full feature set), the prediction performance has been reduced. Moreover, both feature selection methods have performed similarly, even if selected feature subsets differ considerably.

5.3.2. Chile ED dataset

This dataset was provided thanks to the collaboration with Prof. Sebastián Ríos from the University of Santiago de Chile, who was collaborating with the Hospital José Joaquín Aguirre from University of Chile. Different state-of-the-art classification algorithms were used and compared their performance with ensemble approaches. Moreover, different class imbalance addressing methods were tested and new approaches proposed.

Experiment 1, testing class balancing methods

Results

In this section we present the results obtained when trying to predict the readmission risk before 72 hours over the dataset presented in the previous section.

We have tested two data balancing methods: random undersampling (RUS) and random undersampling embedded in a bagging approach. We used the following well-known classification algorithms, implemented in the open source machine learning Python library scikit-learn[2], which has also been used for the rest of the experiments:

1. Decision Tree (DT), setting Gini impurity as splitting criterion
2. Random Forest (RF), setting Gini impurity as splitting criterion and number of estimators=10

The models were evaluated using 10-fold cross-validation, performing 10 independent executions. Accuracy, specificity, sensitivity and AUC were calculated for each execution, so average and standard deviation were computed. In order to compare results in a statistically sound way, we employed an Analysis of Variance (ANOVA) approach.

The following data balancing approaches were compared:

i) Original dataset with its imbalanced class distribution,

ii) Undersampling with random undersampling and

iii) RUSBagging.

Table 5.8 shows the average accuracy, sensitivity, specificity and AUC along with its respective standard deviation, for each method and classifier.

Comparison of classifiers

According to the results shown in Table 5.8 for both classification algorithms, RF achieve significantly better results ($p<0.001$) than DT using the AUC as performance measure. Although

[2] http://scikit-learn.org

DT performs better in the original dataset (anyhow both classifiers perform poorly), when preprocessing and class balancing ensemble approaches are utilized RF performs much better. As shown in Figure 5.7, the AUC is significantly greater for RF when RUSBagging is used, however, sensitivity is sacrificed if compared with DT. Overall, results are poor, however they compare well with the state of the art in readmission prediction [Kansagara2011].

Table 5.8. Mean ± standard deviation of performance metrics for each data balance method and classifier model configuration

method	classifier	accuracy	specificity	sensitivity	AUC
None	DT	.9293 ± .0006	.9599 ± .0006	.0673 ± .0030	.5136 ± .0017
	RF	.9655 ± .0001	.9997 ± .0001	.0012 ± .0003	.5005 ± .0002
RUS	DT	.5578 ± .002	.5574 ± .002	.5674 ± .012	.5624 ± .005
	RF	.6622 ± .0016	.6676 ± .0018	.5086 ± .0096	.5881 ± .0043
RUSBagging	DT	.6530 ± .0011	.6576 ± .0012	.5244 ± .0079	.5910 ± .0037
	RF	.7679 ± .0014	.7796 ± .0015	.4359 ± .0041	.6078 ± .0020

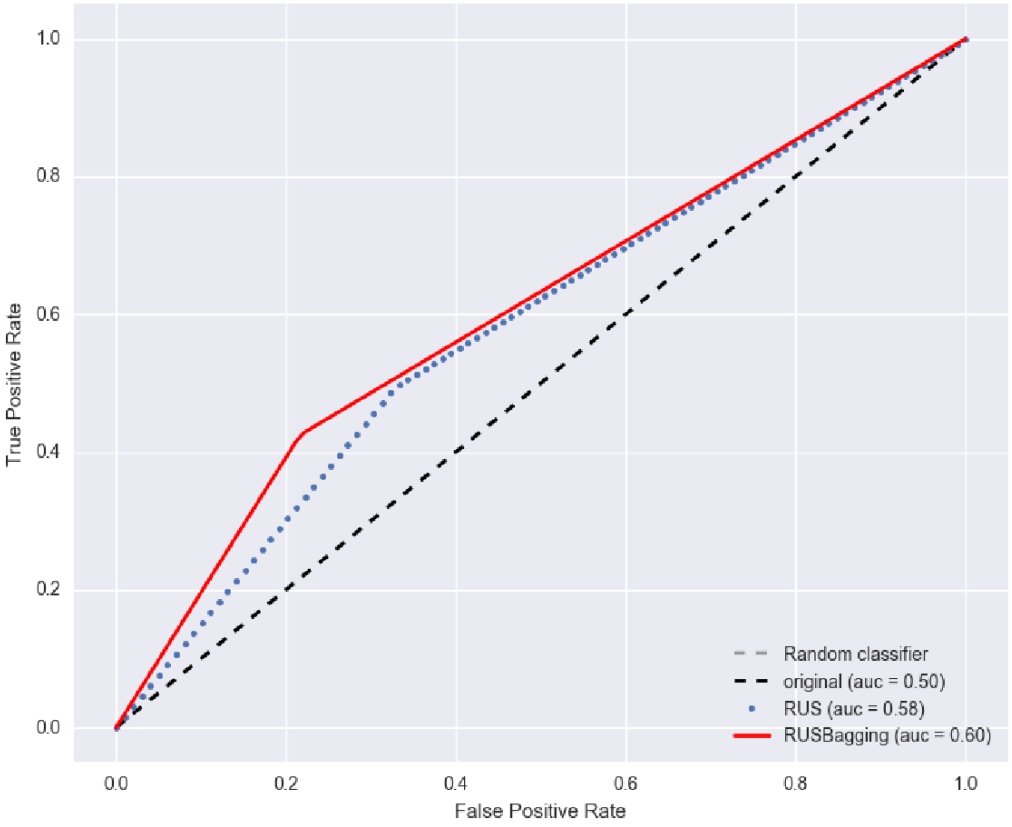


Figure 5.7. ROC curve for DT using undersampling, RUSBagging and original

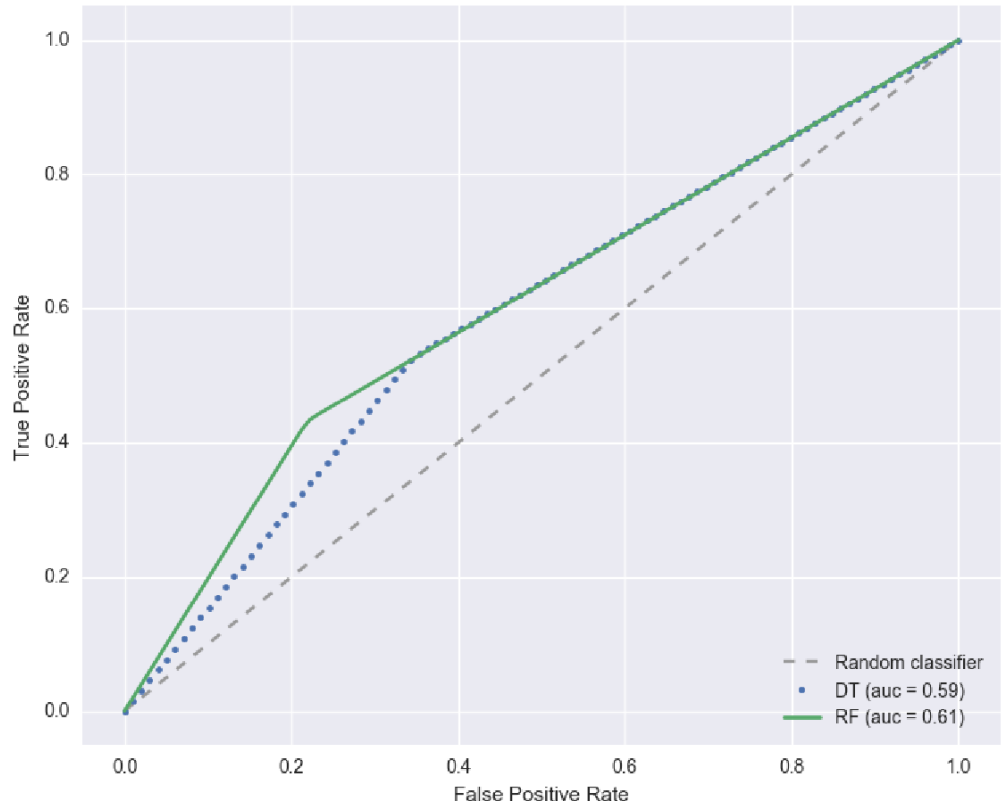


Figure 5.8. ROC curve for DT and RF algorithms using RUSBagging method

The effect of preprocessing and ensemble methods

Several conclusions can be extracted from the results shown in Table 5.8.

- The models trained without modifying the original class distribution were clearly biased towards the majority class. Although accuracy scores were high (>90%), specificity was close to 100% while sensitivity tended to zero. Thus, according to the AUC scores, models performed similar or just slightly better than a random classifier.
- Using random undersampling for class balancing had a direct effect in the performance of the resulting model. Results show that both DT and RF get better AUC scores, 0.56 and 0.58 respectively, and sensitivity increases considerably. However, as could be expected, both accuracy and specificity tend to decrease.
- RUSBagging, which embeds random undersampling within a bootstrap aggregating algorithm, outperforms both previous methodologies. According to the AUC scores, the combination of RUSBagging and Random Forest shows the best performance with a mean of 0.60.
- The performance of the models considering the AUC metric, suggests poor discrimination ability. Nevertheless, a systematic review on risk prediction models for hospital readmission documented similar AUC scores (ranging from 0.50 to 0.70) in most of the studies [Kansagara2011].

Experiment 2, testing AHERF

The goal in this experiment is to test the improvements achieved by AHERF over conventional SVM and RF learning techniques.

Methods

All of the reported experimental results are computed as the average of 50 repetitions of a 10-fold cross-validation approach, where all feature extraction and classification parameters are estimated from the training datasets and applied to the testing datasets as such. We perform a data normalization by the independent computation of the z-score of each input variable given by the expression $z = \frac{x-\mu}{\sigma}$, where x is the input variable, μ is the variable mean estimation, and σ the variable standard deviation estimation. This normalization removes scale effects reducing all variables to the same order of magnitude, and linear shifts. In cross-validation approaches, the μ and σ are estimated on the training data and used as such on the testing data, resulting in some minor inconsistencies if there is any sampling bias.

Model parameter selection.

The following parameters remain to be specified or selected for each combination of data rotation and ensemble of classifiers. All of them are set in the same way for all the cases, because we want to avoid any effect from them in the experimental results.

- L: The number of individual classifiers is set to L = 35 for all experiments.

- Classifier intrinsic parameters: The DT depth is set to 10 in all cases, except for some defaults in scikit-learn. The number of hidden nodes in the ELM is set to min {N/3, 1000}. The SFLN architecture trained by ELM has a single output unit encoding the output of the classifier as an integer value, both for two-class and many-classes datasets.

- K: The number of partitions of the set of features has been set to K =[n/4]. As the effective partitions are random, it is very likely that some of them will be composed of only one vector.

Results

To avoid random number generation bias, we have conducted each execution using a different random number generation seed. Missing values in numerical valued variables (such as glucose level or oxygen saturation) are filled with the arithmetic mean of the variable across the

population. The original dataset is very imbalanced, i.e. the target readmission class samples number is much less than a 0.5% of the dataset. As it is well known, imbalance makes accuracy an unreliable performance measure [Lopez2013]. For instance, a 10-fold cross-validation of the RF classifier upon the entire dataset achieves over 96.2% accuracy, however its average sensitivity is down to 0.4% while specificity reaches 99.8%. The interpretation of the these results is that these RF classifiers are guided by the *a priori* class probability distribution. In essence, RF classification is not very different from assigning all data instances the majority class. The goal in this experiment is to shown the comparative performance of AHERF, therefore we overlook the imbalance problem by building balanced datasets for the computational experiments. The majority class is subsampled to the size of the minority class for each repetition of the cross-validation training process. In our experiment we will consider three different datasets, namely: i) full dataset, ii) paediatric patients and iii) adult patients.

Table 5.9. Accuracy, sensitivity and specificity results (average ± standard deviation) of the classifiers for the different datasets.

		Acc	Sens.	Spec.
Pediatrics	AHERF	78.57±0.47	70.6±0.34	86.55±0.69
	SVM	59±0.17	54.01±0.36	64.04±0.25
	RF	72.72±0.26	64.93±0.52	80.52±0.52
Adults	AHERF	78.17±0.34	72.57±0.33	83.82±0.54
	SVM	65.54±0.6	54.87±0.54	75.23±0.94
	RF	65.54±0.46	55.52±0.63	75.57±0.43
All	AHERF	78.14±0.33	68.02±0.35	88.26±0.6
	SVM	67.78±0.10	44.46±0.10	89.86±0.10
	RF	71.28±0.26	59.56±0.22	82.37±0.46

Table 5.9 shows the average accuracy, sensitivity and specificity along with its respective standard deviation, obtained from the cross-validation experiments. In this table it can be appreciated that sensitivity is much higher than in the reference experiment with the raw unbalanced data, approaching the value of specificity for all the classifier training algorithms, due to the balance of the training dataset. Also, it can be appreciated that AHERF reports results that are significantly better than those of those of SVM and RF ($p < 10{-}6$ in one-sided t-tests using all results of cross-validation folders). Focusing on the sensitivity results, which are more relevant than accuracy and specificity to compare classifier architectures over imbalanced datasets when we are specially concerned by the minority class, we find that AHERF reaches results over or close to 70%, hence it approaches the required performance for real life application. Taking into account that the adult and paediatric populations have quite different statistics, we have performed separate experiments for them, as well as on the entire dataset. It can be appreciated that results on

the separate populations are better than on the entire dataset, which confirms that there are specific discriminant features for these subpopulations. Sensitivity is lower in the paediatric than in the adult population, because the class imbalance is greater in the paediatric dataset than in the adults dataset. Most emergency admissions of children are related to traumatic events that once healed do not relapse. Chronic conditions that are a major cause for readmissions, such as respiratory diseases, are less frequent than in the adult population. More precisely, carrying two-sided t-test in the paediatrics population between sensitivity classifier results, we find that AHERF is significantly ($p<0.0001$) better than SVM and RF, with a performance increase of 22% and 8% respectively. Not surprisingly, RF performance is 15% greater than that of SVM. These differences are bigger if we consider the specificity results measuring success detecting the majority class. If we consider the adult population, we find again that AHERF is significantly better than RF and SVM (two-sided t-test, $p<0.0001$), with a sensitivity performance increase of 23%, while the difference between RF and SVM is not significant. The greater performance increase from AHERF to RF and SVM in the adults population than in the paediatrics population is due to the greater sensitivity of the RF and SVM classifiers to the class imbalance ratio. If we consider the effect on the AHERF we find that there is an increase in sensitivity of 2% from the paediatrics to the adults population, which is barely significant (t-test, $p=0.013$). Pulling together paediatrics and adult population, there is a decrease in sensitivity of AHERF of 5% and 3% relative to the adult and paediatrics results, respectively, due to the fact that discriminant variables are different for each population, so that building a monolithic classifier lose predictive power. The results of AHERF suggest that the approach is promising for a practical implementation of institution specific readmission risk prediction systems.

Experiment 3 testing a new strategy for dealing with highly imbalanced classification problems

Bagging Ensemble Method

Our dataset is highly imbalanced (IR = 28.16), thus we need powerful correcting methods to overcome the bias towards the majority class. Since our dataset has more than 96,000 negative samples, undersampling the majority class may achieve good results, while the risk of discarding crucial information during undersampling is low. We have found that oversampling methods, as SMOTE or ADASYN, perform better in low imbalance ratios. Moreover, we experimentally found that the random generation of samples involving the qualitative variable that specifies the case of the admission gives very bad results. Oversampling qualitative or categorical variables is an open issue not addressed here.

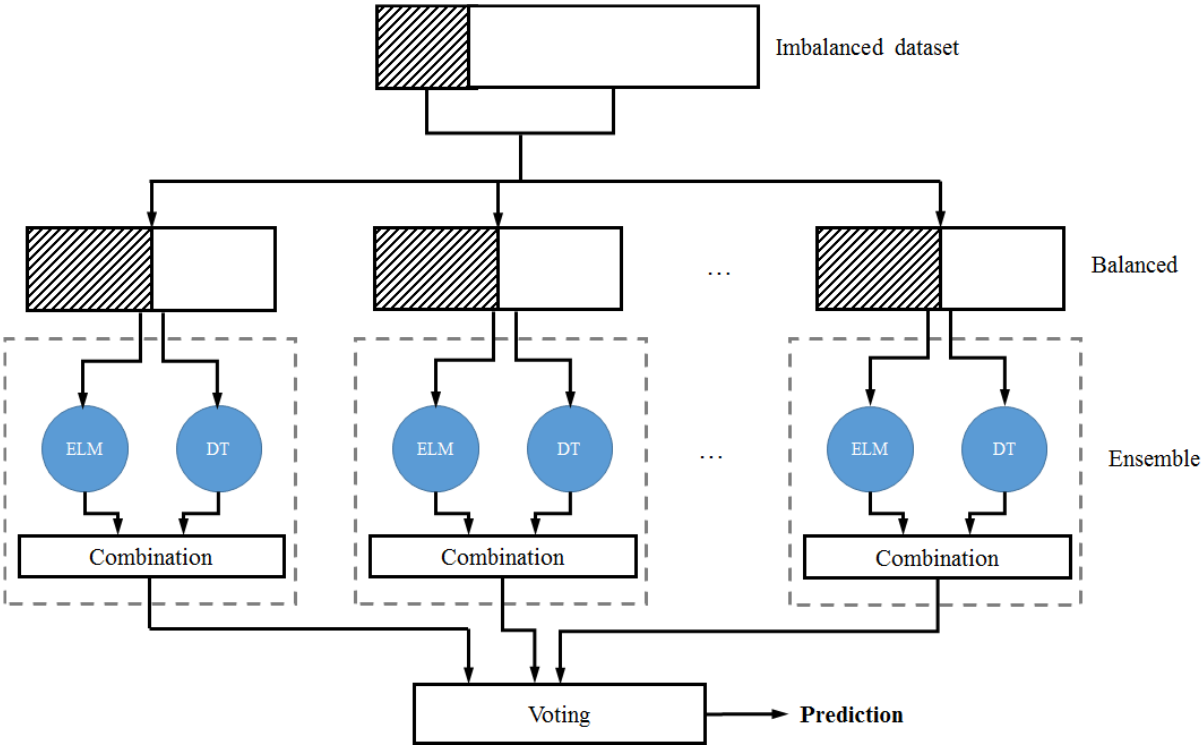


Figure 5.9. Bagging ensemble with resampling

Our method combines a class-balancing preprocessing technique (random undersampling) with bootstrap aggregating, also known as bagging. Bagging consists in creating bootstrapped replicas of the original dataset with replacement (i.e. different copies of the same instance can be found in the same bag), so that different classifiers are trained on each replica. Originally each new dataset or bag maintained the size of the original dataset. Nevertheless, under-bagging and over-bagging

strategies embed a resampling process, so that bags are balanced by means of undersampling or oversampling techniques. To classify an unseen instance, the output predictions of the weak classifiers are collected performing a majority vote in order to produce the joint ensemble prediction. The purpose of this combination is to create a model to classify imbalanced data, improving the generalization capacity without sacrificing overall accuracy. As shown in Figure 5.9, our approach consists in applying a balancing pre-process to each subset obtained from the bootstrap. Following, an ensemble classifier is built, combining ELM and Decision Tree classifiers using soft voting as combination strategy. The black-box nature of ELMs (and ensemble methods in general) is combined with the comprehensibility of a decision tree. Some works [Lin2013] have combined ELM with DT due to its interpretable ability as 'IF-THEN'-like rule generator.

Results

In order to evaluate the effectiveness of our proposed approach, henceforth denoted bagging ensemble, we compare its performance with other well-known classifiers, namely: Naive Bayes, Decision Tree, Random Forest and Extreme Learning Machine. We have evaluated each method using i) the original data distribution, and ii) applying random undersampling (RUS) as a preprocessing technique to achieve a training dataset with balanced a priori class distribution. Our experiments were implemented using the open source machine learning library scikit-learn. All the evaluations were performed using 5-fold cross-validation.

According to the results shown in Table 5.10, it is clear that class imbalance conditions overall performance of the model, regardless of the classifier we use. When original skewed data is employed, high accuracy scores (above 90% in all cases) and fairly poor recall scores are achieved. This behaviour, sometimes referred as 'accuracy paradox', is caused by a high class imbalance that imposes a strong bias towards the majority (normal admission) class. When random undersampling is applied, accuracy decreases and recall increases due to the a priori class probability balancing. Tree-type algorithms (DT and RF) achieve better AUC scores when class balancing techniques are applied (increases of 3.6% and 6.8% respectively). This improvement, on the other hand, does not occur when using Naive Bayes and ELM, which perform similarly in both scenarios.

Table 5.10. Comparison of different machine learning methods (mean ± standard deviation) measured by AUC, recall, specificity, and accuracy. RUS (random undersampling) is applied.

Model		AUC	recall	specificity	accuracy
Bagging Ensemble	-	0.647 ± .01	0.474 ± .04	0.759 ± .00	0.736 ± .03
Naive Bayes	-	0.587 ± .01	0.145 ± .01	0.944 ± .00	0.917 ± .00
	rus	0.589 ± .01	0.211 ± .03	0.894 ± .00	0.869 ± .03
Decision Tree	-	0.517 ± .01	0.071 ± .01	0.959 ± .00	0.929 ± .00
	rus	0.553 ± .00	0.470 ± .02	0.555 ± .00	0.647 ± .01
Random Forest	-	0.559 ± .00	0.001 ± .00	0.999 ± .00	0.965 ± .00
	rus	0.627 ± .00	0.373 ± .01	0.665 ± .00	0.761 ± .01
ELM	-	0.546 ± .02	0.001 ± .00	0.999 ± .05	0.965 ± .00
	rus	0.551 ± .02	0.452 ± .1	0.626 ± 0.00	0.624 ± 0.09

The Area under the ROC curve (AUC) is the most widely used metric to evaluate readmission risk prediction in the literature. According to the results shown in Table 5.10, our bagging-ensemble achieves the best score followed by Random Forest with random undersampling preprocessing. Figure 5.10 shows the ROC curves for different classifiers using random undersampling for data balancing. We can see that bagging-ensemble (red) is the best performing method, followed by random forest (blue dots). The individual classifiers with better sensitivity performance are DT and ELM with class-balancing. This explains why bagging-ensemble has the best sensitivity scores (47.4%). Random Forest and Naive Bayes, on the other hand, score poorly in comparison (37.3% and 21.1% respectively).

When it comes to decision tree classifiers, in our preliminary experiments we have used the default configuration of the CART algorithm implemented in scikit-learn. In that case, the maximum depth of the tree is not specified beforehand, so that tree's depth is set according to a certain termination criterion. In order to analyse the effect of the maximum tree depth in the overall performance of the model we have evaluated several decision trees with different 'maximum tree depth' values. Figure 5.11 shows the AUC scores of decision tree classifiers trained using both original and balanced datasets. Both configurations achieve the best results at a depth of 5-10 and results get worse afterwards, although trends are different. However, when we explore the behaviour of recall scores, we find that classifiers trained with imbalanced dataset achieve poor results as shown in Figure 5.12. Class balancing, on the other hand, improves classifiers' performance to a 55%.

In order to determinate the impact that the number of hidden units of the ELM has in the performance of our bagging ensemble, we have conducted a test consisting of measuring the recall scores of models with different hidden unit values. Figure 5.13 shows a peak at 30 hidden units and a plateau at around 150 units. According to this results, in our tests we have used 30 hidden units.

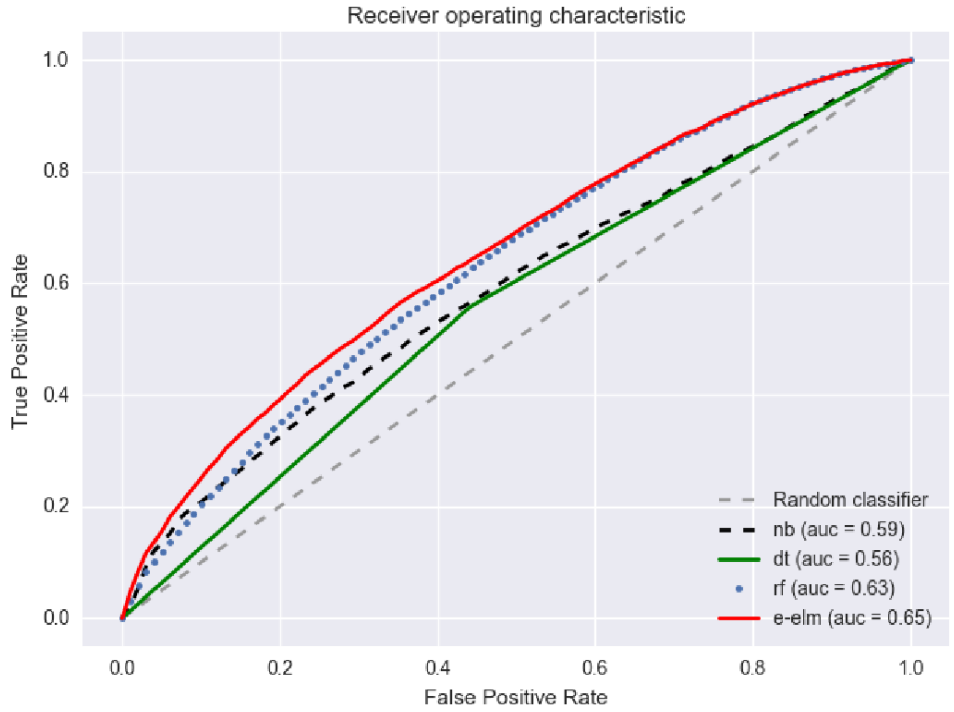


Figure 5.10. Comparison of ROC curves for different methods with random undersampling

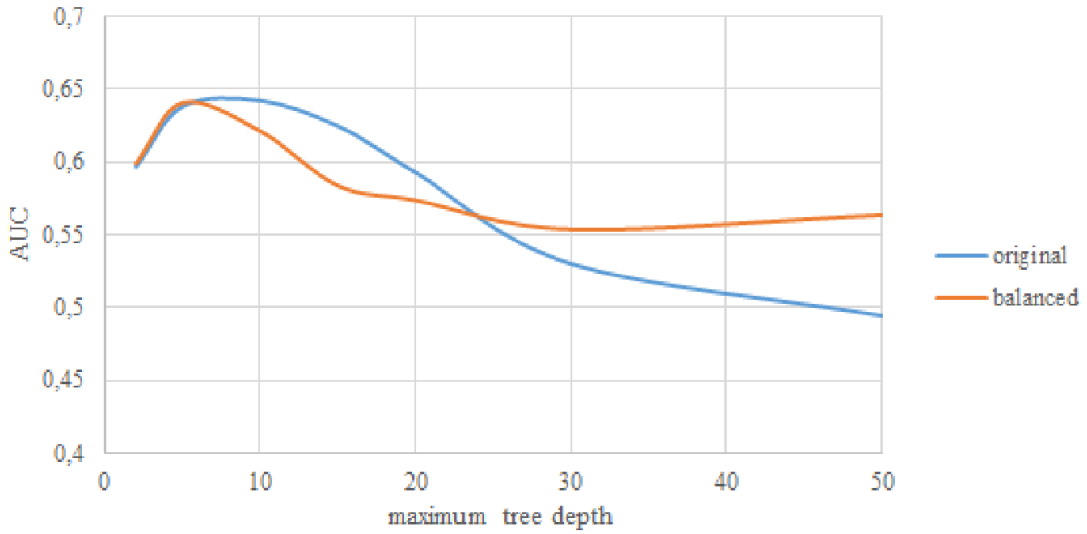


Figure 5.11. AUC versus maximum DT depth.

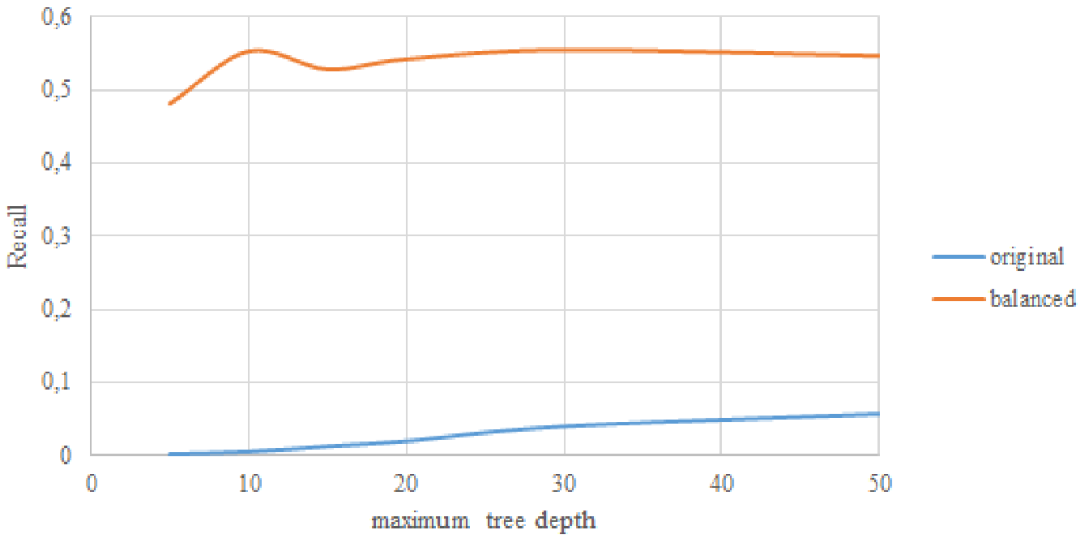


Figure 5.12. Recall versus maximum DT depth

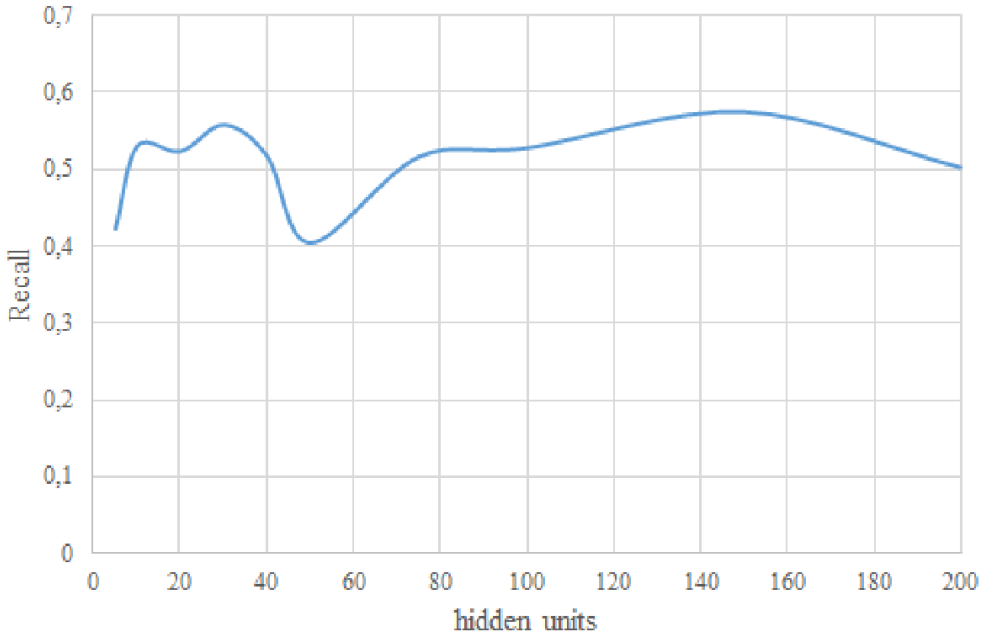


Figure 5.13. Recall versus number of hidden units in the ELM.

5.4. Heart Failure readmission prediction

Work on the problem of readmission risk prediction in heart failure (HF) has been done in a R&D project involving HF patient telemonitoring and predictive modelling (INCAR project funded by the Basque Government by means of HAZITEK 2016 program). We collaborated with cardiologists from OSI Bilbao Basurto, Dr. Nekane Murga and Dr. Vanessa Escolar who gave us access to anonymized and curated EHR and telemonitoring data. A thorough description of the dataset was reported in Chapter 3.

Our approach consisted in experimenting with different feature selection techniques and conventional classification algorithms. We used basal information of the patients collected from hospital's EHR.

5.4.1. Experiment 1: feature selection

Methods

In this experiment our goal is to make a preliminary data analysis in order to figure out which features are more related to the HF readmission risk, using only baseline health status data as reported the dataset description in Chapter 3. The dataset is composed of 60 attributes collected from 119 patients with cardiovascular disease (CVD) from which 30 of them were readmitted within 30 days (if a patient is readmitted more than once, only the first admission is included) and 12 died. We make use of feature subset selection techniques that allow us identifying the most significant variables or groups of variables of our dataset. In this section we will present the results obtained from the application of the following feature selection algorithms to our dataset:

- Correlation-based Feature Selection (CFS)
- Random Forest, embedded FS (RF)
- Sequential Forward Selection + SFS-SVM
- Sequential Backward Selection + SBS-SVM

To avoid the bias that may be introduced by circularity analysis, we carry out the feature selection process independently for each LOO cross-validation iterations, i.e. we carry out 10 feature selection processes.

In order to analyse which features are associated with HF readmission or death, we built classification models using different feature subsets following a wrapper approach. In this models, the outcome was the unplanned readmission or death within 30 days after discharge from HF

hospitalization (0 for not readmitted, 1 for readmitted or dead). The evaluation of the models was made by performing 10 independent executions using leave-one-out accuracy estimation. We used the well-known Random Forest (Gini as splitting criterion and 10 estimators) and SVM (radial basis function kernel, C=1 and gamma=1/number of features) classification algorithms, implemented in the open source machine learning library scikit-learn.

Results

Table 5.11. Mean accuracy and its standard deviation for each classification algorithm and FS method

	none	**CFS**	**RF**	**SFS-SVM**	**SBS-SVM**
RF	.6227 ± .02	.6193 ± .03	.6353 ± .03	.6605 ± .02	.6454 ± .02
SVM	.6471 ± .00	.6471 ± .00	.6471 ± .00	.6639 ± .00	.6639 ± .00

Table 5.11 shows the mean accuracy along with the standard deviation of each model trained with the specified configuration. Results show that wrapper methods (using SVM) outperform other feature selection techniques. However, we observe that our models, regardless of the underlying method they utilize, perform poorly (below 67% accuracy).

In order to assess the stability of the feature selection processes, Table 5.12 shows the list of features that have been selected by each method. For those randomized algorithms the number of times each feature was selected is shown. According to the results shown, several conclusions can be extracted:

- We observe that SBS method tends to be more stable in their feature sets, since the majority of the selected features are present in multiple runs. It is noteworthy that ‘years since first diagnostic’ is a feature that is present in every execution, despite it is not present in the rest of methods. The reason may be related with the hill-climbing algorithm underlying, that is influenced by a local peak at the end part of the feature vector, so that features in this positions are more likely to be selected.
- On the other hand, SFS method selects a greater number of features although many of the selected features are only present in one of the runs.
- There is not a single feature that reaches the total consensus, that is, it is selected by all the methods at least in one run. Nevertheless, urea and pacemaker rhythm are two of the top features in terms of consensus, since they are present in all the FS method groups (i.e. filter, embedded and wrapper) and in many runs.

Figure 5.14 and Figure 5.15 show the ROCs of the SVM and RF, respectively, after SBS-SVM feature selection. Both approaches improve over random choice, but some improvement of RF

over SVM can be appreciated. Nevertheless, the results are far being excellent. Most of the blame goes to the poor informative value of the original variables.

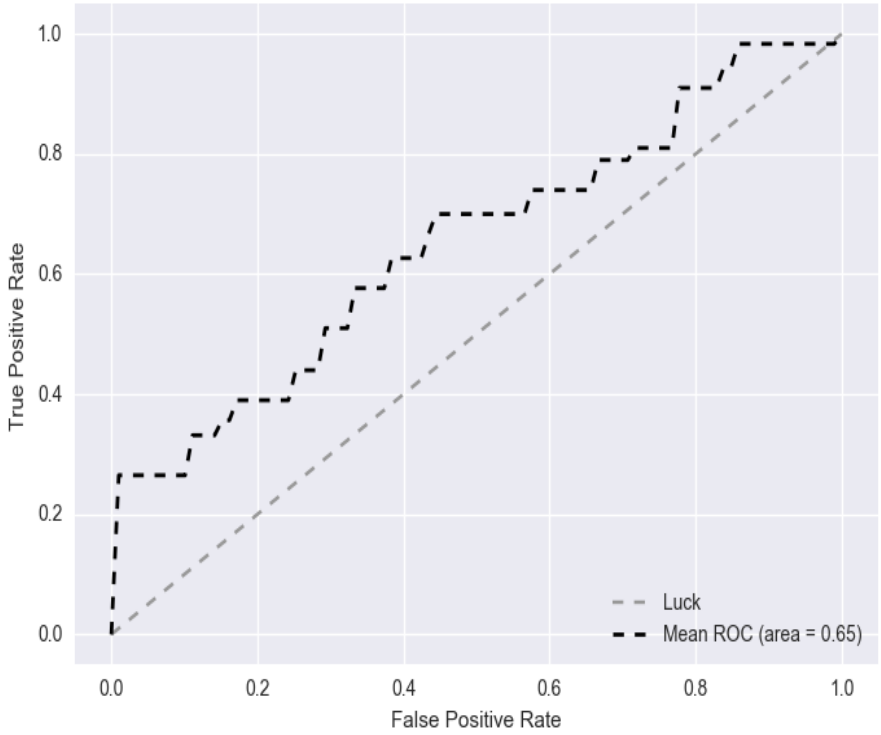


Figure 5.14. Roc curve (SVM + SBS-SVM)

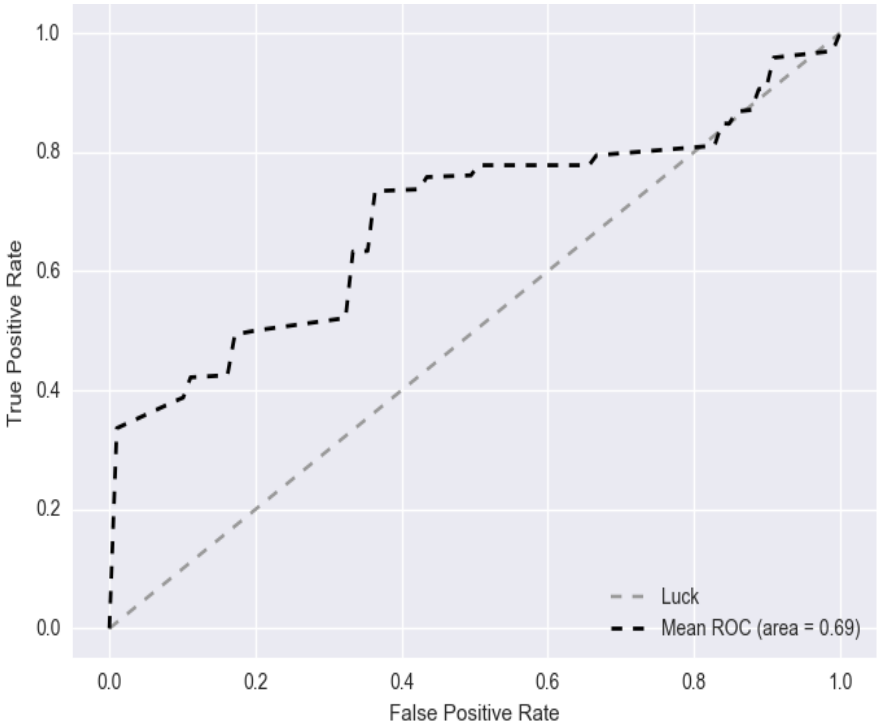


Figure 5.15. ROC curve (RF + SBS-SVM)

Table 5.12. List of variable included in the model by each method and number of times they were selected in the 10 randomized runs

Attribute	CFS	RF	SFS-SVM	SBS-SVM
Gender			1	
Smoker			1	
Weight	x	2	5	
Height		1		
HR		4		
SO2			2	
SBP		5		
Implant-dev			7	
Need oxygen			1	
Urea	x	10		7
Creatinine		4		
Sodium		1	1	
Potassium	x		1	
Hemoglobin			1	
Total cholesterol	x	1		2
HDL cholesterol		2	2	
Triglycerides	x	3		6
Torasemide	x		2	
Thiazide	x			
ACEIs			1	
ARB			3	
Ivabrandine	x			
COPD			1	
Connective tissue disease	x			
Peptic ulcer	x		4	
Diabetes mellitus			2	
Any tumour			1	
Moderate/severe liver disease				1
Metastatic solid tumour				1
Osteoarthritis/arthrosis/spondylitis				1
Osteoporosis	x			3
Sinus rhythm			1	
Atrial fibrillation				4
Pacemaker rhythm	x		8	7
Admission days		1	4	4
Age				2
Years first diagnostic				10

5.4.2. Experiment 2 comparison of classifiers upon complete feature set

Methods

In this experiment we wanted to gain more insights about Basurto dataset and its potential predictive capabilities. Rather than focusing on feature selection, we worked on preliminarily analysing the dataset from a prediction ability point of view by evaluating different model configurations. We first performed a dimensionality reduction process in order to visualize the dataset and get some insights about its linear separability, possible overlaps etc.

Afterwards, we tested different well-known classification algorithms and compare its performance in terms of area under the ROC curve (AUC). In a preliminary phase we compared the following classifiers:

- CART Decision Tree
- Random Forest
- Support Vector Machine (SVM) with radial basis function (RBF)
- SVM with linear kernel

All the classifiers used the default configuration of the parameters as provided by scikit-learn.

In a second phase we compared different resampling techniques for overcoming the class imbalance. In order to do so, we selected the two different algorithms, namely Random Forest and SVM with linear kernel, and evaluated their performance using different oversampling procedures. The following configurations were evaluated:

- Original class distribution
- Random Oversampling (ROS)
- Synthetic Minority Oversampling Technique (SMOTE)
- Adaptive Synthetic Sampling Approach (ADASYN)

In the manner of the classification algorithms, we used the default configuration parameters as provided by the imbalanced-learn module[3] for scikit-learn.

Results

Figure 5.16 and Figure 5.17 show the scatter plot of the dataset after reducing the dimensionality by means of PCA to the 2 and 3 first components respectively. Although, PCA is not a specific technique to find the optimal projection separating the two classes, it is illustrative enough when the classes are well separable. The visualization shows that there is not clear immediate separation between classes.

[3] https://github.com/scikit-learn-contrib/imbalanced-learn

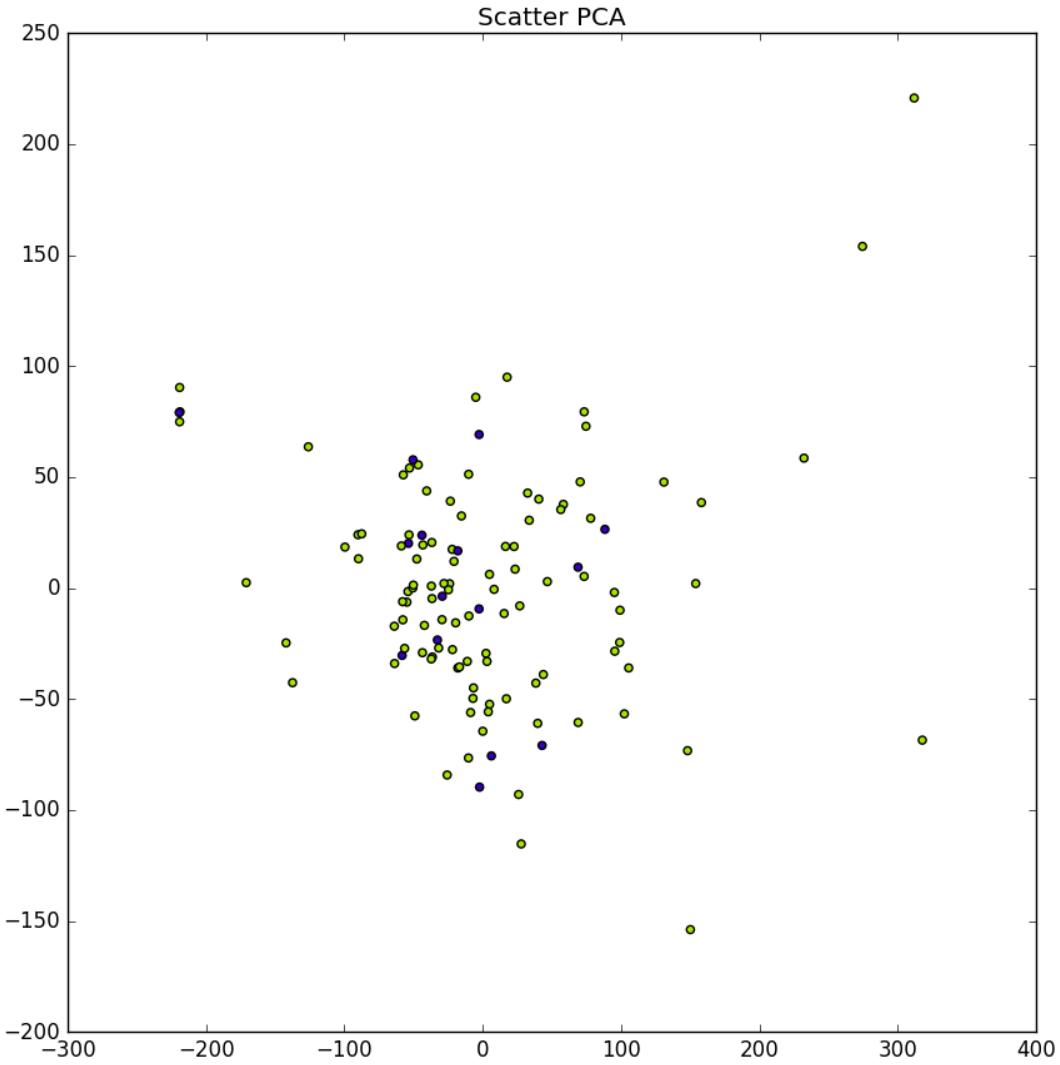


Figure 5.16. Scatter plot of the first 2 components of PCA

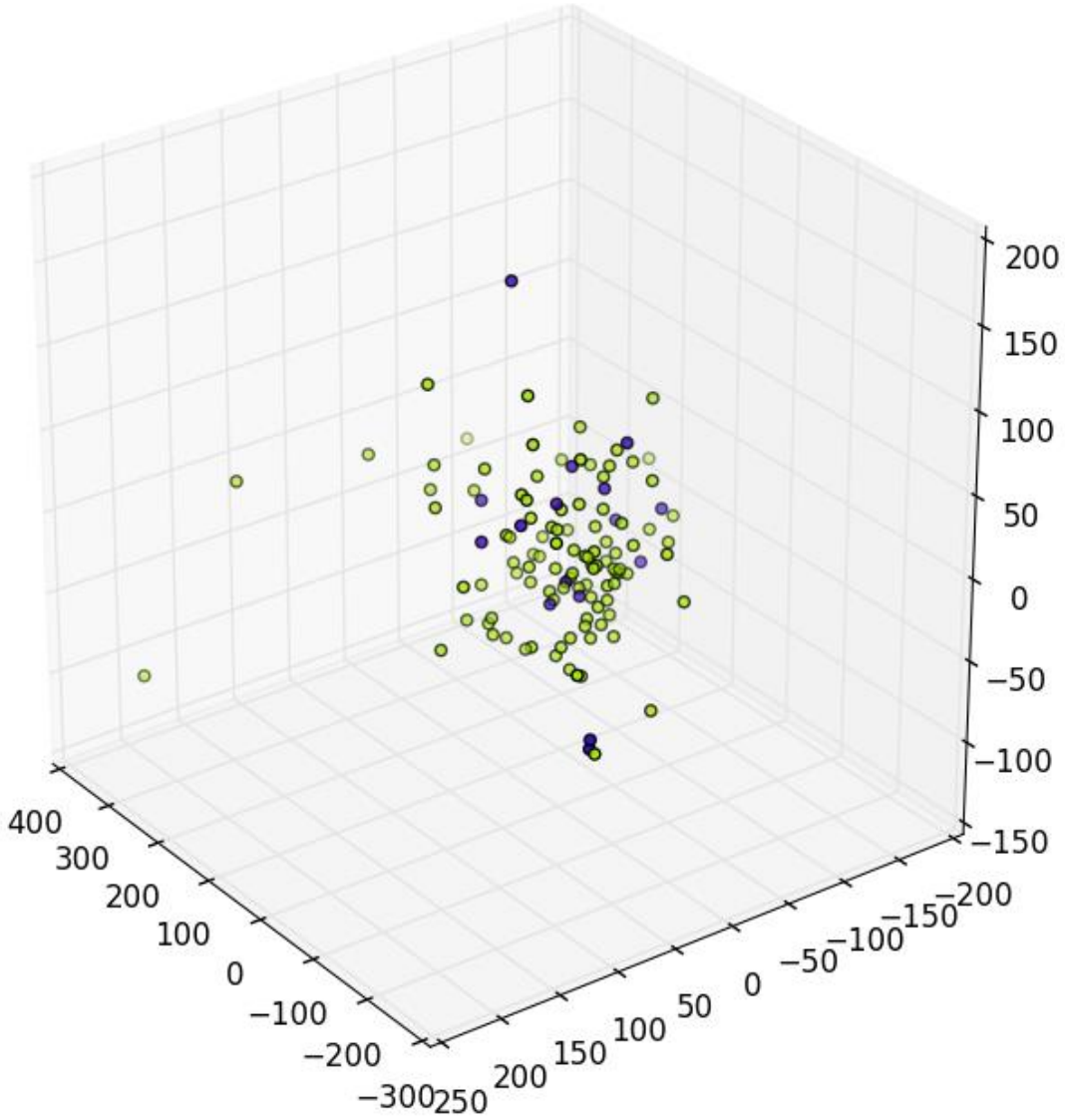


Figure 5.17. 3D Scatter plot of the first 3 components of PCA

Table 5.13 provides the 10-fold cross validation AUC scores of each classifier. In order to reduce the bias produced by the random splitting of the dataset, we repeat each experiment 50 times and average the results obtained (standard deviation is also provided). According to the results we observe that the predictive ability of the models is poor with AUC scores ranging from 0.47 for Decision tree, which performs worst, to 0.58 for SVM with linear kernel, the best performer (yet very low).

Table 5.13. 10-fold cross-validation of AUC over the different classification algorithms

	AUC (mean ± standard deviation)

DT	0.476 ± 0.050
RF	0.500 ± 0.059
SVM (RBF)	0.525 ± 0.039
SVM (linear)	0.587 ± 0.034

It's worth noting that tests are performed using the original dataset, which has a skewed class distribution with an imbalance ratio of 1:4. To minimize the class imbalance problem, we decided to use oversampling methods, since undersampling wasn't feasible due to the small size of the dataset. Table 5.14 and Figure 5.18 show the comparison of AUC over the original distribution and the different oversampling procedures. The adaptive oversampling provides better results, though not statistically significant ($p>0.01$). According to the AUC results, it can be observed that SVM improves more when using oversampling techniques in comparison to random forest. All oversampling procedures perform similarly with small variations that are not statistically significant (ANOVA test). In addition, it can be observed that standard deviation is high. Figure 5.19 shows the ROC curves for the different classifiers for an instance of their execution.

Table 5.14. ROC AUC scores for SVM and RF classifiers with the original data distribution and distributions after different class imbalance correction procedures (mean+-standard deviation)

	Original distribution	**Random oversampling**	**SMOTE**	**ADASYN**
SVM	0.56 +- 0.03	0.65 +- 0.08	0.67 +- 0.05	0.68 +- 0.04
RF	0.53 +- 0.06	0.54 +- 0.08	0.56 +- 0.09	0.58 +- 0.05

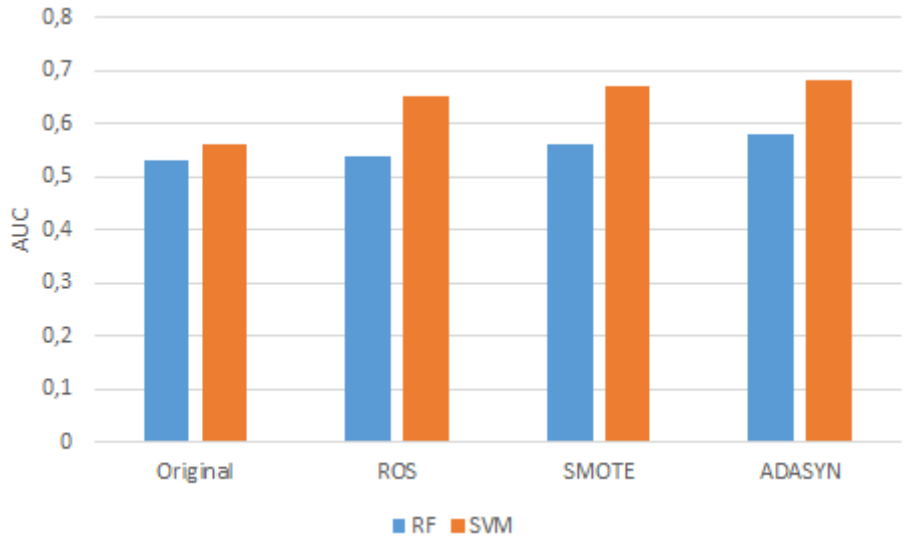


Figure 5.18. AUC comparison for different class distributions

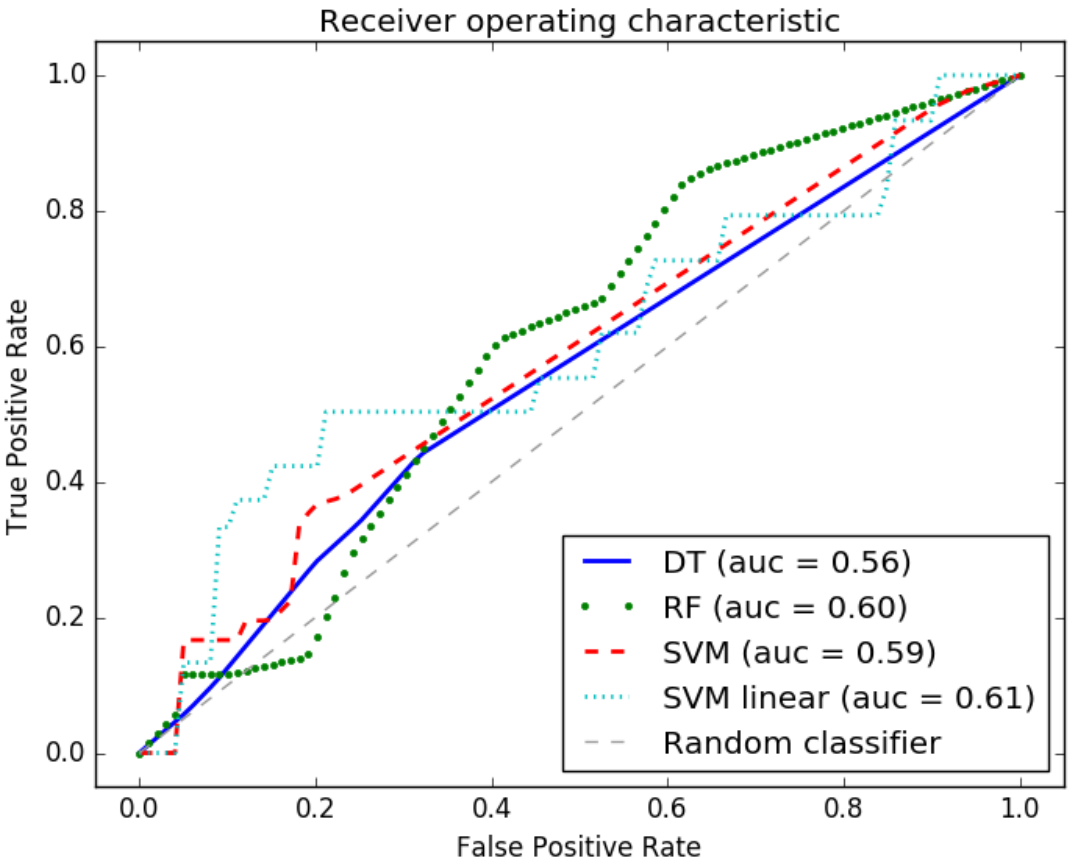
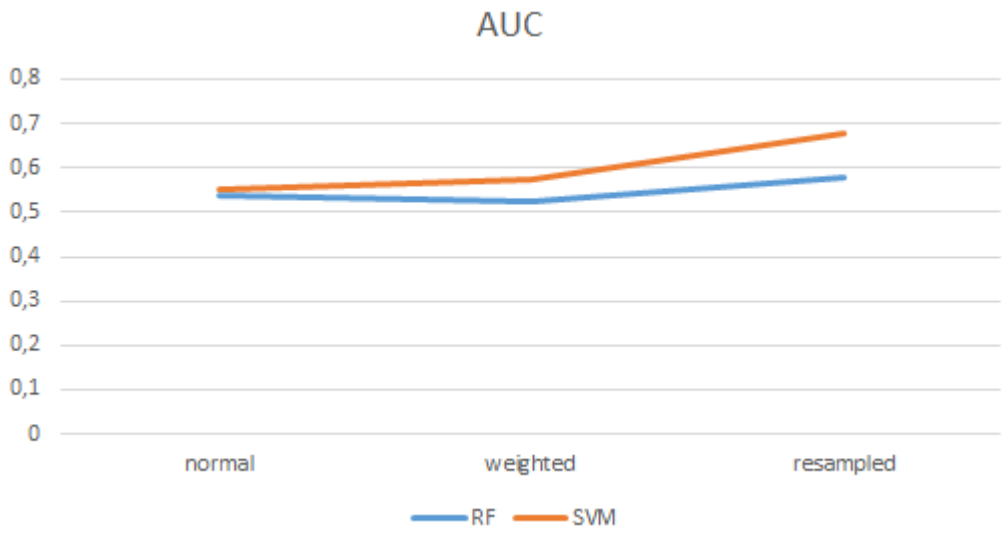
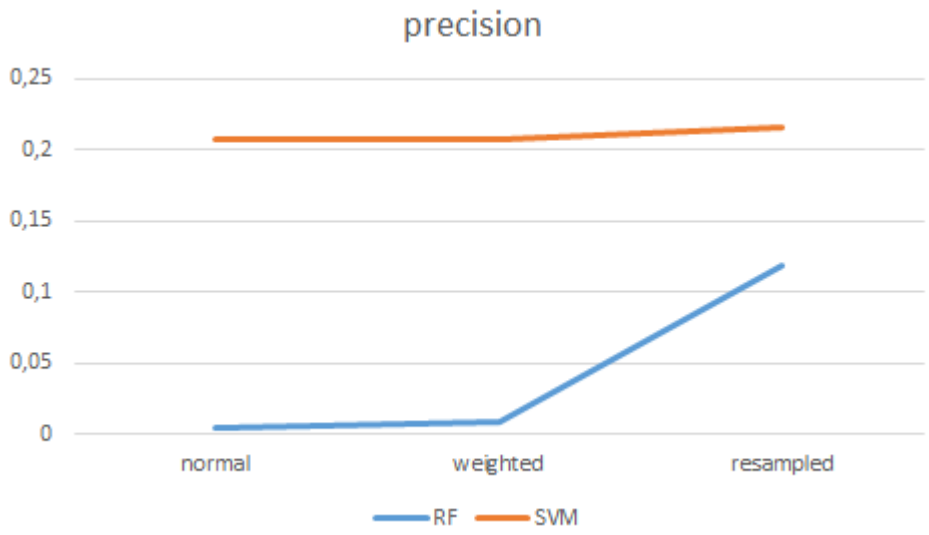


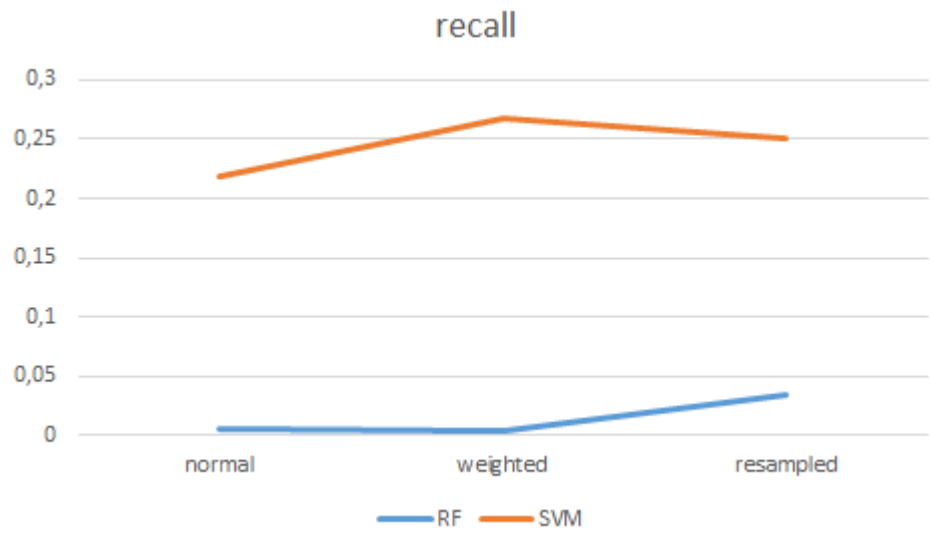
Figure 5.19. ROC plot of different classifiers for an instance of their execution.


AUC
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
normal
weighted
resampled
RF
SVM

(a)


precision
0,25
0,2
0,15
0,1
0,05
0
normal
weighted
resampled
RF
SVM

(b)



(c)

Figure 5.20. Performance comparison of Random Forest and SVM classifiers using normal distribution, weighting and resampling (SMOTE)

In this Chapter we have collected all the experimental results of the Thesis, hence its length. Overall, it can be concluded that the prediction accuracy of readmissions is low in all cases, as has been acknowledged in the literature, almost independently of the classifier building approach used. The poor informative value of the available variables is mostly responsible of the results (primarily administrative and demographic) as well as the very limited clinical inspection at the admission time.

Regarding ED readmission prediction results using several strategies for class balancing have shown some improvement, but not sufficient as to declare the problem satisfactorily solved from a machine learning point of view. Future work must address better strategies for planning data gathering in clinical studies, so that new and more informative variables can be detected.

We have also presented a preliminary study for identifying risk factors associated to unplanned readmission or death, over a HF clinical dataset. Different classification algorithms and feature selection methods were employed in order to increase the prediction ability of the models and reduce their complexity in terms of number of features. Results have shown that sequential (backward or forward) feature selection methods in combination with SVM perform the best in terms of estimated prediction accuracy. Nevertheless, according to the overall poor performance of the models, we hypothesize that baseline status data by itself may not have sufficient predictive capacity. As future work with HF we aim to study the monitored data to further improve the prediction of patient readmission or mortality. Additionally, we aim to develop a system that

incorporates preventive actions. For that, we will develop a patient guidance system in a mobile platform, which will be based on the knowledge obtained from the predictive models, and the preventive actions that clinicians define.

Chapter 6

Conclusions

This Thesis deals with the prediction of patient readmissions in the healthcare system. This issue has been recognized as a key indicator of healthcare quality from both the economical/financial and patient attention points of view, justifying the relevance of the thesis topic. Two specific areas have been addressed, the emergency department (ED) and the monitoring of heart failure (HF) patients. The definition of the appropriate time period in order to consider a patient visit to the hospital as a readmission is a matter of debate and political decisions, even in our restricted study we have to deal with two such thresholds (i.e. 30 and 3 days). Since readmission risk can be moving from one geographical setting to another, recent trends favour the construction of specific prediction models using machine learning techniques and methodologies trying to predict if a patient admission will lead to a readmission. Traditionally, readmission prediction models have been built using a set of classical well-known statistical tools. Although increasingly authors propose machine learning as a way to improve the prediction ability of the models, we encountered several issues that have not been addressed.

An important challenge that we have found attacking the problem is that experimental datasets are heavily class imbalanced, which is challenging for most of the machine learning tools. Consequently, specific methods for dealing with imbalanced datasets can be of great use in this setting. Additionally, we took care of using an appropriate performance measure, such as AUC or sensitivity, because accuracy can be misleading in imbalanced datasets. We have concluded that the precision-recall curve would be more meaningful than the ROC curve, although the *de facto* standard in the field is still the latter.

We have taken special care to carry out pure cross-validated experiments without corrupting test results with training data influences (circular analysis) i.e. by carrying all preprocessing and classifier building exclusively over the training data.

We carried out a detailed analysis of the state-of-the-art approaches addressing the issue of class imbalance. We have worked testing different methods for alleviating the majority class using real life medical datasets and evaluated their effect on model's predictive capabilities. We also contribute an ensemble method that combines resampling with bagging and ensemble of classifiers, which, outperforms other class-balancing procedures, albeit having its limitations.
Nevertheless, the results achieved by all classifier building algorithms are modest, in agreement with most of the literature. The main reason is that the variables are not very informative so that we tried feature selection methods in order to enhance feature discrimination power. There was not any dramatic improvement, even when these procedures usually improve the predictive power of the classifier, by improving the signal to noise ratio and simplifying the search space. This leads us to confirm that from poor quality variables it is not possible to build good feature descriptors.
We have even contributed a new architecture, the AHERF, as a hybrid of ELMs and rotation forests, which produces some improvement in the results and limited robustness against the class imbalance problem, encouraging further experimentation and evolution of this architecture.

For the future work recommendations, the most important one is the realization of data gathering studies including a wider spectrum of variables, so that future computational experiments have a better base for the development of feature selection and classifier building approaches. These studies must incorporate improved data capture methods and devices which facilitate the work of the clinicians and reduce the error and/or the missing data.
Finally, we found that, to the best of our knowledge, there is no readmission dataset publicly available. For this reason, up to now each study makes use of its own healthcare institution's data, what causes a lack of comparability among different studies. In this regard, we made publicly available a synthetic dataset that is a transformation based on real clinical data (due to right holder's permission issues). Nevertheless, we will keep on the effort of making public a real anonymized dataset, so that it can serve as the benchmark for future models.

Appendix A

HF patient telemonitoring program

This appendix is devoted to the description of the telemonitoring program at OSI Bilbao-Basurto which is related to INCAR project (RIS3 EUSKADI SALUD-2016).

Inclusion-exclusion criteria

Inclusion criteria

- Hospitalization or emergency visit due to decompensation of Heart Failure (with need and administration of diuretics) in the previous 6 months, and at least one of the following three conditions.
    - Left ventricular ejection fraction <45% (at least once in the last year or on the last electrocardiogram, if it is older).
    - Left ventricular ejection fraction> 45% but BNP> 400 (or more NT- but BNP> 1500) at least once during the last year.
    - Diagnosis of HF confirmed by a cardiologist.
- Ability to use telemonitoring devices (either by patient or caregiver).
- Existence of telephone line at the patient's home.
- The patient gives written informed consent to use telemonitoring

Exclusion criteria

- Myocardial infarction or percutaneous coronary intervention in the last 3 months or planned.
- Coronary artery bypass graft, valve replacement or correction in the last 6 months.
- Severe comorbidity with life expectancy <12 months.

- Inability to use the devices provided.
- Cognitive inability to participate.
- Denial of written informed consent.

Patient profiling

Left Ventricular ejection fraction (LVEF)

- >50% -> Normal or HF with diastolic dysfunction
- 40-50% -> Intermediate ejection fraction
- 30-40% -> Depressed ejection fraction
- <30% -> Very depressed ejection fraction

(values <15% or >75% are discarded)

Etiology

- Ischemic
- Not ischemic

Cardiac rhythm

- Sinus rhythm
- Atrial fibrillation
- Pacemaker

Evolution

- <1 year
- >1 year

Anemia

Hemoglobin (Hb) <11 -> Yes

Hb>11 -> No

(values Hb>17 or Hb<6 are discarded)

Questionnaire

Questions' answers were encoded to ensure data alignment. For that, the polarity of the questions was modified if necessary in a way that negative answers were given the maximum score and positive answers were given the minimum score.

#	Question	Response Encoding
1	With respect to previous three days, I feel:	Better=0 Same=0 Worse=1
2	Does the medication do me good?	Yes=0 No=1
3	In the last 3 days, have I taken any medication without supervision from my doctor?	No=0 Yes =1
4	Am I following the diet and exercise recommendations given by my doctor and nurse?	Yes =0 No=1
5	In the last 3 days my ankles are:	Better=0 Same=0 Worse=1
6	Can you take walks like previous days?	Yes =0 No=1
7	Do I feel breathless or shortness of breath when I lie in bed?	No=0 Yes =1
8	Do I notice that I have begun to have cough or to expel phlegm?	No=0 Yes =1
9	Have I noticed fatigue at rest?	No=0 Yes =1
10	If fatigue – Can I take walks on flat?	Yes =0 No=1

Appendix B

Systematic Review

The following table contains the data extracted during the systematic review process.

#	Identifier	population	FS Method	Classification algorithm	Readmission rate (%)	No. of instances	Readmission	Discrimination (AUC)
1	abdelrahman2014	HF	Wrapper (final), information gain, gain ratio, symmetrical uncertainty	LR, voting feature intervals (VFI)	19	2787	30-day	0,86
2	alassaad2015	80<	PCA - collinear variables	Cox regression	68	368	12-month	0,71

			removed -> backward elimination					
3	Allaudeen2011	all	univariate GEE	GEE	17	10359	30-day	NR
4	Allen2012	Systolic HF	Stepwise LR	LR	13,3	4584	30-day	0,64
5	allison2014	OPAT	backwards selection LR	LR	26	782	30-day	0,61
6	amalakuhan2012	COPD		RF	47	106	12 month	0,72
7	amarasingham2010	HF	univariate LR and multivariate LR	LR	24,7 (3,1)	1372	30-day (or death)	0,72 (0,86)
8	Au2012	HF	Random Forest	RF	18,77	59652	30-day	0,54-0,61
9	baillie2013	all		NR	14,4	120396	30-day	0,61
10	baltodano2016	ventral hernia repair	univariate LR	LR	4,7	17789	30-day	0,71
11	bergese2017	pediatric ED		Classification Tree, ANN	2,2	28341	120-hour	NR
12	Berman2011	advanced liver disease	univariate -> forward stepwise LR	LR	20	554	30-day	NR
13	betihavas2015	HF	backward elimination Cox	Cox regression	13	280	28-day	0,8
14	Billings2012	all		LR	12,2	576868	30-day	0,7
15	cai2016	all	CBFS with best-first search	Bayesian network	NR	32634	7-day	0,82
16	Coleman2004	65<	backward elimination	LR	NR	1401	30-day	0,77-0,83

			LR					
17	cui2015	all	bivariate	LR	33,7	61926	12 month	0,7
18	Deschodt2015	75<	univariate -> backward LR	LR	18,5-29,1	442	1-month, 3-month	NR
19	Dharmarajan2013	HF, AMI, Pneumonia	Univariate Cox regression	LR	24,8 (HF)	1330157 (HF)	30-day	NR
20	Donze2013	all	univariable LR -> backward elimination LR	LR	22,3	10731	30-day	0,67-0,71
21	dorajoo2017	All	backward elimination LR	LR	45	1291	15-day	0,65
22	Epstein2011	HF, Pneumonia (65<)	univariate -> sequential removal LR	HGLM	11-32 (HF)	234477	30, 60, 90-day	NR
23	fisher2016	In rehabilitation & high risk	univariate	Classification Tree, HGLM	25,3	25908	30-day	0,58-0,69
24	Garrison2013	all	bivariate wilcoxon rank sum, ficher, chi2	LR	30,4	276	30-day	NR
25	Halfon2006	all	univariate -> backward elimination Poisson R.	Poisson regression	5,1 (potentially avoidable)	131809	30-day	0,67-0,72
26	Hao2015	all	variance minimization criterion	Survival RF	NR	211232	30-day	0,72

27	Hasan2010	all	LR	LR	17,5	10946	30-day	0,61-0,65
28	Jencks2009	65< or disabled		Cox regression	19,6	11855702	30-day, 180-day	NR
29	kaur2016	Paediatric ICU	univariate LR -> forward & backward LR	LR	33	256	48-hour	0,61
30	Keenan2008	HF	stepwise selection LR	LR	23,6	567447	30-day	0,61
31	leong2017risk	HF	bivariate	LR	9,8	1475	30-day	0,76
32	lopez2011	64<	forward stepwise LR	LR	1,3	28430	180-day	0,76
33	low2016	Asian adults	univariate LR -> multivariate LR	LR	15,5	74102	30-day	0,78
34	Marcantonio1999	65<	bivariate -> backward elimination LR	LR	50	308	30-day	NR
35	mclaren2016prior	HF	univariate	LR	18	1999	30-day	0,63
36	mcmanus2016	AMI (65<)	PCA-based feature reduction	LR	13,18	804	30-day	0,63
37	Morris2014	ED (60<)	stepwise LR	LR	NR	585888	90-day	NR
38	Nguyen2014	COPD	univariate	GEE	18	4596	30-day	NR
39	Nijhawan2012	HIV	univariate LR	LR	25	2476	30-day	0,72
40	ouanes2012	ICU	univariate	LR	3	3462	7-day	0,74
41	padhukasahasram2015	HF		cox regression,	?	789	?	0,69

				survival forest				
42	Pereira2015	75<	univariate LR -> forward selection LR (& Kaplan-Meier, cox, Gehan or Wilcoxon)	LR, Cox regression	1,8/6,1/10	11521	72-hour, 30-day, 90-day	0,77
43	pugh2014	65<	univariate	GLM	22,7	105450	30-day	0,65
44	Shulan2013	Veterans	multivariate LR (stepwise)	LR	16,15	8718	30-day	0,8
45	Silverstein2008	65<	forward addition & backward elimination LR	LR	11,72	29292	30-day	0,65
46	Singal2013	Cirrhosis	univariate LR -> Multivariate LR	LR	27	836	30-day, 90-day	0,66
47	tsui2015	65<	multivariate LR	LR	7,8	1167521	28-day	0,81
48	turgeman2016	HF	Pearson correlation	ensemble (Boosted C5,0 & SVM)	28	4840	30-day	0,65-0,85
49	vanDiepen2014	cardiovascular ICU	univariate -> stepwise LR	LR	4,4	10799	any	0,799
50	Wallmann2013	Cardiac-related disease	backward elimination LR	LR	4,5	35531	30-day	0,75
51	Walraven2010	all	backward	LR	8	4812	30-day	0,684

			stepping LR					
52	walsh2014	all	LASSO	LASSO, SVM	7,16	92530	30-day	0,68-0,92
53	Wang2012	HF	backward selection -> forward selection Cox	Cox regression	4,2	198640	30-day, 12-month	0,80-0,82
54	Watson2011	HF	univariate & multivariate LR	LR	12,75	729	30-day	0,67
55	Yu2015	HF, AMI, Pneumonia		SVM, Cox regression	18,87	74746	30-day	0,63-0,74
56	Zapatero2012	all	LR	LR	12,4	999089	30-day	NR
57	Zheng2015	HF		SVM, RF	21,63	1641	30-day	NR
58	Mortazavi2016	HF		RF,SVM, Boosting, LR	14,8	1004	30-day, 180-day	0,67
59	Krumholz2016	HF	Random Forest	Cox regression	17,1	1004	30-day	0,62-0,65
60	Bradford2016	HF	univariate	logistic regression	13,3	2420	30-day	0,68
61	Lin2016	65<	univariate	logistic regression	14,6-19,1	39156, 178286	30-day, 1-year	0,64-0,65
62	Corrigan1992	adults		cox regression	30,14	4219	1-year	not reported
63	Vigod2015	acute psychiatric unit	stepwise logistic regression	logistic regression	9,2	65499	30-day	0,63
64	Tulloch2015	psychiatric	stepwise removal LR	cox regression & LR	14,6	7891	90-day	0,65
65	Tabak2017	all	univariate LR	LR	11,9	1195640	30-day	0,69-0,72

66	Spiva2014	all	LR	LR	27,1	598	30-day	0,77

Abbreviations:
NR, Not Reported
HF, Heart failure
ICU, Intensive Care Unit
AMI, Acute Myocardial infarction
ED, Emergency Department
COPD, Chronic Obstructive Pulmonary Disease
HIV, Human Immunodeficiency Virus
OPAT, Outpatient Parenteral Antimicrobial Therapy
RF, Random Forest
SVM, Support Vector Machine
ANN, Artificial Neural Network
LR, Logistic Regression
GLM, Generalized Linear Model
GEE, Generalized Estimating Equation
HGLM, Hierarchical Generalized Linear Model
CPHM, Cox Proportional Hazards Model
LASSO, least absolute shrinkage and selection operator
PCA, Principal Component analysis
CBFS, Correlation-Based Feature Selection

Bibliography

[Greenland1989] Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. American journal of public health, 79(3), 340-349

[Futoma2015] Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. Journal of biomedical informatics, 56, 229-238.

[Ross2008] Ross, J. S., Mulvey, G. K., Stauffer, B., Patlolla, V., Bernheim, S. M., Keenan, P. S., & Krumholz, H. M. (2008). Statistical models and patient predictors of readmission for heart failure: a systematic review. Archives of internal medicine, 168(13), 1371-1386.

[Leppin2014] Leppin, A. L., Gionfriddo, M. R., Kessler, M., Brito, J. P., Mair, F. S., Gallacher, K., et al. (2014). Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials. *JAMA internal medicine*, *174*(7), 1095-1107.

[Brereton2007] Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of systems and software*, *80*(4), 571-583.

[Tranfield2003] Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management*, *14*(3), 207-222.

[Crone2012] Crone, S. F., & Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. International Journal of Forecasting, 28(1), 224-238.

[Jencks2009] Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine*, *360*(14), 1418-1428.

[Chawla2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[Kansagara2011] Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: a systematic review. Jama, 306(15), 1688-1698.

[Lopez2013] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113-141.

[Ponikowski2016] Ponikowski, P., Voors, A. A., Anker, S. D., Bueno, H., Cleland, J. G., Coats, A. J., et al.. (2015). 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. European heart journal, ehw128.

[Mosterd2007] Mosterd, A., & Hoes, A. W. (2007). Clinical epidemiology of heart failure. Heart, 93(9), 1137-1146.

[Krumholz1997] Krumholz, H. M., Parent, E. M., Tu, N., Vaccarino, V., Wang, Y., Radford, M. J., & Hennen, J. (1997). Readmission after hospitalization for congestive heart failure among Medicare beneficiaries. Archives of internal medicine, 157(1), 99-104.

[Urma2017] Urma, D., & Huang, C. C. (2017). Interventions and Strategies to Reduce 30-day Readmission Rates. Hospital Medicine Clinics.

[Kripalani2014] Kripalani, S., Theobald, C. N., Anctil, B., & Vasilevskis, E. E. (2014). Reducing Hospital Readmission: Current Strategies and Future Directions.

Annual Review of Medicine, 65, 471–485. http://doi.org/10.1146/annurev-med-022613-090415

[Balla2008] Balla, U., Malnick, S., & Schattner, A. (2008). Early readmissions to the department of medicine as a screening tool for monitoring quality of care problems. Medicine, 87(5), 294-300.

[Swain2015] Swain, M. J., & Kharrazi, H. (2015). Feasibility of 30-day hospital readmission prediction modeling based on health information exchange data. International journal of medical informatics, 84(12), 1048-1056.

[Kmietowicz2010] Kmietowicz, Z. (2010). Hospitals will be fined for emergency readmissions, says Lansley. BMJ: British Medical Journal (Online), 340.

[CMS2011] Centers for Medicare and Medicaid Services (CMS), HHS. (2011). Medicare program; hospital inpatient prospective payment systems for acute care hospitals and the long-term care hospital prospective payment system and FY 2012 rates; hospitals' FTE resident caps for graduate medical education payment. Final rules. Federal Register, 76(160), 51476.

[Kadi2017] Kadi, I., Idri, A., & Fernandez-Aleman, J. L. (2017). Knowledge discovery in cardiology: A systematic literature review. International Journal of Medical Informatics, 97, 12-32.

[Desai2012] Desai, A. S., & Stevenson, L. W. (2012). Rehospitalization for heart failure. Circulation, 126(4), 501-506.

[Braga2014] Braga, P., Portela, F., Santos, M. F., & Rua, F. (2014). Data mining models to predict patient's readmission in intensive care units. In ICAART 2014-Proceedings of the 6th International Conference on Agents and Artificial Intelligence.

[HSA2011] Critical coverage for heart health: Medicaid and cardiovascular disease (2011). American Heart & Stroke Association, Retrieved from http://www.heart.org/idc/groups/heartpublic/@wcm/@adv/documents/downloadable/ucm_428187.pdf

[Anand2006] Anand, S. S., Razak, F., Davis, A. D., Jacobs, R., Vuksan, V., Teo, K., & Yusuf, S. (2006). Social disadvantage and cardiovascular disease:

development of an index and analysis of age, sex, and ethnicity effects. International Journal of Epidemiology, 35(5), 1239-1245.

[Ceia2002] Ceia F, Fonseca C, Mota T, Morais H, Matias F, De Sousa A, Oliveira AG. Prevalence of chronic heart failure in Southwestern Europe: the EPICA study. Eur JHeart Fail 2002;4:531 – 539

[Riley2009] Riley, J. P., & Cowie, M. R. (2009). Telemonitoring in heart failure. Heart, 95(23), 1964-1968.- Inglis, S. (2010). Structured telephone support or telemonitoring programmes for patients with chronic heart failure. Journal of Evidence-Based Medicine, 3(4), 228-228.

[Cleland2005] Cleland, J. G., Louis, A. A., Rigby, A. S., Janssens, U., Balk, A. H., & Ten-HMS Investigators. (2005). Noninvasive home telemonitoring for patients with heart failure at high risk of recurrent admission and death: The Trans-European Network-Home-Care Management System (TEN-HMS) study. Journal of the American College of Cardiology, 45(10), 1654-1664.

[Lusignan2001] Lusignan, S., Wells, S., Johnson, P., Meredith, K., & Leatham, E. (2001). Compliance and effectiveness of 1 year's home telemonitoring. The report of a pilot study of patients with chronic heart failure. European Journal of Heart Failure, 3(6), 723-730.

[U4H2017] United4Health. Transforming patient experience with telehealth in Europe. Last accessed 2017-02-03. http://united4health.eu/

[WHO2011] World Health Organization, “Global health and ageing,” World Health Organization, Geneva, Switzerland, 2011.

[Carpenter2011] Carpenter CR, Heard K, Wilber S, Ginde AA, Stiffler K, Gerson LW, et al. Research priorities for high-quality geriatric emergency care: medication management, screening, and prevention and functional assessment. Acad Emerg Med 2011 Jun;18(6):644-54.

[Han2009] Han JH, Zimmerman EE, Cutler N, Schnelle J, Morandi A, Dittus RS, et al. Delirium in older emergency department patients: recognition, risk factors, and psychomotor subtypes. Acad Emerg Med 2009 Mar;16(3):193-200.

[Guidelines2014] New guidelines for geriatric EDs: guidance focused on boosting environment, care processes. ED Manag 2014 May;26(5):49-53.

[Friedman2001] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

[Breiman1996] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

[Breiman2001] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[Besga2015] Besga, A., Ayerdi, B., Alcalde, G., Manzano, A., Lopetegui, P., Graña, M., & González-Pinto, A. (2015). Risk factors for emergency department short time readmission in stratified population. BioMed research international, 2015.

[Burges1998] Christopher Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):167–121, 1998.

[Vapnik1998] V. Vapnik. Statistical learning theory. Wiley-Interscience, 1998.

[Quinlan1993] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[Quinlan1986] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[Breiman1984] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth and Brooks, Monterey, CA, 1984

[Huang2006] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: Theory and applications, Neurocomputing 70 (1–3) (2006) 489 – 501.

[Huang2015] G. Huang, G.-B. Huang, S. Song, K. You, Trends in extreme learning machines: A review, Neural Networks 61 (2015) 32 – 48. doi: http://dx.doi.org/10.1016/j.neunet.2014.10.001.

[Huang2011] G. B. Huang, D. H. Wang, Y. Lan, Extreme learning machines: a survey, International Journal of Machine Learning and Cybernetics 2 (2011) 107–122.

[Ayerdi2014] B. Ayerdi, J. Maiora, A. d'Anjou, M. Graña, Applications of hybrid extreme rotation forests for image segmentation, International Journal of Hybrid Intelligent Systems 11 (1) (2014) 13–24.

[Ayerdi2015] B. Ayerdi, I. Marques, M. Graña, Spatially regularized semisupervised ensembles of extreme learning machines for hyperspectral image segmentation, Neurocomputing 149, Part A (2015) 373–386.

[Chyzhyk2015] D. Chyzhyk, A. Savio, M. Graña, Computer aided diagnosis of schizophrenia on resting state fmri data by ensembles of ELM, Neural Networks 68 (2015) 23 – 33. doi: http://dx.doi.org/10.1016/j.neunet.2015.04.002.

[ICS2016] Ayerdi, B., & Graña, M. (2016). Anticipative Hybrid Extreme Rotation Forest. Procedia Computer Science, 80, 1671-1681.

[Schapire1999] R. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Machine Learning 37 (3) (1999) 297–336. doi:10.1023/A:1007614523901.

[Freund1995] Y. Freund, R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: European Conference on Computational Learning Theory, 1995, pp. 37, 23.

[Wolpert1996] D. H. Wolpert, The lack of a priori distinctions between learning algorithms, Neural computation 8 (7) (1996) 1341–1390.

[Wolpert1997] D. Wolpert, W. Macready, No free lunch theorems for optimization, IEEE Trans. on Evol. Comp. 1 (1) (1997) 67 – 82.

[Ditzler1997] Ditzler, G., & Polikar, R. (2013). Incremental learning of concept drift from streaming imbalanced data. ieee transactions on knowledge and data engineering, 25(10), 2283-2301.

[Haixiang2017] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 73, 220 - 239 (2017)

[Sun2009] SUN, Y., WONG, A.K.C., KAMEL, M.S.: Classi_cation of imbalanced data: A review. International Journal of Pattern Recognition and Arti_cial Intelligence 23(04), 687-719 (2009)

[Yang2006] Yang, Q., Wu, X.: 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making 5(04), 597{604 (2006)

[Mazurowski2008] Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D.: Training neural network classi_ers for medical decision making: The e_ects of imbalanced datasets on classi_cation performance. Neural networks 21(2), 427{436 (2008)

[Chawla2003] Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: Smoteboost: Improving prediction of the minority class in boosting. In: European Conference on Principles of Data Mining and Knowledge Discovery. pp. 107{119. Springer (2003)

[Wang2009] Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on. pp. 324{331. IEEE (2009)

[Galar2012] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybridbased approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42(4), 463{484 (2012)

[Borovicka2012] Borovicka, T., Jirina Jr, M., Kordik, P., & Jirina, M. (2012). Selecting representative data sets. In Advances in data mining knowledge discovery and applications. InTech.

[Barandela2003] Barandela, R., Sánchez, J. S., Garcıa, V., & Rangel, E. (2003). Strategies for learning in class imbalance problems. Pattern Recognition, 36(3), 849-851.

[He2008] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on (pp. 1322-1328). IEEE.

[Mani2003] Mani, I., & Zhang, I. (2003, August). kNN approach to unbalanced data distributions: a case study involving information extraction. In Proceedings of workshop on learning from imbalanced datasets.

[Tomek1976] Tomek, I., “Two modifications of CNN,” IEEE Trans. Systems, Man and Cybernetics, vol. SMC-6, Nov. 1976, pp. 769-772.

[Yen2009] Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36(3), 5718-5727.

[Fan1999] Fan, W., Stolfo, S. J., Zhang, J., & Chan, P. K. (1999, June). AdaCost: misclassification cost-sensitive boosting. In Icml (pp. 97-105).

[Sun2007] Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, 40(12), 3358-3378.

[Freund1995] Freund, Y., & Schapire, R. E. (1995, March). A decision-theoretic generalization of on-line learning and an application to boosting. In European conference on computational learning theory (pp. 23-37). Springer Berlin Heidelberg.

[Du2012] Du, P., Xia, J., Zhang, W., Tan, K., Liu, Y., & Liu, S. (2012). Multiple classifier system for remote sensing image classification: A review. Sensors, 12(4), 4764-4792.

[Seiffert2010] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 40(1), 185-197.

[Lin2013] Lin, S.J., Chang, C., Hsu, M.F.: Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction. Knowledge-Based Systems 39, 214{223 (2013)

[AbdelRahman2014] AbdelRahman, S. E., Zhang, M, Bray, B. E., and Kawamoto, K. (2014). A three-step approach for the derivation and validation of high-performing predictive models using an operational dataset: congestive heart failure readmission case study. BMCmedical informatics and decision making, 14(1):41.

[Alassaad2015] Alassaad, A., Melhus, H., Hammarlund-Udenaes, M., Bertilsson, M., Gille-spie, U., and Sundstr¨om, J. (2015). A tool for prediction of risk of rehospitalisation and mortality in the hospitalised elderly: secondary analysis of clinical trial data.BMJ open, 5(2):e007259.

[Allaudeen2011] Allaudeen, N., Vidyarthi, A., Maselli, J., and Auerbach, A. (2011). Re-defining readmission risk factors for general medicine patients.Journal of Hospital Medicine,6(2):54–60.

[Allen2012] Allen, L. A., Tomic, K. E. S., Smith, D. M., Wilson, K. L., and Agodoa, I.(2012). Rates and predictors of 30-day readmission among commercially insured and medicaid-enrolled patients hospitalized with systolic heart failure.Circulation: Heart Failure, 5(6):672–679.

[Allison2014] Allison, G. M., Muldoon, E. G., Kent, D. M., Paulus, J. K., Ruthazer,R., Ren, A., and Snydman, D. R. (2014). Prediction model for 30-day hospital readmissions among patients discharged receiving outpatient parenteral antibiotic therapy. Clinical infectious diseases, 58(6):812–819.

[Amalakuhan2012] Amalakuhan, B., Kiljanek, L., Parvathaneni, A., Hester, M., Cheriy-ath, P., and Fischman, D. (2012). A prediction model for copd readmissions: catching up, catch-ing our breath, and improving a national problem. Journal of Community Hospital Internal Medicine Perspectives, 2(1).

[Amarasingham2010] Amarasingham, R., Moore, B. J., Tabak, Y. P., Drazner, M. H., Clark, C. A., Zhang, S., Reed, W. G., Swanson, T. S., Ma, Y., and Halm, E. A. (2010). An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Medical care , 48(11):981–988.

[Au2012] Au, A. G., McAlister, F. A., Bakal, J. A., Ezekowitz, J., Kaul, P., and van Walraven, C. (2012). Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization. American heart journal, 164(3):365–372.

[Baillie2013] Baillie, C. A., VanZandbergen, C., Tait, G., Hanish, A., Leas, B., French, B., William Hanson, C., Behta, M., and Umscheid, C. A. (2013). The readmission risk flag: Using the electronic health record to automatically identify patients at risk for 30-day readmission. Journal of hospital medicine, 8(12):689–695.

[Baltodano2016] Baltodano, P. A., Webb-Vargas, Y., Soares, K., Hicks, C., Cooney, C. M., Cornell, P., Burce, K., Pawlik, T. M., and Eckhauser, F. (2016). A validated, risk assessment tool for predicting readmission after open ventral hernia repair. Hernia, 20(1):119–129.

[Bergese2017] Bergese, I., Frigerio, S., Clari, M., Castagno, E., De Clemente, A., Ponticelli, E., Scavino, E., and Berchialla, P. (2017). An innovative model to predict pediatric emergency department return visits. Pediatric Emergency Care.

[Berman2011] Berman, K., Tandra, S., Forssell, K., Vuppalanchi, R., Burton, J. R., Nguyen, J., Mullis, D., Kwo, P., and Chalasani, N. (2011). Incidence and predictors of 30-day readmission among patients hospitalized for advanced liver disease. Clinical Gastroenterology and Hepatology, 9(3):254–259.

[Betihavas2015] Betihavas, V., Frost, S. A., Newton, P. J., Macdonald, P., Stewart, S., Carrington, M. J., Chan, Y. K., and Davidson, P. M. (2015). An absolute risk prediction model to determine unplanned cardiovascular readmissions for adults with chronic heart failure. Heart, Lung and Circulation , 24(11):1068–1073.

[Billings2012] Billings, J., Blunt, I., Steventon, A., Georghiou, T., Lewis, G., and Bardsley, M. (2012). Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (parr-30). BMJ open, 2(4):e001667.

[Cai2016] Cai, X., Perez-Concha, O., Coiera, E., Martin-Sanchez, F., Day, R., Roffe, D., and Gallego, B. (2016). Real-time prediction of mortality, readmission, and length of stay us- ing electronic health record data. Journal of the American Medical Informatics Association,23(3):553–561.

[Coleman2004] Coleman, E. A., Min, S.-j., Chomiak, A., and Kramer, A. M. (2004). Posthospital care transitions: patterns, complications, and risk identification. Health services research, 39(5):1449–1466.

[Cui2015] Cui, Y., Metge, C., Ye, X., Moffatt, M., Oppenheimer, L., and Forget, E. L. (2015). Development and validation of a predictive model for all-cause hospital readmissions in winnipeg, canada. Journal of health services research & policy, 20(2):83–91.

[Deschodt2015] Deschodt, M., Devriendt, E., Sabbe, M., Knockaert, D., Deboutte, P., Boonen, S., Flamaing, J., and Milisen, K. (2015). Characteristics of older adults admitted to the emergency department (ed) and their risk factors for ed readmission based on comprehensive geriatric assessment: a prospective cohort study. BMC geriatrics, 15(1):1.

[Dharmarajan2013] Dharmarajan, K., Hsieh, A. F., Lin, Z., Bueno, H., Ross, J. S., Hor-witz, L. I., Barreto-Filho, J. A., Kim, N., Bernheim, S. M., Suter, L. G., et al. (2013). Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. Jama, 309(4):355–363.

[Donze2013] Donzé, J., Aujesky, D., Williams, D., and Schnipper, J. L. (2013). Poten-tially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA internal medicine, 173(8):632–638.

[Dorajoo2017] Dorajoo, S. R., See, V., Chan, C. T., Tan, J. Z., Tan, D. S. Y., Razak, S. M. B. A., Ong, T. T., Koomanan, N., Yap, C. W., and Chan, A. (2017). Identifying potentially avoidable readmissions: A medication-based 15-day readmission risk stratification algorithm. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy.

[Epstein2011] Epstein, A. M., Jha, A. K., and Orav, E. J. (2011). The relationship between hospital admission rates and rehospitalizations. New England Journal of Medicine, 365(24):2287–2295.

[Fisher2016] Fisher, S. R., Graham, J. E., Krishnan, S., and Ottenbacher, K. J. (2016). Predictors of 30-day readmission following inpatient rehabilitation for

patients at high risk for hospital readmission. Physical Therapy, 96(1):62.

[Garrison2013] Garrison, G. M., Mansukhani, M. P., and Bohn, B. (2013). Predictors of thirty-day readmission among hospitalized family medicine patients. The Journal of the American Board of Family Medicine, 26(1):71–77.

[Halfon2002] Halfon, P., Eggli, Y., van Melle, G., Chevalier, J., Wasserfallen, J.-B., and Burnand, B. (2002). Measuring potentially avoidable hospital readmissions. Journal of clinical epidemiology, 55(6):573–587.

[Hao2015] Hao, S., Wang, Y., Jin, B., Shin, A. Y., Zhu, C., Huang, M., Zheng, L., Luo, J., Hu, Z., Fu, C., Dai, D., Wang, Y., Culver, D. S., Alfreds, S. T., Rogow, T., Stearns, F., Sylvester, K. G., Widen, E., and Ling, X. B. (2015). Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the maine healthcare information exchange. PLoS ONE, 10(10):1–15.

[Hasan2010] Hasan, O., Meltzer, D. O., Shaykevich, S. A., Bell, C. M., Kaboli, P. J., Auerbach, A. D., Wetterneck, T. B., Arora, V. M., Zhang, J., and Schnipper, J. L. (2010). Hospital readmission in general medicine patients: a prediction model. Journal of general internal medicine, 25(3):211–219.

[Kaur2016] Kaur, H., Naessens, J. M., Hanson, A. C., Fryer, K., Nemergut, M. E., and Tripathi, S. (2016). Proper development of an early pediatric intensive care unit readmission risk prediction tool. Journal of Intensive Care Medicine , page 0885066616665806.

[Keenan2008] Keenan, P. S., Normand, S.-L. T., Lin, Z., Drye, E. E., Bhat, K. R., Ross, J. S., Schuur, J. D., Stauffer, B. D., Bernheim, S. M., Epstein, A. J., et al. (2008). An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation: Cardiovascular Quality and Outcomes, 1(1):29–37.

[Leong2017] Leong, K. T. G., Wong, L. Y., Aung, K. C. Y., Macdonald, M., Cao, Y., Lee, S., Chow, W. L., Doddamani, S., and Richards, A. M.

(2017). Risk stratification model for 30-day heart failure readmission in a multi-ethnic south east asian community. The American Journal of Cardiology.

[Lopez2011] López-Aguila, S., Contel, J., Farre, J., Campuzano, J., and Rajmil, L. (2011). Predictive model for emergency hospital admission and 6-month readmission. The American journal of managed care, 17(9):e348–57.

[Low2016] Low, L. L., Liu, N., Wang, S., Thumboo, J., Ong, M. E. H., and Lee, K. H. (2016). Predicting 30-day readmissions in an asian population: Building a predictive model by incorporating markers of hospitalization severity. PloS one, 11(12):e0167413.

[Marcantonio1999] Marcantonio, E. R., McKean, S., Goldfinger, M., Kleefield, S., Yurkofsky, M., and Brennan, T. A. (1999). Factors associated with unplanned hospital readmission among patients 65 years of age and older in a medicare managed care plan. The American journal of medicine, 107(1):13–17.

[McLaren2016] McLaren, D. P., Jones, R., Plotnik, R., Zareba, W., McIntosh, S., Alexis, J., Chen, L., Block, R., Lowenstein, C. J., and Kutyifa, V. (2016). Prior hospital admission predicts thirty-day hospital readmission for heart failure patients. Cardiol J, 23:155–162.

[McManus2016] McManus, D. D., Saczynski, J. S., Lessard, D., Waring, M. E., Allison, J.,Parish, D. C., Goldberg, R. J., Ash, A., Kiefe, C. I., Investigators, T.-C., et al. (2016). Reliability of predicting early hospital readmission after discharge for an acute coronary syndrome using claims-based data. The American journal of cardiology, 117(4):501–507.

[Morris2014] Morris, J. N., Howard, E. P., Steel, K., Schreiber, R., Fries, B. E., Lipsitz, L. A., and Goldman, B. (2014). Predicting risk of hospital and emergency department use for home care elderly persons through a secondary analysis of cross-national data. BMC health services research, 14(1):1.

[Nguyen2014] Nguyen, H. Q., Chu, L., Amy Liu, I.-L., Lee, J. S., Suh, D., Korotzer, B., Yuen, G., Desai, S., Coleman, K. J., Xiang, A. H., and Gould, M. K. (2014). Associations between physical activity and 30-day readmission risk in chronic obstructive pulmonary disease. Annals ATS, 11(5):695–705.

[Nijhawan2012] Nijhawan, A. E., Clark, C., Kaplan, R., Moore, B., Halm, E. A., and Amarasingham, R. (2012). An electronic medical record-based model to predict 30-day risk of readmission and death among hiv-infected inpatients. JAIDS Journal of Acquired Immune Deficiency Syndromes, 61(3):349–358.

[Ouanes2012] Ouanes, I., Schwebel, C., Fran ̧cais, A., Bruel, C., Philippart, F., Vesin, A., Soufir, L., Adrie, C., Garrouste-Orgeas, M., Timsit, J.-F., et al. (2012). A model to predict short-term death or readmission after intensive care unit discharge. Journal of critical care,27(4):422–e1.

[Padhukasahasram2015] Padhukasahasram, B., Reddy, C. K., Li, Y., and Lanfear, D. E. (2015). Joint impact of clinical and behavioral variables on the risk of unplanned readmission and death after a heart failure hospitalization. PloS one, 10(6):e0129553.

[Pereira2015] Pereira, L., Choquet, C., Perozziello, A., Wargon, M., Juillien, G., Colosi, L., Hellmann, R., Ranaivoson, M., and Casalino, E. (2015). Unscheduled-return-visits after an emergency department (ed) attendance and clinical link between both visits in patients aged 75 years and over: a prospective observational study. PloS one, 10(4):e0123803.

[Pugh2014] Pugh, J. A., Wang, C.-P., Espinoza, S. E., No ̈el, P. H., Bollinger, M., Amuan, M., Finley, E., and Pugh, M. J. (2014). Influence of frailty-related diagnoses, high-risk prescribing in elderly adults, and primary care use on readmissions in fewer than 30 days for veterans aged 65 and older. Journal of the American Geriatrics Society, 62(2):291–298.

[Shulan2013] Shulan, M., Gao, K., and Moore, C. D. (2013). Predicting 30-day all-cause hospital readmissions. Health care management science, 16(2):167–175.

[Silverstein2008] Silverstein, M. D., Qin, H., Mercer, S. Q., Fong, J., and Haydar, Z. (2008). Risk factors for 30-day hospital readmission in patients ¿ 65 years of age. In Baylor University Medical Center. Proceedings, volume 21, page 363. Baylor University Medical Center.

[Singal2013] Singal, A. G., Rahimi, R. S., Clark, C., Ma, Y., Cuthbert, J. A., Rockey, D. C., and Amarasingham, R. (2013). An automated model using electronic medical record data identifies patients with cirrhosis at high risk for readmission. Clinical Gastroenterology and Hepatology, 11(10):1335–1341.

[Tsui2015] Tsui, E., Au, S., Wong, C., Cheung, A., and Lam, P. (2015). Development of an automated model to predict the risk of elderly emergency medical admissions within a month following an index hospital visit: a hong kong experience. Health informatics journal, 21(1):46–56.

[Turgeman2016] Turgeman, L. and May, J. H. (2016). A mixed-ensemble model for hospital readmission. Artificial Intelligence in Medicine , 72:72–82.

[vanDiepen2014] van Diepen, S., Graham, M. M., Nagendran, J., and Norris, C. M. (2014). Predicting cardiovascular intensive care unit readmission after cardiac surgery: derivation and validation of the alberta provincial project for outcomes assessment in coronary heart disease (approach) cardiovascular intensive care unit clinical prediction model from a registry cohort of 10,799 surgical cases. Critical Care, 18(6):651.

[Walraven2010] van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., Austin, P. C., and Forster, A. J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Canadian Medical Association Journal, 182(6):551–557.

[Wallmann2013] Wallmann, R., Llorca, J., G´omez-Acebo, I., Ortega, ´ A. C., Roldan, F. R., and Dierssen-Sotos, T. (2013). Prediction of 30-day cardiac-related-emergency-readmissions using simple administrative hospital data. International journal of cardiology , 164(2):193–200.

[Walsh2014] Walsh, C. and Hripcsak, G. (2014). The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions. Journal of biomedical informatics, 52:418–426.

[Wang2012] Wang, L., Porter, B., Maynard, C., Bryson, C., Sun, H., Lowy, E., McDonell, M., Frisbee, K., Nielson, C., and Fihn, S. D. (2012). Predicting risk of hospitalization or death among patients with heart failure in the veterans health administration. The American journal of cardiology, 110(9):1342–1349.

[Watson2011] Watson, A. J., O'Rourke, J., Jethwani, K., Cami, A., Stern, T. A., Kvedar, J. C., Chueh, H. C., and Zai, A. H. (2011). Linking electronic health record-extracted psychosocial data in real-time to risk of readmission for heart failure. Psychosomatics, 52(4):319–327.

[Yu2015] Yu, S., Farooq, F., van Esbroeck, A., Fung, G., Anand, V., and Krishnapuram, B. (2015). Predicting readmission risk with institution-specific prediction models. Artificial Intelligence in Medicine, 65(2):89–96.

[Zapatero2012] Zapatero, A., Barba, R., Marco, J., Hinojosa, J., Plaza, S., Losa, J. E., and Canora, J. (2012). Predictive model of readmission to internal medicine wards. European journal of internal medicine, 23(5):451–456.

[Zheng2015] Zheng, B., Zhang, J., Yoon, S. W., Lam, S. S., Khasawneh, M., & Poranki, S. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining. Expert Systems with Applications, 42(20), 7110-7120.

[Mortazavi2016] Bobak J. Mortazavi, et al. Analysis of Machine Learning Techniques for Heart Failure Readmissions. DOI: 10.1161/CIRCOUTCOMES.116.003039

[Krumholz2016] Krumholz, H. M., Chaudhry, S. I., Spertus, J. A., Mattera, J. A., Hodshon, B., & Herrin, J. (2016). Do non-clinical factors improve prediction of readmission risk?: results from the Tele-Hf study. JACC: Heart Failure, 4(1), 12-20.

[Bradford2016] Bradford, C., Shah, B. M., Shane, P., Wachi, N., & Sahota, K. (2016). Patient and clinical characteristics that heighten risk for heart failure readmission. Research in Social and Administrative Pharmacy.

[Lin2016] Lin, K. P., Chen, P. C., Huang, L. Y., Mao, H. C., & Chan, D. C. D. (2016). Predicting Inpatient Readmission and Outpatient Admission in Elderly: A Population-Based Cohort Study. Medicine, 95(16).

[Corrigan1992] Corrigan JM, Martin JB. Identification of factors associated with hospital readmission and development of a predictive model. Health Serv Res. 1992 Apr;27(1):81-101. PubMed PMID: 1563955; PubMed Central PMCID: PMC1069865.

[Vigod2015] Vigod SN, Kurdyak PA, Seitz D, Herrmann N, Fung K, Lin E, Perlman C, Taylor VH, Rochon PA, Gruneir A. READMIT: a clinical risk index to predict 30-day readmission after discharge from acute psychiatric units. J Psychiatr Res. 2015 Feb;61:205-13. doi: 10.1016/j.jpsychires.2014.12.003. PubMed PMID: 25537450. Tulloch2015

[Tabak2017] Tabak YP, Sun X, Nunez CM, Gupta V, Johannes RS. Predicting Readmission at Early Hospitalization Using Electronic Clinical Data: An Early Readmission Risk Score. Med Care. 2017 Mar;55(3):267-275. doi: 10.1097/MLR.0000000000000654. PubMed PMID: 27755391.

[Spiva2014] Spiva, L., Hand, M., VanBrackle, L., & McVay, F. (2014). Validation of a predictive model to identify patients at high risk for hospital readmission. Journal for Healthcare Quality.

[Kotsiantis2006] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190.

[Saeys2007] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.

[Kohavi1997] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1), 273-324.

[Guyon2003] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

[Hall1999] Hall, M. A. (1999). Correlation-based feature selection for machine learning.

[Khan2003] Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. Journal of the Royal Society of Medicine, 96(3), 118-121.