UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MACHINE LEARNING ALGORITHMS TO IMPROVE RISK PREDICTION MODELS IN HEALTHCARE

TESIS PARA OPTAR AL GRADO DE DOCTOR EN SISTEMAS DE INGENIERÍA

PATRICIO ANTONIO WOLFF ROJAS

PROFESOR GUÍA:
SEBASTIÁN A. RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
MANUEL GRAÑA ROMAY
DIEGO MARTÍNEZ CEA
DENIS SAURÉ VALENZUELA

Financiado por CONICYT-PCHA/ Doctorado Nacional/2015-21150115

SANTIAGO DE CHILE
2019

RESUMEN DE LA TESIS PARA OPTAR AL TÍTULO
DE DOCTOR EN SISTEMAS DE INGENIERÍA
POR: PATRICIO ANTONIO WOLFF ROJAS
FECHA: 2019
PROF. GUÍA: SEBASTIÁN A. RÍOS PÉREZ

MACHINE LEARNING ALGORITHMS TO IMPROVE RISK PREDICTION MODELS IN HEALTHCARE

A large number of risk models are currently used in different health institutions. Machine Learning tools can be used to automate these processes and improve the results of these models. These models can be incorporated into clinical processes through systems known as Clinical Decision Support Systems (ML-CDSS). Three different problems were selected, which are part of these ML-CDSS: Hospital Readmissions, ED Triage, and Decompensation of inpatients. One of the most important characteristics of these three problems is that: they focus on assessing the level of risk to make decisions, to adapt the level of care to a predicted risk. The three problems are also characterized by requiring a level of one patient risk, at a specific time. The three selected problems are recognized in the international literature as difficult to resolve, particularly in pediatrics, so there is currently a great interest in research in this area.

This thesis aims to be a methodological contribution in state of the art of Machine Learning algorithms applied to patients' risk problems.

To achieve our goal, a series of operations must be implemented, which considered erroneous data cleaning, labeling data, applying class balancing techniques, testing different classification models and performance evaluation. We have conducted several studies which show that it is possible to improve risk prediction and multi-class classification using an ML approach. In these problems, both the performance evaluation and the labels that allow training the models, are based on clinical outcomes. This allows a larger dataset to be used and guarantees the objectivity of the result by limiting the influence of human judgment. In this thesis, we worked with anonymized data from Exequiel González Cortés pediatric hospital.

The correct use of ML tools improves the predictive result in problems related to patient risk. The excellent results obtained in different evaluation metrics in risk prediction problems allow methodological validation of the ML tools used. Even if they are compared with other knowledge-based and non-knowledge-based methods. This allows enriching the discussion regarding the benefits of these models in real clinical settings. The methodology presented in each problem has, in general terms, similar characteristics and can be used in other CDSS.

RESUMEN DE LA TESIS PARA OPTAR AL TÍTULO
DE DOCTOR EN SISTEMAS DE INGENIERÍA
POR: PATRICIO ANTONIO WOLFF ROJAS
FECHA: 2019
PROF. GUÍA: SEBASTIÁN A. RÍOS PÉREZ

MACHINE LEARNING ALGORITHMS TO IMPROVE RISK PREDICTION MODELS IN HEALTHCARE

Actualmente, se utiliza una gran cantidad de modelos de riesgo en diferentes instituciones de salud. Las herramientas de Machine Learning (ML) pueden ser utilizadas para automatizar estos procesos y mejorar los resultados de estos modelos. Estos modelos pueden incorporarse en procesos clínicos a través de sistemas conocidos como Clinical Decision Support Systems (ML-CDSS). Se seleccionaron tres problemas diferentes, que son parte de estos ML-CDSS: reingresos hospitalarios, triage de urgencia y descompensación de pacientes hospitalizados. Una de las características más importantes de estos tres problemas es que se centran en evaluar el nivel de riesgo para tomar decisiones y de esta forma adaptar el nivel de atención al nivel de riesgo determinado. Los tres problemas también se caracterizan por requerir el nivel de riesgo de un paciente, en un momento específico. Los tres problemas seleccionados son reconocidos en la literatura internacional como difíciles de resolver, particularmente en pediatría, por lo que, actualmente hay un gran interés en investigar en esta área.

Esta tesis pretende ser una contribución metodológica en el estado del arte de los algoritmos de ML aplicados a los problemas de riesgo de pacientes.

Para lograr nuestro objetivo, se debe implementar una serie de operaciones, que incluyen la limpieza de datos erróneos, etiquetado de datos, aplicar técnicas de balanceo de clases, probar diferentes modelos de clasificación y evaluar el desempeño de estos. Se llevaron a cabo varios estudios que muestran que es posible mejorar la predicción del riesgo y la clasificación de varias clases utilizando un enfoque de ML. En estos problemas, tanto la evaluación del desempeño como las etiquetas que permiten entrenar los modelos, se basan en resultados clínicos. Esto permite utilizar un conjunto de datos más grande y garantiza la objetividad del resultado, al limitar la influencia del juicio humano. En esta tesis se trabajó con datos anonimizados del hospital pediátrico Exequiel González Cortés.

El uso correcto de las herramientas de ML permite mejorar el resultado predictivo en problemas relacionados con el riesgo del paciente. Los excelentes resultados obtenidos con diferentes métricas de evaluación en problemas de predicción de riesgos permiten la validación metodológica de las herramientas de ML utilizadas. Incluso si se comparan con otros métodos knowledge-based y non-knowledge-based. Esto permite enriquecer la discusión sobre los beneficios de estos modelos en entornos clínicos reales. La metodología presentada en cada problema tiene, en términos generales, características similares y puede utilizarse en otros CDSS.

a Nadia

Agradecimientos

I would like to personally thank my parents, family, friends, and colleagues, as well as Prof. Sebastián Ríos, Prof. Manuel Graña, the committee members and officials of the University of Chile.

This research was partially funded by Comisón Nacional de Investigación Científica y Tecnológica, Programa de Formación de Capital Humano avanzado (CONICYT-PCHA/Doctorado Nacional/2015-21150115).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 777720.

Tabla de Contenido

1. Introduction **1**
- 1.1. Clinical Decision Support Systems (CDSS) . . . 2
- 1.2. Three Paradigmatic Problems of Patient risk modeling . . . 3
  - 1.2.1. Hospital Readmission . . . 4
  - 1.2.2. Triage . . . 5
  - 1.2.3. Decompensation of inpatients . . . 6
  - 1.2.4. Categorization of the Paradigmatic Problems . . . 7
- 1.3. Objective . . . 7
  - 1.3.1. General Objective . . . 8
  - 1.3.2. Specific Objectives . . . 8
- 1.4. Methodology . . . 8
- 1.5. Structure . . . 9
- 1.6. Contributions List . . . 10
  - 1.6.1. Indexed Journal Papers . . . 10
  - 1.6.2. Other Related Indexed Journal Papers . . . 10
  - 1.6.3. Peer-reviewed Journal Papers and Conference Presentations . . . 11

2. Machine learning readmission risk modeling: a pediatric case study **12**
- 2.1. Introduction . . . 13
- 2.2. Materials and Methods . . . 14
  - 2.2.1. Cohort and dataset . . . 14
  - 2.2.2. Classification methods . . . 15
  - 2.2.3. Classification performance metrics . . . 17
- 2.3. Results . . . 18
- 2.4. Discussion . . . 21
- 2.5. Conclusions . . . 23

3. Setting up standards: A methodological proposal for pediatric Triage machine learning model construction based on clinical outcomes **25**
- 3.1. Introduction . . . 26
- 3.2. Related Work . . . 27
- 3.3. Data Preparation and Experimental Setup . . . 29
  - 3.3.1. Dataset characteristics . . . 30
  - 3.3.2. Current expert knowledge based Triage system . . . 31
  - 3.3.3. Relabeling according to the clinical outcomes . . . 33
  - 3.3.4. Dealing with class imbalance . . . 34

3.3.5. Machine learning models under evaluation . . . 34
3.4. Evaluation framework and best model selection . . . 37
3.4.1. Multi-class ML performance metrics . . . 37
3.4.2. Final evaluation metrics . . . 38
3.5. Experimental results . . . 39
3.5.1. Initial model selection . . . 40
3.5.2. Dychotomic classification into clinical outcomes . . . 41
3.5.3. Comparison against current e-Triage system . . . 41
3.6. Discussion . . . 42
3.7. Conclusions . . . 44

4. A pediatric early warning system machine learning model based on clinical outcomes 46
4.1. Introduction . . . 47
4.1.1. Related Work . . . 47
4.1.2. Clinical Outcomes . . . 47
4.2. Materials and methods . . . 48
4.2.1. Dataset characterization . . . 48
4.2.2. Dealing with Overfitting and class imbalanced . . . 50
4.2.3. Machine learning models under evaluation . . . 51
4.2.4. Evaluation metrics . . . 52
4.3. Results . . . 53
4.4. Discussion . . . 54

5. Final Conclusions 57
5.1. Research aims . . . 57
5.2. First findings . . . 58
5.3. Proposed methodology . . . 58
5.3.1. Datasets . . . 59
5.3.2. Outcome based models . . . 59
5.3.3. Class balance techniques . . . 59
5.3.4. Selected ML tools . . . 60
5.3.5. Evaluation metrics . . . 60
5.4. Results . . . 61
5.5. Applications/implications . . . 61
5.6. Final remarks . . . 62
5.7. Further research . . . 63

Bibliogrfía 64

A. Model Hyperparameter sensitivity analysis in Machine learning readmission risk modeling: a pediatric case study 82
A.1. Describing the models used . . . 82
A.2. Results . . . 82

B. Extension of Machine learning readmission risk modeling: a pediatric case study 85

Índice de Tablas

1.1. Selected problems coded using taxonomy 8

2.1. Descriptive statistics of the dataset. 15
2.2. Diagnostics at discharge accounting for most readmission 16
2.3. Average ± standard deviation Recall (R) performance [%] of SVM, MLP1, MLP2, and NB for decreasing number of folders in the RCV process. no SMOTE = no oversampling correction of class imbalance is done. 19
2.4. Average ± standard deviation Positive predictive value (PPV)[%] of SVM, MLP1, MLP2, and NB for decreasing number of folders in the RCV process. no SMOTE = no oversampling correction of class imbalance is done. 19
2.5. Average ± standard deviation f-score (F) performance [%] of SVM, MLP1, MLP2, and NB for decreasing number of folders in the RCV process. no SMOTE = no oversampling correction of class imbalance is done. 20
2.6. Average ± standard deviation AUC performance of SVM, MLP1, MLP2, and NB for decreasing number of folders in the RCV process. no SMOTE = no oversampling correction of class imbalance is done. 20

3.1. ML-based Triage model Benchmark 29
3.2. Dataset information 31
3.3. Cases an percentage per class in train and test dataset 33
3.4. Machine Learning studied models and their parameter setup 36
3.5. Models performance for the high versus low severity Triage levels. Blue highlights best model *per* performance metric 40
3.6. Diagnostic performance measures for the two high severity clinical outcomes. 41
3.7. Cases assigned per class in preselected models. Expert means the current Triage system at the EGCH. 42

4.1. Dataset information 49
4.2. Vital signs records information (mean±SD) 49
4.3. Diagnostic accuracy measures 53
4.4. AUC (SD) performance by ML model (a) without including Oxygen Support information, (b) Imbalance data for training, (c) without considering Age Range and patient condition, and (d) all variables and balanced dataset. 54

A.1. Machine Learning studied models and their parameter setup 83
A.2. Machine Learning studied models Results 84

B.1. Results 85

Índice de Ilustraciones

1.1. CDSS Classification 3
1.2. CRISP-DM framework (Source CRISP-DM 1.0 http://www.crisp-dm.org) . . 9

2.1. Study design 14
2.2. Average ROCs of machine learning approaches in 5-fold RCV (applying SMOTE class imbalance correction). Solid line corresponds to the ROC mean. . . 18

3.1. Study design schema 30
3.2. EGCH's current rule based e-Triage expert system. 32
3.3. ROC curve of Hospitalization (positive class) versus non-Hospitalization classification. 42
3.4. Proportion of patient with positive outcome assigned per class by ML models, Hospital admission (blue) and death (red). Expert means the current Triage system at the EGCH. 43

4.1. Study design schema 48
4.2. Boxplot of (a) Heart Rate, (b) Respiratory Rate and (c) Systolic blood pressure by age ranges and patient condition 50
4.3. ROC curve 54

B.1. Comparison of ROC curves 86

Chapter 1

Introduction

The concept of risk management has been applied in healthcare since the mid-1970s [155]. Hospitals were always concerned with augmenting safety, but systematic risk management has been considered only in the last two decades [100, 117].

Modern medicine has advanced in the development of more complex treatments and care processes allowing to improve patient care, but also increasing the risk of adverse events and unwanted damage to the patient. [26]. Risks associated with patient care can never be completely eliminated [170] and treatment decisions for physician and patient depend on the perception of risk [55]. Therefore, clinical risk management plays a crucial role in enabling hospitals to enhance patient safety. [170]

Hospitals are considered high-risk organizations [174]. A large number of risk models are currently used in different health institutions. Research in the field of medicine focuses on improving these models using the judgment of experts [127] and new clinical evidence [115]. Nevertheless, "To Err is Human" [51, 102, 134] and there is a significant error-rate associated with manual risk classification, especially in high-workload settings, such as emergency departments [37, 38].

The use of Electronic Health Records (EHR) has been extended to many health institutions during the last decade [120]. EHRs are repositories of clinical information making available a large amount of longitudinal data. Currently, the large amount of accessible scientific and medical information turns biomedicine into a fast growing field [79, 98]. This huge amount of data allows the use of modern Machine Learning (ML) models [167] for patient risk prediction [63], which is a central part of clinical risk management [122]. ML tools can also be used to automate some steps of health care process and to improve the performance of risk prediction [20]. It is rewarding to study the use of sophisticated predictive tools in this field; due to its high health quality impact.

From a clinical viewpoint, systems containing patient risk models provide tools to help clinical decision, reduce variability, provide protocol interventions, improve quality control, and decrease the necessary training time of professionals [133, 165].

Currently, there is scarce evidence that the use of ML tools in real clinical settings may replace the human decision. The process from innovation to routine clinical use is complex [107]. In addition, legal issues of the use of ML tools in health care have been raised [27, 72]. In this scenario, it is still necessary to develop and validate ML tools with the capacity to support important clinical decisions, such as assessing patient risk.

This subject is tremendously interesting because of its great impact and applicability [78]. Likewise, from scientific research viewpoint, it is interesting to explore the application and methodological development of this kind of models in real clinical settings under the constraints and challenges of these environments.

The main applications of ML in health according to [85], that extends the categories presented by [80], can be clustered into the following categories: Administration and delivery: managing health care delivery costs; Clinical decision support; Clinical information; Behavior / consumer; and, Support Information. The concept of patient risk that will be addressed falls in the area of Clinical Decision Support Systems (CDSS), specifically the point-of-care systems defined as *"Computer systems designed to impact clinician decision making about individual patients at the point in time that these decisions are made"* [20].

1.1. Clinical Decision Support Systems (CDSS)

Clinical Decision Support is .[a] process for enhancing health-related decisions and actions with pertinent, organized clinical knowledge and patient information to improve health and healthcare delivery"[125]. The aim of Clinical Decision Support Systems (CDSS) is to help physicians making faster and more reliable clinical decisions. The most common use of CDSS is for addressing clinical needs [19]. This condition reveals the dynamic characteristic of the CDSS. Due to this, there is no single definitive classification. There are different publications such as [18], [152] that claim to categorize the different research in CDSS. A well-studied criterion [20] divides CDSS into two types: a Knowledge-based and Non-knowledge-based. The central difference between this two types is that knowledge-based CDSS does not use any type of artificial intelligence, whereas non-knowledge CDSS does. This work focuses on CDSS of a type known as non-knowledge based.

Based on [19] and [65] figure 1.1 presents a more appropriate classification. This classification is based on the decision task supported by CDSS .

- **Alerts and Risk** This kind of CDSS contains systems that provides real time risk classification alerts and warnings to the medical staff. This category contains CDSS that are based on both continuous (monitoring) and discrete information.
- **Diagnostic Assistance** This kind of CDSS can be used to help clinicians by recommending a possible diagnosis or providing useful information to make a diagnostic decision. It includes diagnostic assistance that uses medical images, laboratory results, among others.
- **Treatment and planning** This kind of CDSS assists clinicians in the plan, therapy, and medical prescriptions. This group contains therapy critic and planning, and pres-

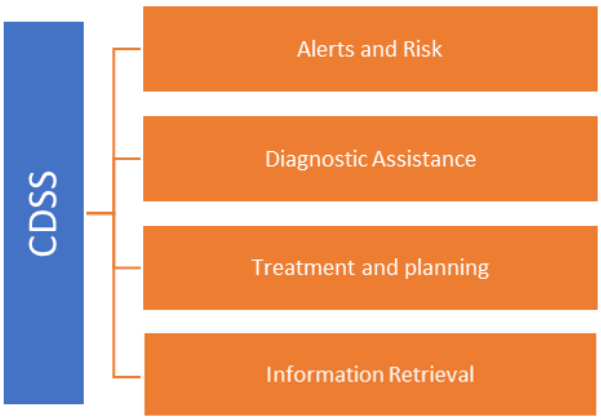

Figura 1.1: CDSS Classification

cription decision support. This category deals with the prescription of medications, as part of the patient's treatment.

- **Information Retrieval** are used to locate and retrieve an appropriate and accurate data that could be used for diagnosis or treatment planning.

Clinical decision support has a long history and is undergoing a renaissance with the advent of new ML techniques [79], giving birth to Machine Learning-based Clinical Decision Support Systems (ML-CDSS). The potential applications of ML-CDSS are manifold. The case studies worked out in this Thesis are a sample of this broad area.

In [153] a Framework for Classifying CDSS is presented using 24 descriptive axes. Some of these axes are used in this thesis to establish similarities and differences between selected problems.

1.2. Three Paradigmatic Problems of Patient risk modeling

The CDSS categories visualized in the figure 1.1 present different research opportunities, but also different challenges. Particularly in this Thesis examples of the first group were addressed, and specifically in predicting patient risk with time discrete information.

In [15] seven groups are presented: High-cost patients, Hospital Readmission, Triage, Decompensation (when a patient's condition worsens), adverse events, and treatment optimization (for diseases affecting multiple organ systems). In this thesis, three of these particular problems were selected: Hospital Readmission, Triage, and Decompensation of inpatients.

The motivation behind this selection is to show the capacity of the ML models to improve the results of the methods currently used by facing problems of similar characteristics, associated with the risk of patients. In this way it is possible to experiment with strategies that

have potential to improve the predictive results at different stages of the patient care processes. Also, these three patient risk problems selected have similar characteristics in terms of:

- The nature of uncertainty involved in patient risk [110]
- Highly class unbalanced data [17, 99]
- They can be formulated as a supervised learning problems [1, 189]
- Sensitivity is an important evaluation metric [184, 124, 3]
- Features and parameters depend on specific population characteristics [5, 57, 164]

One of the most important common characteristics of these three problems is that they focus on assessing the level of risk to make decisions adapting the level of care to the predicted risk. It is fundamental that the proposed solution should be a part of the health care process. The three problems are also characterized by requiring a level of risk assessment at a specific time instant. In other words, risk is assessed for a person at a particular time in the heath care process.

These particular characteristics motivate to focus the investigation on improving risk assessment models through the use of sophisticated ML techniques. It is important to consider that local hospitals does not necessarily have the same conditions for deployment as the places where these models were developed. Conditions like:

- Human and financial Resources
- Available Data
- Demographic and epidemiological characteristics of the populations
- Local politics
- Physical and social environment

Currently, the three selected problems have received a high research interest [73, 142, 181, 104, 67], and are recognized in the international literature as difficult to solve, particularly in pediatrics settings [2, 88, 21]. In the works reported in this thesis we have worked with data from Pediatric Exequiel González Cortés Hospital (EGCH), which is one of only three Pediatric hospitals of Chile. The hospital has approximately 52,000 m2 of surface, 6 Operating Room, 168 wards, 1046 workers, and around 110,000 ER Visits per year.

1.2.1. Hospital Readmission

A hospital readmission is an episode when a patient who had been discharged from a hospital is admitted again after a short time. Hospital readmission have different causes that include many social determinants, such as education, economics, and access to health care [185]. Readmission occurs within a specified time interval for example 7 or 30 days. In our scenario, our approach is to predict the second admission at the time of the first discharge. Hospital readmission is a frequent and expensive problem [6, 50]. It is widely used as an indicator of quality of care [13, 159]. This problem has been widely studied in the world, generally based on the information extracted from the EHR [9, 13].

The problem of predicting hospital readmission is difficult to address because it is multifactorial [46, 185] and the data is highly class imbalanced [7]. The objective is to develop a pediatric readmission prediction method based on the same information used to determine the Diagnosis Related Groups (DRGs) weight.

There is a large number of publications that address the prediction of the risk of patient readmission through the use of statistical techniques and using many types of available data. Logistic Regression (LR) and survival analysis have been the most used models in the literature, however, there is a growing interest in applying ML techniques to this subject. Often, these models exhibit poor predictive performance and would be unsuitable for use in a clinical setting [62]. A systematic review [91] of publications in this area reports AUCs between 0.56 and 0.72 in adults. However, this result may improve [41] if a greater volume of data is considered, such as activities of daily living assistance needed, visual impairment, functional status, or longitudinal data [9, 136].

In our case, to achieve the aim, a series of operations must be implemented, which considered erroneous data cleaning, labeling data, applying class balancing techniques, testing different classification models and evaluating results.

Hospital Readmission is a current problem [9], especially when the possibility of correcting the funding of hospitals based on their DRG production is discussed in Chile. The US experience shows that hospital funding with this characteristic must be corrected by a readmission index [42].

1.2.2. Triage

Triage is an assignment of degrees of urgency to wounds or illnesses of a specific patient to decide the order of treatment of a large number of patients or casualties. Triage are considered a key tool in emergency care process [69]. The motivation of the study of ML models is to support a simple and fast screening method based on the patient's degree and severity of medical need. It has been estimated that 40 % of patients arriving at Emergency Departments (EDs) have non-urgent problems [25]. This leads to overcrowded waiting rooms and long waiting times. As a consequence, patients needing care urgently are in risk of not being treated in time [141].

Triage systems are CDSS [161] that combine individual patient information and triage decision rules to classify patient's urgency [118]. The decision is well defined over specific parameters, such as vital signs, chief complaint, and past medical history [131]. Triage is the first and most critical step toward the early identification of the sick child and the timely delivery of emergency health care [111].

In general, triage systems are based on consensus opinion of experts [76]. The experts design decision trees to support clinical risk assessment or predictions of resource usage to define urgency levels. Triage systems should be simple to use, accurate, rapid, reproducible, and discriminative to avoid potentially dangerous under-triage [81].

In the EGCH a pediatric triage involve rapid recognition of seriously ill or injured children, assigning an acuity rating level, and anticipating appropriate emergency care and referral. In EGCH, as in the state-of-the-art triage tools actually in use, acuity rating levels used to prioritize patients for care go from level 1 (most acute) to level 5 (least acute).

The evaluation of triage systems involves assessment of reliability and validity [118]. Reliability refers to the degree of intraobserver variability and interobserver variability. The validity refers to the degree of triage prediction of "true" urgency. The validity corresponds to the model sensitivity and specificity [124]. It is difficult to evaluate the validity of a triage system. The fundamental problem in conducting studies to validate triage tools is the lack of consensus about the outcome measure [70, 74, 118, 141]. One way to measure validity is to compare the category assessed with a standard value. This includes the resources used and the end result of the triage. Hospitalization, admission to the ICU, proportion of children leaving without being seen by a physician (LWBS), and length of stay (LOS) can be used only as a surrogate markers of the urgency of a situation, but do not represent a perfect criterion standard for triage [75].

ML techniques could increase the consistence of triage classification in EDs [106, 30, 105]. As showed in [105] and [30] Neural Networks have shown better performance in triage prediction in terms of sensitivity. [106] use a combination of Self Organizing Feature Maps (SOMs) and K-Means cluster analysis to examine the emergency triage database.

Building a Triage system in children seems to be more challenging compared to adults and no study has compared international pediatric triage systems in the same group population [2]. Due to this, the EGCH decided to implement its own model in 2013, this model has proved to be valid as a pediatric triage when standard evaluation metrics of the literature are used. Both the ML proposal and the description of this model are described in chapter 3.

1.2.3. Decompensation of inpatients

Some decompensations of inpatients can be predicted using periodic bed-side vital signs observations. Around 85 % of severe adverse events (SAE) are preceded by abnormal vital signs [101], and 59 % within 1 – 4 h before cardiac arrest [4]. A group of models developed with this propose are called Early Warning Systems (EWS). So the bed-side vital signs observation forms the basis of EWS models. EWS have evolved as a means of alerting health professionals to patient clinical decompensation risk.

Currently, there are many different EWS in use in different heath institutions. There are some EWS constructed based on expert opinion, such as the National Early Warning Score (NEWS) [123, 156], Modified Early Warning Score (MEWS) [160] and VitalPAC Early Warning Score (VIEWS) [135]. There are also pediatrics EWS based on expert opinion such as Children's Hospital Early Warning Score (C-CHEWS) [112], Pediatric Early Warning Score (PEWS) [54] and Bedside PEWS [130]. MEWS and ViEWS can be used on non-ICU ward patients with good performance [183].

Others EWS were derived using statistical modeling (Analysis of variance ANOVA, Back-

ward stepwise Regression) such as the Rothman Index [140] and the electronic Cardiac Arrest Risk Triage (eCART) score [36]. [12] Shows EWS generated entirely algorithmic using Decision Tree (DT) analysis. [39] Shows one-class SVM approach using partial AUC to optimize SVM Parameters. Discrete-time logistic regression are also used as an effective and efficient methods to predict adverse clinical outcome [97].

Most of this EWS models were designed to detect deteriorating patients in hospital wards, specifically those at increased risk of: unexpected ICU admission, unplanned return to the operating theatre, or a prolonged length of stay, cardiac arrest, or death. These outcomes are used to create labels in training data, and also, to evaluate model performances.

This wide range of models is due to the fact that each one has different characteristics, such as: the type of patient unit, type of patient (pediatric or adult), the quantity and origin of the used parameters. In our particular case, we will focus in models for pediatric inpatient, in non-ICU wards, focused on vital signs monitoring and others bedside observations.

Possibly, simultaneous use of a Triage system and a model of the risk of decompensation of inpatients can improve the predictive capacity of ED triage [23]. Currently, the EGCH has a pioneering self-developed pediatric EWS model in Chile, however, evaluation and improvement of these models must be developed.

1.2.4. Categorization of the Paradigmatic Problems

In Table 1.1 16 of 24 axes of a Framework for Classifying CDSS [153] were presented to show similarities and differences between the three paradigmatic problems dealt with in this Thesis.

Table 1.1 shows that the selected problems coincide in: Clinical Task, Unit of Optimization, Relation to Point of Care, Reasoning Method, Delivery Format, Delivery Mode, Action Integration effort and Explanation Availability. This shows that three selected problems are similar in terms of context and information delivery, but not in terms of Workflow and the Decision that is supported.

1.3. Objective

We plan to research three paradigmatic problems: Hospital Readmissions, Triage and Decompensation of inpatients. The idea is to improve model results in terms of different performance metrics.

Tabla 1.1: Selected problems coded using taxonomy

	Readmission	**Triage**	**Dec. of inpatients**
Context			
Clinical setting	Inpatient	Outpatient	Inpatient
Clinical Task	Screening	Screening	Screening
Unit of Optimization	Outcomes	Outcomes	Outcomes
Relation to Point of Care	Physician-patient	Clinician-patient	Clinician-patient
Decision Support			
Reasoning Method	ML approach	ML approach	ML approach
Clinical Urgency	Non-urgent	Urgent	Urgent
Recommendation Explicitness	Non-Explicit	Explicit	Explicit
Logistical Complexity	Complex	Non-Complex	Complex
Information Delivery			
Delivery Format	Integrated w/EMR	Integrated w/EMR	Integrated w/EMR
Delivery Mode	Pull	Pull	Pull
Action Integration	Minimal effort	Minimal effort	Minimal effort
Explanation Availability	Non-available	Non-available	Non-available
Workflow			
System User	Physician	Clinician and patient	Clinician
Target Decision Maker	Physician	Clinician	Clinician
Output Intermediary	Non-intermediary	Clinician	Non-intermediary
Workflow Integration	Moderate	Full	Full

1.3.1. General Objective

This thesis aims to be a methodological contribution in state of the art of Machine Learning algorithms applied to patients' risk problems.

1.3.2. Specific Objectives

- Review of the state of the art of different ML techniques applied in three different patient risk problems.
- Benchmarking these techniques using hospital real data.
- Explore and propose different ways to improve models results.
- Propose methodological models will be deployed in real clinical settings.

1.4. Methodology

CRoss-Industry Standard Process for Data Mining (CRISP-DM) [32, 33] is the most used methodology for developing Data Mining and Knowledge Discovery projects. It is actually a "de facto" standard in this area [148]. The CRISP-DM methodology is a hierarchical process model, divided into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. Some of these phases are cyclic, which means that some phases will allow partially or totally revising the previous phases, as shown in

Figure 1.2.

Understanding the business is a phase in which objectives and requirements must be established from a non-technical perspective. This requires an evaluation of the situation from the point of view of resources, requirements, assumptions, constrains, etc.

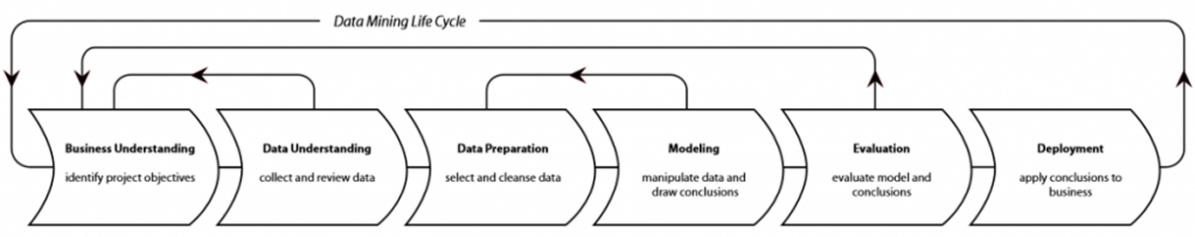

Figura 1.2: CRISP-DM framework (Source CRISP-DM 1.0 http://www.crisp-dm.org)

The Data Understanding phase involves the compilation, description and initial data exploration. This is done in order to comply with an adequate quality verification of the data.

The next CRISP-DM phase corresponds to a selection, cleaning, construction, integration and formatting of data. This phase is called Data Preparation and its objective is to obtain the minable view of the data.

The next phase aims to apply the techniques of data mining to the dataset from the previous stage. This stage is called Modeling and requires the selection of the modeling technique, the design of the evaluation metrics, the construction of a preliminary model and its evaluation in first iterations of the methodology.

The next phase is called Evaluation of the models selected in previous phase and determine if they are useful to the needs of the business. In this stage, the results obtained and their application in real clinical environments must be evaluated.

The final stage is called Deployment and its objective of this phase is to implement the models, integrating them into the decision-making tasks of the organization.

1.5. Structure

This thesis consists of two publications plus a working paper Under Review at International Journal of Medical Informatics. First paper (chapter 2 of this thesis) corresponds to an international publication that shows the results obtained after implementing a prediction model of hospital readmissions in 30 days. This publication shows a methodological proposal that involves the use of different ML models combined with other tools such as: labeling, class balancing and Cross-validation. This corresponds to the first publication that uses the ML approach in solving the problem of pediatric readmission.

The second paper (Chapter 3 of this thesis) corresponds to a publication that addresses the problem of pediatric triage with ML approaches. This publication shows a methodological

proposal to use and evaluate these ML tools from a technical and clinical viewpoint. The results obtained in this publication exceed those presented in other publications in the case of adults and show slightly better results in the case of another single publication in pediatric triage.

The third paper (Chapter 4 of this thesis) shows the development and evaluation of a Pediatric Early Warning System supported by ML approaches. The results obtained in this work outperform the results shown in literature for similar problems. The proposal uses, in general terms, the same methodology shown in the other two previous publications.

In Chapter 5 the First findings are presented, the Proposed methodology is discussed, as well as the Results obtained and the Applications and implications of the results obtained. In this chapter the Final remarks and Further research are also presented.

1.6. Contributions List

The following is the list of academic contributions made during this PhD Thesis. It includes indexed and peer-reviewed journal papers and conference posters, as well as conference presentations.

1.6.1. Indexed Journal Papers

- Wolff P, Graña M, Ríos S & Yarza MB (2019) Machine learning readmission risk modeling: a pediatric case study, BioMed Research International, 2019:9. [178].
  [JCR (2018) **2.197** (Q3 BAM, MRE)]
- Wolff P, Ríos S & Graña M (2019) Setting up standards: A methodological proposal for pediatric Triage machine learning model construction based on clinical outcomes, Expert Systems with Applications, 138:112788 [180].
  [JCR (2018) **4.292** (Q1 CSAI, EEE, ORMS)]
- Wolff P & Ríos S (2019) A pediatric early warning system machine learning model based on clinical outcomes, International Journal of Medical Informatics [Submitted in July 2019]
  [JCR (2018) **2.731** (Q2 MI, CSIS, HSS)]

1.6.2. Other Related Indexed Journal Papers

- Durán G, Rey P, & Wolff P (2017) Solving the operating room scheduling problem with prioritized lists of patients. Annals of Operations Research, 258(2):395–414. [56]
  [JCR (2017) **1.864** (Q2 ORMS)]
- Julio C, Wolff P & Yarza MB (2016) Waiting lists management model based on timeliness and justice. Revista Médica de Chile, 144(6):781–794. [90]
  [JCR (2016) **0.519** (Q4 MGI)]

1.6.3. Peer-reviewed Journal Papers and Conference Presentations

- Wolff P & Ríos S (2019) Predicción de readmisión de pacientes pediátricos mediante aprendizaje supervisado [Accepted] Revista Ingeniería de Sistemas.
- Wolff P, Yarza MV & Ríos S (2018) Predicción de readmisión hospitalaria utilizando data del GRD. In: XXXIII Jornadas Chilenas de Salud Pública. Santiago de Chile.
- Wolff P, Alcaina E & Nalegach ME (2018) Modelo de riesgo de descompensación/deterioro clínico en pacientes hospitalizados. In Poster: XXXIII Jornadas Chilenas de Salud Pública. Santiago de Chile.
- Wolff P, Yarza MV & Ramirez V (2018) Propuesta de Modelo de Asignación de hora médica para lista de espera ambulatoria. In Poster: XXXIII Jornadas Chilenas de Salud Pública. Santiago de Chile.
- Wolff P & Yarza MV (2015) Análisis de la capacidad centrado en la calidad de Servicio. In: XXXII Jornadas Chilenas de Salud Pública. Santiago de Chile.

Chapter 2

Machine learning readmission risk modeling: a pediatric case study

Wolff P, Graña M, Ríos S & Yarza MB.[1]

Background: Hospital readmission prediction in pediatric hospitals has received little attention. Studies have focused on the readmission frequency analysis stratified by disease and demographic/geographic characteristics but there are no predictive modeling approaches, which may be useful to identify preventable readmissions that constitute a major portion of the cost attributed to readmissions.

Objective: To assess the all cause readmission predictive performance achieved by Machine Learning techniques in the emergency department of a pediatric hospital in Santiago, Chile.

Materials: An all cause admissions dataset has been collected along six consecutive years in a pediatric hospital in Santiago, Chile. The variables collected are the same used for the determination of the child's treatment administrative cost.

Methods: Retrospective predictive analysis of 30-day readmission formulated as a binary classification problem. We report classification results achieved with various model building approaches after data curation and preprocessing for correction of class imbalance. We compute repeated cross-validation (RCV) with decreasing number of folders to assess performance and sensitivity to effect of imbalance in the test set and training set size.

Results: Increase in recall due to SMOTE class imbalance correction is large and statistically significant. The Naive Bayes (NB) approach achieves the best AUC (0.65), however the

[1] The following is an unabridged version of the paper published in "BioMed Research International". Please cite this paper as follows: Patricio Wolff, Manuel Graña, Sebastián A. Ríos, and Maria Begoña Yarza, "Machine Learning Readmission Risk Modeling: A Pediatric Case Study", BioMed Research International, vol. 2019, Article ID 8532892, 9 pages, 2019. https://doi.org/10.1155/2019/8532892.The original publication is available at: https://www.hindawi.com/journals/bmri/2019/8532892/

shallow multilayer perceptron has the best PPV and f-score (5.6 and 10.2, resp.). The NB and support vector machines (SVM) give comparable results if we consider AUC, PPV and f-score ranking for all RCV experiments. High recall of deep multilayer perceptron is due to high false positive ratio. There is no detectable effect of the number of folds in the RCV on the predictive performance of the algorithms.

Conclusions: We recommend the use of Naive Bayes (NB) with Gaussian distribution model as the most robust modeling approach for pediatric readmission prediction, achieving the best results across all training dataset sizes. The results show that the approach could be applied to detect preventable readmissions

2.1. Introduction

Hospital readmission is defined as the non-scheduled return of a patient within a short prespecified period of time after hospital discharge. An internationally extended standard period to count a patient return as readmission is 30 days, but it may change for political reasons [91]. In the United States (US), hospital readmission is being used as an indicator of patient care quality. Both public an private funding agencies use this measure to penalize underperforming institutions [121]. It has been argued that up to two thirds of the readmissions are preventable, therefore advances in patient readmission prediction are worth the investment [15, 62]. US policy has inspired similar concerns in other countries, so that readmission analysis and prediction is under consideration worldwide. The data collected in the Electronic Health Record (EHR) is the main information source for the predictive modeling of readmissions, and the analysis of their consequences and structural/organizational causes [15, 151].

Readmission prediction in the case of adult patients has been tackled with diverse statistical approaches [6, 91] such as logistic regression [66, 126], and survival analysis [67]. Recent works favor the application of predictive machine learning approaches, formulating readmission prediction as a binary classification problem [7, 66]. For example, the literature report results from support vector machines (SVM) [62, 189, 44], deep learning [137, 181], Artificial Neural Network [126], and Naive Bayes [171, 151].

Despite this long history of studies about hospital readmission for adult patients, but there are almost no studies devoted to readmission of pediatric patients [121]. In the pediatric case, hospital readmission prediction has been only reported in the setting of emergency department [5, 17] and intensive care units [93]. Few studies report results on both adult and pediatric patients [66], finding lower sensitivity in the pediatric population than in the adult population, due to greater class imbalance in the pediatric datasets. In this paper we report the predictive modeling results over a large cohort of all cause admissions to the emergency department of a pediatric hospital in Santiago, Chile. We tested four modeling applications considering various numbers of folds in a repeated cross-validation approach, achieving results comparable to those reported for adult patient readmissions.

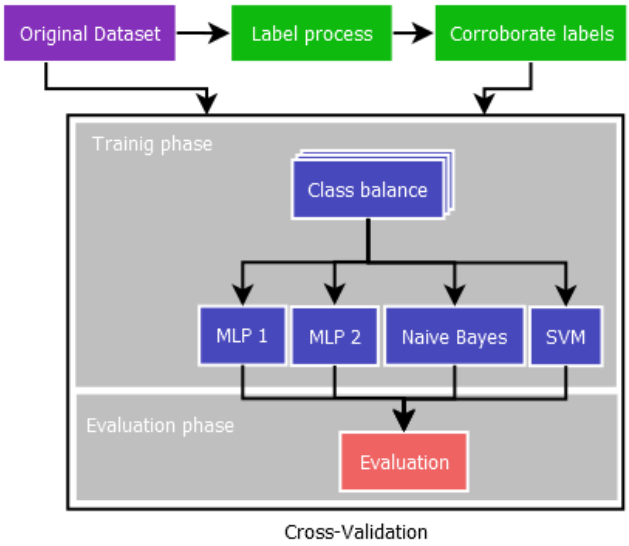

Figura 2.1: Study design

2.2. Materials and Methods

The overall model training and validation process is shown in Figure 2.1. First, the EHR data entries were labeled as readmissions according to the following rules: a) we consider admissions in period of less than 30 days after the previous discharge; b) we discard an admission if it corresponds to programmed treatments such as chemotherapy, or if it is intended for services that are not urgent. We check (corroborate) the correctness of the generated labels by an expert committee; which consisted of two experienced medical doctors and two nurses from the hospital's quality and safety care team. The whole data is then used for validation in a repeated cross-validation (RCV) process with different numbers of folders, we carried out 10-fold, 5-fold, 4-fold and 3-fold RCV. Each cross-validation repetition consists in the following steps: 1) partition of the dataset in the selected number of folds, 2) each fold is alternatively used as the test dataset while the remaining folders are used for model training, 3) average performance measures are computed over all cross–validation folds and repetitions . As illustrated in Figure 2.1, training at each RCV step is preceded by a class balance process carried out on the training dataset. We apply a SMOTE [34] up-sampling procedure using the five nearest neighbors of each minority class sample [7, 66]. The reported results are the average of the 30 repetitions of the CV results. We have published the script of the implementation as open source code for independent examination [179].

2.2.1. Cohort and dataset

The descriptive statistics of the dataset used for the study are summarized in Table 2.1. It contains records of 56,558 admissions with 2106 readmissions in the period from July 2011 to October 2017 at the pediatric Hospital Dr. Exequiel González Cortés in Santiago, Chile. All data has been anonymized for the study. One author (PW) acts as the honest data broker ensuring compliance with data protection regulations. The categories of data available to

Tabla 2.1: Descriptive statistics of the dataset.

Dataset characteristic	
Total number of admissions	56,558
Number of unique individuals	35,064
Percent readmission within 30 days	3.72 %
Number of unique procedures (ICD-10 AM)	1,124
Number of unique diagnoses (ICD-10 AM)	4,370
Variables used in prediction	
Age (years), mean (SD)	5.78 (5.04)
Male (%)	59.2
Public facilities	1
Number of Transfers (SD)	0.61 (0.8)
Length of Stay (days), mean (SD)	3.77 (10.03)

build machine learning based predictors are the following ones:

- Data used by the administrative cost coding system, specifically Age, Sex, Ethnic group, anonymized geographical information (i.e. postal code), Public insurance plan, Principal Diagnosis, Secondary Diagnosis, Tertiary Diagnosis and Main Procedure performed.
- Information about patient's admission: the date of admission, the service in which he/she was admitted, and his/her origin.
- Information on internal transfers: Date/hour, Service of origin and Internal destination.
- Information about the patient's discharge: Discharge date, Service that performs the discharge, and the patient's destination.

Though we have not carried out a detailed statistical survey of the occurrence of readmissions according to specific diagnostics [119], we have been able to identify the diagnostic at discharge accounting for most of readmissions as detailed in Table 2.2. There is a big prevalence of respiratory conditions that can be attributed to pollution events in the city of Santiago.

To improve data quality a manual data curation process was carried out. Identification of admissions that are actual readmissions was carried out automatically. The resulting labeled dataset is heavily class imbalanced. A taxonomy of methods to deal with imbalanced data is presented in the context of readmission prediction is given in [6]. For training, we applied a class balancing technique, specifically a SMOTE [34] on the minority class using five nearest neighbors. We have considered increasing sizes of the balanced training set, leaving the remaining (imbalanced) as the test set.

2.2.2. Classification methods

Several machine learning [177, 84] approaches have been selected for predictive model building . These models have been reported in the literature about readmission prediction for adult patients [6, 91]. We have discarded application of deep learning approaches [71] because the available data is too shallow. There is no spatial information, the time sequences

Tabla 2.2: Diagnostics at discharge accounting for most readmission

Diagnostic	ICD10	%
Viral pneumonia	J129	9.50
Respiratory syncytial virus pneumonia	J121	9.16
Acute bronchitis	J209	3.94
Unspecified gastroenteritis	A090	2.80
Disorders of prepuce	N47	0.90

of readmissions are too short to be exploitable, and the number of variables per patient data entry is too small to generate high dimensional hierarchical representations. Therefore we focus on well known classical methods. The reported applications of deep learning to readmission prediction are restricted to a specific disease, i.e. lupus patients [137], for which there are long clinical histories *per* patient accessible through the EHR, so that the abundance of data allows for the training of deep models.

***Support Vector Machines* [169]** Support Vector Machines (SVM) classifiers are linear discriminant functions built from samples placed at the boundaries of the classes. Their learning algorithm looks for the discriminating hyperplane maximizing its distance to the boundaries belonging to each class, i.e. maximizing the margin of the decision function relative to the class boundary. The parameters that define the solution hyperplane come from the optimization of a quadratic programming problem. When the classes are not linearly separable, then it is possible to project the data into a space of superior dimensionality using the kernel trick [150], so that the transformed dataset becomes linearly separable. The literature shows that SVMs are quite robust against the curse of dimensionality, achieving good results on small datasets of high dimensionality feature vectors. We used LibSVM [31] library for training and estimation of the SVM metaparameters via grid search. Best results were obtained with a Radial Basis Function (RBF) kernel. We have used LibSVM[2] for SVM training.

Multilayer perceptron Multilayer Perceptron (MLP) are the classical feed-forward artificial neural networks (ANN) composed of multiple densely interconnected layers of computational units, aka artificial neurons. The output of each unit is computed as the linear combination of the incoming connection weights and their source units in the previous layer filtered by a non-linear activation function. The classical sigmoid activation function has been replaced by other like the rectified linear activation used in deep learning architectures. The connection weights implement a discriminant function that may take arbitrary shapes. In fact it has been shown that, even with a single hidden layer, an MLP can approximate any function. The connection weights can be learned from data applying the back-propagation algorithm [84].

We have applied two flavors of MLP to pediatric readmission prediction. The first one (denoted MLP1 in the results section) is an auto-tunable implementation, called AutoMLP

[2] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

for short, that performs automatic online model parameter tuning during training process, including the creation of an ensemble of MLPs [144]. The number of maximum training cycles used for the ANN training was 10 equals to the number of generations for AutoMLP training and the number of MLPs per ensemble chosen was 4.

The second (denoted MLP2 in the results section) is a multi-layer feed-forward artificial neural network trained using back-propagation with stochastic gradient descent [71]. The activation function used by the neurons in the hidden layers was a Rectifier function. The MLP2 has two hidden layer, each of 50 neurons. It was trained in 10 epochs using an adaptive learning rate algorithm (ADADELTA) [186] which combine the benefits of learning rate annealing and momentum training to avoid slow convergence. We used the H_20 package[3] for this MLP training and validation [43].

Naïve Bayes method The Naïve Bayes (NB) approach is based on the assumption that the individual features are statistically independent, therefore we approximate the joint probability distribution of a high-dimensional feature vector as the product of the unidimensional distribution probabilities of each feature. In our study we use unidimensional Gaussian probability density models of the independent feature distributions. Training was carried out by straightforward estimation of these unidimensional probability densities.

2.2.3. Classification performance metrics

At each cross-validation fold we compute the confusion matrix and performance metrics derived from it, finally reporting the average of these results. Let us define TP, TN, FP, and FP as true positive, true negative, false positive and false negative counts. Then we compute the Recall (aka sensitivity) as:

$$R = \frac{TP}{TP + FN}, \tag{2.1}$$

Positive predictive value as:

$$PPV = \frac{TP}{TP + FP}, \tag{2.2}$$

and f-score as:

$$F = \frac{2}{1/R + 1/PPV} \tag{2.3}$$

These measures are more informative than the accuracy ($A = \frac{TP+TN}{TP+TN+FP+FN}$) of the successful detection of the minority class (i.e. the readmissions) because the dataset is strongly class imbalanced. The analysis using Receiver Operating Characteristic (ROC) curves has been widely used to compare different binary classifiers. The ROC is a plot of sensitivity versus the false positive rate ($FPR = \frac{FP}{FP+TN}$). It is widely used to compare performances

[3] `https://www.h2o.ai`

of state of art of supervised learning classification methods. Specifically the integral of the ROC, i.e. the Area Under ROC Curve (AUC), is often reported in readmission prediction studies of adult patients [6].

We compute these measures over the test dataset after training the models in an RCV process explained above. At each fold test, the remaining folds are put together as the training dataset. The training dataset is class-balanced using SMOTE [34] with five nearest neighbors on the minority class training samples until we have the same number of samples of each class. However, the test set remains unaffected and heavily imbalanced. One consequence is that small errors in absolute terms (e.g. one misclassified sample) translate into large reductions of the performance measures. The proportion of samples of the minority class in the test dataset depends on the number of folds used for RCV. High number of folds implies big reductions in the number of minority class samples in the test fold, thus increasing its imbalance ratio (the ratio of the majority class sample size to the minority class sample size), which may lead to numerical instabilities of the performance results. For this reason, we have explored the results obtained using a decreasing number of RCV folds.

2.3. Results

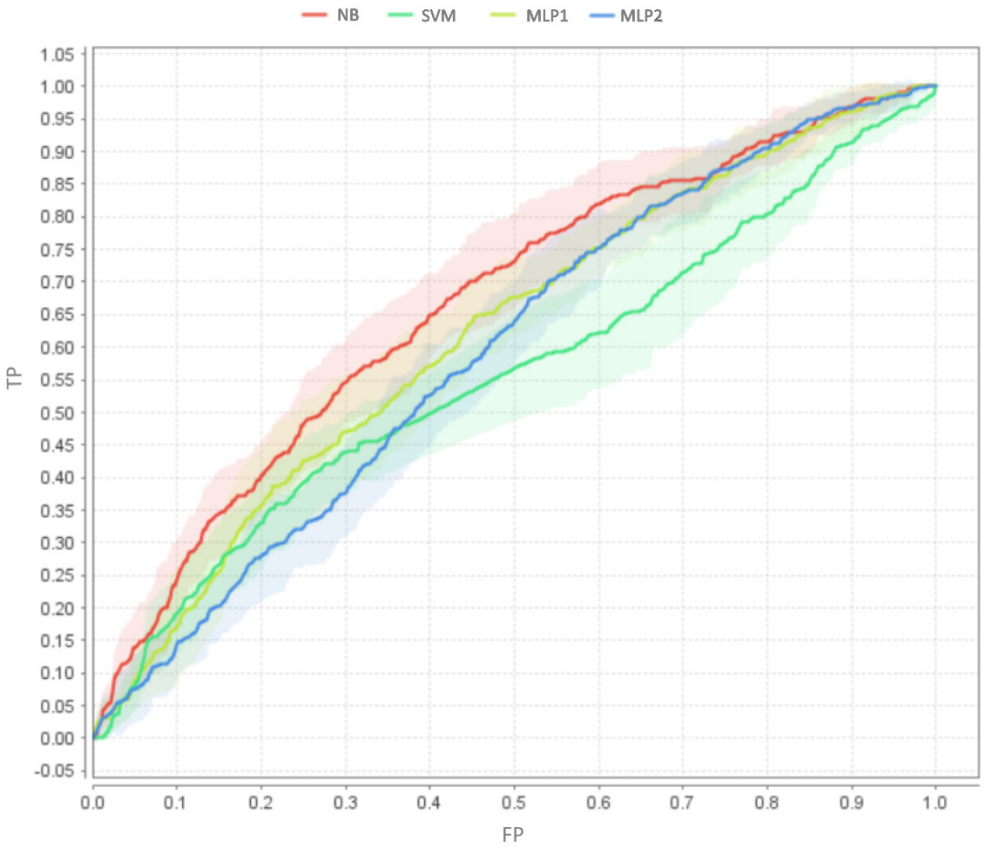

Figura 2.2: Average ROCs of machine learning approaches in 5-fold RCV (applying SMOTE class imbalance correction). Solid line corresponds to the ROC mean.

Tabla 2.3: Average ± standard deviation Recall (R) performance [%] of SVM, MLP1, MLP2, and NB for decreasing number of folders in the RCV process. no SMOTE = no oversampling correction of class imbalance is done.

nfolds	SMOTE			
	SVM	MLP2	MLP1	NB
10	45.63 ±3.35	96.29 ±2.15	59.93 ±5.51	70.8 ±2.68
5	44.64 ±2.69	96.58 ±1.77	61.39 ±6.14	69.8 ±4.97
4	43.83 ±1	95.11 ±1.06	59.87 ±6.29	70.23 ±3.82
3	43.64 ±1.11	96.86 ±0.37	52.8 ±5.24	67.57 ±0.97
	no SMOTE			
	SVM	MLP2	MLP1	NB
10	0.95 ±0.76	27.60 ±11.13	0.00 ±0.00	14.81 ±1.83
5	1.04 ±0.71	33.24 ±8.65	0.00 ±0.00	14.77 ±1.43
4	1.00 ±0.21	29.11 ±13.90	0.00 ±0.00	14.91 ±1.6
3	1.14 ±0.23	30.32 ±17.48	0.00 ±0.00	14.67 ±1.89

Tabla 2.4: Average ± standard deviation Positive predictive value (PPV)[%] of SVM, MLP1, MLP2, and NB for decreasing number of folders in the RCV process. no SMOTE = no oversampling correction of class imbalance is done.

nfolds	SMOTE			
	SVM	MLP2	MLP1	NB
10	5.52 ±0.35	3.92 ±0.09	5.61 ±0.47	5.28 ±0.16
5	5.43 ±0.27	3.98 ±0.1	5.25 ±0.14	5.29 ±0.31
4	5.39 ±0.1	3.99 ±0.01	5.29 ±0.19	5.29 ±0.07
3	5.48 ±0.1	3.94 ±0.03	5.34 ±0.07	5.4 ±0.09
	no SMOTE			
	SVM	MLP2	MLP1	NB
10	42.22 ±29.86	6.23 ±1.53	NA	9.05 ±1.11
5	32.47 ±16.63	5.40 ±0.59	0.00	9.02 ±0.95
4	45.24 ±5.35	6.60 ±1.96	0.00	9.09 ±1.13
3	45.24 ±12.14	6.22 ±0.82	NA	8.90 ±0.89

Tables 2.3, 2.4, 2.5, and 2.6 show the average recall, positive predictive value, f-score, and AUC, respectively, of the machine learning techniques after 30 repetitions of the RCV experiments with varying number of folders, with and without SMOTE class imbalance correction. The effect of the number of folds is negligible. An F- test over the number of folds shows that there is no statistically significant difference ($p>0.1$).

The difference between results due to the use of SMOTE class imbalance correction at model building is largely statistically significat ($p<0.00001$ one sided t-test of PPV, f-score and AUC values almost for all models). For the the results without SMOTE are somehow paradoxical. The PPV grows significatively in some cases (for SVM >40 %), but the recall is extremely low (for SVM <2 %). The interpretation is that the number of cases classified as positive is very small, so that a small number of true positives gives high PPV. For MLP1 we found many instances of NA values due to the lack of positive responses.

Tabla 2.5: Average ± standard deviation f-score (F) performance [%] of SVM, MLP1, MLP2, and NB for decreasing number of folders in the RCV process. no SMOTE = no oversampling correction of class imbalance is done.

	SMOTE			
nfolds	SVM	MLP2	MLP1	NB
10	9.85 ±0.63	7.54 ±0.16	10.23 ±0.8	9.83 ±0.3
5	9.67 ±0.49	7.65 ±0.19	9.67 ±0.26	9.83 ±0.53
4	9.6 ±0.17	7.65 ±0.02	9.71 ±0.23	9.83 ±0.13
3	9.73 ±0.18	7.57 ±0.06	9.69 ±0.07	9.98 ±0.17
	no SMOTE			
	SVM	MLP2	MLP1	NB
10	1.86 ±0.00	9.70 ±1.45	NA	11.23 ±1.37
5	2.04 ±0.00	9.16 ±0.82	NA	11.20 ±1.14
4	1.95 ±0.40	9.62 ±0.75	NA	11.29 ±1.32
3	2.22 ±0.45	9.60 ±0.52	NA	11.08 ±1.23

Tabla 2.6: Average ± standard deviation AUC performance of SVM, MLP1, MLP2, and NB for decreasing number of folders in the RCV process. no SMOTE = no oversampling correction of class imbalance is done.

	SMOTE			
nfolds	SVM	MLP2	MLP1	NB
10	0.597 ±0.022	0.539 ±0.022	0.643 ±0.020	0.654 ±0.014
5	0.587 ±0.010	0.55 ±0.018	0.634 ±0.011	0.653 ±0.014
4	0.585 ±0.008	0.548 ±0.021	0.63 ±0.009	0.655 ±0.008
3	0.584 ±0.009	0.55 ±0.011	0.628 ±0.010	0.653 ±0.011
	no SMOTE			
	SVM	MLP2	MLP1	NB
10	0.495 ±0.020	0.631 ±0.026	0.661 ±0.021	0.656 ±0.014
5	0.481 ±0.019	0.615 ±0.008	0.661 ±0.008	0.658 ±0.007
4	0.473 ±0.004	0.631 ±0.011	0.661 ±0.012	0.659 ±0.008
3	0.471 ±0.007	0.627 ±0.015	0.657 ±0.002	0.658 ±0.009

Let us consider the case when we apply the SMOTE class imbalance correction. Attending to recall (R) in Table 2.3, MLP2 is well above SVM, MLP1, and NB, however, this is at the cost of a high false positive ratio, as demonstrated by the values of the PPV in Table 2.4, which is much lower for MLP2 than for SVM, MLP1, and NB. Figure 2.2 shows the ROC curves for all approaches in the case of RCV with 5 folders.

The f-scores shown in Table 2.5 confirm that SVM, MLP1, and NB improve over MLP2 regardless of RCV number of folders. An F-test carried out over these results confirms ($p<0.01$) that the performance differences between predictive models are statistically significant. Ensuing specific one-sided t-tests comparing each pair of modeling approaches confirms that SVM, MLP1, and NB perform significantly better than MLP2. The AUC results in Table 2.6 confirm that NB is significantly better than the remaining approaches (F-test $p<0.01$, pairwise t-test $p <0.001$). However, the superiority of NB relative to MLP1 is less pronounced (pairwise t-test $p<0.05$). Notice that statistical significance is due also to small standard

deviation of the results, if we consider the mean performance values, we can assert that SVM and NB show comparable performances.

2.4. Discussion

Readmission as a healthcare quality measure Readmissions as a healthcare quality measure has been the subject of strong debate both in adult and pediatric hospital environments [121]. The cost of readmissions within a 365 day period is estimated as $1 billion in United States pediatric hospitals [14], hence the need for focused analysis and predictive tools. There are, however, some studies that question the value of readmissions as a quality of care metric for specific type of patients, e.g. those suffering heart failure [128]. Other studies argue that too much emphasis in readmissions as a measure of the quality of care may lead to an increase of the unequal distribution of resources [91]. There is a need to be precise in the definition of which readmissions are to be penalized. For instance, if there is not distinction between planned and unplanned readmissions, there is a possibility that the hospitals would tend to delay required readmissions after the 30-day limit to avoid financial penalties [10]. It is also well known fact that a small percentage of pediatric patients with chronic conditions and special technological assistance needs account for a big percentage of the actual readmission costs [89]. The emphasis is, therefore, in the identification of the kind of readmission events that can be prevented through special care after discharge, such as phone calls [60].

Quantitative analysis of readmissions in pediatric care Thought readmission prediction has been extensively studied in adult patients, there is very little effort in children hospitals. One reason is that the percentage of admissions that result in readmission is much less frequent event in the pediatric case, in the range 3 % to 5 % on average, that in adult patients, which is close to 17 % on average [62], so it was dismissed in cost analysis studies until recently. To our knowledge, our study is among the first ones applying machine learning techniques to all cause pediatric readmissions. We have only found one similar study with a smaller cohort [17] in an Italian hospital. Recent studies are devoted to the characterization of the readmission events in the pediatric setting. Auger et al. [10] propose a method for the identification of unplanned versus planned readmissions which has many implications in the way readmissions are treated in order to avoid financial penalties. For instance, planned readmissions may be delayed to avert financial penalties. It is also important to identify which pediatric conditions are lead to higher readmission rates, realizing that they may be changing from one institution to another due to local demographic and environmental conditions, for instance some studies found strong dependence of frequency of readmissions on the ethnic, disease, chronic condition, and other demographic information such as the public versus private insurance [89, 129, 28]. Dependency of readmission frequency on clinical and geographic factors for a specific chronic condition (i.e. sickle cells disease) has been reported [113]. On the other hand, shorter length of stay in pediatric hospitals is not a cause for higher readmission rate [119]. Another issue is the impact of the use by the administrations in charge of financial control of the hospital of proprietary algorithms for the detection of preventable readmission detection. Being proprietary, the actual reasoning behind the decision is unknown, and thus it is quite difficult to predict its outcome in order to optimize patient

care and financial management simultaneously [68].

The difficulties are faced when trying to look for agreement among readmission prediction research studies or assessing the significance of a new study are the following:

1. The conditions for readmission are local to the population treated by the hospital. It is unrealistic to apply the same risk assessment/prediction model in two countries with huge differences in life parameters and conditions. Therefore, it is widely recognized that predictive models need to be developed at each site using local data [5, 91].
2. Because hospital readmission is a much less frequent event than no readmission, data used in all reported studies is heavy class imbalanced [17]. In our study, the readmissions account for only $3,7\,\%$ of the samples. Therefore, class balancing techniques are required to avoid model bias towards the majority class [175].
3. Often, EHR data has a lot of errors and missing information due to the stressful conditions of its capture. Moreover, there is no guarantee that the collected variables are indeed the most relevant for the intended prediction. However, it is the only available data for this purpose most of the times. Recent reviews and comparative studies [6, 62, 91] have found that studies on adult readmissions reported low values of area under ROC Curve (AUC aka c-statistic) ranging between 0,56 and 0,72. One way to improve prediction results is to carry out stratified studies, i.e. building specific predictive models for specific patient categories [22].

Class imbalance The readmission rate in our case study is $3,7\,\%$ which is similar to the percentage of readmissions reported in other studies about pediatric readmissions, i..e. 2,6 % in [28]. Class imbalance poses great difficulties both during training and validation. At training time, machine learning approaches are biased towards the majority class, so data preprocessing is required to create balanced training datasets[6, 66]. We choose to up-sample the minority class using SMOTE [34]. Additionally, care must be taken in the selection of the performance metric. Overall accuracy is strongly influenced by the majority class correct classification, therefore we need to use performance measures that take into account the performance regarding the minority class, hence we consider the positive predictive value (PPV), f-score (F), and the area under the ROC (AUC). The cost of false positive decision is much lower than false negatives, therefore we have not considered setting a false positive ratio for all algoriths. The AUC measure has been reported in most predictive studies of readmission. Our top result (AUC=0.655 for NB) is similar to the results already reported for adult readmissions (between 0,56 and 0,72). For a dramatic illustration of the effect of the class imbalance, we report the results without using SMOTE class imbalance correction. We find a huge decrease in recall performance, meaning that the readmission prediction drops drastically relative to the models built upon SMOTE corrected training data, beause of large bias towards the majority class in the non-SMOTE models. The small number of positive predictions lead to some paradoxical results, such as the increase of PPV value relative to the SMOTE models, because the false positive predictions are also very scarce.

Limitations of the study The dataset comes from a single hospital, so results reported need to be assessed with data coming from a network of hospitals in the same country. In-

cluding data from other countries risk the introduction of uncontrollable variations due to diverse data gathering protocols and differences in prevalent morbid conditions. For instance, sickle cell crisis is a costly and frequent readmission condition in USA [68] while it is non-existent in Chile. Therefore, it is quite necessary to carry out local studies in order to assess predictability and preventability instead of importing models from other countries which may be misleading. The existence of EHR data collection, anonymization, and distribution infrastructures in United States, such as the Pediatric Health Information System of the Children's Hospital Association (`https://childrenshospitals.org`) or the Nationwide Readmissions Database (`https://www.hcup-us.ahrq.gov/nrdoverview.jsp`), has favored the realization of studies covering many institutions and large cohorts [89, 68, 119, 28, 14, 129]. We hope that the study in this paper will encourage the creation of similar infrastructures outside United States.

On the practical implementation of the predictive system Reviewers have raised the relevant question of the cost-benefit tradeoff of the implementation of the predictive approach in the clinical practice. In their words, a relevant question is whether it is worth to intervene almost twenty patients in order to reduce the likelihood of one readmission (according to PPV values). From the technical point of view, the system would be implemented as an assistive device, so that the intervention decision is always in the clinician hands. Clinicians have expressed the desire to have some kind of objective reference to help them focus on the risky cases. On the other hand, implementation of a predictive system as described in the paper would give a dichotomy decision. However, there is a gradation of risk underlying this decision, which may be modeled by the *a posteriori* probability estimations computed by the predictive models. In fact, the dychotomic decision is the result of the application of an arbitrary threshold (often 0.5) to these *a posteriori* probability estimations. Future work should be addressing the task of providing a risk gradation to the clinicians, easing the task of targeting really critical cases that need more specific intervention, such as giving detailed training to the parents for child treatment at home, or delaying the child discharge from the hospital. From the administrative point of view, the hospital is increasing the decision assistant tools provided to the clinicians. For instance, there is a tool providing triage recommendations. Therefore, they are definitively in favor of the implementation of the kind of tools described in the paper. Furthermore, the continuous inflow of information and the addition of new variables will allow the improved tuning of the tool. Finally, from the human point of view, any parent will be in favor of the implementation of such tools if they improve somehow the health care quality of their children.

2.5. Conclusions

Following the track of political decisions in United States regarding cost effective quality healthcare, hospital readmissions have become a concern worldwide. There have been many quantitative analysis, mostly for adult patients, including predictive approaches based on machine learning. However, pediatric hospital readmissions have received little attention until recently. One of the lessons learned is that there is much variability between locations so that it is preferable to develop local predictive models than trying to apply models developed upon

foreign country data. Another lesson learned is that it is desirable to have research oriented nationwide data collection and distribution resources that may allow to carry out precise and extensive quantitative analysis.

In this paper we report the results of an all cause predictive modeling study carried out over the anonymized dataset collected over six years of operation in a public pediatric hospital in Santiago, Chile. The amount of data gathered is large for a single site study (56,558 discharges and 2,106 readmissions), but it would be desirable to enlarge it with the contribution of other institutions in Chile. We have applied four predictive methods upon the administrative data used for patient cost estimation. The results are good, achieving a top predictive performance AUC=0.65 that is comparable to other predictive studies on adult patients data. However, this is the result of a dychotomic decision, which puts together mild risk cases with high risk cases. Future work should be addressed to give a more precise quantification of the risk of readmission, allowing to focus more efforts on the riskiest cases.

To our knowledge this is the first such study in Chile, and among the first ones worldwide, devoted to pediatric readmissions. In the future, it will be desirable to have access to a nationwide data repository, in order to be able to derive general models upon which specific policies for optimal cost management maintaining while improving the service quality could be formulated. The inclusion of other data modalities, such as medication, international disease code, laboratory and clinical data would help to extend this study into the so-called phenomics realm, which aims to exploit the big data contained in the EHRs in order to achieve personalized medical recommendations and follow up. Such large data collections would allow also the application of recent breakthrough technologies such as deep learning.

Chapter 3

Setting up standards: A methodological proposal for pediatric Triage machine learning model construction based on clinical outcomes

Wolff P, Ríos S & Graña M [1]

Abstract: Triage is a critical process in hospital emergency departments (ED). Specifically, we consider how to achieve fast and accurate patient Triage in the ED of a pediatric hospital. The goal of this paper is to establish methodological best practices for the application of machine learning (ML) to Triage in pediatric ED, providing a comprehensive comparison of the performance of ML techniques over a large dataset. Our work is among the first attempts in this direction. Following very recent works in the literature, we use the clinical outcome of a case as its label for supervised ML model training, instead of the more uncertain labels provided by experts. The experimental dataset contains the records along 3 years of operation of the hospital ED. It consists of 189,718 patients visits to the hospital. The clinical outcome of 9,271 cases (4.98 %) wa hospital admission, therefore our dataset is highly class imbalanced. Our reported performance comparison results focus on four ML models: Deep Learning (DL), Random Forest (RF), Naive Bayes (NB) and Support Vector Machines (SVM). Data preprocessing includes class imbalance correction, and case re-labeling. We use different well known metrics to evaluate performance of ML models in three different experimental settings: (a) classification of each case into the standard five Triage urgency levels, (b) discrimination of high versus low case severity according to its clinical outcome, and (c)

[1]The following is an unabridged version of the paper published in "Expert Systems with Applications". Please cite this paper as follows: Patricio Wolff, Sebastián A. Ríos, Manuel Graña, "Setting up standards: A methodological proposal for pediatric Triage machine learning model construction based on clinical outcomes", Expert Systems with Applications, Volume 138, 2019, 112788, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2019.07.005.The original publication is available at: http://www.sciencedirect.com/science/article/pii/S0957417419304841

comparison of the number of patients assigned to each standard Triage urgency level against the Triage rule based expert system currently in use at the hospital. RF achieved greater AUC, accuracy, PPV and specificity than the other models in the dychotomic classification experiments. On the implementation side, our study shows that ML predictive models trained according to clinical outcomes, provide better Triage performance than the current rule based expert system in operation at the hospital.

3.1. Introduction

Triage is the assignment of an urgency degree to the wounds or illnesses of a specific patient to decide the order of treatment of a large number of patients. Triage is the first and most critical step when a child enters the Emergency Department (ED). It is necessary to discriminate the child requiring the most immediate care from those that can wait for some time, in order to achieve the timely delivery of emergency health care [111, 131], avoiding under-triage, i.e. assigning a child requiring urgent treatment to a less urgent class, and over-triage, i.e. overestimating the acuity of the patient [81].

Machine learning (ML) models have been proposed to automate the Triage process in order to achieve a simplified quick examination of patients ensuring their timely treatment according to the degree of severity of their condition. It has been estimated that 40 % of patients showing up at EDs have non-urgent problems [25]. This leads to overcrowded waiting rooms and long waiting times. As a consequence, patients with severe urgent care needs are at risk of not being treated on time [141].

In general, automated Triage systems are built up from the consensus opinion of clinical experts [76], which provide the design of urgency level decision trees supporting clinical risk assessment and predictions of resource usage. Triage systems should be simple to apply, accurate, rapid, reproducible, and discriminant to avoid potentially dangerous under-Triage, and costly over-Triage. ML tools have been shown to improve over the results of expert based methods [53, 86, 104].

A review of ML algorithms used to build Triage systems for ED treating adults is summarized in Table 3.1. We found a strong lack of consensus in the research methodology applied in these articles. Critical issues, such as how the datasets were collected and curated, are not explained with sufficient clarity. Also, some papers neither clarify how algorithms were selected or if they have compared several approaches. Papers show no consensus on the metrics used to report the performance results. Papers present an arbitrary selection of precision, F-measure, sensitivity, true positive rate, accuracy, or RMSE as performance results. A general model construction flaw of the papers in Table 3.1 is that none of them took into consideration that datasets have big class imbalance, thus they did not apply any correction strategy for improved model building. Finally, authors did not mention that Triage is a multi-class problem, which is an extremely important aspect when evaluating the final results.

Most studies dealing with the application of ML to the construction of Triage systems try to predict the actual Triage decisions given by the ED staff. However, these Triage labels may

not be the most accurate ones. The clinical outcome of the patient (deceased, transferred, hospitalized) is a more reliable reference to guide the learning process. In a very recent and influential publication, [73] built ML models predicting the patient clinical outcome. Our work follows the same approach. However, we propose some methodological improvements over the work reported in [73]. Specifically, we apply class imbalance correction procedures, and we report results from more comprehensive performance evaluation methods. In this paper we apply methodological best practices for ML model building and validation in order to evaluate the quality of the model in terms of the distribution of the patient outcomes. Additionally, we report a performance analysis of the predictive capacity of clinical outcome in high severity classes.

In this study, we work with the staff of the pediatric tertiary care center Dr. Exequiel González Cortés Hospital (EGCH), serving a population close to 350,000 people in total. At this institution, the Triage involves rapid recognition of seriously ill or injured children, assigning a severity rating level, and anticipating appropriate emergency care and referral. Currently there is an electronic Triage tool in use at EGCH, providing severity rating levels used to prioritize patients for care, from level 1 (most severe) to level 5 (least severe).

This paper reports three computational experiments that confirm the practical value of ML models for pediatric ED Triage:

1. First, we relabel dataset cases with the the standard Triage five levels according to their clinical outcome. We carry out the validation experiments to evaluate ML model performance over the relabeled dataset.
2. Second, we consider specifically the prediction of the clinical outcome as a dychotomic classification on the following discrimination problem instances: (a) death versus non-death, and (b) hospitalization versus non-hospitalization.
3. Third, we propose the classification of dataset cases into the five standard Triage levels. The resulting model is evaluated according to the clinical outcome. Finally, we compare the expert knowledge based Triage system currently used at the hospital against the ML models in terms of the distribution of clinical outcomes over the predicted Triage levels.

The contents of the paper is as follows: Section 2 provides a discussion of the related work. Section 3 discusses the data preparation and experimental setup. Section 4 discusses our actual model evaluation framework for best model selection. Section 5 gives the experimental results. Section 6 provides a discussion of results. Finally, Section 7 gives our conclusions and lines for future work.

3.2. Related Work

As we have mentioned before, Triage is a strongly class imbalanced problem. For example, in EGCH, less than 1 % admissions are assigned an emergency level 1. Indeed, training ML algorithms over small datasets having such big class imbalance poses two big challenges: (a) the correction of the class imbalance, and (b) the correct selection of the performance

metric. In the majority of the studies shown in Table 3.1, dataset sizes are rather small. Five approaches use less than 3000 records, which arguably is not an enough representative sample to reach generalized conclusions. In our study, we have collected a huge dataset over several years of ED operation assuring that we have enough data to train, test and evaluate our results (in fact, ours is - to the best of our knowledge - the biggest dataset in the literature, cf. Table 3.1).

It is noteworthy that only [104] explains in detail the different data selection criteria applied in the study. All other works referred in Table 3.1 leave this important matter unexplained. It is crucial to understand these criteria; specially when manual selection is performed. Data selection may introduce bias in the model performance evaluation, which, of course, will affect the generalization of the model and the quality of the results.

Except for [162], the revised publications do not specify the construction of the evaluation sets (randomized, random, stratified, proportional, etc). Moreover, the research presented by [53], [104] and [158] do not present a comparison of different models and parameter configurations.

Regarding performance metrics, some works [172, 158, 35] use the MAPE or the overall accuracy as a measure of performance of time series models, which is methodologically incorrect for a strongly class imbalanced classification problem. In addition, it is striking that multi-class performance metrics are not reported, when all the revised literature recognizes that the Triage problem is a multi-class classification problem (except [105]).

To the best of our knowledge, there is only one very recent publication on the application of ML for Triage in a pediatric ED [73] despite it is recognised in recent scientific literature that this is a problem where ML may play a big role [49]. Many classification performance metrics have been used for the evaluation of the ML Triage prediction models [73, 104, 190]. Some of them are incorporated as part of our analysis. We note that [104] failed to achieve good recall (sensitivity) of their model for high-severity levels, which is a requirement for a useful screening method.

Other publications on the subject of Triage, such as [29] and [11], were not included in this comparison, because they treat a different problem in terms of the classes identified and the purposes of the research. In the investigation that [29] carried out over pediatric patients, the objective of their research is to find a detection mechanism for low complexity patients, called “Fast track". Thus it is not a study about five-class Triage. In addition, regarding the amount of data used, the number of cases considered is small: 2223 in [11] and 1205 in [29].

The evaluation of a Triage systems involves reliability and validity assessments [118]. Reliability refers to the ratio of intra-observer variability versus inter-observer variability. Validity refers to the degree of success of Triage prediction of a “true” urgency. It is measured by the sensitivity and specificity of the model [124].

A fundamental problem found conducting validation studies of Triage tools is the lack of consensus on a gold standard reference to measure the performance of the Triage system [70, 74, 141, 64]. One way to evaluate a Triage system is to compare its output with a standard cost value (which includes the cost of the use of resources) defined by the experts

Tabla 3.1: ML-based Triage model Benchmark

[Ref]	Dataset	ML Models	Validation Method	Observation
[190]	402	NB, DT	Expert	Five Level Adult Triage
[162]	57,573	ANN, SVM, GA	Outcome	Adult Trauma Triage
[105]	2,000	ANN	Expert	Four Level Adult Triage
[35]	947	ANN, NBN, LR	ESI	N/A
[172]	3,000	ANN,SVR	Expert	Four Level Adult Triage
[53]	25,198	LR	Outcome	Five Level Adult Triage
[104]	172,726	RF	Outcome	Five Level Adult Triage
[158]	537	Fuzzy	Expert	Five Level Adult Triage
[73]	52,037	DL, LR, RF, DT	Outcome	Five Level pediatric Triage
This Article	189,718	DL, NB, RF, SVM	Outcome	Five Level pediatric Triage

(usually nurses or medics). Some authors have used standard risk indexing values, such as the Emergency Severity Index (ESI) [35] model, as the "real" label. However, there is no consensus about the universal applicability of ESI. Besides, clinical outcomes have been widely used for validation of adult Triage models [162, 53, 86, 104] and they have been used also in pediatrics [2, 58, 77, 83].

In this paper, we train multiple ML models on a large real dataset collected from the ED of the EGCH, in order to determine if ML techniques improve over the performance of the currently implemented rule based expert system. In addition, we report multiple performance evaluation metrics, as well as different mechanisms that allow us to improve ML algorithms performance, such as class imbalance correction techniques. Finally, we will compare the best ML model against the current Triage expert system in operation.

3.3. Data Preparation and Experimental Setup

The experimental setup shown on figure 3.1 is the framework for the evaluation of ML based Triage prediction systems. We have evaluated more than eight different ML modeling approaches ranging from regression (such as logit regressions) to probabilistic models (such as Naive Bayes). However, in this article we report results only for the best four approaches: Support Vector Machines (SVM), Deep Learning (DL), Random Forest (RF), and Naive Bayes (NB).

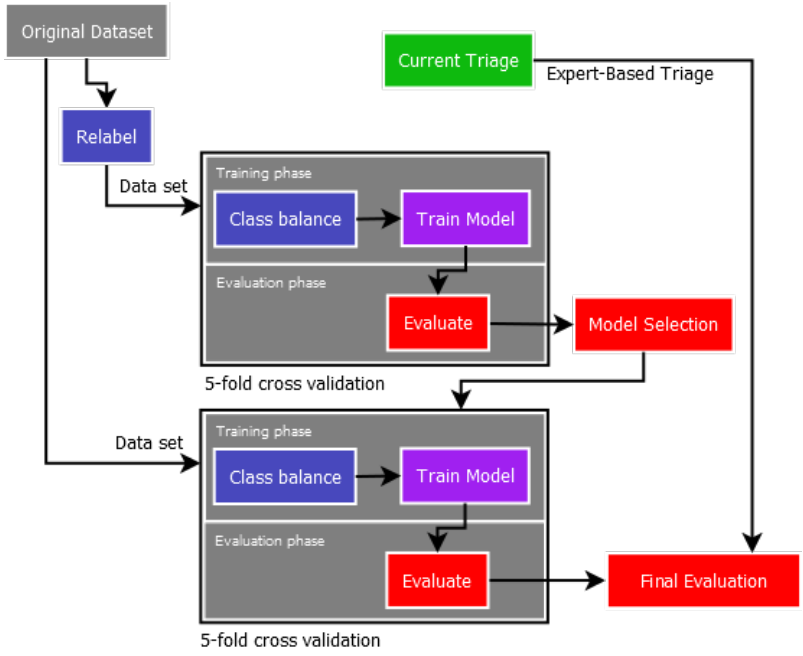

Figura 3.1: Study design schema

In order to perform our experiments, we carried out the tasks of manual data curation, removing cases with inconsistent or missing values, data preprocessing including case relabeling, and class imbalance correction techniques. Then we train the selected models, and, finally, we compare their performance results using several well known metrics. A schema of this setup can be seen in figure 3.1. The upper experimental module (gray colored) corresponds to a multi-class classification problem into the five standard Triage levels. The input to this module is the dataset after being relabeled according to clinical outcomes. The lower experimental module corresponds to the dychotomic classification problem instances (high versus low severity, death versus non-death, hospitalization versus non-hospitalization) where we evaluate the models selected from the results achieved in the upper module. In both experimental modules we use 5-fold Cross-validation (with random stratification according to clinical outcome). The Final Evaluation in the right-bottom corner of the figure corresponds to the comparison between ML approaches and the automated Triage currently working at the EGCH reported in section 3.5.3. We apply a hold-out scheme for this comparison (80 %, 151,774 ED visits for training and 20 %, 37,944 ED visits for testing). We use hold-out instead of cross-validation because the EGCH working Triage system can not be trained on data, so there is no possibility to carry out a k-fold cross-validation on it.

3.3.1. Dataset characteristics

Our study is a single center retrospective cohort study. Records for all ED visits during the study period were retrieved from the logs of the rule based expert system e-Triage currently in operation at EGCH. Anonymized data of pediatric patients ($<$ 18 years) who were admitted for care between August 1, 2014 and October 31, 2017 were included for analysis. Traumatology and planned surgery visits were not included. All erroneous or incomplete in-

formation records were deleted. Finally, our cleaned dataset (see table 3.2) contained 189,718 ED visits, only 9,271 of them became a hospital admission (4,89 %).

Tabla 3.2: Dataset information

General	
ED visits - number	189,718
Age - median ± SD	2.9 ± 4.03
Female	89,319 [47.07 %]
Vital signs	
Temperature °C - median ± SD	37 ± 0.87
Heart rate - median ± SD	123 ± 27.85
Respiratory Rate - median ± SD	32 ± 4.88
Oxigen saturation % - median ± SD	98 ± 1.64
VAS - median ± SD	2 ± 1.55
LOC (not alert)	438 [0.23 %]
Outcomes	
Death	24 [0.01 %]
Admission Sev. 1	4,304 [2.26 %]
Admission Sev. 2	1,123 [0.59 %]
Admission Sev. 3	3,844 [2.03 %]
Procedures	10,024 [5.28 %]
Surgical procedure	1,346 [0.71 %]

To identify the visits that became a hospital admission, we use the international standarized cost management system (DRG). Information on deaths and procedures performed on patients were also obtained from the administration cost management system, because the hospital did not have an Electronic Health Record (EHR) system at the time of collecting the data.

3.3.2. Current expert knowledge based Triage system

The main motivation for the local construction and improvement of an automated Triage system is the fact that the parameters and variables that determine the emergency level are highly depend on the local characteristics of the population on which it is applied. Our hospital has different characteristics from the hospitals where conventional commercial structured Triage systems were developed (hospitals located in first world countries), namely: human and financial resources, data available on the patients, local policies, educational level of the population, and epidemiological profile, and many others.

The EGCH's current five-level Triage system is implemented as a decision tree (illustrated in figure 3.2) that allows to determine the emergency level based on the chief complaint, vital signs, and the answers to some additional questions. This expert knowledge based electronic Triage system has been implemented in the hospital since August 2015. It is applied to 100 % of the ED visits since then. The Triage emergency level assigned to each patient is determined

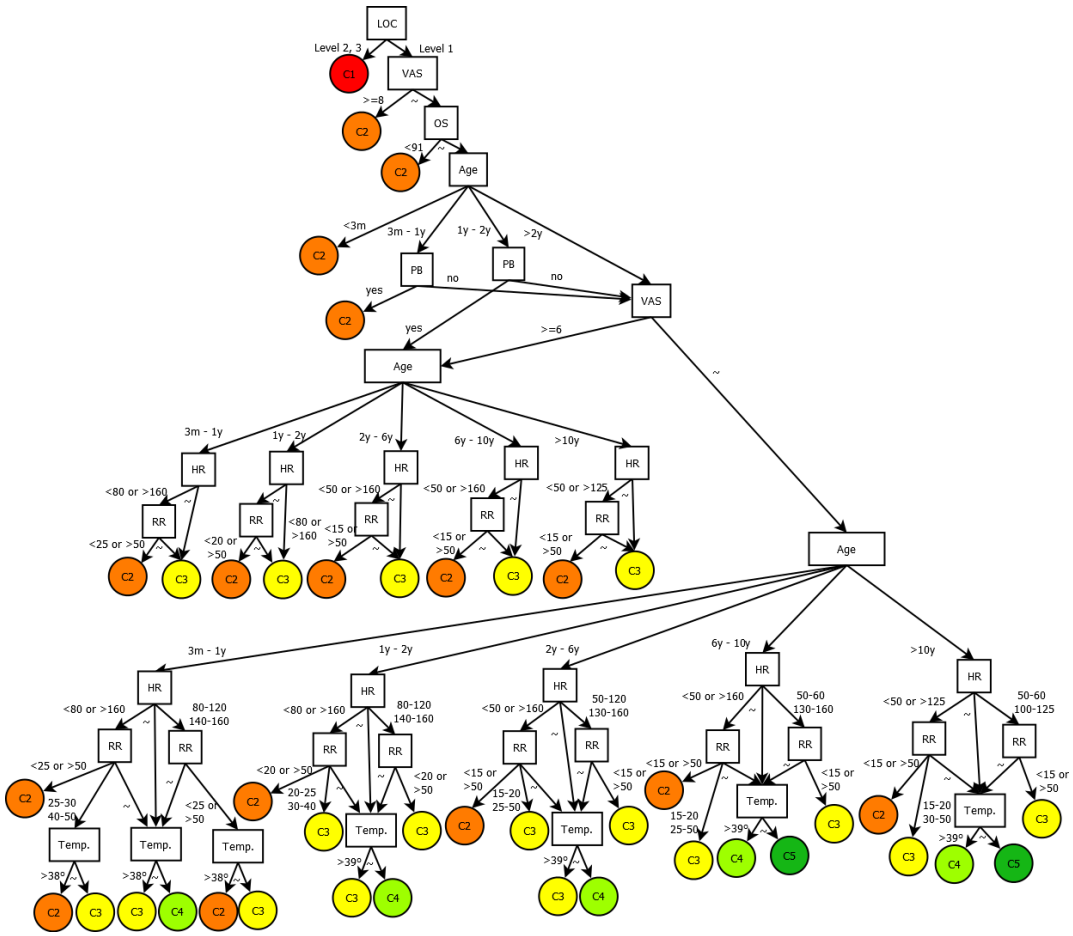

Figura 3.2: EGCH's current rule based e-Triage expert system.

by a structure of qualitative decisions, and quantitative thresholds for the values of hearth rate (HR), level of consciousness (LOC), respiratory rate (RR), body temperature (Temp.), pain Visual Analogue Scale (VAS), oxygen saturation (OS), and age of the patient. Local experts whose criteria was used to built the system were selected among the members of the ED on the basis of their experience and expertise.

At the time of EGCH's current ED Triage system development, the committee of experts was formed by ten pediatricians and three nurses, who defined the basic structure of the decision tree and the value of the decision thresholds at each node, according to the international literature and their own expert judgment. In the initial stages of the implementation, adjustments were made to these threshold values to improve the results of the classification in terms of accuracy and sensitivity.

The Triage implementation allowed to improve the care process in the hospital, to reduce the waiting times of the patients, to incorporate risk assessment in the clinical practice, among other advantages. Although the results obtained are tremendously positive, it is essential to advance improving the performance of Triage level assignment as a function of the clinical outcome. The performance of the Triage will be measured by the clinical outcome prediction, and the number of ED visits assigned by emergency level in the results section.

3.3.3. Relabeling according to the clinical outcomes

We need to realize three case relabeling processes. Firstly, according to [2], we reformulate the classification into a two class problem, where the high and low severity classes are deduced from the actual case Triage labels. Secondly, according to [73], we guide the learning process by the clinical outcomes instead of the actual Triage emergency level given by the expert system. Hence we need to relabel each case accordingly.

Thirdly, we carry out an independent relabeling into five categories for training of ML algorithms to produce an outcome-guided classification into the conventional Triage five categories. The relabeling rules are as follows:

1. $C1$ label was assigned to patients with death or hospitalization severity level 3 (IR-DRG severity index);
2. $C2$ label was assigned patients Hospitalized with severity level 2 and 1 (IR-DRG severity index);
3. Non hospitalized patients with emergency procedure were assigned to $C3$ label;
4. $C4$ label was assigned to non-hospitalized patients, without ED procedure, younger than eight years and without fever;
5. finally, the rest of the patients were labelled as $C5$.

These new labels let us train our machine learning models to classify ED visits based on objective information, i.e. final clinical outcome, while allowing us to assign patients to five classes, as required by hospital rules.

Table 3.3 shows the distribution of labels after a partition of our dataset into train and test datasets, were it can be appreciated that class $C3$ is under-represented, and that class $C5$ may also be underrepresented relative to class $C4$ which is the majority class. This distribution of classes is due to the fact that the hospitalization rate, the mortality rate and the amount of outpatient procedures in the hospital, in the study period is smaller than in other studies from the literature. For example, for adult patients the hospital admission rate (class $C2$ above) reported by [53], [86], and [61] is 14 %, 12 %, and 20.3 %, respectively, while in pediatric populations this rate is of the order of 6 % [58], much higher than in our re-labeled dataset.

Tabla 3.3: Cases an percentage per class in train and test dataset

Cat.	Train cases	[%]	Test cases	[%]
C1	3,518	2.32	865	2.28
C2	3,899	2.57	1,067	2.81
C3	515	0.34	159	0.42
C4	115,639	76.19	29,092	76.69
C5	28,212	18.59	6,752	17.80

3.3.4. Dealing with class imbalance

A major difficulty faced designing ML based Triage systems is the high class imbalance ratio of the datasets. In general, this is due to the fact that high-acuity events are infrequent. Nowadays, there is a number of techniques that allow to tackle the class-imbalance problem. A taxonomy of methods to achieve robust ML model building with imbalanced data is presented in [6]. Specifically we have used SMOTE [34] to augment the minority classes. In the SMOTE algorithm, new samples of the minority class are generated by random convex interpolation between k randomly picked samples from the minority class. Random convex interpolation means that the new sample is computed as a polynomial of degree one of the reference samples whose coefficients are in the interval [0,1]. In multi-class datasets, each minority class is treated independently. Bootstrapping techniques were also tested to achieve class balance, but in this particular study, they showed lower performance than SMOTE in all selected classification models. The use of SMOTE allowed us to have a balanced train dataset of $578,195$ patient cases. Randomized subsampling of this balanced dataset retaining $20\,\%$ of it($115,635$ patient cases) was used for SVM training, in order to have affordable processing time.

3.3.5. Machine learning models under evaluation

The selection of ML models will depend on: those that present better results, the characteristics of the problem, and the state of the art. Four state of the art multi-class classification algorithms were preselected. Specifically we have used R implementations of the following supervised learning approaches.

Support Vector Machine (SVM)

One of the most important supervised learning methods is Support Vector Machine (SVM) [169]. SVM can be used to address regression, binary and multi-class classification problems. In the case of classifiers based on SVM, we look for a hyperplane that divides the feature space and maximizes the distance between groups of feature vectors belonging to a class, with respect to the feature vectors belonging to another class. The two classes may not be linearly separable. In such cases, feature vectors are previously projected to a space of superior dimensionality using the kernel trick [150]. The parameters that define the solution hyperplane are obtained solving a quadratic programming problem. SVMs require several training cases, but they depend on few parameters. The literature shows that SVMs are not sensitive to the size of training datasets [169], in other words, they are robust to the dimensionality curse. In our case we use multi-class version of C-SVC with two different kernels: linear kernel and Radial Basis Function (RBF) kernel, cf. table 3.4 for parameter details.

Deep Learning (DL)

Deep Learning (DL) approaches are considered state of the art, based on the excellent results obtained in different pattern recognition tasks[71]. To achieve good results, DL require a large volume of data for their training. The progress of these techniques is also enhanced by the creation of open source libraries and software freely available to companies and researchers. We use H2O version 3.8.2.6. [43] which is an available library that supports Deep Learning using a multi-layer feedforward artificial neural network trained with back-propagation of the error using stochastic gradient descent approach. We have used both Rectifier and Tanh as alternative activation function of the neurons in the hidden layers. We explore three different network topologies: (1) with two hidden layers constituted of 50 neurons and 25 neurons, respectively; (2) with two same size hidden layers of 50 neurons; and (3) with three hidden layers constituted of 50 neurons, 50 neurons and 25 neurons respectively. All the architectures were trained in 10 epochs (as showed in table 3.4). Adaptive learning rate algorithm (ADADELTA) [186] was used to combine the benefits of learning rate annealing and momentum training to avoid slow convergence. Other recommended parameter values used were $epsilon = 10\text{e} - 8$ and $rho = 0{,}99$.

Naive Bayes (NB)

Naïve Bayes methods (NB) is a simple yet very effective method in the machine learning toolbox. Its main idea is based on the naïve assumption that the individual features are statistically independent, so that we may approximate the joint probability of a D-dimensional feature vector as a product of D probabilities of the 1-dimensional features. In our case we use two different Naïve Bayes methods, based on parametric Gaussian probability density estimation, and non-parametric kernel density estimators, to model the likelihood density function of the features.

In kernel density estimation we use greedy search to set the kernel bandwidth. We test 10, 100 and 1000 kernels, with minimal bandwidth of 0,1 in all cases. We also test full estimation mode with bandwidth selection heuristic and fix (as shown in table 3.4). We use a Laplace correction to prevent high influence of zero probabilities.

Random Forest (RF)

Random Forest (RF) [24] is an ensemble of random trees trained on bootstrapped sub-sets of the dataset, each of which is built on a bootstrap sample of the training dataset using a subset of randomly selected variables.

Each node of a tree represents a splitting rule for one specific attribute. Gain ratio, Gini index, and Information gain were chosen as a criterion to select attributes for splitting. Maximal depth used was 20 and we test 10, 100 and 1000 trees (as shown in table 3.4).

Tabla 3.4: Machine Learning studied models and their parameter setup

ML Alg.	**Descrip.**	**Descrip. Cont**	**Name**
Deep Learning	Rectifier Act. Func.	Neurons 50-25	DL1
		Neurons 50-50	DL2
		Neurons 50-50-25	DL3
	Tanh Act. Func.	Neurons 50-25	DL4
		Neurons 50-50	DL5
		Neurons 50-50-25	DL6
Naive Bayes	Gaussian d-estimator		NB1
	Greedy d-estimator	10 kernels	NB2
		100 kernels	NB3
		1000 kernels	NB4
	full d-estimator	Heuristic	NB5
		Fix	NB6
Random Forest	Gain-ratio split criterion	10 trees	RF1
		100 trees	RF2
		1000 trees	RF3
	Gini-index split Criterion	10 trees	RF4
		100 trees	RF5
		1000 trees	RF6
	Info-gain split Criterion	10 trees	RF7
		100 trees	RF8
		1000 trees	RF9
Support Vector Machine	RBF kernel	$C = 0{,}1$ & $\gamma = 0$	SVM1
		$C = 0{,}1$ & $\gamma = 0{,}1$	SVM2
		$C = 10$ & $\gamma = 0$	SVM3
		$C = 10$ & $\gamma = 0{,}1$	SVM4
	Linear kernel	$C = 0{,}1$	SVM5
		$C = 10$	SVM6

3.4. Evaluation framework and best model selection

Performance evaluation metrics in multi-class classification problems are fundamental in assessing the quality of learning methods and compare performance of different models. As we shown in Table 3.1, not many studies take this important issue into consideration. Using the wrong performance metric might lead to erroneous results, which is especially critical in hospital context. Thus we aim to recommend several ML performance metrics that should be used for reporting results in future research studies regarding Triage prediction models. For example, it is common in highly imbalanced class problems to report very high accuracy over 90 % or even close to 100 %, while recall of minority class is very low (e.g. 30 %); which in a hospital context means that, many people (i.e. 70 %) with high emergency risk is given a low risk Triage level. Thus, reporting only accuracy or precision (as most publications in Table 3.1 do) is not enough to asses model quality, in fact, some articles report only the success predicting the majority class, which is uninteresting and misleading.

3.4.1. Multi-class ML performance metrics

In figure 3.1 we represent in red boxes two evaluation phases labelled "Evaluate". In the following we explain the performance metrics used in these phases.

Many different measures have been defined in the literature for imbalance multi-class performance. However, there is not gold-standard performance metrics that can be applied to all types of multi-class problems [94]. In this subsection we select and present several well known performance metrics used to evaluate our models. We borrow the notation from [59].

- Overall Accuracy (Acc.) is the most common and simplest measure to evaluate the degree of right predictions of a classification model.

$$Acc = \frac{\sum_{\mathrm{i}=1}^{m} \sum_{j=1}^{c} f(\mathrm{i}, j) C(\mathrm{i}, j)}{m}, \tag{3.1}$$

  were $f(\mathrm{i}, j)$ represents the actual probability that case i belongs to class j, m denotes the number of examples, c the number of classes, and $C(\mathrm{i}, j)$ is 1 if j is the predicted class for i.

- The Cohen's Kappa ($Kappa$) [40] measure is frequently used as a performance measure in multi-class literature, and recently is being used in emergency Triage prediction research [45].

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}, \tag{3.2}$$

  were $P(A) = Acc$, and $P(E)$ is defined as follows:

$$P(E) = \frac{\sum_{k=1}^{c} (\sum_{j=1}^{c} \sum_{\mathrm{i}=1}^{m} f(\mathrm{i}, j) C(\mathrm{i}, j) \cdot \sum_{j=1}^{c} \sum_{\mathrm{i}=1}^{m} f(\mathrm{i}, j) C(\mathrm{i}, k))}{m^2} \tag{3.3}$$

- According to [184] it is recommended to use Triage-weighted Kappa (TWK) when we are trying to predict the Triage label as an ordinal variable [168].This metric corresponds

to a Cohen's weighted Kappa whose weights are determined by:

$$W_{ij} = (1 - ((i - j)^2/(c-1)^2))/(i/j) \tag{3.4}$$

for Over-triage and,

$$W_{ij} = (1 - ((i - j)^2/(c-1)^2))/(i/j)^2 \tag{3.5}$$

for Under-triage, where c represents the number of categories, i represents the category rated by a predictive model, and j represents the category rated by a reference standard (in our case the result of relabeling described above).

- Mean F-measure (MFM) has been widely used for multi-class problems. It is computed decomposing the multi-class problem into several binary classification problems. It is calculated as follows:

$$MFM = \frac{1}{c} \cdot \sum_{j=1}^{c} \frac{2 \cdot recall(j) \cdot prec.(j)}{recall(j) + prec.(j)}, \tag{3.6}$$

were $recall(j)$ and $prec.(j)$ are recall and precision per class j,

$$recall(j) = \frac{\sum_{i=1}^{m} f(i,j)C(i,j)}{m_j}, \tag{3.7}$$

and,

$$prec.(j) = \frac{\sum_{i=1}^{m} f(i,j)C(i,j)}{\sum_{i=1}^{m_j} C(i,j)}, \tag{3.8}$$

- We average the binary classification results to obtain Macro-recall ($recall_M$) and Macro-precision ($prec_M$). First we estimate the binary performance measures separately for each class, independently. Second, we compute the average of the obtained measures [157]:

$$recall_M = \frac{1}{c} \sum_{j=1}^{c} \frac{\sum_{i=1}^{m} f(i,j)C(i,j)}{m_j}, \tag{3.9}$$

and

$$prec_M = \frac{1}{c} \sum_{j=1}^{c} \frac{\sum_{i=1}^{m} f(i,j)C(i,j)}{\sum_{i=1}^{m_j} C(i,j)}, \tag{3.10}$$

The selection of the best ML model will be made by virtue of the performance measured by these multi-class performance metrics.

3.4.2. Final evaluation metrics

To conclude this section on performance metrics, we explain the "Final Evaluation" block in Figure 3.1.

Dychotomic classification instances

We consider two instances of dychotomic classification.

- On the one hand, following the approach of [2] patients were divided into high severity (Triage levels 1, 2) and low severity (Triage levels 3, 4, 5). We train the ML models to discriminate these two levels of severity, computing sensitivity and specificity. Given this two metrics we compute several well studied performance metrics of high severity level detection.
- On the other hand, [73] poses directly the dychotomic classification problems of death (or intensive care unit (ICU) use) vs. non death (and non ICU use), and hospitalization vs. non hospitalization.

The evaluation of ML models in these dychotomic classification problems is carried out separately in terms of their sensitivity, specificity, Area Under ROC (AUC), Accuracy, PPV and NPV. To avoid overfitting and to have more reliable results, 5-fold Cross-Validation was used.

In this dichotomized analysis it is also possible to include the ROC Curve and diagnostic accuracy measures, which are recommended as evaluation metrics in this type of problems by [184].

Patients assigned by Triage level

The number of patients assigned per class, also known as 'fingerprint', is commonly used (as table or graphic) to visualize the results of a Triage model. Is important to analyze this representation because it may provide a standardized descriptor of the ED and the hospital [45]. The idea is to compare the assignation per Triage level, to ensure a consistent distribution in terms of resources and local demand characteristics. Numerically, we measure similarity between two distributions computing the maximum distance between their cumulative distributions, known as Kolmogorov–Smirnov statistic test (KS-test) [92], were $F(x)$ is a cumulative distribution function.

$$D_n = \max_x |F_n(x) - F(x)| \tag{3.11}$$

3.5. Experimental results

In this section we present the results obtained following our proposed methodology. We discuss all performance metrics together to establish the best ML algorithm for pediatric Triage problem.

3.5.1. Initial model selection

We have tested twenty seven different parameter configurations of Deep Learning, Naive Bayes, Random Forest and Support Vector Machines, as shown in Table 3.4. Table 3.5 summarizes five multi-class performance metrics for all these ML algorithm configurations on the high versus low emergency level relabelled dataset according to [2].

The average training times per model was as follows: for DL, 13 min; for NB, 2 min; for RF, 30 min; and for SVM, 105 min. For comparison purposes the models were run on the same computer, powered by an Intel Core i7-8750H up to 4.1 GHz processor with 8 GB DDR3 of RAM.

Tabla 3.5: Models performance for the high versus low severity Triage levels. Blue highlights best model *per* performance metric

	ACC ± SD[%]	Kappa ± SD	Recall ± SD[%]	Prec ± SD[%]	TWK ± SD	MFM ± SD
DL1	94.2 ± 0.060	0.835 ± 0.002	44.10 ± 0.070	46.67 ± 0.300	0.66 ± 0.003	0.447 ± 0.011
DL2	94.2 ± 80.030	0.839 ± 0.001	45.65 ± 0.200	46.33 ± 0.190	0.66 ± 0.003	0.458 ± 0.013
DL3	94.2 ± 10.080	0.837 ± 0.002	45.10 ± 0.210	51.02 ± 1.670	0.67 ± 0.003	0.458 ± 0.020
DL4	94.1 ± 40.040	0.835 ± 0.002	45.02 ± 0.750	49.13 ± 0.740	0.66 ± 0.003	0.457 ± 0.011
DL5	94.3 ± 90.020	0.840 ± 0.001	45.73 ± 0.530	46.24 ± 0.200	0.67 ± 0.003	0.458 ± 0.015
DL6	94.1 ± 70.070	0.835 ± 0.002	43.42 ± 0.220	46.83 ± 0.530	0.66 ± 0.003	0.439 ± 0.008
NB1	80.8 ± 70.380	0.577 ± 0.005	47.41 ± 0.870	38.61 ± 0.110	0.44 ± 0.004	0.408 ± 0.023
NB2	78.8 ± 10.090	0.561 ± 0.001	46.74 ± 0.590	44.47 ± 0.470	0.40 ± 0.004	0.437 ± 0.022
NB3	91.0 ± 30.010	0.769 ± 0.000	49.68 ± 0.690	46.03 ± 0.180	0.60 ± 0.003	0.465 ± 0.017
NB4	77.8 ± 10.200	0.554 ± 0.003	51.04 ± 1.280	42.45 ± 0.200	0.39 ± 0.004	0.422 ± 0.012
NB5	34.4 ± 90.260	0.218 ± 0.001	44.65 ± 0.590	42.99 ± 0.190	0.27 ± 0.002	0.325 ± 0.015
NB6	77.8 ± 10.290	0.554 ± 0.005	51.04 ± 1.570	42.45 ± 0.260	0.39 ± 0.004	0.422 ± 0.012
RF1	92.1 ± 20.140	0.795 ± 0.004	49.78 ± 0.570	44.11 ± 1.980	0.62 ± 0.003	0.452 ± 0.011
RF2	92.4 ± 10.150	0.801 ± 0.004	49.69 ± 0.500	43.85 ± 1.230	0.62 ± 0.003	0.454 ± 0.009
RF3	92.4 ± 10.140	0.801 ± 0.003	49.69 ± 0.260	44.68 ± 2.440	0.62 ± 0.003	0.454 ± 0.009
RF4	81.0 ± 00.170	0.603 ± 0.003	52.03 ± 0.270	42.78 ± 0.150	0.41 ± 0.004	0.437 ± 0.014
RF5	81.8 ± 30.440	0.615 ± 0.005	52.19 ± 0.970	42.85 ± 0.250	0.42 ± 0.004	0.438 ± 0.013
RF6	82.4 ± 50.150	0.624 ± 0.002	52.22 ± 0.380	42.96 ± 0.150	0.43 ± 0.004	0.440 ± 0.015
RF7	82.8 ± 40.320	0.630 ± 0.006	51.89 ± 0.770	42.95 ± 0.290	0.44 ± 0.004	0.440 ± 0.012
RF8	82.2 ± 40.210	0.621 ± 0.003	52.29 ± 0.360	42.76 ± 0.070	0.43 ± 0.004	0.438 ± 0.014
RF9	82.2 ± 50.290	0.621 ± 0.005	52.23 ± 0.410	42.77 ± 0.230	0.43 ± 0.004	0.433 ± 0.013
SVM1	71.4 ± 30.130	0.473 ± 0.002	46.19 ± 0.340	38.65 ± 0.520	0.28 ± 0.004	0.367 ± 0.010
SVM2	78.1 ± 40.070	0.559 ± 0.001	49.8 ± 0.180	39.21 ± 0.130	0.38 ± 0.004	0.394 ± 0.014
SVM3	81.2 ± 40.100	0.335 ± 0.002	26.89 ± 0.140	37.08 ± 0.270	0.25 ± 0.005	0.285 ± 0.001
SVM4	50.4 ± 30.170	0.258 ± 0.001	38.61 ± 0.700	40.60 ± 0.070	0.16 ± 0.003	0.285 ± 0.008
SVM5	79.7 ± 70.160	0.580 ± 0.002	51.99 ± 1.440	41.42 ± 0.270	0.42 ± 0.004	0.311 ± 0.008
SVM6	79.7 ± 90.120	0.581 ± 0.003	52.00 ± 0.960	41.45 ± 0.110	0.42 ± 0.004	0.427 ± 0.009

According to the results presented in Table 3.5, we selected three model configurations for further experimentation. These models are:

- DL with Rectifier activation function and 50-50-25 hidden layer topology ($DL3$), selected on the basis of its higher macro precision performance;
- DL with Tanh activation function and 50-50 neurons hidden layer topology ($DL5$), selected on the basis of its higher Accuracy and kappa performance;

- NB with 100 kernels of 0.1 bandwidth ($NB3$), selected on the basis of its mean F-measure performance;
- and finally, RF with 100 trees using information gain as attributes split criteria($RF8$), on the basis of its superior macro recall performance;

SVM were tested with different configurations of kernel functions, and C and gamma parameter settings. However, SVM results were consistently worse than the three selected models described in this section.

3.5.2. Dychotomic classification into clinical outcomes

In this section we report the performance of our selected ML models on clinical outcome prediction according to [73]. Table 3.6 summarizes the achieved results on the the prediction of the two main outcomes: hospitalization and death. Figure 3.3 shows the ROC curves for the hospitalization outcome. The bold lines correspond to the average ROC.

Tabla 3.6: Diagnostic performance measures for the two high severity clinical outcomes.

Outcome (Hospitalization)

	AUC (SD)	Acc(SD)[%]	Sens(SD)[%]	Spec(SD)[%]	PPV(SD)[%]	NPV(SD)[%]
DL3	0.79(0.005)	61.5(3.26)	81.0(3.12)	56.9(4.73)	30.5(1.53)	92.9(0.54)
DL5	0.77(0.014)	56.4(3.89)	80.0(2.23)	54.6(5.12)	29.2(2.21)	92.2(0.61)
NB3	0.78(0.010)	75.5(0.36)	64.7(1.79)	78.1(0.13)	40.5(0.69)	90.5(0.44)
RF8	0.80(0.006)	79.4(0.40)	61.9(1.51)	83.4(0.56)	46.4(0.76)	90.4(0.32)

Outcome (Death)

	AUC (SD)	Acc(SD)[%]	Sens(SD)[%]	Spec(SD)[%]	PPV(SD)[%]	NPV(SD)[%]
DL3	0.79(0.084)	65.8(8.48)	71.4(12.60)	65.8(8.52)	0.5(0.11)	99.9(0.04)
DL5	0.78(0.072)	80.7(4.43)	65.9(10.12)	80.8(4.44)	0.8(0.29)	99.9(0.03)
NB3	0.77(0.154)	61.3(4.72)	72.7(21.72)	61.3(4.70)	0.4(0.16)	99.9(0.09)
RF8	0.86(0.062)	85.3(13.44)	66.7(12.08)	85.4(13.49)	2.3(1.63)	99.9(0.03)

3.5.3. Comparison against current e-Triage system

Our last computational results refer to the comparison of the selected ML models against the e-Triage system currently in operation at the EGCH. This comparison is made in terms of the distribution of cases per Triage emergency level. We use the third dataset relabeling presented in section 3.3.3 for the ML model training. We compute the KS test against the ideal case distribution discussed in the literature [45]. Table 3.7 gives the achieved results. It is important to analyze this characteristic because it provides a "fingerprint" or standardized descriptor of the ED and the hospital [45] that is easily accepted by the medical staff.

Finally, the results of each approach are presented graphically in figure 3.4. We present the percentage of patients *per* class admitted for hospitalization (blue plots), and dead (red

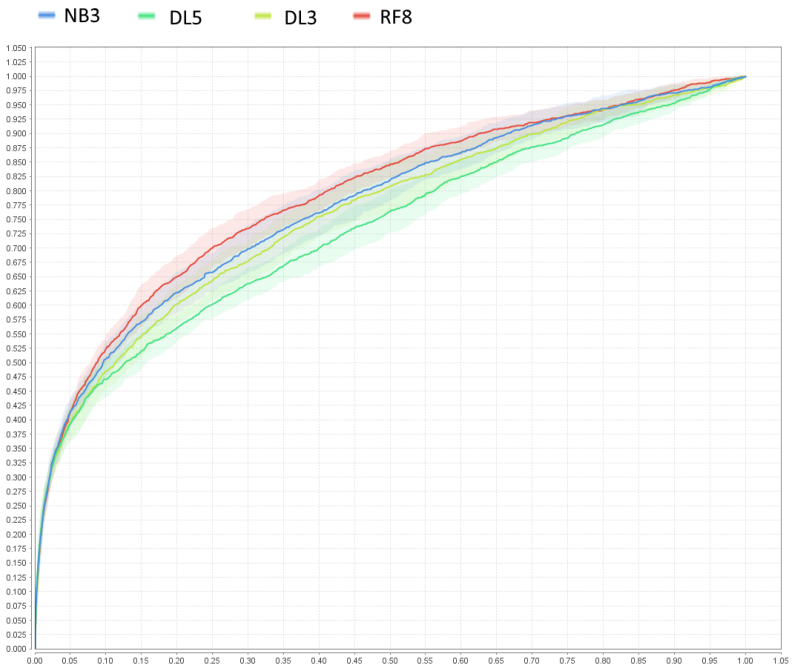

Figura 3.3: ROC curve of Hospitalization (positive class) versus non-Hospitalization classification.

Tabla 3.7: Cases assigned per class in preselected models. Expert means the current Triage system at the EGCH.

	C1	C2	C3	C4	C5	KS-test
DL3	513	71	0	30,272	7,079	4347
DL5	622	2	0	30,148	7,163	4307
NB3	1,779	426	417	28,630	6,683	3489
RF8	3,984	1,902	574	24,452	7,023	3868
Expert	116	2,911	1,904	22,832	10,172	-

plots), under the class assignment done by the ML models and the current expert-based e-Triage at EGCH.

3.6. Discussion

We have tested several ML algorithms with various parameter configurations for Triage category prediction, finding out that there is not a single model that provides the best values on all the considered performance metrics. Results on a relabeled dataset are summarized in table 3.5. Naive Bayes (NB) achieves the best results in terms of Mean F-measure (0,465), while Random Forest ($RF8$) model reports the highest $recall_M$ (0,523). The Deep Learning models ($DL3$ and $DL5$) showed excellent performance in accuracy (0,943), Kappa (0,840), Triage-weighted Kappa (0,67) and $prec_M$ (0,510). Surprisingly, SVM did not surpass the other models in terms of their performance in the metrics selected for the evaluation. These

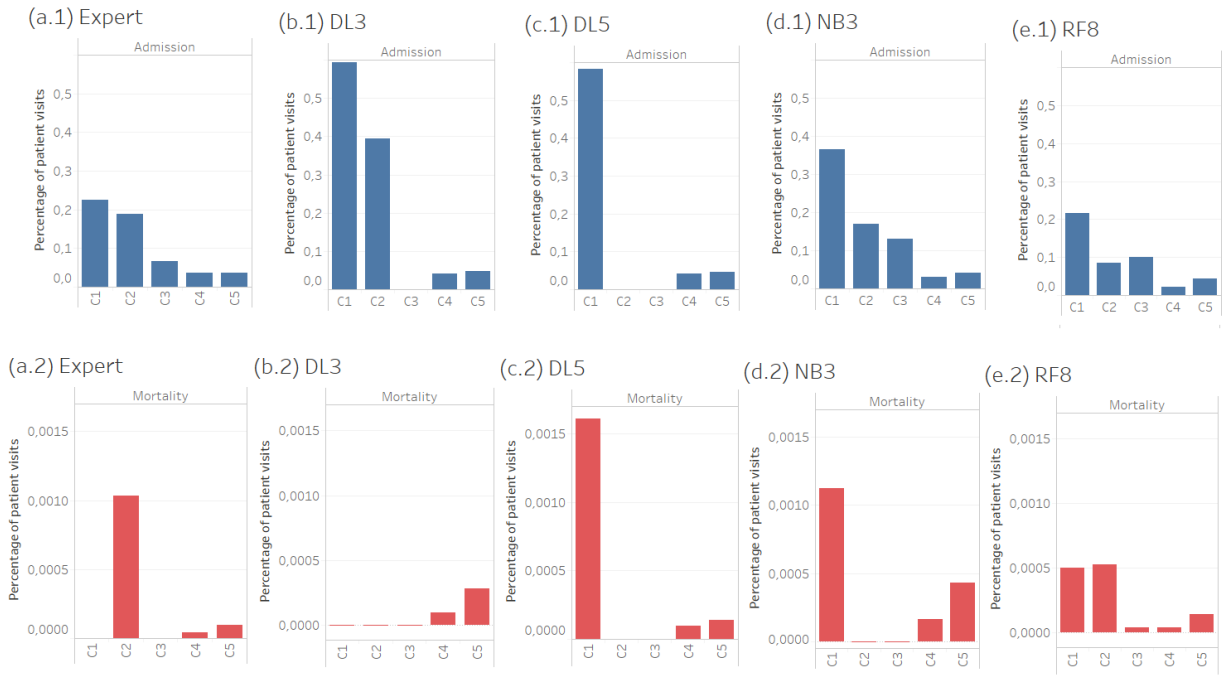

Figura 3.4: Proportion of patient with positive outcome assigned per class by ML models, Hospital admission (blue) and death (red). Expert means the current Triage system at the EGCH.

results serve to preselect the best four ML models which are compared against the current rule based expert Triage system in operation at the hospital.

Model validation is not an easy task in clinical decision problems. Each hospital must establish its strategic goals before carrying out model validation. This way, it will be possible to use a combination of well-known performance evaluation metrics to evaluate if the proposed systems meet these strategic goals. For example, evaluation may be defined in terms of allocation of patients by category (fingerprint), or the predictive capacity detection of critical clinical outcome (high severity detection performance).

The results of the tested approaches are also expressed graphically by plotting the proportion of patient with a specific clinical outcome *per* class, as can be seen in figure 3.4. In this figure, we appreciate the improved result of the algorithms $DL3$ and $DL5$ assigning hospitalized patients to the $C1$ class, compared to the current Triage. Similarly, a better response of the algorithm $NB3$ is also observed regarding the proportion of patients hospitalized by categories of high severity clinical outcome in $C1$, compared to the current Triage. On the other hand, $DL3$ shows a poor result in clinical outcome prediction. Finally, $RF8$ shows non-outstanding results in both cases.

The KS-test statistic is used as a metric to determine the distance between two distributions. It gives us a quantitative basis to determine which model produces patient categorization with greatest similarity respect to a desired distribution. The model whose assignation of patients by category is most similar to the desired distribution is $NB3$, but the KS distance of $RF8$ is not significantly higher. The response of DL models was not good regarding

this KS metric (cf. table 3.7). This result counts against the use of DL models to replace the current e-Triage system. It is important to note that the clinical environment imposes the following constraint: proposed Triage prediction models might not alter significantly the distribution of allocations per class, due to resource limitations. Our results do not violate this constraint, but there is plenty of room to improve the construction mechanism of the desired distribution.

Although, the prediction of the high severity classes was a central objective of the study, it is relevant to compare the results obtained with traditionally studied Triage models in pediatric patients. This mechanism that simplifies the multi-class problem in one of binary classification, cannot be considered as the only measure of evaluation of Triage models, but its simplicity and objectivity motivate to incorporate it. It is possible to observe that Random forest ($RF8$) improves other models in terms of AUC, PPV, accuracy, and specificity in predicting hospital admissions in the severe classes (C1 and C2).

Naive Bayes (NB) is the most robust approach in terms of patient assignation per class (see table 3.7). The excellent results of DL models in terms of Accuracy, Kappa and Triage-weighted Kappa in the most acute classes (as show table 3.5) do not necessary mean that DL models achieves a good result in term of final evaluation three metrics. Random forests ($RF8$) provide good results in assignation of patients by category in terms of KS-test distance to expert-based class distribution. It is of special interest to note that Random Forest AUC in high severity prediction improved over all traditional models in a recent pediatric validation of Triage models [2] and our results are similar to the best models presented in [73].

3.7. Conclusions

A lot of research has been conducted on the application of Machine Learning (ML) techniques for Triage prediction in the context of adult patient care, however, we research conducted in the applications of ML to the modeling of pediatric Triage has been scarce until very recent publications. Moreover, reviewing the literature on Triage prediction, we discover a big lack of methodological standards in terms of how to construct datasets, select the data, and to pre-process it, and the lack of consensus on the performance metrics used for the presentation of results. One goal of this paper was to establish methodological best practices in order to develop future algorithms on the Triage and screening problems in healthcare. This paper presents a study design detailing how our research was conducted. This methodology may be transferred to other multi-class classification problems where there is a record of final outcomes after the decision, which allows the evaluation of the performance of new combinations of new (and more objective) labels to train machine learning models.

Because of the nature of the problem, obtaining a huge data set to build and evaluate the algorithms is a strong barrier for researchers. After several years working with a pediatric hospital clinical staff, we were able to build a good quality dataset with $189,718$ pediatric ED visits over a period of three years. This is a large dataset compared to any published study, even for adults.

The problem of classifying patients at the very beginning of the emergency care process, i.e. admittance time, is a difficult problem, closely related to the readmission prediction [178]. The pediatric clinical environment presents specific challenges for the development of rapid screening methods in the ED for several reasons. It is not always possible to have access to all the information about the child state, since it is the tutor who refers the history and the symptoms. Often, he/she does not possess all the relevant information, because he does not know it or does not remember. Also, the symptoms of pain or discomfort require sophisticated instruments to capture the intensity perceived by an infant or child who does not have a clear language to express it quickly and clearly. This, among other things, motivates the development and constant improvement of the mechanisms of Triage incorporating all available tools.

From the model viewpoint, our results show a successful experimentation of different machine learning techniques who have a recent interest in scientific research and has a huge potential to be used in real clinical settings as a Triage decision support tool. We have tested several predictive ML tools, finding that Naive Bayes (NB) is the most robust approach in terms of assignation per class and sensibility of death outcome. Additionally, we found that Random Forest presented a better AUC, accuracy, PPV and specificity in clinical outcome predictive capacity of high severity classes than the other models. This result outperforms our expert-based models and traditional Triage models validated recently in pediatric patients.

Currently, there is a worldwide tendency to talk about how machine learning and artificial intelligence will replace human experts in different important tasks. However, this is very difficult to achieve in real clinical context. Moreover, even from the legal and ethical point of view, it is necessary to understand the decision makers' responsibility taken and the lack of knowledge of the real potential that this type of tool has in order to improve the quality of patient care. Despite this, we have made progress in understanding what is happening in this kind of socio-technical environment, through electronic health records, there is still a long way to go.

Chapter 4

A pediatric early warning system machine learning model based on clinical outcomes

Wolff P & Ríos S [1]

Background and Objective: Some pediatric inpatients' decompensation can be predicted using periodic bedside vital signs observations. A group of models developed with this proposal is called Pediatric Early Warning Systems (PEWS). PEWS can be constructed using Machine Learning (ML) techniques. The aim of this study is to develop an ML-based PEWS to predict unplanned transfer from a general hospital ward to an intensive care unit (ICU), within 8 hours of a given vital sign observation.

Methods: This study was performed with 178,970 pediatric bedside vital sign observations from 4,104 patients. We tested 25 different configurations for Multilayer perceptron (MLP), Naive Bayes (NB), Random Forest (RF) and Support Vector Machine (SVM) considering different parameterizations. We use 10-fold Cross-Validation with stratified sampling and SMOTE technique to deal with overfitting and class imbalanced dataset.

Results: RF performs better ($p < 0{,}005$) regarding AUC (0,898), Acc (92,4 %) and specificity (97,94 %) than the other models in our experiments. Naive Bayes (NB) is the most robust approach in terms of Sensitivity. Receiver Operating Characteristic (ROC) curve is also presented for the selected 4 best approaches.

Conclusions: Our results show that ML algorithms can outperform ICU transfer prediction AUC compared to expert-based traditional PEWS and other recent ML approaches. The use of manually-collected vital signs and the inclusion of pediatric patient condition information (i.e. sleep, awake and crying) and age ranges show an improvement in prediction

[1]The following is an up-to-date (to the time of writing) version of the paper "A pediatric early warning system machine learning model based on clinical outcomes", please do not cite this paper without authorization.

performance traditional outcome, as well as the supplemental oxygen support information.

4.1. Introduction

Some inpatients' decompensation can be predicted using periodic bed-side vital signs observations. Around 85 % of severe adverse events (SAE) are preceded by abnormal vital signs [101], and 59 % within 1 – 4 hours before cardiac arrest [4]. The group of models developed to predict decompensations is called Early Warning Systems (EWS). Currently, the bed-side vital signs observation are the basis of all reported EWS models. EWS has evolved to alert health professionals regarding potential clinical decompensation.

4.1.1. Related Work

Currently, there are many different EWS in use internationally. There are some EWS built based on experts' perspectives, such as the National Early Warning Score (NEWS) [123, 156], Modified Early Warning Score (MEWS) [160] and VitalPAC Early Warning Score (VIEWS) [135]. There are also pediatric EWS based on suggestions from experts such as Children's hospital Early Warning Score (C-CHEWS) [112], Pediatric Early Warning Score (PEWS) [54] and Bedside PEWS [130]. MEWS and ViEWS can be used on non-ICU ward patients with good results [183].

Other EWS had been derived using statistical modeling (Analysis of variance ANOVA, Backward stepwise Regression) such as the Rothman Index [140] and the electronic Cardiac Arrest Risk Triage (eCART) score [36]. Badriyah et al. shows a Machine Learning based (ML-based) EWS using Decision Tree (DT) analysis [12]. Clifton et al. shows a one-class Support Vector Machine (SVM) [39]. Discrete-time logistic regression was also used as an effective and efficient method to predict adverse clinical outcome [97].

This wide range of models use different features, such as patient unit, patient age group (pediatric or adult) and the quantity and origin of the used observations. In our case, the objective was to enhance pediatric patient models within the hospital general ward, focused on manually-collected vital sign measurements and other bedside observations.

4.1.2. Clinical Outcomes

Most of these EWS were designed to detect patient deterioration in hospital general wards, specifically those at increased risk of: unplanned ICU transfer [116], unplanned return to the operating theatre, a prolonged stay, cardiac arrest, or death.

Rather than relying only on expert opinion, we assessed our ML-based EWS performance through the discriminative ability to predict unplanned intensive care unit (ICU) transfer, within at least 8 hours of a given vital sign observation. This definition is well founded on the

fact that 8 hours is a reasonable time to react by medical staff in case of a decompensation alarm, it offers (as it will be seen in section 4.3) a better predictive capacity (see section 4.3). Our selected outcome (unplanned ICU transfer) is related to other outcomes (such as cardiac arrest). We select unplanned ICU transfer as the only outcome, mainly due to the low frequency of the death episodes and cardiac arrest events within a general hospital ward. The prolonged stay was not used, because it is a less objective outcome that could lead to errors in the training processes and model evaluation.

Rubin et al. approach of a pediatric ML-based EWS shows an AUC performance for the ensembled model of 0.840 [143]. Watkinson et al. result is shown after a review of 23 EWS (including these authors' data) whose AUC performance does not exceed 0.868 in the outcomes early detection (in the 12, 24 and 48 hours prior to the outcome occurrence) [173].

4.2. Materials and methods

Different steps were performed to obtain our results. The overall model training and validation processes are shown in figure 4.1.

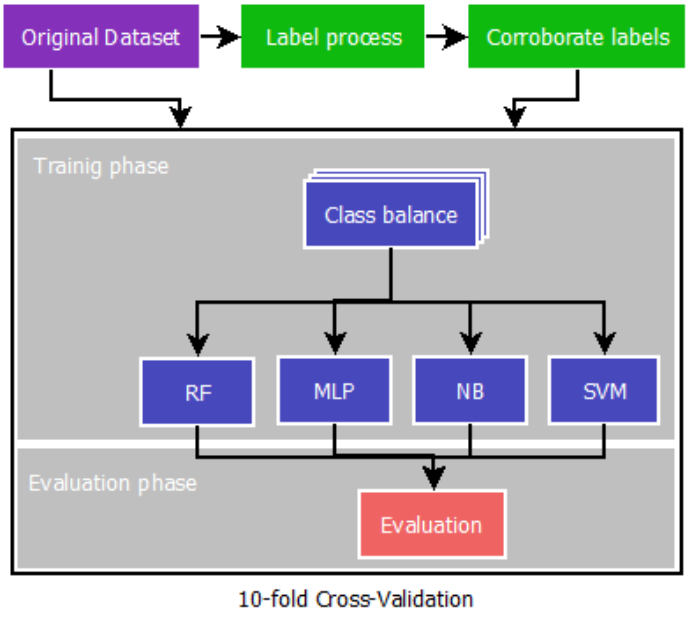

Figura 4.1: Study design schema

4.2.1. Dataset characterization

We carried out a single public center retrospective cohort study. Records for all vital sign measures during the study period were retrieved from the currently local operating system. We include 4,104 pediatric patients (< 18 years) who were discharged between January 1, 2018 and December 31, 2018. All erroneous or incomplete data records were discarded.

Finally, our cleaned dataset (see table 4.1) contained 178,970 manually collected records. Only 681 of the collected data points are generated 8 hours prior to ICU transfer (0.38 %).

Tabla 4.1: Dataset information

General		
No. of Patients		4,104
ICU transfers		203
Females		1,793 (43.7 %)
Vital signs Records		178,970
Records per Age Range	0 - 1m	10,239 (5.7 %)
	1m - 3m	12,737 (7.1 %)
	3m - 1y	30,584 (17.1 %)
	1y - 2y	22,303 (12.5 %)
	2y - 4y	22,218 (12.4 %)
	4y - 8y	27,107 (15.1 %)
	8y - 10y	13,336 (7.5 %)
	>10y	40,446 (22.6 %)
Patient condition	Awake	109,533 (61.2 %)
	Sleep	57,709 (32.2 %)
	Crying	11,728 (6.6 %)
Oxygen Support		
Airway	Natural airway	177,403 (99.12 %)
	Tracheotomy	1,549 (0.87 %)
	Endotracheal tube	18 (0.01 %)
Supplemental Oxygen Support		45,395 (25,4 %)

In most adult EWS age range is not a central feature. In pediatric case it is relevant to know the patient age range to determine the heart rate (HR), respiratory rate (RR) and systolic blood pressure (SBP) risk levels, since the normal vital signs and out of range reference values of these groups are different (see in the table 4.2). The age ranges were defined based in Pediatric advanced life support (PALS) [48].

Tabla 4.2: Vital signs records information (mean±SD)

Range	HR [bpm] ± SD	T [C] ± SD	RR [bpm] ± SD	SBP [mmHg] ± SD	SAT [%] ± SD
0 - 1m	141.13 ±14.50	36.69 ±0.41	39.32 ±7.39	92.80 ± 9.94	98.57 ±1.94
1m - 3m	135.87 ±15.77	36.63 ±0.44	37.13 ±6.92	96.02 ± 10.47	98.59 ±1.94
3m - 1y	127.73 ±17.23	36.53 ±0.49	33.98 ±6.98	98.95 ± 10.46	97.96 ±2.83
1y - 2y	120.79 ±18.15	36.52 ±0.50	30.86 ±6.65	100.81 ± 10.20	97.79 ±2.42
2y - 4y	115.47 ±18.84	36.53 ±0.54	27.92 ±5.83	99.74 ± 10.80	97.80 ±2.11
4y - 8y	101.21 ±18.94	36.52 ±0.55	24.64 ±5.36	100.89 ± 10.35	97.78 ±2.98
8y - 10y	96.09 ±19.00	36.56 ±0.60	23.55 ±4.84	103.01 ± 10.17	97.87 ±2.87
>10y	88.84 ±16.97	36.55 ±0.52	21.41 ±4.84	107.61 ± 11.86	97.89 ±2.88

Additionally, it is possible to see the difference between the medians and quartiles between age ranges by the Boxplot for the cases of the HR, RR, and SBP (see figure 4.2). With respect to the age range and patient condition:

- a statistically significant decrease is shown in the case of RR and HR ($p <$ 0,00001 using pairwise t-test in 92.9 % of pairs) and
- a significantly increase in SBP ($p <$ 0,01 using pairwise t-test in 73.8 % of pairs), can be also observed in figure 4.2.

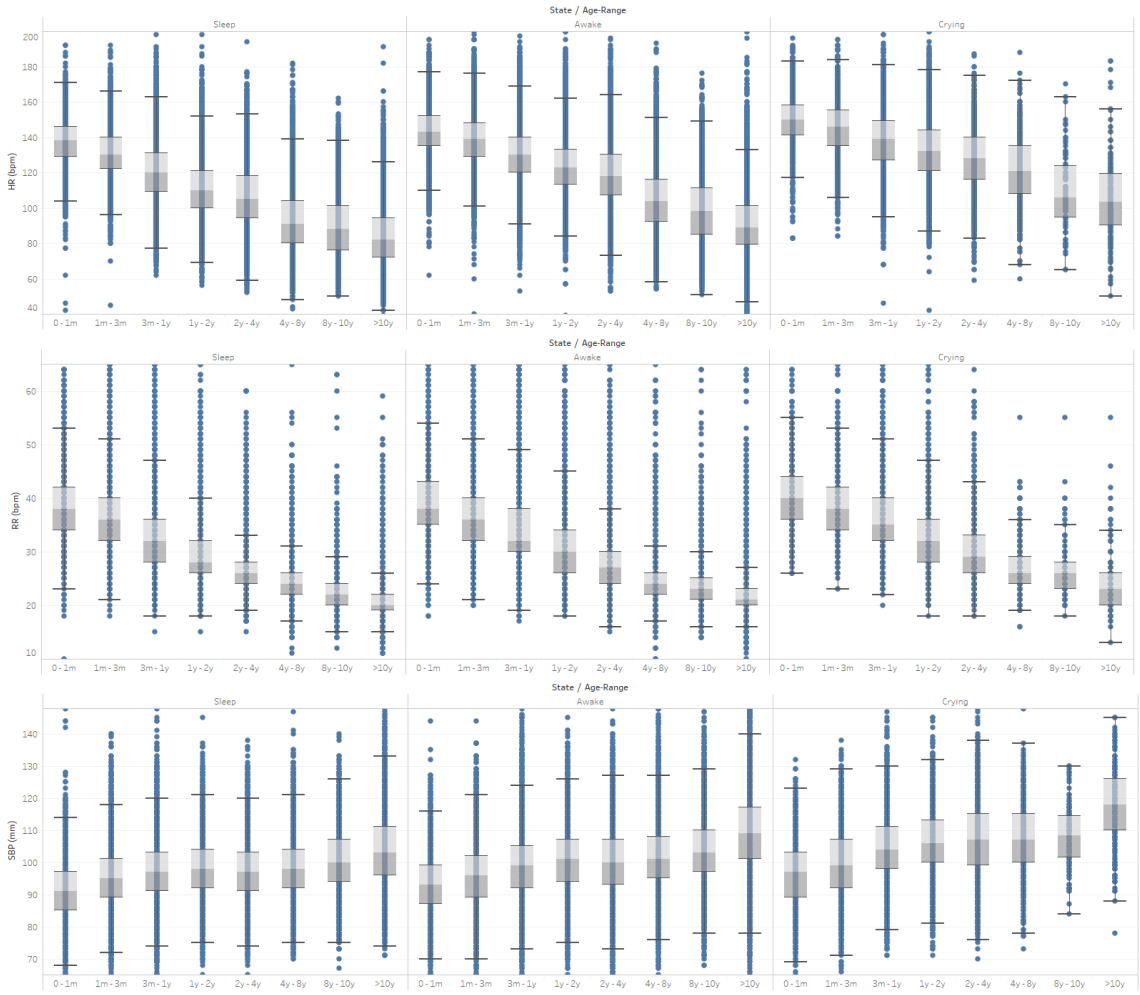

Figura 4.2: Boxplot of (a) Heart Rate, (b) Respiratory Rate and (c) Systolic blood pressure by age ranges and patient condition

To identify the vital sign measurement prior to ICU transfer, the local Electronic Health Record (EHR) system was used. In this way, positive labels were generated. All these labels were manually reviewed by a local committee of medical experts to avoid including errors in the dataset. This review was very time consuming, but necessary to ensure the data base quality.

The features selected for the model are the following: HR, level of consciousness (LOC), RR, body temperature (Temp.), oxygen saturation (SAT), Diastolic blood pressure (DBP), SBP and patient age range. Also, information about patient condition (sleep, awake and crying) and about Supplemental Oxygen Support (Airway, O2 in L/min, etc.) were used.

4.2.2. Dealing with Overfitting and class imbalanced

To avoid overfitting, a 10-fold Cross-Validation with stratified sampling were performed. The Cross-validation technique consists of repeating and calculating the arithmetic mean

of the results obtained by the model on different partitions of the data. This problem is characterized by the fact that, in general, the available data is highly unbalanced. To avoid the negative effect of this imbalance in the learning process, different strategies can be implemented [108]. One strategy that has shown good results is to use the Synthetic Minority Over-sampling Technique (SMOTE) during the training phase [34]. Random Oversampling was also tested. The SMOTE tool consists of creating new samples of the minority class by random convex interpolation between k randomly picked samples (neighbors) from the minority class. Random convex interpolation means that the new sample is computed as a degree polynomial, one of the reference samples whose coefficients are in the interval [0,1]. The up-sampling technique is subsequently performed to the split dataset and only applies to the training set, as it was presented in a previous work [178].

4.2.3. Machine learning models under evaluation

Differently supervised classification methods can be used and compared to solve this problem. Four methods were selected, based on their good results in classification problems.

Multilayer perceptron (MLP)

Multilayer perceptron (MLP) is a broad set of machine learning methods of different characteristics. They are based on the theory of artificial neural networks (ANN). MLP is considered an important method, because of the good results obtained in different classification problems. H2O library [43] was used to perform a multi-layer feedforward artificial neural network, trained for classification with back-prop and stochastic gradient descent [84]. The classical sigmoid activation function of neurons in the hidden layers has been replaced by others like the Hyperbolic Tangent (Tanh) activation used in Deep Learning architectures [71]. Different neuron topologies were trained, finding the best result in 4 hidden layers topology with 50-50-25-5 neurons respectively. Adaptive learning rate algorithm (ADADELTA) [186] was used to combine the benefits of momentum and learning rate annealing to avoid slow convergence in the training process. Models were trained considering 10 epochs, $epsilon = 10e{-}8$ and $rho = 0{,}99$. We used the H20 package (https://www.h2o.ai) for training process.

Naive Bayes (NB)

Naive Bayes (NB) is a simple, and very effective method of machine learning. Its central concept is the naive assumption of independence of the individual features. This allows us to approximate the probability of a D-dimensional feature vector as a D probabilities product of 1-dimensional feature vectors. In this case, Gaussian probability densities estimation was used to model the likelihood density function of the features. A Laplace correction to prevent the high influence of zero probabilities was also used.

Random Forest (RF)

The central concept of Random Forest (RF) [24] is a random trees assembly. These trees are trained on bootstrapped sub-sets, each of which is built on a bootstrap sample of the training dataset using a subset of randomly selected variable parameters. Each tree node represents one specific attribute splitting rule. In our test, Gini index was identified as the best criterion to select attributes for splitting. Maximal depth used was 8 and 1,000 trees were trained.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the most important groups of supervised learning methods [169]. SVM can be used to address regression and classification problems. In the case of binary classifiers, SVM tries to find a hyperplane that divides the feature space in two and maximizes the distance between groups of feature vectors belonging to a class, with respect to the feature vectors belonging to another class. The parameters that define the solution hyperplane are obtained solving a quadratic programming problem. The two classes divided by the hyperplane may not be linearly separable. In these cases, either the kernel trick was used [150] or penalizing an error, in a band around the hyperplane (soft margin) determined by an additional parameter called C (C-SVM). A C-SVM with Radial Basis Function (RBF) kernel and $C = 0{,}1$) were selected in this case.

4.2.4. Evaluation metrics

In problems like these, it is central to define beforehand the metric that will be used to measure the model performance. There are many techniques to evaluate the model results. There are many well-studied evaluation metrics, while the most commonly metrics used in binary classification are: Accuracy (Acc), Sensitivity, Specificity and Positive predictive value (PPV).

The Accuracy (Acc) metric will be determined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Sensitivity (Sens) and Specificity (Spec) are obtained using:

$$Sens = \frac{TP}{TP + FN}, \tag{4.2}$$

and,

$$Spec = \frac{TN}{TN + FP}, \tag{4.3}$$

Finally, the values of PPV are obtained using:

$$PPV = \frac{TP}{TP + FP}, \tag{4.4}$$

where true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) are case counts in a confusion matrix.

The analysis using Receiver Operating Characteristic (ROC) curves has been widely used to compare different binary classifiers. ROC curve graphs the Sensitivity versus the 1-Specificity. To identify the trade-off between the Sensitivity and Specificity, the Area Under ROC Curve (AUC) is performed.

4.3. Results

Twenty-five different configurations were tested considering different parameterizations: 6 for Multilayer perceptron (MLP); 1 for Naive Bayes (NB) 9 for Random Forest (RF) and 9 for Support Vector Machine (SVM). In the case of MLP, different activation functions (Tanh and Rectifier) and different topologies of the hidden layers (50-50-5, 50-25-5, 50-50-25-5 neurons) were tested. In the RF approach, different tree numbers (10, 100, 500) and different splitting criteria (Gain ratio, Gini index, and Information gain) were tested. In the case of SVM, 3 kernels (linear, polynomial and Radial Basis Function) and different values of C were tested (0,1 , 1 and 10). For each approach, the configuration that showed the best performance in AUC was selected. In table 4.3, five performance metrics are shown for each selected approach.

Tabla 4.3: Diagnostic accuracy measures

	AUC (SD)	Acc(SD)[%]	Sens(SD)[%]	Spec(SD)[%]	PPV(SD)[%]
MLP	0.867(0.021)	97.50(0.17)	53.97(4.64)	97.66(0.17)	7.67(0.93)
NB	0.825(0.031)	92.14(0.22)	65.27(3.71)	92.24(0.22)	2.93(0.20)
RF	0.898(0.013)	97.79(0.10)	54.42(5.92)	97.94(0.10)	8.66(0.79)
SVM	0.844(0.051)	92.40(0.55)	63.73(7.04)	92.51(0.55)	3.17(0.45)

The results of each approach are graphically presented in ROC curves (see figure 4.3). The figure 4.3 also shows: the ROC thresholds and the area between the Pessimistic ROC and the Optimistic ROC for RF.

It is possible to observe the effect of performing data balance strategy in the training process, as well as the relevance of variables such as: the defined Age Range, patient's condition (awake, sleep and crying), and oxygen support information. The AUC performance metric obtained without considering these conditions are presented in the table 4.4.

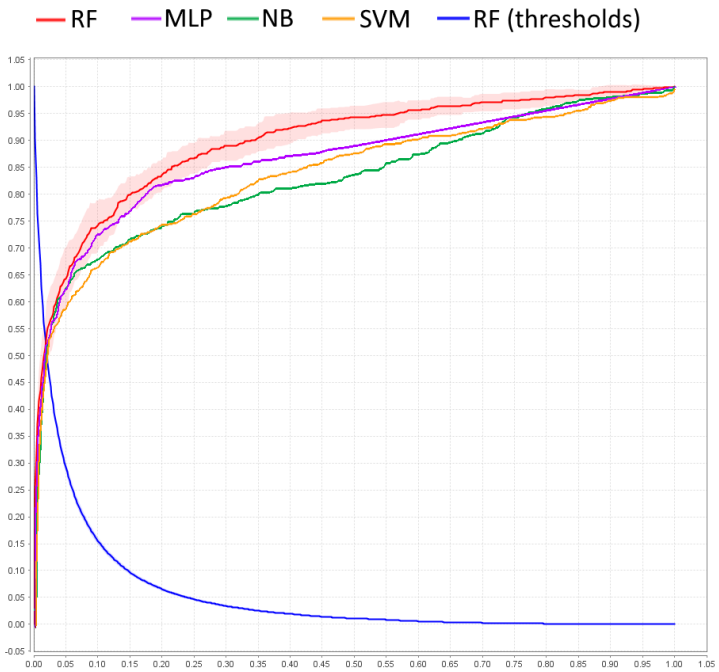

Figura 4.3: ROC curve

Tabla 4.4: AUC (SD) performance by ML model (a) without including Oxygen Support information, (b) Imbalance data for training, (c) without considering Age Range and patient condition, and (d) all variables and balanced dataset.

	(a)	(b)	(c)	(d)
MLP	0.744(0.046)	0.820(0.024)	0.840(0.024)	0.867(0.021)
NB	0.787(0.029)	0.833(0.017)	0.841(0.029)	0.826(0.033)
RF	0.824(0.038)	0.877(0.036)	0.853(0.019)	0.898(0.013)
SVM	0.804(0.034)	0.486(0.124)	0.824(0.049)	0.844(0.051)

4.4. Discussion

Four ML tools with different configurations to perform a pediatric EWS have been tested. None of the models tested simultaneously improves all the metrics. Results of these four models are summarized in table 4.3.

RF showed a better AUC, Acc, PPV, and specificity than the other models in classification experiments ($p < 0{,}005$ using pairwise t-test). NB showed a lower variability in Sensitivity ($p < 0{,}001$ using one-side t-test).

Graphically, the ROC curve (see figure 4.3) allows us to observe a superior performance in practically the whole thresholds spectrum.

From the clinical viewpoint, a high sensitivity prediction method is preferred to avoid a false negative errors. However, Accuracy is also a desirable metric that allows you to handle the trade of between a false negative errors and overload the medical staff. RF simultaneously

shows a high Accuracy and PPV ratio. In general, EWS models have more than one level of detection through the use of regression or scoring methods. This allows the EWS to have different levels of sensitivity and specificity. In this case, it is possible to use the built model by varying the bias settings and in this way to establish levels based on the expected results.

As it can be seen in table 4.4, the outstanding results can be explained based on the simultaneous use of information on patient oxygen support, age range and condition. In addition, it is possible to see that in the case of RF, MLP, and SVM the positive effect in classification performance of class balancing technique.

It can also be observed that RF is more stable than NB and SVM in terms of its standard deviation ($p < 0{,}001$ using F-test) while RF is less pronounced than MLP.

The inclusion of oxygen support information allows us to significantly improve the performance (in AUC $p < 0{,}05$ using pairwise t-test) of all ML models, as in the case of scoring methods and centile-based unsupervised learning, as shown in a previous work [173].

It is interesting to note a largely statistically significant RF performance improvement ($p < 0{,}00001$ in one-side t-test) when the patient age range and the patient condition are included.

RF performance in AUC exceeds all reported AUC of models in recent studies in EWS, for both adult and pediatric populations [143, 173]. Based on the most recent publications on ML and non-ML models, the EWS presented here exceeds all other models when the different calculated AUCs are compared. This result is achieved with the incorporation of specific features in the ICU transfer prediction problem, in combination with widely studied strategies in the machine learning field. Our results show that ML algorithms can improve ICU transfer prediction Acc compared to expert-based traditional EWS.

A significant contribution of this publication is to show the positive effect in model performance if strategies are used for the class balancing prior to the learning models training. The use of tools such as Cross-Validation allows us to avoid overfitting and to have more reliable results. The inclusion of information about patient condition (sleep, awake and crying) and age ranges of pediatric population in EWS shows an improvement in traditional outcome prediction performance, as well as the information of supplemental oxygen support.

One of the important limitations of this study is that the data set comes from a single hospital, so the results reported should be evaluated with data from other pediatric hospitals. Another important limitation is that it should be compared with other models using the same data. In addition, the study is limited in the number of clinical variables available that are used for prediction, such as capillary refill. An important consideration in this work is that each observation of vital signs was analyzed independently of the rest, and not as a part of a temporal sequence. Including this feature may allow to improve the results, but it hinders the construction of the training and test sets and limits the number of models that can be tested.

In this work we try to provide knowledge about the real potential that this type of tool has. We have succeeded in understanding the dynamics within this socio-technical environ-

ment. However, there is still a long way ahead. Even when ML models are able to outperform an expert based algorithm or local experience, we are still far from being able to completely replace their knowledge with any of the ML models already developed. However, the developed models could be integrated into the hospital EHR setting to support the medical staff to increase the frequency of vital signs monitoring, in specific conditions.

Currently, there is a worldwide preoccupation on how artificial intelligence (AI) will replace human experts in different important tasks. In our opinion this is very difficult to achieve in real clinical settings. Moreover, it is necessary to understand that algorithms might help to take a better decision, however, there is a huge responsibility decision makers assumed with their patients which prevent us from fully automate these processes.

Chapter 5

Final Conclusions

5.1. Research aims

Healthcare risk may be approached from several perspectives such as clinical risk management, identification [154], and stratification. This thesis has been focused on clinical risk management. Particularly, we aimed to improve the results of three risk models based on Machine Learning tools. These models can be incorporated into clinical decision making processes embedded into systems generically known as Machine Learning Clinical Decision Support Systems (ML-CDSS). There is a wide spectrum of clinical issues that may be tackled with ML-CDSS. We have focused our research effort on three kinds of issues based on their common characteristics: Readmissions, Triage, and Decompensation of inpatients. These three problems require assessing the level of risk to guide the clinical decision that is made, in order to adapt the care level to the predicted risk. In this sense, it is essential that the risk assessment should be a part of the care process. Another common characteristic of the three problems is that they require a level of risk at a specific time. Finally, another commonality is that the risk determined by each model refers to an individual patient, not to a whole population.

ML methods used in risk prediction have common characteristics such as: Highly class unbalanced data; Non-unique label; Sensitivity as the relevant performance evaluation metric; and, a strong dependence of modeling features and parameters on specific population characteristics.

Although we have addressed separately the three problems in different publications, they belong to this compendium. Furthermore, there is a shared body of strategies used to address each problem. In other words, our experience of dealing with one problem has contributed to improve the solution of the next considered problem.

The models selected to deal with in each case study differ as follows: each of them must be used in different parts of the care process; the number of predicted class in each case is different; the origin and type of data used in each problem are not the same. In addition, different performance evaluation metrics are used to report results in each case of study.

5.2. First findings

One of our first findings was the scarcity of scientific research on pediatric patients, probably because the construction of predictive models for children seems to be more challenging than for adults [2].

Another important finding of the international scientific literature review carried out is the wide variety of decisions that can be supported by computerized systems [19, 65]. This demonstrates a non-recent interest, which has grown in recent years, due to multiple factors such as: Availability of datasets; availability of software packages that facilitate the development and testing of different models; advances in processing capacity; and recent interest of clinicians in the advantages of using these models.

Although we found previous research that addresses problems such as triage and decompensation of inpatients, the use of these models is not always correct. In addition, in some cases, the performance evaluation metrics reported were inappropriate or incomplete for the assessment of the claimed results. In this Thesis, we describe the correct model validation methodology, and, in addition, we provide a large number of performance evaluation metrics.

A large number of investigations report model evaluation based on human expert judgment [76, 158, 190, 123, 156, 160, 135, 105]. Although this strategy is widely spread, it may not be the most appropriate [16]. In the three problems tackled in this Thesis, both the performance evaluation metrics and the labels that allow training the models, are based on clinical outcomes. This allows us to use a larger dataset and guarantees the result objectivity, by limiting the influence of subjective human judgment.

Many studies on ML-CDSS focus their research efforts on different hospital problems, such as identifying and managing care for high-cost patients [163, 8], and analyzing clinical data and inferring a diagnostics [182]. The characteristics of the selected models focus on enriching the quality of patient care and safety, by setting as a central objective the determination of the patient level of risk in order to adapt the level of care. In this way, our research results can help to prevent treatment complications to the patient. The focus on risk reduction is a central component in the Person-centered health care model defined by the hospital [82, 146].

5.3. Proposed methodology

The methodology followed addressing each problem was detailed in-depth in each publication. The versatility of the tools selected for these three cases allowed us to solve problems with great importance for hospitals and the safety of care. Some final reflections are presented regarding common characteristics of the methodology used in approaching these cases.

5.3.1. Datasets

The informatization level of hospitals in our country is lower than that observed in international research [136]. Although there are national strategies to reverse this situation, up to now they have not shown the expected results. This affects the availability of data to carry out high-level research in the Chilean population since available clinical data is scarce and unreliable. This project was feasible due to the high commitment of the management team in charge until 201 of the Exequiel González Cortés Hospital (EGCH) 8, concurrent with the informatization process that the hospital underwent since 2015.

Both the commitment of the hospital management and the computerization process of the EGCH, allowed to gather massive datasets containing anonymous information encompassing vital signs records, and data from the EHR and cost accounting systems. This large volume of data allowed training models that are quite sensitive to the data sample size, such as artificial neural networks, with a large number of labeled cases at different levels of risk.

5.3.2. Outcome based models

A key focus on security in all industries is the investigation of the link between the causes and the risks that underlie faulty outcomes of the industrial process [154], guiding the development of predictive models, which is a structured approach to reduce the occurrence of these negative outcomes. In this Thesis, three examples of how to address different problems with similar characteristics are presented, using clinical outcomes, both for model training and performance evaluation.

The labeling or relabeling (in the case of triage) of data samples constructed using clinical outcomes, allowed a more objective approach to solve the problem, than in the case of approaches where the data sample label was established on the basis of human expert judgment. This was clearly observed in the problems of triage and decompensation of inpatients. In addition, this data labeling approach allows for greater training and test dataset, since in general, the manual labeling processes are very time-consuming. Although this approach has been followed previously dealing with some modeling problems using crowd-sourcing systems and noisy labels [187, 188], it is a problem solving traversal characteristic of great relevance in this study.

5.3.3. Class balance techniques

One of the relevant findings of the literature review is the low use of class imbalance correction strategies. Especially when high-risk classes prevalence is very small. In the case studies considered in this Thesis, we use minority class upsampling to improve the results especially when the data sets are unbalanced. Different methods were tested finding that SMOTE allowed improving the results. When dealing with decompensation of inpatients and readmission, the effect of carrying out the class balancing strategy before training is assessed separately. In both cases, it is shown that applying this technique has a significant

contribution to improve the results of trained predictive models.

5.3.4. Selected ML tools

Four classifier building algorithms were selected on the basis of their performance results, the characteristics of the problem, and the state of the art. Namely, they are Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes (NB), and Random Forest (RF). These models allow multi-class and dichotomous classification. Other models such as Gradient Boosting Trees (GBT) and Logistic Regression (LR) were also tested, but not reported in the published papers.

When presenting and discussing each study, we report the configuration and parameterization of each predictive model. The inclusion of four different models per study on average is in striking contrast with the literature. We have explored a wide range of alternatives, however we select for presentation the best performing models, in order not to overwhelm the reader. In addition, available open source libraries were used in each case study, so the experiments are reproducible for researchers with access to the data. Sharing the data was always a longing of the author of this Thesis, however, by national bioethical regulations, this was not possible.

5.3.5. Evaluation metrics

The proposed methodology not only provides strategies to improve the performance of the models but also to achieve a better assessment of their ability to be deployed in real clinical settings. This is why we propose the "fingerprint.ºf the chapter 3 and the PPV of the chapter 2. These metrics are not commonly used as a central tool for the evaluation of classification models, but incorporating them allows enriching the discussion about the applicability of these models.

In the studies about hospital readmission and decompensation of inpatients, the use of standard metrics for the evaluation of dichotomous models is proposed. From the point of view of the results of the models, it is relevant to evaluate the cost-benefit of a model whose implementation does not leave us with a hypersensitive, but imprecise model. The metrics selected when dealing with these problems allow for a thorough discussion, not only of the capabilities of the model but also about their potential to be implemented in real clinical settings. More precisely, the ability to deal with the costs associated with the adequacy of the level of care, when necessary, in proportion to a possible over-categorization.

In the case study of the triage into five categories, we make an important methodological contribution regarding the evaluation of these models. This was achieved by incorporating traditional metrics into multi-class classification models with specific evaluation metrics to evaluate emergency triage. The evaluation of results presented as part of the methodology allows us not only to compare the power of machine learning methods but also to show their power against other knowledge-based models. In this case, dichotomized analysis is also

included in order to compare the results obtained with the results obtained in traditional expert-based models.

5.4. Results

Overall, Random Forest consistently showed a better predictive capacity. In the case of pediatric readmission, RF results are not presented in the main investigation. However, this result is presented in a subsequent investigation included as an annex to this thesis (see Annex A).

In the case of pediatric readmission, it was difficult to compare the results obtained with other investigations, since no previous experiences of the application of ML tools in readmission of pediatric patients were found. It is recognized that the AUC of other models may be higher in the case of readmission prediction of adult patients. However, direct comparisons to other studies are difficult because of different study designs, incomplete definitions of cohorts and outcomes, restrictions on disease-specific cohorts, or use of data unavailable [136].

In the case of pediatric triage, the result presented exceeds the AUC of the high-level risk (C1 and C2) hospital admission prediction of all previous research, only matched by [73] published in the same year as our paper.

In the case of the prediction of decompensation of inpatients, the AUC of the dichotomous label is shown to be far superior to a wide range of research shown in recent publications [173].

Finally, we want to emphasize the discussion included in each publication regarding the best available metrics and how they express the best result of the model, enriching the discussion regarding the benefits of these models to face each of the tasks considered in this Thesis. The excellent results obtained in different evaluation metrics in risk prediction problems allow methodological validation of the ML tools used, even if they are compared with other knowledge-based and non-knowledge-based methods.

5.5. Applications/implications

The applicability of the studied predictive models in real clinical settings may vary. Although, the three models are designed to be part of the care process, therefore validation by the clinical teams in charge is essential. Further research is required to validate these models from a clinical perspective. An important challenge is to determine what characteristics of systems make them effective in supporting particular types of clinical decisions [114]. More precisely, to establish which is the best metric, and the decision thresholds that make the use of an ML model valid in the health care process. In other words, the most important success factor for CDSS is to make them fit into daily clinical workflow [95, 138].

There is a research line of legal affairs in the use of ML tools in healthcare [27, 72]. Before an effective implementation can be achieved in hospitals, medico-legal responsibility will require further attention [147]. Locally, in spite of the validity of the model, the clinicians consider the ability to audit the implementation as fundamental. This means that the model must be fully documented and its behavior must be clearly explainable in front of an external referee [79]. This consideration is raised in the event of a medico-legal dispute. For this reason, Regression models and Simple Decision Trees have a better chance of being implemented in real clinical contexts, regardless of lower predictive power, compared to models considered black-box (difficult to interpret)[52].

During the development and improvement of the expert-based model described in the publication of the pediatric triage, the central level decided to implement the ESI model in all Chilean hospitals. At this moment, the commitment of the hospital management team was fundamental, not only the validity studies. This team was able to defend the use of the local expert-based triage, through international experience and the benefits of this model.

In the case of the decompensation of inpatients risk model, there is less local regulation about the type of model that should be used, in contrast with emergency triage. However, in this case, clinicians also require validated and auditable methods. This last condition favors the use of scoring models, widely disseminated for this task, as shown in Chapter 4. The models that analyze the variables separately have the advantage of allowing the system to report to the operator the cause of the calculated risk. Locally, not only the identification of the risk is assessed, but also the possibility of having an automated method to determine what are the characteristics that generate a modification in the risk assessment. This limits the feasibility of using only ML tools that showed better results.

The prediction of readmissions has as a central objective to prevent them from happening, by supporting the clinical decision at the time of discharge. In the discharge process, all the information that is required for the readmission prediction is available. At that moment it is when the prediction can support hospital staff by recommending, for example, post-discharge special care, caregiver education or postponement of discharge. If the cost of these strategies is affordable, then it is possible to consider implementing models such as the one presented.

5.6. Final remarks

The correct use of ML tools allows improving the predictive result in problems related to patient risk. This was shown in three different examples, in which different ML tools were used (in the state of the art). The methodology presented in each problem has, in general terms, similar characteristics and can be used in other CDSS.

The progress of research on patient risk allows a better understanding of the care process. The development of these tools and the study of their impact on real clinical settings showed excellent results, even when compared with mechanisms considered standard in the clinical literature. In addition, this research requires a deep understanding of how CDSS interacts with their operators in complex socio-technical environments. The success of the proposal is

due to a clinical point of view of the problem and the opportunities found to improve the way in which health organizations work.

The proposed models are a contribution to the validation of ML methods as outstanding performance tools. The outcome perspective (for labeling or relabeling), both in training and in evaluation, proved to be an improvement from the performance viewpoint. The relabeling was used to improve the results obtained when an expert label was used for training. The expert triage labeling was built by a group of clinicians, who do not necessarily label cases with the same criteria. The result of this is a noisy label that trained models with poor performance. The same happened in the case of EWS, where a label constructed with unplanned transfers to ICU was used.

The risk characteristic addressed in this thesis is a non-measurable concept, there is no ground-truth in triage and pediatric EWS problems allowing to determine which tool is better than another. This research supports the standardization of evaluation metrics in the three problems studied.

In our country, there is a tacit commitment adopted by the authorities: "children first". However, one of the first findings of this thesis is the lack of application of ML models in the problems described for the pediatric population. In addition, it is recognized in the literature that it is a difficult problem to address and therefore it is outstanding in this research to validate our hypotheses in pediatrics.

5.7. Further research

Understanding the business is probably the least documented, structured and researched element of the CRISP-DM methodology. In this thesis, becoming familiar with the business fundamentals took about 9 years (prior to this work), due to the concurrent development of other research projects, process assessment and modeling, and technological implementation in public hospitals. This accelerated the development of the Thesis, although the author acknowledges that much remains to be learn.

The problems selected in this thesis are part of a wide range of problems described in the literature, grouped under the name of CDSS. The proposed approach can be used to develop models that solve the wide range of CDSS as an extension of this work. As for example, in other areas of the hospital such as Decision support in mechanical ventilation [109]; Decompensation risk in chronic outpatients [103]; Medication error prediction [149]; Severity scoring and mortality prediction [132, 87, 47, 176]. In addition, it is possible to extend this research to other populations such as adult patients, if the data is available.

A limitation of this work is that it was carried out in a single hospital, where it was feasible to have the necessary data for the training and validation of the models. Future work must be addressed to validate these studies in a multi-centre setting, in order to evaluate whether the results and conclusions are altered or not.

Some approaches such as [136] raise the use of Deep Learning tools, considering the longi-

tudinal data available in the EHR. This data was not fully available since the EHR is still in the implementation phase. Having more data will allow other supervised learning methods to be incorporated into the study. A greater amount of information such as active diagnoses will allow research in subgroups of patients, such as chronic patients or other specific pathologies.

The feature selection is a crucial stage in the ML process. It was addressed in each case study based on expert knowledge and scientific literature. Consistently selected as a part of an entity schema of the framework presented in [166]. Nevertheless, there are statistical techniques for feature selection that have shown good results in other medical applications [145, 139] and can be incorporated into future research. The most appropriate method of feature selection will depend on the data, problems characteristics, computational speed and accuracy [177].

Currently, the validation of the model construction method is not sufficient to allow the use of the model in real clinical settings. For this reason, its application in the future depends on the existence of more clinically-relevant studies [96], such as this one, that validates accurate and reliable methodologies and models. With this Thesis, we aim to lower the barriers to implementing machine learning methods in real clinical settings, and, in this way, contribute to enrich the quality and safety of patients.

Bibliografía

[1] María M. Abad-Grau, Jorge Ierache, Claudio Cervino, and Paola Sebastiani. Evolution and challenges in the design of computational systems for triage assistance. *Journal of Biomedical Informatics*, 41(3):432 – 441, 2008. Computerized Decision Support for Critical and Emergency Care.

[2] Kanokwan Aeimchanbanjong and Uthen Pandee. Validation of different pediatric triage systems in the emergency department. *World journal of emergency medicine*, 8(3):223–227, 2017.

[3] Mari Akre, Marsha Finkelstein, Mary Erickson, Meixia Liu, Laurel Vanderbilt, and Glenn Billman. Sensitivity of the pediatric early warning score to identify patient deterioration. *Pediatrics*, 125(4):e763–e769, 2010.

[4] Lars W. Andersen, Won Young Kim, Maureen Chase, Katherine M. Berg, Sharri J. Mortensen, Ari Moskowitz, Victor Novack, Michael N. Cocchi, and Michael W. Donnino. The prevalence and significance of abnormal vital signs prior to in-hospital cardiac arrest. *Resuscitation*, 98:112–117, Jan 2016.

[5] Arkaitz Artetxe, Borja Ayerdi, Manuel Graña, and Sebastian Rios. Using anticipative hybrid extreme rotation forest to predict emergency service readmission risk. *Journal of Computational Science*, Feb 2017.

[6] Arkaitz Artetxe, Andoni Beristain, and Manuel Graña. Predictive models for hospital readmission risk: A systematic review of methods. *Computer Methods and Programs in Biomedicine*, 164:49 – 64, 2018.

[7] Arkaitz Artetxe, Manuel Graña, Andoni Beristain, and Sebastián Ríos. Balanced training of a hybrid ensemble method for imbalanced datasets: a case of emergency department readmission prediction. *Neural Computing and Applications*, pages 1–10, 2017.

[8] A. S. Ash, Y. Zhao, R. P. Ellis, and M. Schlein Kramer. Finding future high-cost cases: comparing prior cost versus diagnosis-based methods. *Health services research*, 36(6 Pt 2):194–206, Dec 2001.

[9] Awais Ashfaq, Anita Sant'Anna, Markus Lingman, and Sławomir Nowaczyk. Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics*, 97:103256, 2019.

[10] Katherine A. Auger, Emily L. Mueller, Steven H. Weinberg, Catherine S. Forster, Anita Shah, Christine Wolski, Grant Mussman, Anna J. Ipsaro, and Matthew M. Davis. A validated method for identifying unplanned pediatric readmission. *The Journal of Pediatrics*, 170:105 – 112.e2, 2016.

[11] Dhifaf Azeez, Mohd Alauddin Mohd Ali, Kok Beng Gan, and Ismail Saiboon. Comparison of adaptive neuro-fuzzy inference system and artificial neutral networks model to categorize patients in the emergency department. *SpringerPlus*, 2(1):416, Aug 2013.

[12] Tessy Badriyah, James S. Briggs, Paul Meredith, Stuart W. Jarvis, Paul E. Schmidt, Peter I. Featherstone, David R. Prytherch, and Gary B. Smith. Decision-tree early warning score (dtews) validates the design of the national early warning score (news). *Resuscitation*, 85(3):418–423, Mar 2014.

[13] Charles A. Baillie, Christine VanZandbergen, Gordon Tait, Asaf Hanish, Brian Leas, Benjamin French, C. William Hanson, Maryam Behta, and Craig A. Umscheid. 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, 2013.

[14] Naomi S. Bardach, Eric Vittinghoff, Renée Asteria-Peñaloza, Jeffrey D. Edwards, Jinoos Yazdany, Henry C. Lee, W. John Boscardin, Michael D. Cabana, and R. Adams Dudley. Measuring hospital quality using pediatric readmission and revisit rates. *Pediatrics*, 132(3):429–436, 2013.

[15] David W. Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131, 2014. Habla de redmision, ram y triage.

[16] Arie Ben-David and Eibe Frank. Accuracy of machine learning models versus "hand crafted" expert systems – a credit scoring case study. *Expert Systems with Applications*, 36(3, Part 1):5264 – 5271, 2009.

[17] Ilaria Bergese, Simona Frigerio, Marco Clari, Emanuele Castagno, Antonietta De Clemente, Elena Ponticelli, Enrica Scavino, and Paola Berchialla. An innovative model to predict pediatric emergency department return visits. *Pediatric Emergency Care*, page 1, Oct 2016.

[18] Amy Berlin, Marco Sorani, and Ida Sim. A taxonomic description of computer-based clinical decision support systems. *Journal of Biomedical Informatics*, 39(6):656–667, Dec 2006. taxonomia de berlin.

[19] Eta S. Berner. Clinical Decision Support Systems: State of the Art. *Agency for Healthcare Research and Quality*, 9(69), 06 2009.

[20] Eta S. Berner and Tonya J. La Lande. *Overview of Clinical Decision Support Systems*, pages 1–17. Springer International Publishing, Cham, 2016.

[21] Jay G. Berry, Sara L. Toomey, Alan M. Zaslavsky, Ashish K. Jha, Mari M. Nakamura,

David J. Klein, Jeremy Y. Feng, Shanna Shulman, Vincent W. Chiang, William Kaplan, Matt Hall, and Mark A. Schuster. Pediatric readmission prevalence and variability across hospitals. *JAMA*, 309(4):372–380, Jan 2013.

[22] Ariadna Besga, Borja Ayerdi, Guillermo Alcalde, Alberto Manzano, Pedro Lopetegui, Manuel Graña, and Ana González-Pinto. Risk factors for emergency department short time readmission in stratified population. *BioMed Research International*, 2015:Article ID 685067, 2015.

[23] Bente Bilben, Linda Grandal, and Signe Sovik. National early warning score (news) as an emergency department predictor of disease severity and 90-day survival in the acutely dyspneic patient – a prospective observational study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 24(1):80, 2016.

[24] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[25] J. C. Brillman, D. Doezema, D. Tandberg, D. P. Sklar, K. D. Davis, S. Simms, and B. J. Skipper. Triage: Limitations in predicting need for emergent care and hospital admission. *Annals of Emergency Medicine*, 27(4):493–500, 1996.

[26] Matthias Briner, Oliver Kessler, Yvonne Pfeiffer, Theo Wehner, and Tanja Manser. Assessing hospitals' clinical risk management: Development of a monitoring instrument. *BMC Health Services Research*, 10(1):337, 2010.

[27] Steven H. Brown and Randolph A. Miller. *Clinical Decision Support (Second Edition)*, chapter Chapter 26 - Legal and Regulatory Issues Related to the Use of Clinical Software in Health Care Delivery, pages 711–740. Academic Press, Oxford, 2014.

[28] Emily M. Bucholz, James C. Gay, Matthew Hall, Mitch Harris, and Jay G. Berry. Timing and causes of common pediatric readmissions. *The Journal of Pediatrics*, 200:240–248.e1, Sep 2018.

[29] William Caicedo-Torres, Gisela García, and Hernando Pinzón. A machine learning model for triage in lean pediatric emergency departments. In Manuel Montes y Gómez, Hugo Jair Escalante, Alberto Segura, and Juan de Dios Murillo, editors, *Advances in Artificial Intelligence - IBERAMIA 2016*, pages 212–221, Cham, 2016. Springer International Publishing.

[30] William Caicedo-Torres, Gisela García, and Hernando Pinzón. *A Machine Learning Model for Triage in Lean Pediatric Emergency Departments*, page 212–221. 2016.

[31] Chih-chung Chang and Chih-jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2:1–39, 2013.

[32] P Chapman. *CRISP-DM 1.0: Step-by-step Data Mining Guide*. SPSS, 2000.

[33] P Chapman, R Kerber, J Clinton, T Khabaza, T Reinartz, and R Wirth. The crisp-dm process model. *The CRIP–DM Consortium*, 310(C):91, 1999.

[34] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[35] Seifu J. Chonde, Omar Ashour, David A. Nembhard, and Gül E.Okudan Kremer. Model comparison in emergency severity index level prediction. *Expert Systems with Applications*, 40(17):6901–6909, 7 2013.

[36] Matthew M. Churpek, Trevor C. Yuen, Christopher Winslow, Ari A. Robicsek, David O. Meltzer, Robert D. Gibbons, and Dana P. Edelson. Multicenter development and validation of a risk stratification tool for ward patients. *American Journal of Respiratory and Critical Care Medicine*, 190(6):649–655, 2014.

[37] David A. Clifton, David Wong, Susannah Fleming, Sarah J. Wilson, Rob Way, Richard Pullinger, and Lionel Tarassenko. Novelty detection for identifying deterioration in emergency department patients. In Hujun Yin, Wenjia Wang, and Victor Rayward-Smith, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2011*, pages 220–227, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[38] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko. Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *IEEE Journal of Biomedical and Health Informatics*, 18(3):722–730, May 2014.

[39] Lei Clifton, David a. Clifton, Peter J. Watkinson, and Lionel Tarassenko. Identification of patient deterioration in vital-sign data using one-class support vector machines. *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, (ii):125–131, 2011.

[40] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[41] Eric A. Coleman, Sung-joon Min, Alyssa Chomiak, and Andrew M. Kramer. Posthospital care transitions: Patterns, complications, and risk identification. *Health Services Research*, 39(5):1449–1466, Oct 2004.

[42] Larry A. Allen Colleen K. McIlvennan, Zubin J. Eapen. Hospital readmissions reduction program. *Circulation*, 131(20):1796–1803, 2012.

[43] D. Cook. *Practical Machine Learning with H2O: Powerful, Scalable Techniques for Deep Learning and AI*. O'Reilly Media, 2016.

[44] Shaoze Cui, Dujuan Wang, Yanzhang Wang, Pay-Wen Yu, and Yaochu Jin. An improved support vector machine-based diabetic readmission prediction. *Computer Methods and Programs in Biomedicine*, 166:123 – 135, 2018.

[45] Mohammed Dalwai, Katie Tayler-Smith, Michèle Twomey, Masood Nasim, Abdul Qayum Popal, Waliul Haq Haqdost, Olivia Gayraud, Sophia Cheréstal, Lee Wallis, and Pola Valles. Inter-rater and intrarater reliability of the south african triage scale in low-

resource settings of haiti and afghanistan. *Emergency Medicine Journal*, 35(6):379–383, 2018.

[46] Michele D'Apuzzo, Geoffrey Westrich, Chisa Hidaka, Ting Jung Pan, and Stephen Lyman. All-cause versus complication-specific readmission following total knee arthroplasty. *The Journal of bone and joint surgery. American volume*, 99(13):1093–1103, Jul 2017. 28678122[pmid].

[47] Hamid R. Darabi, Daniel Tsinis, Kevin Zecchini, Winthrop F. Whitcomb, and Alexander Liss. Forecasting mortality risk for patients admitted to intensive care units using machine learning. *Procedia Computer Science*, 140:306 – 313, 2018. Cyber Physical Systems and Deep Learning Chicago, Illinois November 5-7, 2018.

[48] Allan R. de Caen, Marc D. Berg, Leon Chameides, Cheryl K. Gooden, Robert W. Hickey, Halden F. Scott, Robert M. Sutton, Janice A. Tijssen, Alexis Topjian, Élise W. van der Jagt, Stephen M. Schexnayder, and Ricardo A. Samson. Part 12: Pediatric advanced life support: 2015 american heart association guidelines update for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*, 132(18 Suppl 2):S526–S542, Nov 2015.

[49] Maria Clara de Magalhães-Barbosa, Jaqueline Rodrigues Robaina, Arnaldo Prata-Barbosa, and Claudia de Souza Lopes. Validity of triage systems for paediatric emergency care: a systematic review. *Emergency Medicine Journal*, 34(11):711–719, 2017.

[50] Kumar Dharmarajan, Angela F Hsieh, Zhenqiu Lin, Héctor Bueno, Joseph S Ross, Leora I Horwitz, José Augusto Barreto-Filho, Nancy Kim, Susannah M Bernheim, Lisa G Suter, et al. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *Jama*, 309(4):355–363, 2013.

[51] Molla S Donaldson, Janet M Corrigan, Linda T Kohn, et al. *To err is human: building a safer health system*, volume 6. National Academies Press, 2000.

[52] Patrick Doupe, James Faghmous, and Sanjay Basu. Machine learning for health services researchers. *Value in Health*, 22(7):808 – 815, 2019.

[53] Andrea Freyer Dugas, Thomas D. Kirsch, Matthew Toerper, Fred Korley, Gayane Yenokyan, Daniel France, David Hager, and Scott Levin. An electronic emergency triage system to improve patient distribution by critical outcomes. *The Journal of Emergency Medicine*, 50(6):910 – 918, 2016.

[54] Heather Duncan, James Hutchison, and Christopher S. Parshuram. The pediatric early warning system score: A severity of illness score to predict urgent medical need in hospitalized children. *Journal of Critical Care*, 21(3):271–278, Sep 2006.

[55] Nick Dunn. Practical issues around putting the patient at the centre of care. *Journal of the Royal Society of Medicine*, 96(7):325–327, Jul 2003. 12835443[pmid].

[56] Guillermo Durán, Pablo A. Rey, and Patricio Wolff. Solving the operating room

scheduling problem with prioritized lists of patients. *Annals of Operations Research*, 258(2):395–414, Nov 2017.

[57] Nasim Farrohknia, Maaret Castrén, Anna Ehrenberg, Lars Lind, Sven Oredsson, Håkan Jonsson, Kjell Asplund, and Katarina E. Göransson. Emergency department triage scales and their components: a systematic review of the scientific evidence. *Scandinavian journal of trauma, resuscitation and emergency medicine*, 19:42–42, Jun 2011.

[58] A Fernandez Landaluce, Jose I Pijoan, Santiago Mintegi, and Javier Benito. Evaluacion de la escala canadiense de triaje pediatrico en un servicio de urgencias de pediatria europeo. *Emergencias: Revista de la Sociedad Espanola de Medicina de Urgencias y Emergencias pags*, 22(5):355 – 360, 01 2010.

[59] C. Ferri, J. Hernández-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. *Pattern Recogn. Lett.*, 30(1):27–38, January 2009.

[60] Renee Flippo, Elizabeth NeSmith, Nancy Stark, Thomas Joshua, and Michelle Hoehn. Reduction of 30-day preventable pediatric readmission rates with postdischarge phone calls utilizing a patient- and family-centered care approach. *Journal of Pediatric Health Care*, 29(6):492 – 500, 2015.

[61] Hiraku Funakoshi, Takashi Shiga, Yosuke Homma, Yoshiyuki Nakashima, Jin Takahashi, Hiroshi Kamura, and Masatomi Ikusaka. Validation of the modified japanese triage and acuity scale-based triage system emphasizing the physiologic variables or mechanism of injuries. *International journal of emergency medicine*, 9(1):1, Jan 2016.

[62] Joseph Futoma, Jonathan Morris, and Joseph Lucas. A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, 56:229–238, 2015.

[63] Panagiota Galetsi, Korina Katsaliaki, and Sameer Kumar. Big data analytics in health sector: Theoretical framework, techniques and prospects. *International Journal of Information Management*, 50:206 – 216, 2020.

[64] Sashikumar Ganapathy, Joo Guan Yeo, Xing Hui Michelle Thia, Geok Mei Andrea Hei, and Lai Peng Tham. The singapore paediatric triage scale validation study. *Singapore medical journal*, 59(4):205–209, Apr 2018.

[65] Amit X. Garg, Neill K. J. Adhikari, Heather McDonald, M. Patricia Rosas-Arellano, P. J. Devereaux, Joseph Beyene, Justina Sam, and R. Brian Haynes. Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes A Systematic Review. *JAMA*, 293(10):1223–1238, 03 2005.

[66] Asier Garmendia, Manuel Graña, Jose Manuel Lopez-Guede, and Sebastian Rios. Predicting patient hospitalization after emergency readmission. *Cybernetics and Systems*, 48(3):182–192, 2017.

[67] Asier Garmendia, Manuel Graña, Jose Manuel Lopez Guede, and Sebastian Rios. Neural and statistical predictors for time to readmission in emergency departments: a case

study. *Neurocomputing*, in press, 2019.

[68] James C. Gay, Rishi Agrawal, Katherine A. Auger, Mark A. Del Beccaro, Pirooz Eghtesady, Evan S. Fieldston, Justin Golias, Paul D. Hain, Richard McClead, Rustin B. Morse, Mark I. Neuman, Harold K. Simon, Javier Tejedor-Sojo, Ronald J. Teufel, J. Mitchell Harris, and Samir S. Shah. Rates and impact of potentially preventable readmissions at children's hospitals. *The Journal of Pediatrics*, 166(3):613 – 619.e5, 2015.

[69] Nicki Gilboy, Paula Tanabe, and Debbie A. Travers. The emergency severity index version 4: Changes to esi level 1 and pediatric fever criteria. *Journal of Emergency Nursing*, 31(4):357–362, 2005.

[70] S W Goodacre, M Gillett, R D Harris, and K P Houlihan. Consistency of retrospective triage decisions as a standardised instrument for audit. *Journal of accident and emergency medicine*, 16(5):322–324, 1999.

[71] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. urlhttp://www.deeplearningbook.org.

[72] Kenneth W. Goodman. *Ethical and Legal Issues in Decision Support*, pages 131–146. Springer International Publishing, Cham, 2016.

[73] Tadahiro Goto, Jr Camargo, Carlos A., Mohammad Kamal Faridi, Robert J. Freishtat, and Kohei Hasegawa. Machine Learning–Based Prediction of Clinical Outcomes for Children During Emergency Department Triage Machine Learning–Based Prediction of Pediatric Outcomes in Emergency Department Triage Machine Learning–Based Prediction of Pediatric Outcomes in Emergency Department Triage. *JAMA Network Open*, 2(1):e186937–e186937, 01 2019.

[74] Serge Gouin, Jocelyn Gravel, Devendra K. Amre, and Sylvie Bergeron. Evaluation of the paediatric canadian triage and acuity scale in a pediatric ed. *American Journal of Emergency Medicine*, 23(3):243–247, 2005.

[75] Jocelyn Gravel, Eleanor Fitzpatrick, Serge Gouin, Kelly Millar, Sarah Curtis, Gary Joubert, Kathy Boutis, Chantal Guimont, Ran D. Goldman, Alexander S. Dubrovsky, and et al. Performance of the canadian triage and acuity scale for children: A multicenter database study. *Annals of Emergency Medicine*, 61(1):27–32.e3, 2013.

[76] Jocelyn Gravel, Serge Gouin, Ran D. Goldman, Martin H. Osmond, Eleanor Fitzpatrick, Kathy Boutis, Chantal Guimont, Gary Joubert, Kelly Millar, Sarah Curtis, and et al. The canadian triage and acuity scale for children: A prospective multicenter evaluation. *Annals of Emergency Medicine*, 60(1):71–77, 2012.

[77] Jocelyn Gravel, Sergio Manzano, and Michael Arsenault. Validity of the canadian paediatric triage and acuity scale in a tertiary care hospital. *CJEM*, 11 1:23–8, 2009.

[78] Robert A. Greenes, David W. Bates, Kensaku Kawamoto, Blackford Middleton, Jerome Osheroff, and Yuval Shahar. Clinical decision support models and frameworks: Seeking

to address research issues underlying implementation successes and failures. *Journal of Biomedical Informatics*, 78:134 – 143, 2018.

[79] Cosima Gretton. *Trust and Transparency in Machine Learning-Based Clinical Decision Support*, pages 279–292. Springer International Publishing, Cham, 2018.

[80] Peter Groves, Basel Kayyali, David Knott, and Steve Van Kuiken. The "big data" revolution in healthcare: accelerating value and innovation. *McKinsey Global Institute*, (January):1–22, 2013.

[81] H. Haas. Outils de triage aux urgences pédiatriques. *Archives de Pédiatrie*, 12(6):703–705, Jun 2005.

[82] Jakob Håkansson Eklund, Inger K. Holmström, Tomas Kumlin, Elenor Kaminsky, Karin Skoglund, Jessica Höglander, Annelie J. Sundler, Emelie Condén, and Martina Summer Meranius. "same same or different?.ª review of reviews of person-centered and patient-centered care. *Patient Education and Counseling*, 102(1):3–11, 2019.

[83] Bhakti Hansoti, Alexander Jenson, Devin Keefe, Sarah Stewart De Ramirez, Trisha Anest, Michelle Twomey, Katie Lobner, Gabor Kelen, and Lee Wallis. Reliability and validity of pediatric triage tools evaluated in low resource settings: a systematic review. *BMC Pediatrics*, 17(1):37, Jan 2017.

[84] S Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, 1998.

[85] Rebecca Hermon and Patricia Williams. Big data in healthcare: What is it used for. *Proceedings of the 3rd Australian eHealth Informatics and Security Conference*, page 40–49, 2014.

[86] Jeremiah S. Hinson, Diego A. Martinez, Paulo S. K. Schmitz, Matthew Toerper, Danieli Radu, James Scheulen, Sarah A. Stewart de Ramirez, and Scott Levin. Accuracy of emergency department triage using the emergency severity index and independent predictors of under-triage and over-triage in brazil: a retrospective cohort analysis. *International Journal of Emergency Medicine*, 11(1):3, Jan 2018.

[87] Meng Hsuen Hsieh, Meng Ju Hsieh, Chin-Ming Chen, Chia-Chang Hsieh, Chien-Ming Chao, and Chih-Cheng Lai. Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units. *Scientific Reports*, 8(1):17116, 2018.

[88] Claus Sixtus Jensen, Hanne Vebert Olesen, Hanne Aagaard, Marie Louise Overgaard Svendsen, and Hans Kirkegaard. Comparison of two pediatric early warning systems: A randomized trial. *Journal of Pediatric Nursing*, 44:e58–e65, Jan 2019.

[89] Berry JG, Hall DE, Kuo DZ, and et al. Hospital utilization and characteristics of patients experiencing recurrent readmissions within children's hospitals. *JAMA*, 305(7):682–690, 2011.

[90] Cristian Julio, Patricio Wolff, and María Vegoña Yarza. Modelo de gestión de listas de

espera centrado en oportunidad y justicia. *Revista médica de Chile*, 144:781 – 787, 06 2016.

[91] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission. *JAMA*, 306(15):1688, Oct 2011.

[92] Marvin Karson. Handbook of methods of applied statistics. volume i: Techniques of computation descriptive methods, and statistical inference. volume ii: Planning of surveys and experiments. i. m. chakravarti, r. g. laha, and j. roy, new york, john wiley; 1967, $9.00. *Journal of the American Statistical Association*, 63(323):1047–1049, 1968.

[93] Harsheen Kaur, James M. Naessens, Andrew C. Hanson, Karen Fryer, Michael E. Nemergut, and Sandeep Tripathi. Proper: Development of an early pediatric intensive care unit readmission risk prediction tool. *Journal of Intensive Care Medicine*, 33(1):29–36, Jan 2018.

[94] Thomas Kautz, Bjoern M. Eskofier, and Cristian F. Pasluosta. Generic performance measure for multiclass-classifiers. *Pattern Recognition*, 68:111–125, 2017. Exported from https://app.dimensions.ai on 2018/11/19.

[95] Mohamed Khalifa. Clinical decision support: Strategies for success. *Procedia Computer Science*, 37:422–427, 2014.

[96] Jong Taek Kim. Application of machine and deep learning algorithms in intelligent clinical decision support systems in healthcare. *Journal of Health & Medical Informatics*, 09(05), 2018.

[97] Patricia Kipnis, Benjamin J. Turk, David A. Wulf, Juan Carlos LaGuardia, Vincent Liu, Matthew M. Churpek, Santiago Romero-Brufau, and Gabriel J. Escobar. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the icu. *Journal of Biomedical Informatics*, 64:10–19, 2016.

[98] Anton Kocheturov, Panos M. Pardalos, and Athanasia Karakitsiou. Massive datasets and machine learning for computational biomedicine: trends and challenges. *Annals of Operations Research*, 276(1):5–34, May 2019.

[99] Joon-myoung Kwon, Youngnam Lee, Yeha Lee, Seungwoo Lee, Hyunho Park, and Jinsik Park. Validation of deep-learning-based triage and acuity score using a large national dataset. *PLOS ONE*, 13(10):1–10, 10 2018.

[100] Chiozza Maria Laura and Plebani Mario. *cclm*, volume 44, chapter Clinical Governance: from clinical risk management to continuous quality improvement, page 694. 2019 2006. 6.

[101] Marie Danielle Le Lagadec and Trudy Dwyer. Scoping review: The use of early warning systems for the identification of in-hospital patients at risk of deterioration. *Australian Critical Care*, 30(4):211–218, 2017.

[102] Lucian L Leape and Donald M Berwick. Five years after to err is human: what have we learned? *Jama*, 293(19):2384–2390, 2005.

[103] Soo-Kyoung Lee, Youn-Jung Son, Jeongeun Kim, Hong-Gee Kim, Jae-Il Lee, Bo-Yeong Kang, Hyeon-Sung Cho, and Sungin Lee. Prediction model for health-related quality of life of elderly with chronic diseases using machine learning techniques. *Healthcare informatics research*, 20(2):125–134, Apr 2014.

[104] Scott Levin, Matthew Toerper, Eric Hamrock, Jeremiah S. Hinson, Sean Barnes, Heather Gardner, Andrea Dugas, Bob Linton, Tom Kirsch, and Gabor Kelen. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Annals of Emergency Medicine*, 71(5):565 – 574.e2, 2018.

[105] Wen-tsann Lin, Yung-tsan Jou, Yih-chuan Wu, and Yuan-du Hsiao. Data mining applied to the predictive model of triage system in emergency department. 7(6):834–841, 2013.

[106] W.T. Lin, Y.C. Wu, J.S. Zheng, and M.Y. Chen. Analysis by data mining in the emergency medicine triage database at a taiwanese regional hospital. *Expert Systems with Applications*, 38(9):11078–11084, Sep 2011.

[107] Joseph Liu, Jeremy C. Wyatt, and Douglas G. Altman. Decision tools in health care: focus on the problem, not the solution. *BMC Medical Informatics and Decision Making*, 6(1):4, 2006.

[108] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013.

[109] Christopher A. Lovejoy, Varun Buch, and Mahiben Maruthappu. Artificial intelligence in the intensive care unit. *Critical care (London, England)*, 23(1):7–7, Jan 2019.

[110] Spyros Makridakis, Richard Kirkham, Ann Wakefield, Maria Papadaki, Joanne Kirkham, and Lisa Long. Forecasting, uncertainty and risk; perspectives on clinical decision-making in preventive and curative medicine. *International Journal of Forecasting*, 35(2):659 – 666, 2019.

[111] Peter A. Maningas, Derek A. Hime, and Donald E. Parker. The use of the soterion rapid triage system in children presenting to the emergency department. *Journal of Emergency Medicine*, 31(4):353–359, 2006.

[112] Mary C. McLellan and Jean A. Connor. The cardiac children's hospital early warning score (c-chews). *Journal of Pediatric Nursing*, 28(2):171–178, Apr 2013.

[113] Jefferson E. McMillan, Emily R. Meier, Jeffrey C. Winer, Megan Coco, Mary Daymont, Sierra Long, and Brian R. Jacobs. Clinical and geographic characterization of 30-day readmissions in pediatric sickle cell crisis patients. *Hospital Pediatrics*, 5(8):423–431, 2015.

[114] Stephanie Medlock, Jeremy C Wyatt, Vimla L Patel, Edward H Shortliffe, and Ameen Abu-Hanna. Modeling information flows in clinical decision support: key insights for enhancing system effectiveness. *Journal of the American Medical Informatics Association*, 23(5):1001–1006, 02 2016.

[115] Anthony F Milano. Evidence-based risk assessment. *JOURNAL OF INSURANCE MEDICINE-NEW YORK-*, 33(3):239–250, 2001.

[116] Alison H. Miles, Michael C. Spaeder, and David C. Stockwell. Unplanned icu transfers from inpatient units: Examining the prevalence and preventability of adverse events associated with icu transfer in pediatrics. *Journal of pediatric intensive care*, 5(1):21–27, Mar 2016.

[117] JC Misson. A review of clinical risk management. *Journal of quality in clinical practice*, 21(4):131—134, December 2001.

[118] Henriëtte A. Moll. Challenges in the validation of triage systems at emergency departments. *Journal of Clinical Epidemiology*, 63(4):384–388, 2010.

[119] Rustin B. Morse, Matthew Hall, Evan S. Fieldston, Denise M. Goodman, Jay G. Berry, James C. Gay, Marion R. Sills, Rajendu Srivastava, Gary Frank, Paul D. Hain, and Samir S. Shah. Children's hospitals with shorter lengths of stay do not have higher readmission rates. *The Journal of Pediatrics*, 163(4):1034 – 1038.e1, 2013.

[120] Naiara Muro, Eider Sanchez, Carlos Toro, Eduardo Carrasco, Sebastián A. Ríos, Frank Guijarro, and Manuel Graña. Experience-based electronic health records. *Cybernetics and Systems*, 47(1-2):126–139, 2016.

[121] Mari M. Nakamura, Sara L. Toomey, Alan M. Zaslavsky, Jay G. Berry, Scott A. Lorch, Ashish K. Jha, Maria C. Bryant, Alexandra T. Geanacopoulos, Samuel S. Loren, Debanjan Pain, and Mark A. Schuster. Measuring pediatric hospital readmission rates to drive quality improvement. *Academic Pediatrics*, 14(5, Supplement):S39 – S46, 2014. Advances in Children's Healthcare Quality: The Pediatric Quality Measures Program.

[122] Ziad Obermeyer and Ezekiel J. Emanuel. Predicting the future - big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216–1219, Sep 2016. 27682033[pmid].

[123] Royal College of Physicians. *National Early Warning Score (NEWS) - Standardising the assessment of acute-illness severity in the NHS. Report of a working party*. Number July. 2012.

[124] Pia Olofsson, Martin Gellerstedt, and Eric D. Carlstrom. Manchester triage in sweden - interrater reliability and accuracy. *International Emergency Nursing*, 17(3):143–148, 2009.

[125] Jerry Osheroff, Jonathan Teich, Donald Levick, Luis Saldana, Ferdinand Velasco, Dean Sittig, Kendall Rogers, and Robert Jenders. *Improving outcomes with clinical decision support: an implementer's guide*. HIMSS Publishing, 2012.

[126] Kenneth J Ottenbacher, Pam M Smith, Sandra B Illig, Richard T Linn, Roger C Fiedler, and Carl V Granger. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. *Journal of Clinical Epidemiology*, 54(11):1159 – 1165, 2001.

[127] Harry Otway and Detlof von Winterfeldt. Expert judgment in risk analysis and management: process, context, and pitfalls. *Risk analysis*, 12(1):83–93, 1992.

[128] Ambarish Pandey, Harsh Golwala, Haolin Xu, Adam D. DeVore, Roland Matsouaka, Michael Pencina, Dharam J. Kumbhani, Adrian F. Hernandez, Deepak L. Bhatt, Paul A. Heidenreich, Clyde W. Yancy, James A. de Lemos, and Gregg C. Fonarow. Association of 30-day readmission metric for heart failure under the hospital readmissions reduction program with quality of care and outcomes. *JACC: Heart Failure*, 4(12):935 – 946, 2016.

[129] Kavita Parikh, Jay Berry, Matt Hall, Grant M. Mussman, Amanda Montalbano, Joanna Thomson, Rustin Morse, Karen M. Wilson, and Samir S. Shah. Racial and ethnic differences in pediatric readmissions for common chronic conditions. *The Journal of Pediatrics*, 186:158 – 164.e1, 2017.

[130] Christopher S Parshuram, James Hutchison, and Kristen Middaugh. Development and initial validation of the bedside paediatric early warning system score. *Critical Care*, 13(4):R135, 2009.

[131] Vimla L. Patel, Lily A. Gutnik, Daniel R. Karlin, and Martin Pusic. Calibrating urgency: Triage decision-making in a pediatric emergency department. *Advances in Health Sciences Education*, 13(4):503–520, 2008.

[132] Romain Pirracchio, Maya L Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, and Mark J van der Laan. Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study. *The Lancet Respiratory Medicine*, 3(1):42–52, Jan 2015.

[133] Luciano M. Prevedello, Ali S. Raja, Ivan K. Ip, Aaron Sodickson, and Ramin Khorasani. Does clinical decision support reduce unwarranted variation in yield of ct pulmonary angiogram? *The American journal of medicine*, 126(11):975–981, Nov 2013.

[134] Peter J Pronovost, James I Cleeman, Donald Wright, and Arjun Srinivasan. Fifteen years after to err is human: a success story to learn from. *BMJ Quality & Safety*, 25(6):396–399, 2016.

[135] David R. Prytherch, Gary B. Smith, Paul E. Schmidt, and Peter I. Featherstone. Viewstowards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, 81(8):932–937, 2010.

[136] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson,

Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018.

[137] Bhargava K Reddy and Dursun Delen. Predicting hospital readmission for lupus patients: An rnn-lstm-based deep-learning methodology. *Computers in Biology and Medicine*, 101:199 – 209, 2018.

[138] A. Núñez Reiz, M.A. Armengol de la Hoz, and M. Sánchez García. Big data analysis and machine learning in intensive care units. *Medicina Intensiva (English Edition)*, 2019.

[139] Beatriz Remeseiro and Veronica Bolon-Canedo. A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, 112:103375, 2019.

[140] Michael J Rothman, Steven I Rothman, and Joseph Beals. Development and validation of a continuous measure of patient condition using the electronic medical record. *J. Biomed. Inform.*, 46(5):837–848, 2013.

[141] J Roukema, E W Steyerberg, A van Meurs, M Ruige, J van der Lei, and H A Moll. Validity of the manchester triage system in paediatric emergency care. *Emergency Medicine Journal*, 23(12):906–910, 2006.

[142] Jonathan Rubin, Cristhian Potes, Minnan Xu-Wilson, Junzi Dong, Asif Rahman, Hiep Nguyen, and David Moromisato. An ensemble boosting model for predicting transfer to the pediatric intensive care unit. *International Journal of Medical Informatics*, 112:15 – 20, 2018.

[143] Jonathan Rubin, Cristhian Potes, Minnan Xu-Wilson, Junzi Dong, Asif Rahman, Hiep Nguyen, and David Moromisato. An ensemble boosting model for predicting transfer to the pediatric intensive care unit. *International Journal of Medical Informatics*, 112:15 – 20, 2018.

[144] Sebastian Ruder. An overview of gradient descent optimization algorithms. Technical report, arXiv:1609.04747, 2016.

[145] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 08 2007.

[146] Maria J. Santana, Kimberly Manalili, Rachel J. Jolley, Sandra Zelinsky, Hude Quan, and Mingshan Lu. How to practice person-centred care: A conceptual framework. *Health expectations : an international journal of public participation in health care and health policy*, 21(2):429–440, Apr 2018.

[147] Shihab Sarwar, Anglin Dent, Kevin Faust, Maxime Richer, Ugljesa Djuric, Randy Van Ommeren, and Phedias Diamandis. Physician perspectives on integration of artificial intelligence into diagnostic pathology. *npj Digital Medicine*, 2(1):28, 2019.

[148] S Hanumanth Sastry and Prof M S Prasada Babu. Implementation of crisp methodology for erp systems. *International Journal of Computer Science Engineering (IJCSE)*, 2(5):203–217, 2013.

[149] Gordon D Schiff, Lynn A Volk, Mayya Volodarskaya, Deborah H Williams, Lake Walsh, Sara G Myers, David W Bates, and Ronen Rozenblum. Screening for medication errors using an outlier detection system. *Journal of the American Medical Informatics Association*, 24(2):281–287, 01 2017.

[150] B. Schölkopf. Learning with kernels. *Journal of the Electrochemical Society*, 129(November):2865, 2002.

[151] Khader Shameer, Kipp W Johnson, Alexandre Yahi, Riccardo Miotto, Li Li, Doran Ricks, Jebakumar Jebakaran, Patricia Kovatch, Partho P. Sengupta, Sengupta Gelijns, Alan Moskovitz, Bruce Darrow, David L David, Andrew Kasarkis, Nicholas P Tatonetti, Sean Pinney, and Joel T Dudley. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: A case-study using mount sinai heart failure cohort. In *Pacific Symposium on Biocomputing 2017*, pages 276–287, 2017.

[152] Ida Sim and Amy Berlin. A framework for classifying decision support systems. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, (Figure 1):599–603, 2003. taxonomia de berlin.

[153] Ida Sim and Amy Berlin. A framework for classifying decision support systems. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2003:599–603, 2003.

[154] M.C. Emre Simsekler, Ayse P. Gurses, Brian E. Smith, and Al Ozonoff. Integration of multiple methods in identifying patient safety risks. *Safety Science*, 118:530 – 537, 2019.

[155] Balbir Singh and M Habeeb Ghatala. Risk management in hospitals. *International journal of innovation, management and technology*, 3(4):417, 2012.

[156] Gary B. Smith, David R. Prytherch, Paul Meredith, Paul E. Schmidt, and Peter I. Featherstone. The ability of the national early warning score (news) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, 2013.

[157] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427 – 437, 2009.

[158] Mahsa Dehghani Soufi, Taha Samad-Soltani, Samad Shams Vahdati, and Peyman Rezaei-Hachesu. Decision support system for triage management: A hybrid approach using rule-based reasoning and fuzzy logic. *International Journal of Medical Informatics*, 114:35 – 44, 2018.

[159] Mihaela S Stefan, Penelope S Pekow, Wato Nsa, Aruna Priya, Lauren E Miller, Dale W Bratzler, Michael B Rothberg, Robert J Goldberg, Kristie Baus, and Peter K Lindenauer. Hospital performance measures and 30-day readmission rates. *Journal of*

general internal medicine, 28(3):377–385, 2013.

[160] C P Subbe, M Kruger, P Rutherford, and L Gemmel. Validation of a modified early warning score in medical admissions. *QJM: monthly journal of the Association of Physicians*, 94(10):521–6, Oct 2001.

[161] Felipe Tabares, Jhonatan Hernandez, and Ivan Cabezas. Architectural design of a clinical decision support system for clinical triage in emergency departments. In Andrés Solano and Hugo Ordoñez, editors, *Advances in Computing*, pages 267–281, Cham, 2017. Springer International Publishing.

[162] Douglas A. Talbert, Matt Honeycutt, and Steve Talbert. A machine learning and data mining framework to enable evolutionary improvement in trauma triage. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 348–361, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[163] Suzanne Tamang, Arnold Milstein, Henrik Toft Sørensen, Lars Pedersen, Lester Mackey, Jean-Raymond Betterton, Lucas Janson, and Nigam Shah. Predicting patient 'cost blooms' in denmark: a longitudinal population-based study. *BMJ Open*, 7(1), 2017.

[164] Lionel Tarassenko, David A. Clifton, Michael R. Pinsky, Marilyn T. Hravnak, John R. Woods, and Peter J. Watkinson. Centile-based early warning scores derived from statistical distributions of vital signs. *Resuscitation*, 82(8):1013 – 1018, 2011.

[165] G. E. Thomsen, D. Pope, T. D. East, A. H. Morris, A. T. Kinder, D. A. Carlson, G. L. Smith, C. J. Wallace, J. F. Jr Orme, and T. P. Clemmer. Clinical performance of a rule-based decision support system for mechanical ventilation of ards patients. *Proceedings. Symposium on Computer Applications in Medical Care*, pages 339–343, 1993.

[166] Truyen Tran, Wei Luo, Dinh Phung, Sunil Gupta, Santu Rana, Richard Lee Kennedy, Ann Larkins, and Svetha Venkatesh. A framework for feature extraction from hospital medical data with applications in risk prediction. *BMC Bioinformatics*, 15(1):1–9, 2014.

[167] Tjeerd van der Ploeg, Peter C Austin, and Ewout W Steyerberg. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC medical research methodology*, 14(1):137, 2014.

[168] Ineke van der Wulp and Henk F. van Stel. Adjusting weighted kappa for severity of mistriage decreases reported reliability of emergency department triage systems: a comparative study. *Journal of Clinical Epidemiology*, 62(11):1196–1201, Nov 2009.

[169] Vladimir N Vapnik. Statistical learning theory. *Adaptive and learning Systems for Signal Processing, Communications and Control*, 2:1–740, 1998.

[170] Charles Vincent. *Patient safety*. John Wiley & Sons, 2011.

[171] Milan Vukicevic, Sandro Radovanovic, Ana Kovacevic, Gregor Stiglic, and Zoran Obradovic. *Improving Hospital Readmission Prediction Using Domain Knowledge Based*

Virtual Examples, pages 695–706. 2015.

[172] Shen-Tsu Wang. Construct an optimal triage prediction model: A case study of the emergency department of a teaching hospital in taiwan. *Journal of Medical Systems*, 37(5):9968, Aug 2013.

[173] Peter J. Watkinson, Marco A.F. Pimentel, David A. Clifton, and Lionel Tarassenko. Manual centile-based early warning scores derived from statistical distributions of observational vital-sign data. *Resuscitation*, 129:55 – 60, 2018.

[174] Karl E Weick and Kathleen M Sutcliffe. *Managing the unexpected: Resilient performance in an age of uncertainty*, volume 8. John Wiley & Sons, 2011.

[175] Gary M. Weiss. Mining with rarity: A unifying framework. *SIGKDD Explor. Newsl.*, 6(1):7–19, June 2004.

[176] Stephen F. Weng, Luis Vaz, Nadeem Qureshi, and Joe Kai. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PloS one*, 14(3):e0214365–e0214365, Mar 2019.

[177] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[178] Patricio Wolff, Manuel Graña, Sebastián A. Ríos, and Maria Begoña Yarza. Machine Learning Readmission Risk Modeling: A Pediatric Case Study. *BioMed Research International*, 2019:9, 2019.

[179] Patricio Wolff, Manuel Graña, Sebastián A. Ríos, and María Begoña Yarza. RapidMiner code for a pediatric case of readmission risk modeling. https://doi.org/10.5281/zenodo.2597686, March 2019.

[180] Patricio Wolff, Sebastián A. Ríos, and Manuel Graña. Setting up standards: A methodological proposal for pediatric triage machine learning model construction based on clinical outcomes. *Expert Systems with Applications*, 138:112788, 2019.

[181] Cao Xiao, Tengfei Ma, Adji B. Dieng, David M. Blei, and Fei Wang. Readmission prediction via deep contextual embedding of clinical concepts. *PLOS ONE*, 13(4):1–15, 04 2018.

[182] Juri Yanase and Evangelos Triantaphyllou. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, page 112821, 2019.

[183] Shun Yu, Sharon Leung, Moonseong Heo, Graciela J Soto, Ronak T Shah, Sampath Gunda, and Michelle Gong. Comparison of risk prediction scoring systems for ward patients: a retrospective nested case-control study. *Critical Care*, 18(3):R132, 2014.

[184] Joany M. Zachariasse, Daan Nieboer, Rianne Oostenbrink, Henriëtte A. Moll, and Ewout W. Steyerberg. Multiple performance measures are needed to evaluate triage systems in the emergency department. *Journal of Clinical Epidemiology*, 94:27–34, Feb 2018.

[185] Syed Nabeel Zafar, Adil A. Shah, Christine Nembhard, Lori L. Wilson, Elizabeth B. Habermann, Mustafa Raoof, and Nabil Wasif. Readmissions after complex cancer surgery: Analysis of the nationwide readmissions database. *Journal of Oncology Practice*, 14(6):e335–e345, 2018. PMID: 29894662.

[186] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

[187] J. Zhang, X. Wu, and V. S. Shengs. Active learning with imbalanced multiple noisy labeling. *IEEE Transactions on Cybernetics*, 45(5):1095–1107, May 2015.

[188] L. Zhao, G. Sukthankar, and R. Sukthankar. Incremental relabeling for active learning with noisy crowdsourced annotations. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 728–733, Oct 2011.

[189] Bichen Zheng, Jinghe Zhang, Sang Won Yoon, Sarah S. Lam, Mohammad Khasawneh, and Srikanth Poranki. Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, 42(20):7110 – 7120, 2015.

[190] Dror Zmiri, Yuval Shahar, and Meirav Taieb-Maimon. Classification of patients by severity grades during triage in the emergency department using data mining methods. *Journal of Evaluation in Clinical Practice*, 18(2):378–388, 2010.

Annex A

Model Hyperparameter sensitivity analysis in Machine learning readmission risk modeling: a pediatric case study

Machine learning readmission risk modeling: a pediatric case study objective is to assess the all cause readmission predictive performance achieved by Machine Learning techniques in the emergency department of a pediatric hospital in Santiago, Chile. In this annex we present additional results to Machine learning (ML) readmission risk modeling: a pediatric case study. In particular, the results of testing different hiperparameters are presented in all ML models used for the calculation of risk of readmissions in a pediatric hospital. We report classification results achieved with various model building approaches after data curation and preprocessing for correction of class imbalance.

A.1. Describing the models used

The ML models selected in that publication were: Naive Bayes, Support Vector Machines, and Multilayer Perceptron. The models are evaluated according to the characteristics presented in the table A.1. We compute repeated cross-validation (RCV) with 5 folders to assess performance of this models.We apply a SMOTE up-sampling procedure using the five nearest neighbors of each minority class sample. The reported results are the average (and standard deviation) of the test RCV results.

A.2. Results

The evaluation was made based on: AUC, recall, PPV and f1-score. the results obtained are shown in the table A.2

ML Alg.	**Descrip.**	**Descrip. Cont**	**Name**
Multilayer Perceptron 1	4 Esemble MLPs	10 cycles, 10 gen.	MLP 1-1
		100 cycles, 10 gen.	MLP 1-2
		10 cycles, 100 gen.	MLP 1-3
	8 Esemble MLPs	10 cycles, 10 gen.	MLP 1-4
		100 cycles, 10 gen.	MLP 1-5
		10 cycles, 100 gen.	MLP 1-6
Multilayer Perceptron 2	Rectifier Act. Func.	Neurons 50-25-5	MLP 2-1
		Neurons 50-50-5	MLP 2-2
		Neurons 50-50-25-5	MLP 2-3
	Tanh Act. Func.	Neurons 50-25-5	MLP 2-4
		Neurons 50-50-5	MLP 2-5
		Neurons 50-50-25-5	MLP 2-6
Naive Bayes	Gaussian d-estimator		NB1
	Greedy d-estimator	10 kernels	NB2
		100 kernels	NB3
		1000 kernels	NB4
	full d-estimator	Heuristic	NB5
		Fix	NB6
Support Vector Machine	RBF kernel	$C = 0{,}1$ & $\gamma = 0$	SVM1
		$C = 0{,}1$ & $\gamma = 0{,}1$	SVM2
		$C = 10$ & $\gamma = 0$	SVM3
		$C = 10$ & $\gamma = 0{,}1$	SVM4
	Linear kernel	$C = 0{,}1$	SVM5
		$C = 10$	SVM6
		$C = 1$	SVM7
		$C = 0$	SVM8

Tabla A.1: Machine Learning studied models and their parameter setup

	AUC (SD)	recall (SD) [%]	f-measure (SD) [%]	PPV (SD) [%]
MLP 1-1	0.634 (0.011)	61.39 (6.14)	9.67 (0.26)	5.25 (0.14)
MLP 1-2	0.625 (0.004)	50.85 (5.93)	10.05 (0.32)	5.59 (0.24)
MLP 1-3	0.618 (0.170)	47.77 (5.45)	10.03 (0.56)	5.61 (0.32)
MLP 1-4	0.634 (0.008)	52.9 (4.88)	10.28 (0.51)	5.70 (0.32)
MLP 1-5	0.617 (0.011)	48.57(3.11)	10.20 (0.40)	5.70 (0.29)
MLP 1-6	0.620 (0.011)	40.84(3.76)	10.39 (0.37)	5.96 (0.19)
MLP2-1	0.553 (0.033)	96.35 (3)	7.57 (0.24)	3.94 (0.13)
MLP2-2	0.530 (0.054)	86.77 (22.47)	7.28 (0.39)	3.81 (0.16)
MLP2-3	0.612 (0.057)	71.36 (35.78)	8.59 (0.00)	4.57 (0.00)
MLP2-4	0.602 (0.027)	92.78 (3.56)	8.01 (0.39)	4.19 (0.22)
MLP2-5	0.593 (0.032)	91.03 (5.2)	8.21 (0.53)	4.30 (0.30)
MLP2-6	0.654 (0.004)	91.31 (2.02)	8.60 (0.35)	4.51 (0.19)
NB1	0.654 (0.015)	69.19 (4.98)	9.82 (0.38)	5.29 (0.22)
NB2	0.655 (0.014)	34.53 (12.53)	12.34 (1.31)	7.89 (1.33)
NB3	0.667 (0.011)	42.78 (3.43)	11.87 (0.39)	6.90 (0.28)
NB4	0.666 (0.006)	7.77 (0.72)	7.78 (0.66)	7.77 (0.72)
NB5	0.669 (0.015)	76.73 (5.3)	9.80 (0.49)	5.24 (0.30)
NB6	0.485 (0.018)	24.83 (1.77)	6.55 (0.52)	3.77 (0.31)
SVM1	0.607 (0.014)	39.13 (2.59)	11.25 (0.66)	6.57 (0.38)
SVM2	0.469 (0.015)	99.9 (0.12)	7.41 (0.02)	3.85 (0.01)
SVM3	0.572 (0.015)	42.31 (2.28)	9.19 (0.48)	5.16 (0.27)
SVM4	0.465 (0.011)	68.09 (2.17)	6.32 (0.21)	3.31 (0.11)
SVM5	0.590 (0.076)	80.39 (4.13)	8.85 (1.04)	4.68 (0.57)
SVM6	0.587 (0.071)	62.74 (16.14)	8.75 (1.35)	4.78 (0.91)
SVM7	0.524 (0.080)	59.59 (8.12)	7.60 (1.38)	4.06 (0.75)
SVM8	0.534 (0.125)	62.36 (24.8)	7.37 (2.32)	3.92 (1.21)

Tabla A.2: Machine Learning studied models Results

Annex B

Extension of Machine learning readmission risk modeling: a pediatric case study

The idea of this annex is to extend the results of the research presented in [178], testing other supervised learning methods well studied in the scientific literature. In [178] the results of the prediction of hospital readmission in pediatric patients were presented, using Support Vector Machine (SVM), Naive Bayes (NB) and Artificial Neural Networks (ANN). In this study, class balancing and 5-fold Cross-validation techniques were used, finding that the best AUC ($p < 0{,}001$) was obtained with the Naive Bayes approach (0,655).

In this annex, two models based on Decision Trees (Random Forest and Gradient Boosted Trees) and one based on logistic regression were trained and tested. In addition, the same dataset, performance metrics and tools (SMOTE and 5-fold Cross-validation) described in the previous work were used

Table B.1 shows the results obtained with the 3 models used in this work and including the best of the results shown in the previous work.

Tabla B.1: Results

	AUC (SD)	recall (SD) [%]	f-score (SD) [%]	PPV (SD) [%]
Prev. work [178]	0.653 (0.014)	69.80 (4.97)	9.83 (0.53)	5.29 (0.31)
Random Forest	0.683 (0.009)	24.07 (1.67)	11.94 (0.9)	7.94 (0.62)
Gradient Boosted Trees	0.682 (0.013)	57.65 (5.52)	11.56 (0.6)	6.43 (0.31)
Logistic Regression	0.668 (0.010)	68.90 (1.65)	10.27 (0.17)	5.55 (0.09)

In figure B.1 the ROC curve obtained by the 3 models presented in this annex is presented, in addition to the best result obtained in the previous work, which corresponds to Naive Bayes.

The AUC values obtained in this annex for RF show a higher classification behavior (Pairwise t-test $p < 0.004$) than the result obtained in the previous work. In the case of

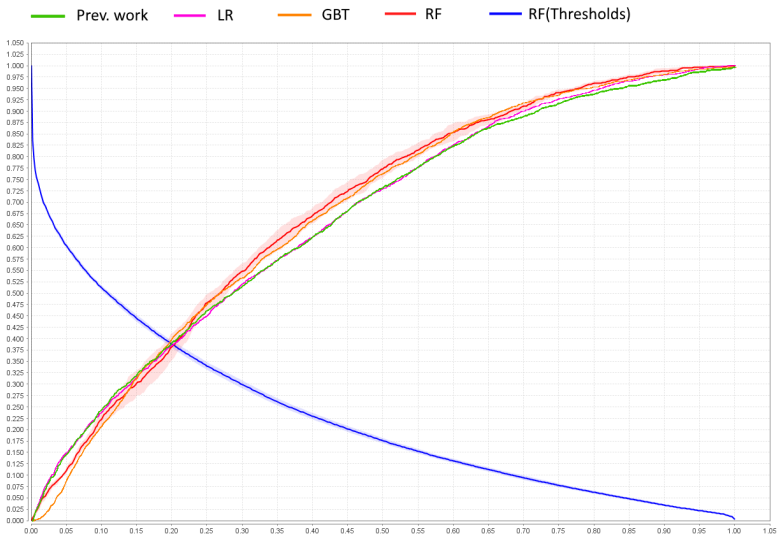

Figura B.1: Comparison of ROC curves

GBT the result in AUC is higher, but less significant (Pairwise t-test $p < 0.008$), due to the variability of the result expressed in a larger standard deviation. Both RF and GBT show better results ($p < 0.025$ and $p < 0.07$) in AUC compared to LR. On the other hand, the result obtained in PPV shows that the RF model is significantly superior to the result shown by NB ($p < 0.0002$) and GBT ($p < 0.003$).

The ROC curve graphically shows a slightly higher result in RF and GBT models with respect to the ROC curve previously obtained for readmission prediction. The ROC curves presented are above the best ROC curve presented previously. This intrinsically shows that the results based on decision trees in this particular problem show better behavior than SVM and ANN.