Expert Systems With Applications 223 (2023) 119864

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

ML models for severity classification and length-of-stay forecasting in emergency units

Jonathan Moya-Carvajal [a], Francisco Pérez-Galarce [b], Carla Taramasco [c], César A. Astudillo [d], Alfredo Candia-Véjar [e,*]

[a] Magister en Gestión de Operaciones, Faculty of Engineering, Universidad de Talca, Los Niches km. 1, Curicó, Chile
[b] Department of Computer Science, Pontificia Universidad Católica de Chile, Santiago, Chile
[c] Faculty of Engineering, Universidad Andrés Bello, Chile
[d] Department of Computer Science, School of Engineering, Universidad de Talca, Los Niches km. 1, Curicó, Chile
[e] School of Civil Engineering, Universidad Finis Terrae, Avenida Pedro de Valdivia 1509, Providencia, Santiago, Chile

ARTICLE INFO

Keywords:
Length-of-stay prediction
Applied machine learning
Text embeddings
Emergency units
Explanaible artificial intelligence

ABSTRACT

Length-of-stay (LoS) prediction and severity classification for patients in emergency units in a clinic or hospital are crucial problems for public and private health networks. An accurate estimation of these parameters is essential for better planning resources, which are usually scarce. Although it is possible to find several works that propose traditional Machine Learning (ML) models to face these challenges, few works have exploited advances in Natural Language Processing (NLP) on Spanish raw-text vector representations. Consequently, we take advantage of those advances, incorporating sentence embeddings in traditional ML models to improve predictions. Moreover, we apply a strategy based on SHapley Additive exPlanations (SHAP) values to provide explanations for these predictions. The results of our case study demonstrate an increase in the accuracy of the predictions using raw text with a minimum preprocessing. The precision increased by up to 2% in the classification of the patient's post-care destination and by up to 8% in the prediction of LoS in the hospital. This evidence encourages practitioners to use available text to anticipate the patient's need for hospitalization more accurately at the earliest stage of the care process.

1. Introduction

The collapse of emergency units (EUs) is a latent problem for health decision-makers. This issue severely affects the whole health system by reducing the quality of care and patients' welfare and increasing operative costs. Moreover, it has a higher risk of death for the patients (Jo et al., 2015). Specifically, the quality of care decreases exponentially with the waiting time in EUs. This problem has several sources, e.g., limited resources for planning (beds, wards, and teams) and unpredictable demand peaks. Most of them related to an underlying uncertainty for the planning.

The purpose of a EUs is to provide immediate health care to any patient who consults spontaneously or is referred by another facility in the network. Typically, an EU care team diagnoses and treats people at risk of life-threatening, chronic disease decompensation, relieving pain, and treating emergencies that cannot be postponed. They decide which patients continue their treatment as inpatients or outpatients.

In emergency units, patients' total LoS is composed of the time in care and the waiting time between stages of the care process. Within these waiting times, both the waiting time for medical care and the waiting time for hospitalization is the most critical. In this research, we focus on the latter. Long waiting times for hospitalization beds in the EU come from three leading causes: (i) the low rotation and availability of hospital beds, forcing staff to use beds and stretchers to observe emergency patients (Hoot & Aronsky, 2008) (when a patient requires observation), (ii) the high demand for hospital beds, which patients additionally use from different sources, such as patients referred from the EU, other units within the same hospital, and another hospital in the same healthcare network and outpatients (Marfil-Garza et al., 2018). The high demand for beds is due, in general, to the increase of chronic diseases in the country and, in the last two years, to the rise of communicable diseases (e.g., COVID19). The origin of the patients does not imply an increase in demand. (iii) the delay in the medical discharge of a hospitalized patient may affect the availability of hospital beds. In summary, all the causes mentioned above generate a low availability of hospital beds and, consequently, a delay in the hospitalization of

* Corresponding author.
E-mail addresses: jmoya@ucm.cl (J. Moya-Carvajal), fjperez10@uc.cl (F. Pérez-Galarce), carla.taramasco@unab.cl (C. Taramasco), castudillo@utalca.cl (C.A. Astudillo), acandia@uft.cl (A. Candia-Véjar).

https://doi.org/10.1016/j.eswa.2023.119864
Received 22 July 2022; Received in revised form 10 March 2023; Accepted 10 March 2023
Available online 15 March 2023

Nomenclature

LoS	Length-of-stay
ML	Machine Learning
NLP	Natural Language Processing
SHAP	SHapley Additive exPlanation
EUs	Emergency Units
ICD 9	International classification of Diseases version 9
AUC ROC	Area Under the Receiver Operating Characteristic curve
SOM	Self Organizing Maps
RST	Rough Set Theory
ANN	Artificial Neural Networks
MLP	Multilayer Perceptron
DSS	Decision Support Systems
UMLS	Unified Medical Language System
CT	Computed Tomography
BOW	Bag of Words
PCA	Principal Components Analysis
TF-IDF	Term Frequency-Inverse Document Frequency
MSE	Mean Squared Error
MAE	Mean Absolute Error
ICU	Intensive Care Unit
LIME	Local interpretable model-agnostic explanations
CNN	Convolutional Neural Networks
CDSS	Clinical Decision Support Systems
P.C.S.	Primary Care Service
HGT	Blood Sugar Level
LCFA	Chronic Airflow Limitation
ECG	Whether the patient had an EKG
FR	Breathing Rate
SATO2	Blood Oxygen Saturation
DM	Presence of Diabetes Mellitus
EVA	Airway Evaluation

patients, and this is a crucial bottleneck. In EUs, the triage stage aims to categorize the level of risk and thus prioritize the wait for emergency care, there is still no certainty as to the patient's diagnosis or whether they will require hospitalization. Nonetheless, the patient's severity level is determined according to their vital signs. Considering the above-mentioned, determining the need for hospitalization and the LoS would allow more efficient bed management to streamline procedures. In this context, ML models have been widely used to provide accurate predictions. Parva, Boostani, Ghahramani, and Paydar (2017) presented a set of situations such as classifying patients in Triage, identifying patients who are genetically predisposed to certain diseases, and developing decision support systems for disease diagnosis. Then, various of models have been used in medical applications, such as prediction, classification, regression, and clustering algorithms (Azeez, Gan, Ali, & Ismail, 2015; Ceglowski, Churilov, & Wassertheil, 2005; Sariyer, Öcal T., & Cepe, 2019). Each of these articles addresses the use of ML tools to solve problems specific to a particular type of disease or focuses only on one of the stages of the care process. In recent years, increasingly complex and effective ML and deep learning models have been developed for ML. However, few provide explanations for these models, and some are black box ML models. In healthcare, it is essential to be sure that the results can be trusted, as a mistake can have unfortunate consequences for the well-being of patients. For this reason, we include model interpretability as part of the methodology for implementing ML models in healthcare as a decision support tool. Surprisingly, few authors consider the interpretability of their results as part of their study, so this article aims to contribute from two points of view to the state of the art of ML in healthcare. First, by using NLP techniques to process free text and include interpretability in our results in order to make it easier for health professionals to use these models in their usual diagnostic procedures.

We propose using raw text and traditional features generated in the different stages of the patient care process in an EUs which are saved in the hospital's electronic clinical record. There are free-text fields within the electronic records in which nurses and admissions staff enter information that could be of great relevance when categorizing patients through triage and diagnosing them at the care stage. These free-text fields are, for example, the reason for consultation, where the patient indicates what motivated his or her talk to the health service, and the ailments that the patient describes during the registration and triage processes. This questionnaire is carried out by health personnel to identify information relevant to the diagnosis. It mainly consists of inquiring about the patient's symptoms and perceptions regarding the consultation and whether they have any diseases, allergies, or other irregularities that are important in determining the patient's treatment. Another text field is generated by the treating physician, where they record treatment indications for patients after diagnosing them. In this text field, the physician indicates the medications to be administered, for how long, and in what quantity. Both areas are usually discarded from the study data set due to the complexity of understanding and adapting this information. It is generally presented unstructured and usually introduces technicalities, acronyms, and typing errors that are challenging to handle.

Our research proposes applying models and algorithms of ML to predict emergency patients' severity levels and hospitalization needs in the early stages of care. For this purpose, two datasets from an EU and a hospitalization unit were processed, including records corresponding to the year 2018. By using this information, an experimental dataset of 7848 cases was constructed. In addition, unstructured text fields, usually discarded from this type of study, will be incorporated into the classification model. We propose a model-agnostic method, incorporating sentence embedding to improve predictions, considering the severity classification and LoS prediction. Also, we enriched the data by integrating event description records written in Spanish obtained from EUs in the triage step. Our methodology integrates explainability techniques and an ad-hoc meaningful discovery experiment to facilitate understanding the model and feature relevance. Our experiments improve when including text descriptions compared to models using standard features in severity classification and LoS prediction, encouraging practitioners to extract this valuable information from text records.

The remainder of the article organizes as follows: Section 2 provides a literature review about ML models in EUs for severity classification and LoS prediction. Subsequently, Section 3 defines a sequential prediction methodology based on the evolution of the patient during the care process. Then, Section 4 details the experimental results considering classification and regression models. Finally, Section 5 specifies the conclusions and main findings.

2. Literature review

The following literature review is divided three fold. In Section 2.1, papers that have addressed patient severity estimation through ML models are presented. In Section 2.2, we discuss papers that have proposed methods for predicting patients' LoS in hospital units. Section 2.3 provides articles that have used text information to make predictions in healthcare contexts and EUs. Finally, in Section 2.4, we present a summary of the main interpretability approaches for ML models.

2.1. Severity classification

In recent decades, with the increasing availability of information in the health information systems, multiple research studies have been developed to support the process of classifying the severity of patients, commonly called triage. Until some time ago, triage was performed intuitively, relying on the experience of the examiner. Over time, patient triage methodologies have emerged around the world, using the patient's vital signs as the basis for decision-making (St George, 1992). Classification methodologies such as the *Emergency Severity Index* (Jafari-Rouhi, Sardashti, Taghizadieh, Soleimanpour, & Barzegar, 2013) and the *Manchester Triage System* (Soler, Gómez Muñoz, Bragulat, & Álvarez, 2010) are widely used. In both methodologies, vital signs are contrasted with acceptance ranges that allow the examiner to categorize the patient's level of severity (St George, 1992).

Various authors have focused on improving this patient classification procedure using ML models to make more precise and reduce the need for extensive experience to classify patients correctly. Lin et al. (2010) addressed the correction of abnormal triage classifications based on attributes with incorrect values. For this purposed, the authors propose the use of the *k-means* algorithm and decision trees. They took a sample of 501 erroneous triage records and reclassified them using an alternative to the usual triage methodology. The researchers applied clustering to the data to group them into similar records and then used decision trees to establish rules for analyzing each cluster's records based on patient pressure, temperature, and pulse rate. Based on these results, they developed patient classification rules using the vital signs recorded during the triage process. It is an interesting result, however these rules have to be updated periodically.

Camilloni et al. (2010) and Seymour et al. (2013) applied linear and logistic regressions, respectively, to predict patient severity. Both articles presented vital signs measured in the triage stage as necessary sources of information, such as oximetry, pulse, heart rate, and blood pressure. Camilloni et al. (2010) showed that the regression models proposed can accurately predict the patient's final state. For this purpose, they analyzed about 264,000 records, which were characterized according to the International classification of Diseases version 9 (ICD 9) Centers for Medicare and Medicaid Services and National Center for Health Statistics (2017). The results obtained by the authors showed that the use of linear regressions for the data set studied provide a more accurate classification of the severity of the patients concerning the Injury Severity Score classification system, showing an increase from 0.66 to 0.77 in the Area Under the Receiver Operating Characteristic curve (AUC ROC) for traffic accidents.

Lin, Wu, Zheng, and Chen (2011) proposed the use of Rough Set Theory (RST), which is a technique that obtains hidden information in the data, no matter how messy, after a previous classification using clustering techniques such as k-means and Self Organizing Maps (SOM). In this way, the authors identified the data that presented insufficient or confusing information. They then applied RST using ROSE2" software to this data. The software identified patterns and classification rules that simplified the necessary attributes, which facilitated the categorization of this subset of data. The authors then applied the rules learned to estimate the severity category of the patients with the preprocessed data.

Mathew and Obradovic (2012) presented an adaptation of the distributed decision tree model, based on the id3 algorithm. The article shows the use of distributed data from nine U.S. hospitals, obtained from the *National Inpatient Sample*. The main contribution here was to developing a methodology for predictions with little data, less than one hundred records per hospital. Seymour et al. (2013) proposed, in addition to the use of vital signs as primary biomarkers, the use of specific biomarkers corresponding to the most frequent diseases that represent a greater risk for patients, such as cardiac diseases. The biomarkers were entered into the regression model by simulating their behavior, obtaining accurate results when reclassifying frequent diseases in the EU.

Azeez, Ali, Gan, and Saiboon (2013) used artificial neural networks (ANN) to predict the severity level of patients. It is proposed to use information from the anamnesis to make predictions of patient's severity index. To process this information, the author suggested the use of the CEDIS system (Grafstein, Unger, Bullard, Innes, et al., 2003), which classifies ailments, by assigning them numeric values that facilitate their use a fuzzy neural network, an adaptive neuro-fuzzy interference system, is proposed in the article. This neural network allows the aggregation of expert judgments through the logical operators *IF-THEN*, being able to improve the results of a standard multilayer perceptron (MLP). An MLP is a generalization of the neural network. It groups a defined set of simple perceptrons, incorporating layers of hidden neurons for the representation of nonlinear functions.

Azeez et al. (2015) presented prediction models based on decision trees, with a reduced amount of data. To increase the efficiency of their models, they developed preprocessing tools for unbalanced data. For this purpose, the authors designed several experiments with different amounts of data generated by resampling, generating data with 150% of the resampled data, increasing the amount of data up to 500%. Then, with the data not used in training, they calculated the accuracy of each experiment, which ranged between 85% and 88%. The method used was randomized resampling and obtained better results by resampling the data at 300% and 700 trees in the structure.

Hong, Haimovich, and Taylor (2018) applied deep learning, XGBoost, and logistic regressions, to predict the severity category of patients, developing three datasets with information from EU records. The first dataset contains information on patients' reports in the registration stages and the vital signs acquired in triage. The second dataset includes historical records of each patient, including a history of their vital signs, laboratory tests, among others. Finally, the third dataset corresponds to the union of the previously described datasets to develop predictions with the totality of the available data. Using this dataset offered the best results in terms of sensitivity and specificity for all the tested algorithms. The authors also proposed post-processing of the models to experiment on two relevant aspects. The first one was to omit some variables from the data because they do not contribute much information to the result, managing to reduce the dimensionality of the models and without notably impacting the accuracy of the result. They also tested the influence of reducing the size of the data sample. The method that delivered the best results was XGBoost, with a sensitivity index of 0.83 and a specificity index of 0.85.

Raita et al. (2019) compared logistic regression with gradient boosted random forest and neural networks to identify which prediction method gave the best results. Gradient boosted random forest presented better results both on patient classification and the prediction of hospitalization. It showed sensitivity and specificity values superior to the other models studied. In this type of study, it is common to find incomplete or very disordered data, complicating their treatment and predicting variables.

There are several applications of ML models that allow us to make predictions or classify patients according to a set of data. In health, its development has shown an important growth, improving the capabilities of disease detection, patient classification, among other functionalities. Recently, several authors have made use of text processing techniques with the aim of retrieving useful information from medical records, using ML models to learn more about the care processes and the importance of the features fed to the models. NLP techniques have been used in several areas such as linguistics in the detection of words in translation tools, audio-to-text transformation and vice versa, sentiment analysis, and spam detection in emails. In health, there is an inexhaustible source of free medical text developed during and for the purpose of carrying out patient care processes in the various areas of medicine. In the following section we will review some research linking the use of NLP techniques with medical records generated in the patient care process, with various applications.

2.2. LoS prediction

The use of ML has also been considered in predicting the LoS in a hospital. Tu and Guerriere (1992) presented the application of a neural network to determine the LoS of coronary surgery patients in Canada during the years 1990 and 1991. The authors describe the experience by commenting that the network achieved a root mean square error of 0.056 on the test set and 0.0564 with a collection of 28,000 records, demonstrating the ability of ANNs as a predictor of the LoS in a hospital. Jiang, Qu, and Davis (2010) analyzed patients with the four most prevalent chronic diseases in those over 65 years old, who, as mentioned in the article, have a higher tendency to stay longer in hospital. Four ML models were applied to evaluate the effectiveness of each in predicting the LoS. The following models were compared: logistic regression, MLP, decision trees, and a combination of them. The results showed that the mixture achieved the best results. In addition, age and chronic disease were reported as the most valuable predictors within the dataset available.

Marfil-Garza et al. (2018) used multivariate regression and logistic regression tools to determine the variables corresponding to characteristics of hospitalized patients that have the most significant influence on the duration of their stay in the hospital. They concluded that the variables that have the most significant impact on hospitalization time are gender, with men staying longer in hospital, age, older adults, and low socioeconomic level.

Sariyer et al. (2019) developed a patient classification process and applied the methodology for predicting LoS to each subgroup. The subgroups were formed based on similarities between patient diagnoses described by the dataset, relying on the international ICD10 classification. The authors used MLP, random forest, decision trees, and logistic regression. The results were evaluated in terms of sensitivity, specificity, and accuracy. Logistic regression and MLP were the most efficient prediction models.

In summary, over time, there have been numerous efforts to support the decisions associated with the classification of patients and to study the reasons and causes of the time required for hospitalization. For this purpose, ML tools have been used and have reported interesting results, generally better than the traditional methodologies, i.e. score systems based on expert knowledge (Wilding & Evans, 2017), for making such decisions. Moreover, there is free-text information generated at various stages of the care process that, in practice, are of great importance when classifying a patient or deciding to request hospitalization. Therefore, we will analyze how to integrate this free-text information into the ML models studied, specifically with records that represent the description of the event, which details the symptoms that motivated the patient to come to the EU.

2.3. NLP applied to health prediction problems

Demner-Fushman, Chapman, and McDonald (2009) presented a review of the contributions that NLP has made to the field of medicine. NLP tools are an essential part of decision support systems (DSS) used for more than 40 years with good results. The research divides these systems into two active and passive tools. The active ones correspond to systems that develop automatically with existing information previously registered. On the other hand, the passive support tools require a user who enters new data into the system. They also developed a classification in general-purpose systems, such as Linguistic String Project, Medical Language Processor, and MedLEE, among others. These systems transform narrative physician records, without great specification, into a structured form, with the support of controlled, specialized vocabulary (Demner-Fushman et al., 2009). Another classification presented by the authors is specialized systems, among which they mention clinical event monitors that identify adverse events in patient discharge records. Radiological report processors are another common use for specialized NLP systems. The Special Purpose Radiology Understanding System, Natural language Understanding System and Symbolic text processor are examples of this classification of systems SymText. They use radiological reports to detect patterns, for example, the presence or absence of acute pneumonia bacteria in pulmonary radiographs.

Another type of specialized system is those that process emergency department reports. The authors comment on the experience of Chapman, Fiszman, Dowling, Chapman, and Rindflesch (2004), who developed free text mapping of emergency patients records, with the software MetaMap, of emergency patient records, seeking to identify acute lower respiratory syndrome in them. They used three methods that related the records to the Unified Medical Language System (UMLS). In results, the authors reported an accuracy of 0.72, which is acceptable considering the limitations indicated, based on the handling of the software and the lack of keywords, for certain records, in the UMLS system.

Claster, Shanmuganathan, and Ghotbi (2008) developed a methodology for obtaining information from medical records for patients undergoing radiology. Aware of the effect of repeated exposure to ionizing radiation, the authors sought to identify in the records, through text mining, the keywords in the records of indications for computed tomography(CT) scans, symptomatic description of the patients, and comments on the results of such examinations. So, this would make it possible to classify the cases in which the test was necessary and those in which it was not through a correlation analysis between words related to favorable and unfavorable cases regarding the need for the test.

Bacchi et al. (2020) applied prediction algorithms to the LoS and referral destination of patients by obtaining information from free-text medical records, extracting the information using NPL techniques such as tokenization, transforming words to word stems, and then applying data mining algorithms to predict the variables described. In their results, they showed that by mixing the data obtained through NPL and algorithms such as neural networks, decision trees, and logistic regressions. They were able to predict with an accuracy close to 0.88 the LoS of patients, under a sample of 313 patients, with a test segment of 15%.

Recently, research has emerged that attempts to make use of NLP techniques to improve predictions of relevant events in medicine. We will now review some of the articles that have experimented with this aspect.

Amunategui, Markwell, and Rozenfeld (2015) proposed two methodologies to transform the free-text into numeric data useful for training ML models. The first method proposed consists of manually selecting a bag of words (BOW), which the authors consider important for predicting patient hospitalization. Then, they use Word2vec, which is a classic approach to vectorially represent words (Mikolov, Chen, Corrado, & Dean, 2013), to obtain the most similar words. The second method consisted of performing clustering of the vectors obtained by Word2vec, with the complete set of text, in 50 clusters generated by k means. The resulting vectors in each method were used to train an adaptive boost model, obtaining an AUC ROC of 0.63 and 0.61 in each respective experiment.

Sterling, Patzer, Di, and Schrager (2019) worked the prediction of patient hospitalization, using free text obtained from notes generated by nurses during the patient review. They used the BOW technique to process the text, which consists of counting the frequency with which certain words occur in the documents analyzed. Then, using principal components analysis (PCA), they reduce the dimensionality of the text vector obtained by BOW and use it to develop predictions by classifying patients considering as an outcome a dichotomous variable indicating whether the patient is hospitalized or discharged to home care. To measure the model's effectiveness, the authors propose the AUC ROC curve as a metric, with the values 0.737, 0.740, and 0.687 with three sets of text used independently.

Chen et al. (2020) proposed a methodology to explore the potential of NLP processing in LoS prediction, using a mixture of structured data and physician records during patient examination. To do so, the authors developed a preprocessing of the text logs, removing punctuation marks, stop words, and capitalization. Then, a frequency-inverse document frequency (TF-IDF) is used, they created a vector representation of the text by counting the words present in the text and normalizing the data, obtaining a data matrix useful to feed ML models. Their results show a good performance of the models provided by text data, with an average mean squared error (MSE) of 3 h and an average mean absolute error (MAE) of 1.5 h.

Bacchi et al. (2020) proposed a methodology to use free text as the primary input for training classification and regression models to estimate the LoS of patients in an emergency unit. To transform the text, the authors preprocessed the text by reducing conjugated verbs to word stems, i.e., words such as improve, improvement, improving, are reduced to improv. In addition, they eliminated stop words and negations such as painful and not painful. After that, the authors develop arrays with the frequency count of the obtained word stems. With the processed text plus numeric data, the authors developed classification and regression experiments to determine the LoS of patients. For classification models, they obtained accuracy results of 0.82 with neural networks. For regression models, they got an MAE of 2.9 and an MSE of 16.8.

Bardak and Tan (2021) developed a methodology based on the use of convolutional neural networks using data from three different text sources to improve predictions of patient's LoS in the EU. They developed a preprocessing of the text using an entity extraction model from the text, identifying seven different entities. Subsequently, they made use of several embedding models such as Word2vec, Doc2vec, and Fastext, which together with convolutional networks developed four prediction scenarios to predict: patient's in-hospital mortality in intensive care unit (ICU) mortality, a LoS of more than three days and a LoS of more than seven days. They used Area Under the Receiver Operating Characteristics (AUC-ROC) as evaluation metric, obtaining mean values of 85, 86, 69, and 71, respectively, for each scenario. After analyzing the results, the authors emphasized the importance of the interpretability of results in ML and how these can be an excellent precedent to establish relationships between vital markers, as analyzed variables and the results obtained.

2.4. Interpretability for ML models in healthcare

Interpretability and explicability are two fundamental concerns for the application of the ML models in health. Interpretability is *the degree to which a human can consistently predict the model's result* (Kim, Khanna, & Koyejo, 2016). It is the ability to understand the behavior of the results of a ML model, i.e., to understand why a prediction model made such predictions or classifications. Explicability is instead about understanding how ML models work. Thus, explainability is about being able to explain in human terms how a ML model works.

Stiglic et al. (2020) developed a review of the main approaches to the use of interpretability techniques applied to the understanding of ML models in various areas of healthcare. Some of the techniques mentioned by the authors in interpretability techniques in healthcare are SHAP, Model Understanding through Subspace Explanations, Local interpretable model-agnostic explanations (LIME), and graph neural networks. The authors commented about the importance of interpretability in ML models for treating physicians and end-users of the models who could see the interpreted results, trends, and information about patients or diseases that might not be easy to see. For example, the authors mentioned a case in which SHAP was used to interpret predictions for the prevention of hypoxemia during surgery, which increased by 15% the anesthesiologist's anticipation of hypoxemia events.

ElShawi, Sherif, Al-Mallah, and Sakr (2020) proposed a collection of outcome explainability techniques for ML models in healthcare, briefly describing six of them. The methods described are LIME, SHAP, and Anchors.

LIME proposed in Ribeiro, Singh, and Guestrin (2016) is a technique that trains local surrogate models to explain individual predictions. Local surrogate models are easily interpretable models used to describe individual predictions of classical ML models. In this way, the predictions of surrogate models are explained using a dataset corresponding to permuted samples of the original data plus forecasts generated by the ML model to be analyzed. The results of these surrogate models are weighted by closeness to the original data in the hope of achieving a model that can be compared with the ML model that was initially employed.

Anchors proposed in Ribeiro, Singh, and Guestrin (2018) is a local explanation technique based on rules of the LIME technique. In Anchors, variables are reduced to certain conditions called anchors used to generate the prediction. Anchors considers the original data set, and then it generates an anchor. Anchors are constructed based on If-Then rules to find the features of the input data responsible for the prediction. Anchors start with an empty rule, and at each iteration, the rule is expanded with a feature such that the new rule has the highest estimated accuracy. To select the best rule at each iteration, the KL-LUCB algorithm is used.

Subsequently, the authors performed experiments to compare the techniques by measuring their five metrics: identity, stability, separability, similarity, and execution time. The results highlight SHAP, in terms of execution time and, in terms of identity, MAPLE, a model proposed by Plumb, Molitor, and Talwalkar (2019), that combine local linear modeling techniques along with a dual interpretation of random forests, and . The authors concluded that the importance of interpretability lies mainly with clinicians, who are still unfamiliar with these prediction techniques and therefore reluctant to consider their results as valid information. They also point out that more and more robust models are designed to obtain better accuracy but are very complex to understand. Hence the importance of including interpretability techniques to improve the interpretation of results.

Jia, McDermid, Lawton, and Habli (2021) developed an investigation regarding the importance of explainability in the development and analysis of ML projects. To do so, they start by describing the need for security in the development of engineering processes applied to health systems, pointing out the differences between verification of the tool with its correct implementation and its validation, emphasizing the results obtained. Subsequently and with the development of a case study, they implemented a model of convolutional neural networks (CNN), logistic, regression, support vector machine, random forest and decision trees, in two instances for predicting patient readiness for extubation so as to avoid the negative side effects of mis-timed extubation. The results obtained by CNN were analyzed using DeepLIFT which is a model-specific explainable AI method for deep NNs, obtaining a ranking of features from the most important for prediction to the least important. Finally, and with the results obtained, the authors present the arguments and proofs that guarantee that the applied models provide an acceptable safety in the case that it is implemented in a health system.

Amann, Blasimme, Vayena, Frey, and Madai (2020) developed a multidisciplinary analysis of the importance of the relevance of explainability for medical AI from a technological, legal, medical and patient point of view. To this end, the authors developed an analysis from two aspects. First, they delve into the relevance of explainability in clinical decision support systems (CDSS), from the technological, legal, medical, and patient perspectives. technological, legal, medical, and patient perspective, identifying the main ethical implications involved in the use of these systems for health decision support. From the technological point of view, the authors mention that there are cases in which ML models generate false positives and damage the confidence that health professionals can have in their implementation as a decision support

tool, they also mention that the use of explainability techniques, allow developers to identify these They also mention that the use of explainability techniques allows developers to identify these types of errors before AI tools go through the clinical validation and certification process. In legal terms, the authors comment on the rigor with which sensitive data must be acquired, stored, transported, processed and analyzed, complying with the respective legal regulations for the type of data being considered in the analysis. From a medical point of view, the authors mention that it is possible to compare AI-based CDSS with laboratory tests where ML models are represented as a black box unlike laboratory tests where the biochemical reactions that trigger the results are clearly known. However, AI-based CDSSs have steadily improved their performance and the use of explainability techniques has favored the acceptance of these systems by clinicians. However, these systems must still be subjected to rigorous clinical validation and certification systems.

From the patient's perspective, the authors mention that patients should be aware of the use of these systems and they should be approved by them to be used in their medical treatment, making them aware of the risks involved in the use of this type of CDSS. In this way, explainability could improve the understanding of the results of a CDSS, both by physicians and patients to correctly evaluate the possible treatments to be used from the diagnosis.

Finally, with all the analyses performed, the authors present the ethical implications of the use of ML and AI-based tools in healthcare, indicating that they are based on four key principles: autonomy, beneficence, nonmaleficence, and justice. As for autonomy, it is represented by informed consent, which is an autonomous authorization, usually written, with which the patient grants a physician permission to perform a given medical act. Beneficence translates into making use of AI-based tools, only with the aim of improving the patient's service quality, improving diagnoses and response times to the needs of care.

3. Data and methodologies

In this section, all the steps of our methodological approach are presented. First, in Section 3.1, we present the three dependent variables of our study in detail. Then, in Section 3.2, the methods used in each stage are presented. Finally, in Section 3.3, the implementation details are described. Fig. 4 presents a graphical abstract that illustrates the methodology used in this research.

3.1. Data set

According to Santelices and Santelices (2017), Chile is one of the countries with the highest rates of emergency care per 1000 inhabitants, with 571 visits per 1000 inhabitants per year, surpassing countries such as the United States with 445 visits per 1000 inhabitants per year, Canada with 440 visits per 1.000 inhabitants per year, and the United Kingdom with 360 visits per 1.000 inhabitants per year. Chile ranks 99th in the number of hospital beds with 2.1 beds per 1000 inhabitants, below Brazil, which has 2.3 beds per 1000 inhabitants, and below Uruguay, which has 2.5 beds per 1000 inhabitants.

The Hospital San Juan de Dios de Curicó is a hospital facility of high technical complexity. It is the base of the health network of the province of Curicó and part of the health care network of the Maule Health Service in which participating hospitals are in the towns of Molina, Teno, Linares, Cauquenes, and Parral, among others. According to data provided by hospital workers, in 2018, this EU had 17 observation beds plus 14 examination couches, where outpatient care is provided to pediatric and adult patients. In addition, 92,338 emergency consultations were performed that year. The Curicó Hospital provided two patient care records, corresponding to the hospital EU and the hospitalization unit, obtained during 2018. One corresponded to the records of the consulting patients of the EU during 2018, and another one presented the hospitalized patients that indicated their diagnosis and days of hospitalization. Both datasets were combined, obtaining a unified dataset with which, after the described preprocessing tasks, the available data sample is reduced to 7848 records. datasets were provided anonymized by the hospital's statistics department.

The experimentation was conducted in three stages of patient care: categorization, diagnosis of the patient's destination, and the hospitalization time for those patients who required it. Each of these stages was analyzed independently, although data from previous steps were used for the experimentation in later stages, i.e., for stages such as diagnosis.

3.1.1. Target variables

Severity category (step 1). It is carried out in the first stage of the care process, in which information is requested from the patient. The patient's registration data, the reason for consultation, vital signs, and medical history are available at this stage. This information makes it possible to classify patients according to severity level using data mining tools. The severity category is data of the ordinal categorical type. Fig. 1 presents the proportion in which the different severity categories of patients registered in the EU during 2018 are distributed. Originally there were five levels of severity, level 5 being the level corresponding to patients who did not present an emergency and did not have to attend an EU but a health facility of lower complexity. The records associated with this type of patient were those with the most significant missing data. When pre-processing and data cleaning tasks were carried out, they were discarded due to their limited usefulness. The variables with the least amount of missing data were selected. Some variables with missing data were imputed, but when the missing data percentage was bigger than 60%, the variable was deleted from the experiments.

Patient destination (step 2). After being diagnosed, the patient is treated according to the doctor's indications and placed under observation for some time. After analyzing the evaluation of the patient's treatment, the patient's possible destination is determined: discharge, hospitalization, referral to a health institution of higher complexity, or referral to a health institution of lower complexity, among others. Data describing the possible destinations of the patient corresponds to the categories. Fig. 2 indicates the possible patient destination after being diagnosed by a physician. Among the possible destinations are hospitalization, a category of particular interest to us in this research since it is directly related to the third step of analysis that we proposed, estimating the LoS of hospitalized patients. A Primary Care Service (P.C.S.) is a health entity of lesser complexity, generally located in residential areas, that functions as an intermediary between patients and the E.U., resolving minor ailments or stabilizing patients with a high severity level and transferring them by ambulance to the E.U.

LoS (step 3). In this stage, there is information on registration and triage, in addition to the severity index classification and diagnosis obtained in previous steps. The need for hospitalization is a dichotomous variable; if it is positive, the length of hospitalization should be predicted as a numeric variable. Fig. 3 shows a histogram of patient hospitalization time records, corresponding to the year 2018 provided by the hospital. By joining the datasets corresponding to emergency records plus the records corresponding to the hospital's inpatient unit, a large percentage of patients with zero days of hospitalization appeared. These corresponded to patients who were either discharged or referred to another health facility. The rest of the data corresponded to the original records in the dataset of the hospitalization unit and these were mainly distributed between 1 and 40 days of stay. There are records with a longer LoS; however, these were so infrequent that they were atypical data in the records.

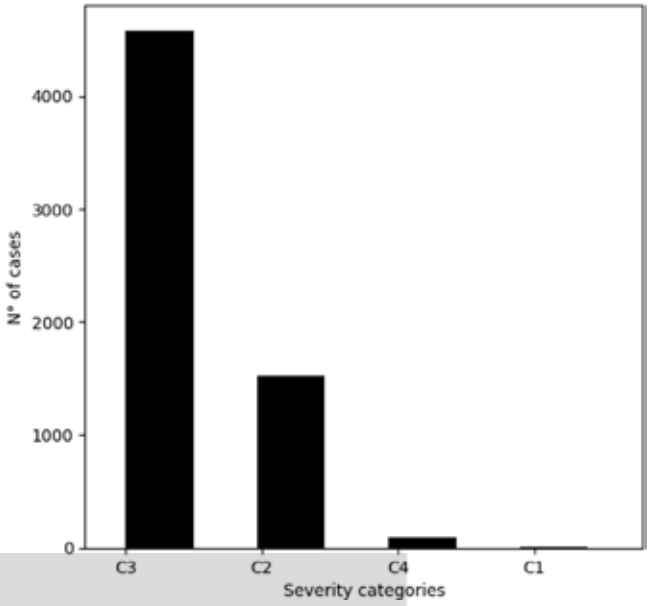

Fig. 1. Target variable stage 1. Severity index category where C1 represents the highest severity level and C4 is used to classify the lowest severity level.

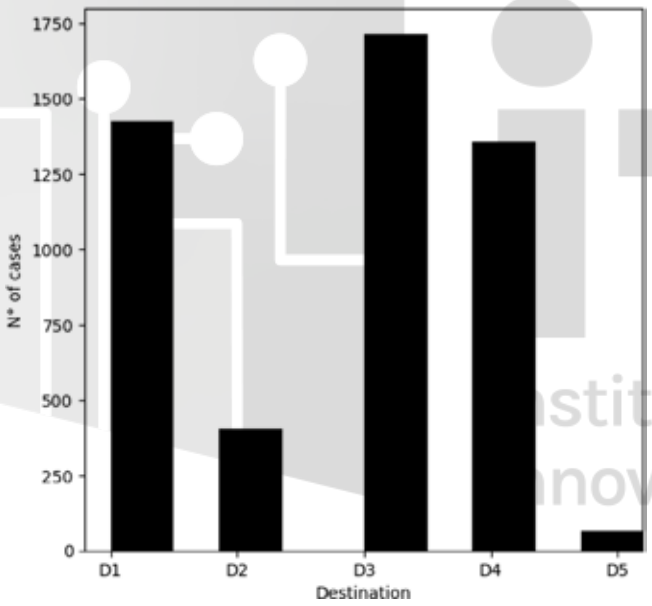

Fig. 2. Target variable stage 2. Destination type category after disgnostication. D1 is home, D2 is another destination, D3 is a primary care service, D4 correspond to being hospitalized and D5 be transferred to another hospital.

3.1.2. Standard features

Data was divided into two parts. Registration data, which indicates patient information before any medical intervention, and Triage data, where vital sign measurement indicators are obtained. However, these indicators, such as blood pressure, respiratory rate, are the ones that mainly present missing data. Then, in the care stage, records without diagnosis and patient data with incomplete records were eliminated due to desertion from the system during the various stages of care.

Triage data corresponds to data recorded up to the triage stage, among them: age, severity index, prognosis, HGT (blood sugar level), LCFA (chronic airflow limitation), ECG (Whether the patient had an EKG), Glasgow (measures the patient's level of consciousness), breathing rate (FR), blood oxygen saturation (SATO2), sex (SEX), airway evaluation (EVA) and presence of diabetes mellitus (DM). Table 1 describes each variable's type of data, in addition to the classification/prediction stage used.

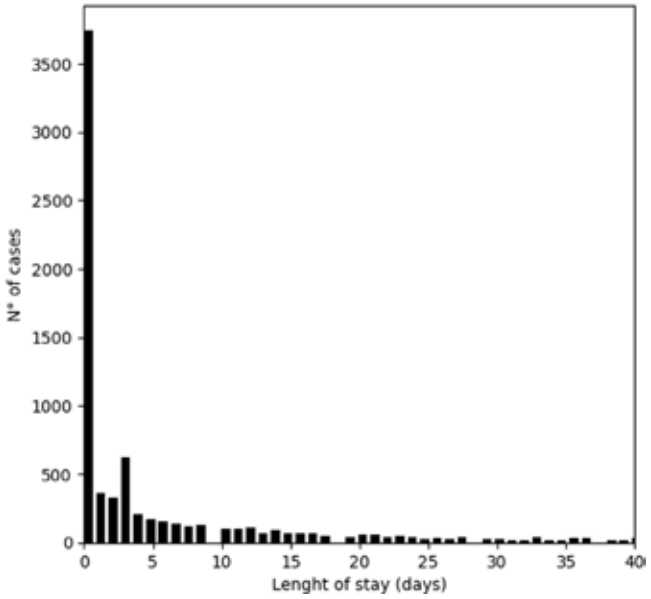

Fig. 3. Target variable stage 3, Length of stay in hospitalization unit.

Care data corresponds to data recorded up to the care stage diagnosis, represented by a code from the international classification of diagnoses CIE10 and destination: discharge, hospitalization, and referral to another facility, among others. More information on the variables is presented in Table 2. Both datasets have a unique ID per query, so we decided to link them into a single dataset containing the records of the patients seen in the EU and the records associated with patients hospitalized during the study period.

The severity of the patient is determined in terms of the hospital's triage index, the possible diagnosis considering the ICD10 coding provided by Centers for Medicare and Medicaid Services and National Center for Health Statistics (2017), the need for a hospital bed, and the expected LoS, in days, in case of hospitalization.

3.1.3. Raw text from event description

We use the free-text data of text records annotated by technicians and nurses when carrying out the triage process. In such records, health professionals indicate why the patient comes to the EU and mention whether they have comorbidities or pre-existing diseases that may affect their current health status. The text records are incomplete, poorly written, or have acronyms and medical technicalities unknown to us. If processed, it would be necessary to go through the records individually to decipher each sentence's relevant elements using common words. However, apart from being complex, this task would require the support of someone with sufficient medical knowledge to interpret the records, which would take a long time. Table 3 presents some text records in the:Event description feature.

3.2. Methods

Our method, whose graphical abstract is presented in Fig. 4, is divided into three parts. The first step considers data processing, cleaning, imputation, and transforming, including the raw-text representation using Word2vec. The second phase focuses on the training model and related decisions (e.g., hyperparameter tuning, oversampling, and dimensionality reduction). Lastly, step three presents a strategy to provide model-agnostic explanations for ML predictions.

Table 1
Description of data available in the dataset.

Variable	Minimum	Maximum	Mean	Mode	Type	Variable	Step
Age	15	126	42	–	Numeric	Independent	1-2-3
HGT	11	600	145	–	Numeric	Independent	1-2-3
FR	3	99	18.54	–	Numeric	Independent	1-2-3
SATO2	96	100	97.10	–	Numeric	Independent	1-2-3
LCFA	–	–	–	–	Binary	Independent	1-2-3
Glasgow	3	15	14.94	–	Numeric	Independent	1-2-3
SEX	–	–	–	–	Categorical	Independent	1-2-3
DM	–	–	–	–	Binary	Independent	1-2-3
EVA	–	–	–	–	Binary	Independent	1-2-3
Diagnosis	–	–	–	–	Categorical	Independent	1-2-3
Category	1	5	–	3	Numeric	Target	1
Destiny	–	–	–	–	Categorical	Target	2
LoS	–	–	–	–	Numeric	Target	3

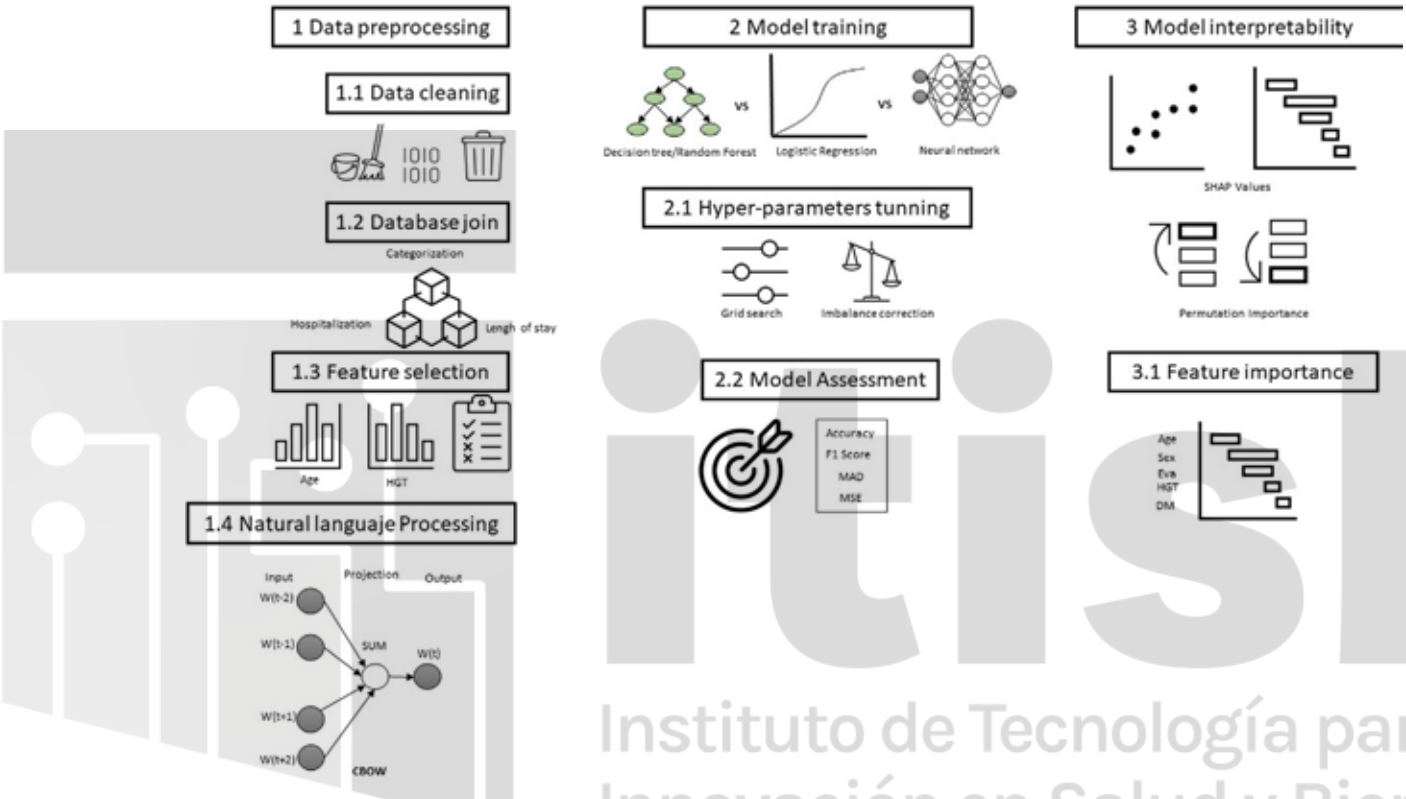

Fig. 4. Graphical Abstract presents all the activities developed in the proposed methodology divided into three stages. (1) Data preprocessing, (2) Model training and (3) Model interpretability.

Table 2
Sources of information during the care process.

Stage	Information
Registration	Name, age, sex and hometown
Triage	Vital signs and Event description
Patient care	Patient diagnosis and indication for treatment
Observation	Indication of treatment according to the patient's evolution.
Hospitalization	Indication for hospitalization
Medical discharge	LoS

3.2.1. Word embedding generation

A big step in NLP was generated by Word2vec, an algorithm developed by Mikolov et al. (2013) to vectorially represent words and position them in a multidimensional space that respects the similarity between them. Mikolov et al. (2013) trained the algorithm with billions of words, significantly exceeding the volumes of information used for training by previously developed algorithms such as N-gram or NNLM.

To get representations for each Spanish words w from the anamnesis process using Word2vec model $g : w \rightarrow \mathbf{x}_w^{1\times 300} \in \mathbb{R}$, we used a model trained with 2.6 billion Spanish words (Cañete, 2019). The corpus consists of Spanish words obtained from various sources, including Spanish Wikipedia entries, Open subtitles, and TED talk subtitles, among other sources.

By preprocessing a subset of data composed of the event description, a record details the ailments that motivated the patient to go to an EU and whether the patient has pre-existing conditions that could influence the patient's categorization and diagnosis. A vector representation $\mathbf{x}_w^{1\times 300} \in \mathbb{R}$ of each word w was obtained together, which formed the sentence contained in the event description field. Each of the vector representations of a sentence was averaged to receive an average vector of the sentence as follows:

$$\bar{\mathbf{x}}_s = \frac{1}{l_w} \sum_{w}^{l_w} \mathbf{x}_w,$$

where l_w is the sentence length. This representations is used to train regression and classification models.

Table 3
Example of a free-text record corresponding to the variable: Event description, taken at random from 130.021 text records. Recorded in the patient registration process and completed during the triage process. The words in bold are those that could be vectorized. The row under each expression corresponds to the translation (e.g. 1E is the English translation for 1S) of words recognized by the embedding generator (in bold type).

#	Description
1S	***TRAÍDO*** *X* ***CARABINEROS*** *PARA* ***CONSTATACIÓN*** *DE* ***LESIONES*** . + **ALCOHOLEMIA** ***NO APLICA*** *LA* ***EVALUACIÓN*** *DE* ***SIGNOS VITALES***.
1E	{ brought, police, injury, breathalyzer test. } { vital signs, evaluation, not apply}
2S	***TRAÍDA*** *POR* ***INGESTA MEDICAMENTOSA, POLIFARMACIA. PCTE GLASGOW*** *PTS.****PACIENTE RECHAZA EVALUACIÓN DE SUS SIGNOS VITALES.***
2E	{ brought, medication intake, polypharmacy } {GLASGOW 8, patient, rejects, Vital signs, evaluation}
3S	***PACIENTE DERIVADO DESDE LICANTEN*** *POR* ***DERRAME PLEURAL IZQUIERDO***
3E	{Patient, referred, Licanten, left, pleural, effusion.}
4S	*PCTE CON FX DE* ***TIBIA DERECHA***. ***INMOVILIZADA CON YESO***. *RX OK. PARA EV POR TMT*
4E	{ Right tibia, inmobilized, cast }

3.2.2. Data preprocessing

Having two datasets, we selected all the structured variables related to the predictions we wanted to provide from them. For the severity classification, in stage 1, we chose all the data from the patient's vital signs record in the triage process. We identified the missing data for each variable, selected those with sufficient quantity to perform imputation, and discarded the rest. The second dataset, from hospitalization, had variables that matched those chosen in the EU dataset, so we merged them and discarded the duplicate variables, unifying the datasets. We imputed the missing values of the selected numeric variables by the mean with the unified datasets, while the most frequent category did it with the categorical variables.

For categorization, the patient's severity category was predicted using information from the patient registry, which included: age, sex, commune, temperature, and pressure, among others. A number represents each piece of information. In the case that a variable may take multiple values, as with the "sex" category, binary variables were created to designate whether or not a record belonged to a specific category (dummy variables). In this way, the available data was adapted to represent a number for use in the previous prediction algorithms. In addition, standardization and normalization were applied to the dataset to avoid disparity in the scale on which the data was expressed.

We used PCA to reduce the dimensionality of the data. PCA is a technique that allows us to transform a set of variables into another set of smaller dimensions but with an equivalent amount of information. This set of new variables is called principal components, and they are not correlated with each other. They are ordered according to the size of the variance and in descending order. In this way, the first components describe most of the variance of the original data. Therefore, they are the most useful for the analysis; the rest with minimum variance values can be discarded for their lack of contribution.

3.2.3. Model training

The categories presented a significant imbalance, with classes representing a percentage close to 2% of the complete data set. In contrast, the most frequent ones accounted for close to 40% of the entire data set. For this reason, we apply SMOTE-TOMEK to improve the effectiveness of the predictions. This is because classes such as life-threatening (C1) or non-severe (C5) are very rare and therefore with few records in the dataset. For this reason we experimented by performing predictions with balanced data using subsampling, oversampling and SMOTE-TOMEK methods. The technique that generated the best results after class balancing was SMOTE-TOMEK, using class balancing as a strategy. The technique that generated the best results after balancing the classes was SMOTE-TOMEK, using as a strategy to balance the minority classes with respect to the majority classes, balancing them to approximately 3200 records for each class.

Table 4
ML models implemented at each prediction stage.

Step	Step 1 Severity level	Step 2 Patient destination	Step 3 LoS in hospitalization
Type of target	Categorical	Categorical	Numeric
Model	Decision trees Random Forests Logistic regression MLP	Decision trees Random Forests Logistic regression MLP	Decision trees Random Foreest Lasso Regression MLP

To improve the quality of predictions, we use columns with text written by the health professionals who assisted the patient in the different stages. The information in these columns could not be treated as a categorical variable because of the complexity of categorizing such text since there would be practically no matches. For this reason, it is not typically considered in this type of experiment. We used Word2vec, which allowed us to transform each word to compose the text records into a vector, and thus each sentence and the complete record as an average vector of such vectors. It is to say, each word composing the sentence describing the event was transformed into a vector, and all the vectors generated for a sentence were averaged into a single average vector representing the complete sentence. These vectors of 300 components were added to the previously processed numeric data, making predictions accurate. The Tables 5, 6 and 9 demonstrate the improvement in the quality of the results.

To improve the performance of the ML models used in the experiments (see Table 4), we decided to apply grid search to optimize the search for the appropriate hyperparameters for each model. The models and their respective hyperparameters are described below. The experiments were conducted using the following models:

Logistic Regression: It is primarily a data analysis technique, although, due to its versatility, it is widely used to study the relationship of a dependent variable, dichotomous or multinomial. It has one or more independent variables, measuring its sign, whether the dependence is direct or inverse, and determining the probability of an event occurring as a function of the independent variables. When applying cross-validation grid search (5-fold) to adjust the hyperparameters, we considered varying the value of parameter C in a range of (0.1,1,.10). In addition, we defined three alternatives for the regularization of L1, L2, and elastic net.

Decision Trees: Decision trees are statistical models that allow the prediction of data analytics based on their classification according to

specific characteristics or properties or regression through the relationship between different variables to predict the value of another. We considered modifying the following to train the model: the splitting evaluation criterion, considering Gini and entropy. In addition we considered a range of values for the tree depth, which varied between 3 and 21. Finally, we defined the minimum number of objects by leaf between 2 and 100.

Random Forest: This approach is a combination of predictive decision trees. Each tree evaluates the independent variables by binary tests at each node, constituting the tree's branches. The grid search considered the space previously described for decision trees. We also included the number of estimators ranging from 100 to 500.

Multi Layer Perceptron (MLP): A directed acyclic graph (DAG) defines a mapping between Euclidean spaces. The nodes and their connections (weights) in this DAG emulate the behavior of neural networks in the brain. This model can be used for classification and regression tasks. Each node or neuron performs simple activation operations over a linear combination of weights and inputs. The parameters adjusted by the grid search strategy are the following: the activation function (tanh and relu functions), optimizer (SGD and Adam), penalty parameter (0.0001 and 0.05), learning rate (constant and adaptive). Three different structures were evaluated for the MLP, three hidden layers of 50 neurons, three hidden layers with 50, 100, and 50 neurons, with one hidden layer of 100 neurons.

Lasso Regression: A traditional regression that minimizes the residual error squared, with an additional component focused on shrinking its parameters towards zero, restricts the regression coefficients. This model sacrifices bias to reduce the variance of prediction. Moreover, the reduction in the number of predictors simplifies the parameter interpretation. The grid search selects the value for this penalty parameter, evaluating the following set of alternatives $\{0.005, 0.02, 0.03, 0.05, 0.06\}$.

3.2.4. Model assessment

This section presents the metrics for each step (classification and regression) and the validation strategies. For classification steps (step 1 and step 2), we have used the following metrics:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}},$$

which measures the percentage of cases that the model got right.

$$\text{F1-score} = 2\left(\frac{precision \cdot recal}{precision + recal}\right),$$

which combines the precision and recall measures into a single value. This is practical because it makes it easier to compare the combined performance of accuracy and completeness between various solutions. F1 is calculated by taking the harmonic mean between precision and recall.

For regression, we have selected the following metrics:

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2,$$

$$MAD = \frac{1}{n}\sum_{t=1}^{n} |e_t|,$$

where $e_t = y_t - \hat{y}_t$, being y_t and $\hat{y}_t$ the real and predicted value, respectively. Minimum, maximum, and R^2 were also considered to assess the prediction quality.

We applied a cross-validation (5 folds) grid search to validate these metrics and select hyperparameters. The parameters used for each model are presented in Section 3.2.3.

3.2.5. Model interpretability

We apply interpretability techniques to understand the performance of the applied ML models and the behavior of the variables used in predicting and classifying response variables. A key concept in these techniques is an additive feature attribution method. It has an explanation model that is a linear function of binary variables. It is assumed that f is the original prediction model and g is the explanation model. The focus is local methods designed to explain a prediction $f(x)$ based on a single input x. Typically, explanation models use simplified inputs x' that map to the original inputs through a mapping function $x = h_x(x')$. The aim of the local methods is to ensure $g(z')$ approximates $f(h_x(z'))$ whenever z' approximates x'.

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i',$$

where $z' \in \{0,1\}^m$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$.

According to the definition above, methods with explanation models attribute an effect to each feature. Summing the effects of all feature attributions approximates the original model's output $f(x)$. Several current methods, like LIME and classic Shapley Value Estimation, follow this approach. SHAP (SHapley Additive exPlanations) method satisfies three crucial properties: Local accuracy, Missingness, and Consistency. A significant result establishes that for a given simplified input mapping h_x, there is only one possible additive feature attribution method, see Lundberg and Lee (2017a).

SHAP values of a conditional expectation function of the original model were proposed as a suitable measure of feature importance. SHAP values provide the unique additive feature importance measure that satisfies the above properties and uses conditional expectations to define simplified inputs. This definition of SHAP values uses a simplified input mapping $h_x(z') = z_S$, where S is the set of non-zero indexes in z', and z_S has missing values for features not in the set S. Since most models cannot handle arbitrary patterns of missing input values, $f(z_S)$ is approximated with $E[f(z)|z_S]$. Given the complexity of the exact computation of SHAP values, it can be approximated by combining insights from current additive feature attribution methods.

Another technique we used was Permutation Importance, which is a technique that allows inspecting ML models by removing a variable from the dataset and then retraining the model to evaluate the impact on some metric such as accuracy or R^2. To avoid retraining the estimator, the variable is replaced by random noise, i.e., the variable is still there but no longer contains valuable information. This method works if the noise is drawn from the same distribution as the original values of the features. This random noise is obtained by alternating the values of the same variable to lose efficiency in the model's training.

3.3. Implementation

The proposed method was implemented using Python 3.7. The following libraries are the most crucial for reproducing our results:

Scikit-Learn: methods for preprocessing, ML models (e.g., Logistic Regression, RF, Lasso Regression, Multilayer perceptron, PCA), cross-validation methods, and metrics for assessing models. (Pedregosa et al., 2011)
Pandas: Library used for reading and managing datasets (McKinney et al., 2011)
Shap: Package which contains functions for providing explainability to predictions (Lundberg & Lee, 2017b)
Gensim: Library used for processing the free-text data (Rehurek & Sojka, 2010)
Spanish embeddings: The Spanish word embeddings used were pretrained by Cañete (2019)

Lastly, the source code will be available at https://github.com/accepted/MLM-EUs

Table 5
Step 1: Patient categorization. Baseline models were trained using standard features.

	Baseline models		Combined features		Word2vec features	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Decision tree	0.84	0.82	0.85	0.83	0.74	0.64
Random forest	0.80	0.75	0.74	0.63	0.74	0.63
Logistic regression	0.63	0.68	0.65	0.69	0.29	0.34
MLP	0.83	0.81	0.79	0.77	0.74	0.63

Table 6
Step 2. Determine patient destination. Baseline models were trained using standard features.

	Baseline models		Combined features		Word2vec features	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Decision tree	0.38	0.36	0.38	0.37	0.36	0.25
Random forest	0.39	0.28	0.37	0.26	0.37	0.26
Logistic regression	0.27	0.34	0.28	0.32	0.20	0.24
MLP	0.38	0.29	0.40	0.31	0.35	0.18

4. Results and analysis

In this section, we present the results obtained from the different classifications and prediction models. Specifically, in Sections 4.1–4.3 we provide the results obtained by employing the automatic learning models studied at each stage of patient care. In 4.4, we identify which variables of the data set contribute significantly to the results obtained and the specific impact of the text vectorized using Word2vec. Moreover, we explain some latent meanings for essential Word2vec features.

4.1. Severity classification

As can be seen in Table 5, there is an increase in the performance of the implemented algorithms when adding to the training and testing dataset a set of text processed by vectorization. Even using only processed text data, it is possible to obtain acceptable results. The best results in terms of accuracy are obtained by applying the decision tree model, with an accuracy of 0.84 with the available numeric data and increasing to 0.85 by including vectorized text data. The F1 score value also increased from 0.82 to 0.83 by including text data in the training of the models. For the combined data column, 300 extra variables are added to the original dataset, corresponding to the components of the average vector obtained from the "Event description". With this enhanced dataset, the models were re-trained. We observed increased accuracy and F-1 scores for the Decision Tree and Logistic Regression.

4.2. Patient destination prediction

As seen in Table 6, the results are less accurate than those obtained in the previous stage due to the poor quality and the little usable information available in the records corresponding to this stage. The results obtained for Logistic Regression and MLP show a minimal but reasonable improvement when vectorized text records are included in the prediction, even considering that they correspond to text referring to the previous attention stage. In addition, to accelerate the time spent executing the models, PCA was performed on the vectors transformed from the text, reducing their dimensionality from 300 components to only 32. Thus, it reduced the computational need of the experiment, reduced the waiting times for results, and slightly increased the performance indicators of the models, as shown in Table 6. The best results were obtained at this stage by applying the MLP model, getting a precision of 0.38. By using vectorized text and reducing its dimensionality by PCA (see Table 7), the accuracy of the results increased to 0.40. It is worth mentioning that the vectorized text with dimensionality reduction significantly improved the F1 score value of the MLP.

Table 7
Step 2. Comparison of patient destination prediction, without text and with PCA text. Baseline models were trained using standard features.

	Baseline models		PCA-Word2vec features	
	Accuracy	F1-Score	Accuracy	F1-Score
Decision tree	0.38	0.36	0.39	0.36
Random forest	0.39	0.28	0.38	0.27
Logistic regression	0.27	0.30	0.29	0.32
MLP	0.38	0.29	0.40	0.35

4.3. LoS prediction

The third proposed prediction stage corresponds to the length of hospitalization of patients. In this stage, confirmation of the diagnosis was added to the hospitalization time, which is the variable to predict. Table 8 presents the results of the prediction of LoS in the hospital for each trained ML model.

At this stage, we also reduced the dimensionality of the average vector obtained by transforming the event description field using Word2vec, reducing the execution time but not significantly impacting the results. Table 9 presents the results of predicting the LoS in the hospital by applying a dimensionality reduction of the combined dataset using PCA.

There was still a tendency to obtain less accurate results due to the lost information in the attention process. Although new information was generated at each stage of the process, this information was textual, unstructured, and challenging to process. For this reason, we have left this information out of the analysis. There was also a tendency to improve the results by including free text from the anamnesis generated in the triage process. The best results were obtained using the decision tree model with a coefficient of determination of 0.16, which improved with the inclusion of vectorized text, reaching a coefficient of 0.23.

In conclusion, based on our results for the three stages of the patient care process studied in this article, unstructured text processed with NLP techniques improved the results of the predictions for the best model found. The following section explores the individual contribution of the components of the vectors resulting from the application of NLP techniques in the results.

4.4. Relevance of features

In this section, we discuss the relevance of each feature on the predictions using interpretability techniques. First, in Section 4.4.1, results from the Permutation Importance method are provided. Then, in Section 4.4.2 feature relevance is estimated using the SHAP technique shown. These techniques are applied to the MLP model for the three stages considered in our paper.

Table 8
Step 3. Comparison of the time prediction for patient hospitalization. Baseline models were trained using standard features.

	Baseline models				Combined features			
	R^2	Max	MAD	MSE	R^2	Max	MAD	MSE
Decision tree	0.16	191	8.73	298.43	0.23	190	7.19	249.88
Random forest	0.14	193	9.17	304.74	0.18	194	7.82	264.29
Lasso	0.12	196	9.53	311.56	0.20	196	8.57	256.94
MLP	0.12	198	9.04	310.26	0.02	201	10.12	315.70

Table 9
Step 3: Comparison of predicted patient hospitalization time, without text and with PCA text. Baseline models were trained using standard features.

	Baseline models				PCA-Word2vec features			
	R^2	Max	MAD	MSE	R^2	Max	MAD	MSE
Decision tree	0.16	191	8.73	298.44	0.23	190	7.18	249.52
Random forest	0.14	192	9.17	304.74	0.18	194	7.81	264.28
Lasso	0.12	196	9.53	311.55	0.19	195	8.72	261.85
MLP	0.12	198	9.04	310.26	0.00	206	8.71	322.14

Table 10
Results of the application of the Permutation Importance to the MLP model at each stage of the care process. Baseline models were trained using standard features.

Patient classification		Destination prediction		Days of hospitalization	
Weight	Feature	Weight	Feature	Weight	Feature
0.0562 ± 0.0051	EVA	0.0309 ± 0.0020	CAT_C3	0.1354 ± 0.0231	DM_N
0.0331 ± 0.0045	SATO2	0.0256 ± 0.0071	Desc_31	0.1322 ± 0.0267	DM_S
0.0133 ± 0.0016	DM_S	0.0243 ± 0.0042	Desc_169	0.0928 ± 0.0049	PAC_EDAD
0.0122 ± 0.0021	GLASGOW	0.0236 ± 0.0045	CAT_C2	0.0482 ± 0.0052	Desc_206
0.0116 ± 0.0030	FR	0.0224 ± 0.0068	Desc_80	0.0433 ± 0.0027	Desc_114
0.0067 ± 0.0029	Desc_237	0.0202 ± 0.0048	Desc_95	0.0290 ± 0.0116	EVA
0.0056 ± 0.0027	PAC_EDAD	0.0201 ± 0.0031	DM_S	0.0246 ± 0.0088	Desc_89
0.0049 ± 0.0027	Desc_187	0.0201 ± 0.0037	Desc_292	0.0240 ± 0.0061	Desc_221
0.0048 ± 0.0028	Desc_130	0.0201 ± 0.0048	Desc_258	0.0232 ± 0.0089	Desc_42
0.0044 ± 0.0036	Desc_77	0.0196 ± 0.0070	Desc_106	0.0224 ± 0.0061	Desc_158

4.4.1. *Use of permutation importance for feature analysis*

This method of feature importance analysis evaluates the impact of removing any of the variables from the data set on the model's performance indicators. Features presented in Table 10 were described in Table 1. Features denoted with the prefix "DESC" correspond to components of the 300-dimensional vectors obtained by applying Word2vec to the event description variable.

As we can see in Table 10, the EVA variable and SATO2 of the vectorized text are the principal responsible for the results obtained using the MLP model.

It should be noted that the EVA (airway assessment) variable is important in obtaining results in the three stages of the care model studied. Another feature frequently appears is the saturation oxygenation level (SATO2). Altered oxygenation levels are directly related to the patient's category and condition during the care process. Another aspect to consider is the great importance of diabetes mellitus (DM) in predicting the LoS of patients in hospitalization. There is likely a direct relationship between this disease's existence and the patient's hospitalization. At least one text-embedding feature is observed within the top-five most important features in the three stages.

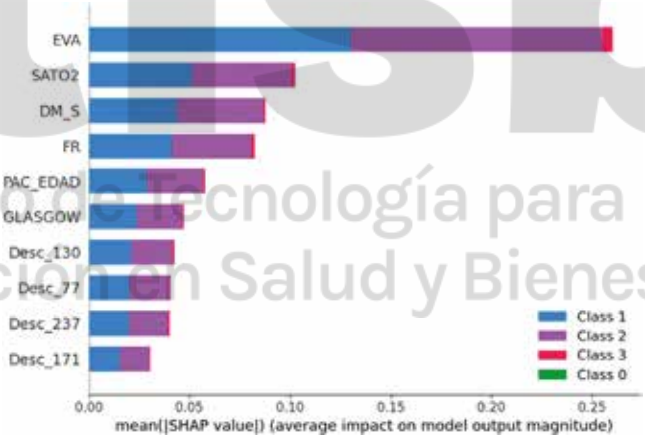

Fig. 5. SHAP values applied to MLP modeling in step 1. Color bar represents the class and length represents the impact magnitude. Class 1 = category C3, Class 2 = category C2, and Class 3 = category C4.

4.4.2. *Use of SHAP values for feature analysis*

When applying SHAP values to analyze the contribution of text in the prediction of severity categories using MLP, we identified that the variable EVA reappears as one of the most relevant variables in the classification of patient severity, together with some components of vectorized words.

In Fig. 5, we can see that the EVA and SATO2 features appear again as the most important in classifying patients according to severity level, obtaining a more significant contribution according to the average impact on model output magnitude. We can also see that such variables significantly influence the three most represented categories of patients (Class 1, Class 2, and Class 3). Furthermore, we highlight features FR, PAC_EDAD, GLASGOW, Desc_130, Desc_77, Desc_237, and Desc_171 as good predictors for severity classification.

In Fig. 6 explaining stage 2 shows that the descriptors have a significant impact on the outcome. Each column presents through the colors the magnitude in which the result is described from the variable.

In Fig. 7, we can see a different visualization to those described in the previous stages. This is because, unlike in stages 1 and 2, in stage 3, we try to predict the patient's LoS through regression models. In other words, the data presented in the figure represents how the features

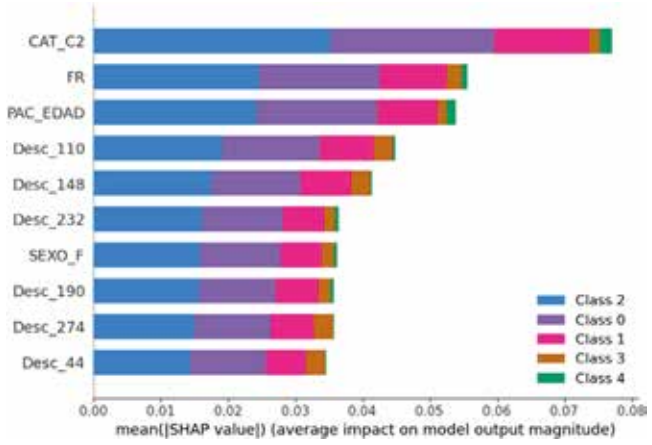

Fig. 6. SHAP values applied to MLP modeling in stage 2. Color bar represents the class and length represents the impact magnitude.

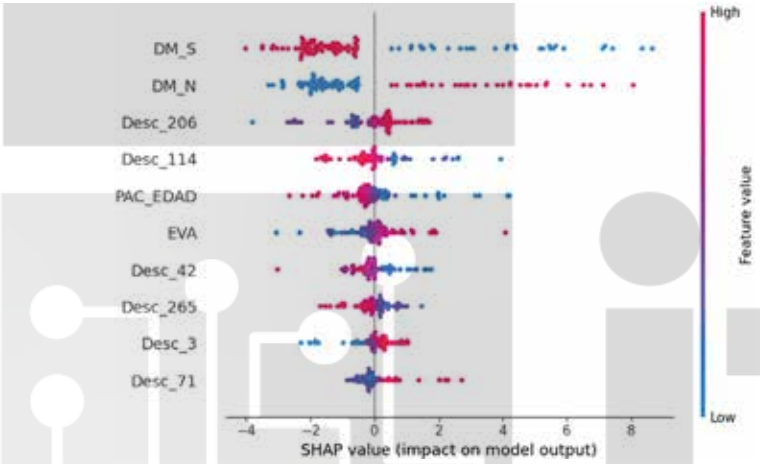

Fig. 7. SHAP values applied to MLP model in stage 3.

influenced the LoS calculation and not necessarily the classification of patients. In this figure, six out of 10 features are text descriptors and identified as being relevant for determining the models' outcome. Moreover, we can see that the features DM, EVA, and patient age have a significant impact on the results of the MLP model. Specifically, DM-S, which denotes the presence of diabetes mellitus in the patient, generates a high impact on the outcome, unlike DM-N, which indicates the absence of the disease in the patient.

4.5. Meaning discovery of relevant dimensions

In previous sections, we noted that some components of sentence embeddings appeared as essential features in obtaining results from ML models. Because of that, we propose understanding the implicit meaning of these vectors through modifications δ in the value along the specific relevant dimension. Firstly, we get the embedding of the word *paciente* (patient), then we apply the before commented shift, obtaining the words presented in Table 11.

When looking at Table 11, we observe that when δ is small, the near words are closely related to concepts of preoperation and postoperation. Then, as the component's value increased, new words gradually appeared with similar contexts. This experiment provides an alternative for discovering and understanding the meaningful of explanations provided by SHAP.

Table 11
Modification of the value of component No. 277 for the vector representation of the word "patient". The original Spanish word and the English translation are presented original/translation.

δ	Most near words	
−6	diagnosticador/diagnostician	postoperatorias/postoperative
−4	diagnosticador/diagnostician	postoperatorio/postoperative
−2	pacientes/patients	postoperatorio/postoperative
origin	pacientes/patients	preoperatorio/preoperative
2	pacientes/patients	preoperatorio/preoperative
4	preoperatorio/preoperative	pacientes/patients
6	clonazepane/clonazepano	clonazepan/clonazepan

5. Conclusions

We propose using sentence embeddings to improve prediction performances using information from EUs. Moreover, our approach considers explanations using SHAP values and suggests a procedure to discover meaning from relevant dimensions of sentence embeddings. From a practical point of view, this method offers an alternative to easily incorporate text in predictive models in UEs, improving the quality of predictions and, consequently, preparing better planning of resources. Our methodology can be considered model-agnostic since both the embedding generator and the classifier can be replaced.

After the experiments performed and the result obtained, we can conclude that the contribution of including free text as training data for prediction and classification models, when available, is an extra effort considering classical ML methodologies. However, when datasets present missing or minor information, each valid data becomes crucial for obtaining good quality predictions and classifications. In our experiments, the vectorized text was shown to contribute enough information for prediction, improving the results to determine the severity category of patients at an acceptable level.

Another relevant aspect is the importance of interpretability in ML. Although the models implemented in this research are not the most complex in this field of study, it is difficult to understand some of them (e.g., MLP) and why the training data allowed us to obtain results such as those described. That is why explainability techniques provide transparency to the results, give researchers and practitioners more confidence in their results, and allow physicians or end-users to understand the behavior of relevant data in treating a patient.

Future research could consider an in-depth study of the meaning of relevant dimensions from this type of sentence embedding. Moreover, this methodology could be applied using the most recent and ad-hoc approaches to represent text from EUs. Finally, we also encourage the application of this method using other sources of textual information, e.g., anamnesis of patients.

CRediT authorship contribution statement

Jonathan Moya-Carvajal: Conceptualization, Data curation, Methodology, Software, Visualization, Writing – original draft. **Francisco Pérez-Galarce:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Carla Taramasco:** Methodology, Validation, Writing – review & editing. **César A. Astudillo:** Methodology, Validation, Writing – review & editing. **Alfredo Candia-Véjar:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We are grateful to four anonymous reviewers for their valuable comments that help to improve the paper.

F. Pérez-Galarce acknowledges support from ANID, through Scholarship Program/Doctorado Nacional/2017-21171036.

We thank the project FONDECYT Regular 1201787-Multimodal Machine Learning approach for detecting pathological activity patterns in elderly people and ANID - MILENIO - NCS2021_013.

References

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for Artificial Intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, URL https://pubmed.ncbi.nlm.nih.gov/33256715/.

Amunategui, M., Markwell, T., & Rozenfeld, Y. (2015). Prediction Using Note Text: Synthetic Feature Creation with word2vec. arXiv:1503.05123.

Azeez, D., Ali, M. A. M., Gan, K. B., & Saiboon, I. (2013). Comparison of adaptive neuro-fuzzy inference system and artificial neutral networks model to categorize patients in the emergency department. *SpringerPlus*, *2*(1), 416.

Azeez, D., Gan, K. B., Ali, M. A., & Ismail, M. S. (2015). Secondary triage classification using an ensemble random forest technique. *Technology and Health Care*, *23*(4), 419–428. http://dx.doi.org/10.3233/THC-150907.

Bacchi, S., Gluck, S., Tan, Y., Chim, I., Cheng, J., Gilbert, T., et al. (2020). Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study. *Internal and Emergency Medicine*, (Ml), http://dx.doi.org/10.1007/s11739-019-02265-3.

Bardak, B., & Tan, M. (2021). Improving clinical outcome predictions using convolution over medical entities with multimodal learning. *Artificial Intelligence in Medicine*, *117*(July 2020), Article 102112. http://dx.doi.org/10.1016/j.artmed.2021.102112, arXiv:2011.12349.

Camilloni, L., Rossi, P. G., Farchi, S., Chini, F., Borgia, P., & Guasticchi, G. (2010). Triage and Injury Severity Scores as predictors of mortality and hospital admission for injuries: A validation study. *Accident Analysis and Prevention*, *42*(6), 1958–1965.

Cañete, J. (2019). Spanish Word Embeddings. http://dx.doi.org/10.5281/zenodo.3255001.

Ceglowski, A., Churilov, L., & Wassertheil, J. (2005). Knowledge discovery through mining emergency department data. *00*, In *Proceedings of the annual Hawaii international conference on system sciences* (C), (p. 142).

Centers for Medicare and Medicaid Services, National Center for Health Statistics, & Centers for Medicare and Medicaid Services and National Center for Health Statistics (2017). ICD-10-CM Official Guidelines for Coding and Reporting FY2018. *2018*, 1–117, URL https://www.cms.gov/medicare/coding/icd10/downloads/icd10cm-guidelines-2015.pdf.

Chapman, W. W., Fiszman, M., Dowling, J. N., Chapman, B. E., & Rindflesch, T. C. (2004). Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. *Medinfo*, *11*(Pt 1), 487–491.

Chen, C. H., Hsieh, J. G., Cheng, S. L., Lin, Y. L., Lin, P. H., & Jeng, J. H. (2020). Early short-term prediction of emergency department length of stay using natural language processing for low-acuity outpatients. *American Journal of Emergency Medicine*, *38*(11), 2368–2373. http://dx.doi.org/10.1016/j.ajem.2020.03.019.

Claster, W., Shanmuganathan, S., & Ghotbi, N. (2008). Text mining of medical records for radiodiagnostic decision-making. *Journal of Computers*, *3*(1), 1–6. http://dx.doi.org/10.4304/jcp.3.1.1-6.

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, *42*(5), 760–772. http://dx.doi.org/10.1016/j.jbi.2009.08.007.

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2020). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, (June), 1–18. http://dx.doi.org/10.1111/coin.12410.

Grafstein, E., Unger, B., Bullard, M., Innes, G., et al. (2003). Canadian emergency department information system (CEDIS) presenting complaint list (Version 1.0). *Canadian Journal of Emergency Medicine*, *5*(1), 27–34.

Hong, W. S., Haimovich, A. D., & Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning. *PLoS One*, *13*(7), Article e0201016.

Hoot, N. R., & Aronsky, D. (2008). Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of Emergency Medicine*, *52*(2), 126–136. http://dx.doi.org/10.1016/j.annemergmed.2008.03.014, URL http://www.ncbi.nlm.nih.gov/pubmed/18433933.

Jafari-Rouhi, A. H., Sardashti, S., Taghizadieh, A., Soleimanpour, H., & Barzegar, M. (2013). The emergency severity index, version 4, for pediatric triage: a reliability study in Tabriz Children's Hospital, Tabriz, Iran. *International Journal of Emergency Medicine*, *6*(1), 36.

Jia, Y., McDermid, J., Lawton, T., & Habli, I. (2021). The role of explainability in assuring safety of machine learning in healthcare. arXiv:2109.00520.

Jiang, X., Qu, X., & Davis, L. (2010). Using data mining to analyze patient discharge data for an urban hospital. In *DMIN 2010, International conference on data Mining; 2010 Jul 12-15; Las Vegas, NV* (March), (pp. 139–144).

Jo, S., Jeong, T., Jin, Y. H., Lee, J. B., Yoon, J., & Park, B. (2015). ED crowding is associated with inpatient mortality among critically ill patients admitted via the ED: post hoc analysis from a retrospective study. *The American Journal of Emergency Medicine*, *33*(12), 1725–1731.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in Neural Information Processing Systems*, *29*.

Lin, W. T., Wang, S. T., Chiang, T. C., xin Shi, Y., yu Chen, W., & min Chen, H. (2010). Abnormal diagnosis of Emergency Department triage explored with data mining technology: An Emergency Department at a Medical Center in Taiwan taken as an example. *Expert Systems with Applications*, *37*(4), 2733–2741. http://dx.doi.org/10.1016/j.eswa.2009.08.006.

Lin, W. T., Wu, Y. C., Zheng, J. S., & Chen, M. Y. (2011). Analysis by data mining in the emergency medicine triage database at a Taiwanese regional hospital. *Expert Systems with Applications*, *38*(9), 11078–11084. http://dx.doi.org/10.1016/j.eswa.2011.02.152.

Lundberg, S., & Lee, S.-I. (2017a). A unified approach to interpreting model predictions. arXiv:1705.07874.

Lundberg, S. M., & Lee, S.-I. (2017b). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Marfil-Garza, B. A., Belaunzarán-Zamudio, P. F., Gulias-Herrero, A., Zuñiga, A. C., Caro-Vega, Y., Kershenobich-Stalnikowitz, D., et al. (2018). Risk factors associated with prolonged hospital length-of-stay: 18-year retrospective study of hospitalizations in a tertiary healthcare center in Mexico. *PLoS One*, *13*(11), Article e0207203.

Mathew, G., & Obradovic, Z. (2012). Distributed privacy preserving decision support system for predicting hospitalization risk in hospitals with insufficient data. *2*, In *Proceedings - 2012 11th international conference on machine learning and applications, ICMLA 2012* (pp. 178–183). IEEE, http://dx.doi.org/10.1109/ICMLA.2012.180.

McKinney, W., et al. (2011). Pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*(9).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *1st international conference on learning representations, ICLR 2013 - workshop track proceedings* (pp. 1–12). arXiv:1301.3781.

Parva, E., Boostani, R., Ghahramani, Z., & Paydar, S. (2017). The necessity of data mining in clinical emergency medicine; A narrative review of the current literatrue. *Bulletin of Emergency and Trauma*, *5*(2), 90–95.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Plumb, G., Molitor, D., & Talwalkar, A. (2019). Model agnostic supervised local explanations. arXiv:1807.02910.

Raita, Y., Goto, T., Faridi, M. K., Brown, D. F., Camargo, C. A., & Hasegawa, K. (2019). Emergency department triage prediction of clinical outcomes using machine learning models. *Critical Care*, *23*(1), 64.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2939672.2939778.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *32*, In *Proceedings of the AAAI conference on artificial intelligence*. (1), URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.

Santelices, E., & Santelices, J. L. (2017). Descripción y análisis del sistema de red de urgencia (RDU) en Chile. Recomendaciones desde una mirada sistémica. *Revista Médica Clínica Las Condes*, *28*(2), 186–198.

Sariyer, G., Öcal T., C., & Cepe, G. (2019). Use of data mining techniques to classify length of stay of emergency department patients. *Bio-Algorithms and Med-Systems*, *15*(1), 1–15. http://dx.doi.org/10.1515/bams-2018-0044.

Seymour, C. W., Cooke, C. R., Wang, Z., Kerr, K. F., Yealy, D. M., Angus, D. C., et al. (2013). Improving risk classification of critical illness with biomarkers: a simulation study. *Journal of Critical Care*, *28*(5), 541–548.

Soler, W., Gómez Muñoz, M., Bragulat, E., & Álvarez, A. (2010). El triaje: herramienta fundamental en urgencias y emergencias. *33*, In *Anales Del Sistema Sanitario de Navarra* (pp. 55–68). SciELO Espana.

St George, D. (1992). Nurse triage in accident and emergency departments. *BMJ (Clinical Research Edtiton)*, *304*(6838), 1379. http://dx.doi.org/10.1136/bmj.304.6838.1379-a.

Sterling, N. W., Patzer, R. E., Di, M., & Schrager, J. D. (2019). Prediction of emergency department patient disposition based on natural language processing of triage notes. *International Journal of Medical Informatics*, *129*, 184–188. http://dx.doi.org/10.1016/j.ijmedinf.2019.06.008.

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), 1–13. http://dx.doi.org/10.1002/widm.1379, arXiv:2002.08596.

Tu, J. V., & Guerriere, M. R. (1992). Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. In *Proceedings / the ... annual symposium on computer application [Sic] in medical care. Symposium on computer applications in medical care* (pp. 666–672).

Wilding, D., & Evans, K. (2017). Predicting length of stay for acute medical admissions using the ALICE score: a simple bedside tool. *Acute Medicne*, *16*(2), 60–64.