Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 210 (2022) 248–253

Procedia Computer Science

www.elsevier.com/locate/procedia

The 3rd International Workshop on Healthcare Open Data, Intelligence and Interoperability (HODII)
October 26-28, 2022, Leuven, Belgium

Predictive analytics for hospital discharge flow determination

Mariana Faria[a], Agostinho Barbosa[b], Tiago Guimarães[a], João Lopes[a], Manuel Santos[a*]

[a]*ALGORITMI/LASI Research Center, University of Minho, Portugal*
[b]*Centro Hospitalar do Tâmega e Sousa, Portugal*

Abstract

In recent years, hospitals around the world are faced with large patient flows, which negatively affect the quality of patient care and become a crucial factor to consider in inpatient management. The main objective of this management is to maximize the number of available beds, using efficient planning. Intensive Care Units (ICU) are hospital units with a higher monetary consumption, and the importance of indicators that allow the achievement of useful information for a correct management is critical. This study allowed the prediction of the Length of Stay (LOS) based on their demographic data, information collected at the time of admission and clinical conditions, which can help health professionals in conducting a more assertive planning and a better quality service. The results obtained show that Machine Learning (ML) models, using demographic information simultaneously with the patient's pathway, as well as clinical data, drugs, tests and analysis, introduce a greater predictive ability for LOS.

Keywords: Length of Stay; Machine Learning; Predictive Analytics

1. Introduction

Healthcare organizations are in a period of great transformation in their management processes and in the quality of care provided to patients. Alongside this reflection, the constant growth of their costs has been evidenced, which highlights the high need to optimize the management of resources in a more effective and intelligent way [1].

* Corresponding author.
E-mail address: mfs@dsi.uminho.pt

1877-0509
10.1016/j.procs.2022.10.145

Intensive Care Units (ICUs), which specialize in providing care to critically ill patients, stand out as the most demanding and costly in terms of expenditure and use of resources, largely due to the high level of daily unpredictability in the type of patients present.

Besides this, the vast volume of information stored by hospitals is also well known, which makes obtaining useful information for decision making increasingly complex and challenging. However, once computerized, useful insights can be obtained by applying Data Mining (DM) techniques, applications that, in Health, can be used to evaluate the effectiveness of a treatment [2], the management of hospital resources [3], the reduction of health insurance fraud [4], among others.

For these reasons, in this study is intended to predict the Length of Stay (LOS) of a patient in the Multipurpose Intensive Care Unit (UCIP), corresponding to the number of days that a patient remains hospitalized. Thus, Health professionals will be able to estimate the flow of future discharges, allowing the optimization of hospital resources, reducing costs.

2. Background

2.1. Resources Planning in Hospital Settings

Worldwide hospitals are faced with continuous overcrowding, mostly caused by the lack of beds, which leads to, postponement of scheduled hospitalizations, both for surgeries, as well as any other pathological treatment, and the increased risk of contracting any disease [5]. However, less frequently, the opposite is also true, where the needs are less than the number of beds available, leading to unnecessary costs. Thus, bed management is an issue present in hospitals, which, when executed incorrectly, causes a mismatch between means and human resources, thus implying considerable costs.

It should also be noted that hospitals nowadays generate a large volume of data in their various interactions with patients, making it increasingly difficult to understand them. In this way, the use of DM techniques, in recent years, has been a potentiating factor to transform the valuable knowledge stored in useful information [6], which positively promotes the decision making of the hospital. That said, if it is possible to accurately predict the LOS in the specialties to which the patient is admitted, it becomes feasible to assist professionals in developing more assertive planning.

2.2. Related Works

The application of ML techniques to the prediction of LOS is increasingly notorious [6], allowing an optimization of the planning and management of hospital resources.

When the focus shifts to Intensive Care Units (ICU), the LOS becomes an even more important metric, given the difficulty present in these units regarding the control of the patient's condition, as well as the fact that at any time a patient may be admitted there. As a result, some studies have been made to predict the LOS in these units. An example of this, Veloso et al. [1] estimated LOS in an intensive care unit based on two different DM classification approaches. The approach that considered admission data and clinical data (vital signs and laboratory results) collected in real time from the patient proved to be more accurate for predicting LOS to find the most likely time a patient would be discharged, as opposed to the approach based on admission and physiological patient variables collected during the first 24 hours of hospitalization. It is noteworthy that in this context the Decision Tree (DT) technique presented an accuracy of 75% and sensitivity of 87%. This study has shown that at the ICU level, LOS prediction models at the moment of patient admission are not interesting. The models that adapt in real time to the LOS, according to the patient's conditions, thus making it possible to predict patient discharge hourly, are highlighted [E1].

Also in an ICU, Gholipour et al. [7] applied the Artificial Neural Network technique to predict the LOS of patients with traumatic injuries in a hospital in Tabriz, Iran. This was based on clinical and biochemical patient analysis data collected between October 2006 and October 2009. The results achieved with this model were quite satisfactory in the three metrics used, Sensitivity 75%, Specificity 96.3% and Accuracy 93.3% [E2].

Focusing on the Covid-19 pandemic, the management of hospital resources proved to be an even greater challenge. Dina A. Alabbad et al. [8] proceeded to predict the LOS in the intensive care unit at King Fahad University Hospital, Saudi Arabia. The data source used includes a total of 47 attributes referring to information collected at the time of admission, as well as information regarding the patient's clinical history (e.g. age, gender, fever, comorbidities, laboratory results). As ML algorithms, Random Forest (RF) and Gradient Boosting (GB) were applied, highlighting RF with an accuracy of 94.1%, compared to 88.1% for GB. Regarding the definition of classes, 9 possibilities were taken into consideration: class 1 is for patients who did not require intensive care, class 2 for patients staying in the service in question for less than 1 day, and the remaining classes, 3 to 9, were used to represent the various periods of days, 1-5, 6-10, 11-15, 16-20, 21-25, 26-30, and more than 30, respectively. This study demonstrated that the variables, age, C-Reactive protein and days on oxygen support are the main factors implicated in the target set [E3].

From the analysis of the State-of-the-Art presented, it is possible to verify the inherent importance of the patient's LOS, revealing itself to be a crucial indicator for hospital management. With the use of these models, a better planning of hospital resources is provided, allowing an increase in the number of beds available for new admissions and the reduction of waiting lists.

Overall, it can be seen that good results in predicting a patient's LOS are related to the use of demographic information. It is denoted that little attention is paid to the whole pathway already traveled since the moment of admission to the inpatient service, as well as information regarding drugs administered to the patient, and analyses performed, thus dealing with a theme little studied in detail.

The present problem will cross-reference all this information to analyze the implications on LOS. For this purpose, a multi-class approach will be followed, where the LOS intervals are not defined previously for the identification of classes, but all the possibilities of remaining residence times are considered, adopting a more informative approach.

3. Materials and Methods

For this research, two methodologies were considered: DSR (Design Science Research), composed of six activities [9], which are: Problem Understanding (1), Suggestion (2), Development (3), Evaluation (4), Conclusion (5) and Communication (6). This methodology has become increasingly popular in Information Systems, as it allows the acquisition of new technical and scientific knowledge from the design of innovative artifacts to solve a practical problem in a specific context.

To support Data Mining (DM) techniques, the methodology adopted was CRISP-DM (*Cross Industry Standard Process for Data Mining*), a methodology that describes the DM project life cycle, consisting on a set of 6 phases [10]: Business Understanding (1), Data Understanding (2), Data Preparation (3), Modeling (4), Evaluation (5), and Implementation (6). The present work only covers the first 5 phases since the implementation can be developed in future work.

Python programming language was used for the analysis of data sources. These were made available by the Centro Hospitalar do Tâmega e Sousa (CHTS), containing admissions, discharges, transfers between specialties, diagnoses, exams, analyses, drugs and surgeries performed in the inpatient service, and also patient demographics.

4. Case Study

This section presents the processes and decisions made in the first 5 phases of the CRISP-DM methodology. As previously mentioned, the present study aims to predict the patient's LOS in an inpatient service of the CHTS.

4.1. Business Understanding

The objective of this research is to find alternatives that lead to an optimization of efficiency in the planning and management of hospital beds, through the prediction of the LOS. The UCIP was chosen, given its inherent unpredictability, since at any moment a patient may enter in this unit, as well as the difficulty related to the control

of the patient's condition. In this way, a more adequate planning of the number of available beds is possible, once the time that a certain patient will remain hospitalized is known.

4.2. Data Understanding

The data provided by the CHTS is restricted to a time interval of 5 years, from 2017 to 2021, containing data regarding admissions, discharges, transfers between specialties, diagnoses, tests and analyses performed, drugs administered in the inpatient service, and also patient demographics.

4.3. Data Preparation

first phase consisted of selecting the attributes that aim to demonstrate a higher predictive value for the models. New attributes were derived, where age was constructed according to the date of birth. At the level of a clinical record, attributes referring to the number of services visited, previous hospitalizations, exams and tests performed, and medications administered by the patient from the time of admission to the specialty to the performance of complementary diagnostic means in a day were designed, identifying how many were new and how many were repeat requests. At a global level, an attribute called affluence was created to characterize the number of patients admitted to the UCIP at the time of admission, as well as a variable called record type, identifying the time of admission.

Using the discharge date, we defined the target, which represents the number of days of hospitalization remaining in this specialty. As for the definition of the classes, these correspond to the possible values for the defined target, except for the first class that includes all records when the remaining LOS is between 1 and 3 days. The multiple data sources were integrated to originate a compact data source with all the necessary cross information to serve as input for the developed models.

Due to irrationally high values in terms of dwell time, it was necessary to treat the target variable by assuming the existence of outliers. As for data formatting, we coded the data so that they could be processed by ML techniques present in the Scikit-learn library.

4.4. Modelation

A variety of classification techniques were applied, namely Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and Gradient Boosting (GB). In terms of sampling and validation of results, the Cross Validation K-fold (CV) technique was used, since it allows the use of all data for training and testing, providing greater reliability in the models developed [11].

Given the presence of an imbalance of classes in the defined target, it was necessary to use an oversampling method to balance the respective classes.

This study took into consideration two distinct scenarios, the first corresponding to the prediction of LOS at the time of patient admission to the UCIP (C1) and the second to the prediction of LOS both at the time of patient admission and at the time of analysis, tests or drug administration (C2). This resulted in 8 models (2 scenarios x 4 techniques).

4.5. Evaluation

To verify whether the developed models meet the defined objectives, it is crucial to evaluate them. To compare the performance of all algorithms, it was necessary to select the following evaluation metrics: Accuracy (AC), Precision (PC), Recall (RC), F1-Score (F1), Kappa (KP) and Area Under the ROC Curve (AUC). This selection is based on the type of approach adopted, as well as on all the knowledge acquired from the study of the works already applied under the same context. Success criteria were defined for each defined evaluation metric:

- AC, PC, RC, F1, AUC $\geq$ 85%;
- KP $\geq$ 80%.

5. Results And Discussion

Table 4 presents the best results achieved for each scenario for the defined target.

Table 1. ML Table Results

Model	Metric	C1	C2
DT	AC	90.85	93.14
	PC	90.78	93.38
	RC	90.33	93.14
	F1	89.84	92.78
	KP	88.72	92.76
	AUC	99.94	99.97
RF	AC	91.66	94.44
	PC	92.02	94.60
	RC	91.66	94.44
	F1	91.38	94.38
	KP	90.27	94.13
	AUC	99.94	99.96
KNN	AC	82.92	86.70
	PC	83.24	87.52
	RC	82.92	86.70
	F1	81.14	86.50
	KP	80.07	85.96
	AUC	99.51	99.82
GB	AC	91.92	94.46
	PC	92.21	94.62
	RC	91.92	94.46
	F1	91.70	94.40
	KP	90.57	94.15
	AUC	99.94	99.97

Second scenario (C2) shows the best performance, meaning that the predictive models achieve better performances when predict the moment of patient admission introducing new data to the algorithms, namely analysis, exams and drugs.

Moreover, among the various algorithms implemented in the two scenarios, the GB technique stands out from the others, since it combines a set of "weak" models iteratively [12], with the aim of increasingly shortening the error obtained in the previous model [13]. Moreover, this type of model performs well with categorical input variables [14].

6. Conclusions

This study aims to portray the ability to predict the LOS of a patient, upon admission, as well as during the patient's clinical process. This prediction was supported by applying ML techniques, based on patient demographics and clinical records for the period from 2017 to 2021. To achieve the respective predictive results, two different scenarios were considered, with the application of four types of ML techniques. Of these combinations, GB in C2 was the one that, given the defined evaluation metrics, achieved the best predictive results with an accuracy of 94.5%. The importance of introducing data regarding complementary diagnostic tests performed during the patient's hospitalization is highlighted, since all ML techniques performed better in C2 than in C1.

Thus, it is expected that the application of the developed model may bring several benefits to CHTS, since the accurate and prior knowledge of the LOS in this unit will allow a more adequate discharge planning, and a more optimized management of the number of beds, leading to a better quality of care provided to the patient.

In terms of future work, it is expected to proceed with the implementation of the ML model at CHTS along with an Adaptive Business Intelligence (ABI) system, extended to all inpatient service specialties [15].

Acknowledgements

The work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: DSAIPA/DS/0084/2018.

References

[1] R. Veloso *et al.*, «Real-time data mining models for predicting length of stay in Intensive Care Units», *KMIS 2014 - Proceedings of the International Conference on Knowledge Management and Information Sharing*, n. Dm, pp. 245–254, 2014, doi: 10.5220/0005083302450254.

[2] H. C. Koh e G. Tan, «Data mining applications in healthcare.», *Journal of healthcare information management : JHIM*, vol. 19, n. 2, pp. 64–72, 2005, doi: 10.4314/ijonas.v5i1.49926.

[3] M. K. Obenshain, «Application of Data Mining Techniques to Healthcare Data», *Infection Control & Hospital Epidemiology*, vol. 25, n. 8, pp. 690–695, 2004, doi: 10.1086/502460.

[4] N. Priya, C. Anuradha, R. Kavitha, e D. Vimala, «Analysing data mining applications in healthcare sector», *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, n. 9 Special Issue 3, pp. 1119–1122, 2019, doi: 10.35940/ijitee.I3242.0789S319.

[5] Randhawa & Humayun, «Reasons of Overcrowding in Emergency Department», em *Journal of the Society of Obstetrics and Gynaecologists of Pakistan*, 1.ª ed., vol. 8, 2018.

[6] J. Boyle *et al.*, «Predicting emergency department admissions», *Emergency Medicine Journal*, vol. 29, n. 5, pp. 358–365, 2012, doi: 10.1136/emj.2010.103531.

[7] C. Gholipour, F. Rahim, A. Fakhree, e B. Ziapour, «Using an artificial neural networks (ANNS) model for prediction of intensive care unit (ICU) outcome and length of stay at hospital in traumatic patients», *Journal of Clinical and Diagnostic Research*, 2015, doi: 10.7860/JCDR/2015/9467.5828.

[8] Dina A. Alabbad , Abdullah M. Almuhaideb, Shikah J. Alsunaidi, Kawther S. Alqudaihi, Fatimah A. Alamoudi, Maha K. Alhobaishi, Naimah A. Alaqeel, Mohammed S. Alshahrani, «Machine learning model for predicting the length of stay in the intensive care unit for Covid-19 patients in the eastern province of Saudi Arabia», 14 de abril de 2022.

[9] V. Vijay, B. Kuechler, e S. Petter, «Design Science Research in Information Systems», n. 1, pp. 1–66, 2012, doi: 1756-0500-5-79 [pii]\r10.1186/1756-0500-5-79.

[10] C. Pete *et al.*, «Crisp-Dm 1.0», *CRISP-DM Consortium*, p. 76, 2000.

[11] G. James, D. Witten, T. Hastie, e R. Tibshirani, *Springer Texts in Statistics An Introduction to Statistical Learning - with Applications in R*. 2013.

[12] J. H. Friedman, «Stochastic gradient boosting», *Computational Statistics and Data Analysis*, vol. 38, n. 4, pp. 367–378, 2002, doi: 10.1016/S0167-9473(01)00065-2.

[13] M. Kuhn e K. Johnson, *Applied Predictive Modeling with Applications in R*. 2013.

[14] Tomonori Masui, «All You Need to Know about Gradient Boosting Algorithm − Part 1. Regression», jan. 2020.

[15] Michalewicz, Zbigniew & Schmidt, Martin & Michalewicz, Matthew & Chiriac, Constantin. (2007). Adaptive Business Intelligence. 10.1007/978-3-540- 32929-9.