Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 231 (2024) 433–438

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Workshop on Healthcare Open Data, Intelligence and Interoperability (HODII)
November 7-9, 2023, Almaty, Kazakhstan

Prediction of nosocomial infections associated with surgical interventions

Diogo Fernandes[a], Sara Cardoso[b], João Miranda[b], Júlio Duarte[a,*], Manuel Filipe Santos[a]

[a]ALGORITMI/LASI Research Center, University of Minho, Guimarães, Portugal, [b] Hospital da Senhora da Oliveira, Guimarães

Abstract

Nosocomial infections represent an ongoing challenge to healthcare quality and patient safety, negatively impacting clinical outcomes and increasing the burden on healthcare systems. Thus, controlling this type of infection plays a very important role in ensuring a better quality of life for patients. Although the control and prevention measures for these infections are well defined, their signaling and detection is carried out manually and sometimes late, which compromises the health status of patients and everyone around them. In this context, this study emerged with the aim of exploring the potential of data mining techniques to predict the occurrence of nosocomial infections, with a specific focus on infections associated with surgical interventions.

Using datasets for the period between 2018 and 2022, sourced from a Portuguese hospital and duly anonymized to protect patient privacy, several classification algorithms and data balancing techniques were analyzed to deal with the uneven nature of the data and the presence of minority classes. Among the algorithms and balancing techniques used, it was found that the Random Forest algorithm combined with the Oversampling technique showed superior performance in identifying cases of nosocomial infections associated with surgical interventions.

The results of this study highlight the importance of collaboration between medicine and technology, indicating that the integration of data mining techniques can prove to be valuable tools to improve clinical decision-making and infection management in surgical context.

Keywords: Classification; CRISP-DM; Data Mining; Nosocomial Infections; Prediction; Surgical Interventions.

* Corresponding author. Tel.: +351 253510319; fax: +351 253510300.
E-mail address: jduarte@di.uminho.pt

10.1016/j.procs.2023.12.230

1. Introduction

Nosocomial infections, also known as hospital-acquired infections, are defined as infections acquired in the hospital or healthcare institution that first manifest 48 hours or more after admission or within 30 days after discharge. They are unrelated to the reason for hospitalization and are not present or incubating at the time of admission [1].

These infections have become a major public health problem worldwide and are among the leading causes of mortality and morbidity in hospitals, leading to longer patient stays, sometimes requiring readmission, and increased hospital costs [2].

According to the World Health Organization, about 7% of patients in developed countries and 10% in emerging and developing countries acquire at least one type of nosocomial infection, and about 10% of these patients die as a direct result of the infection [3]. In the European Union, according to the European Centre for Disease Prevention and Control, more than 3.2 million people are affected by healthcare-associated infections each year, resulting in at least 37000 deaths [4].

These numbers prove the existence and the seriousness of the problem, which makes it necessary to act to reduce the number of infections and, consequently, the number of deaths, hospital costs and other complications. This is the context of this project, which aims to use data mining techniques to create predictive models capable of predicting the occurrence of infections in patients undergoing surgical procedures, thus helping not only to prevent and control infections, but also to improve the care provided. In the area of nosocomial infections, surgical infections are the third most common type of infection (18,3%), surpassed only by pneumonia (21,4%) and urinary tract infections (18,9%), which reinforces the need to act against this type of infection [5].

2. State of Art

2.1. Data Mining in Healthcare

Healthcare institutions collect and generate data every day, which creates large and complex datasets, making it difficult to analyze them to support decision making. Therefore, because there is a need to have an efficient methodology to detect unknown and valuable information in the data, data mining has become increasingly popular in this specific area.

With the application of data mining, it is possible to collect several benefits that will help in several aspects, such as providing professionals with faster, more informed and accurate decisions, thus providing an improvement in efficiency and productivity, which allows improving the treatment provided to the patient and reducing the cost of the same. It is also applied in other situations in the health area, such as effective management of hospital resources, hospital classification, better customer relations, hospital infection control, recognition of high-risk patients, among others.

However, there are still some difficulties that complicate the success of the data mining process, such as obtaining relevant and quality data. It is difficult to obtain accurate and complete health data due to its complexity and heterogeneity, as it is collected from different sources such as medical reports and conversations with patients. This challenge proves to be the most significant for data mining in the healthcare area, as data quality is a major factor in the process and no useful results can be obtained without quality data. Another difficulty that stands out is data sharing, as healthcare institutions and patients themselves are not willing to share health data due to privacy concerns [6].

2.2. Data Mining in Nosocomial Infections

Data mining can be used in healthcare institutions to create predictive models that make forecasts using real data, thereby improving patient care and reducing the costs associated with patient treatment. By analyzing the information contained in databases, it is possible to prevent future infections by gaining knowledge that helps to identify possible occurrences. In recent years, several studies have been carried out with the aim of reducing the number of nosocomial infections by applying data mining techniques to predict these same infections.

A study carried out by Álvaro Rocha in 2015, presented the results of the application of predictive models to real clinical data, where models with sensitivities greater than 91,90% were obtained through the application of Support Vector Machines and Naïve Bayes techniques. Thus, this study showed that by analyzing the information present in the databases of health institutions, it is possible to prevent hospital infections and obtain knowledge that helps to predict possible occurrences [7].

In another study, carried out by Daniel Silva in 2019, text mining and machine learning methods were used to predict and detect nosocomial infections at the surgical site. Methods such as textual descriptions of surgeries and postoperative patient records were used. Several algorithms and processing strategies with prediction and detection objectives were tested in this study, namely Logistic Regression, Naïve Bayes, Nearest Centroid, Random Forest, Stochastic Gradient Descent and Support Vector Classification. In terms of prediction, the best result was obtained by the Stochastic Gradient Descent method with 79,7% ROC-AUC. In terms of detection, the best result was obtained by Logistic Regression with 80,6% ROC-AUC [8].

Lastly, in 2011, Mary Gerontini conducted a study where data mining techniques were used to predict antibiotic sensitivity and nosocomial infections. The results obtained showed quite high prediction rates for the three algorithms applied. However, the Support Vector Machines algorithm stood out, which showed slightly better results, with accuracy, sensitivity and F-measure rates of 97,8% [9].

3. Practical Work

In the practical work carried out, the CRISP-DM methodology was used, as it is the most appropriate methodology used in projects involving data mining. Thus, the phases that form part of this methodology were followed [10].

3.1. Business Understanding

The first phase, Business Understanding, was to clearly understand the existing problems and needs, as well as the objectives to be achieved. For this, having the help of a health professional in the area was essential. The health professional explained the procedures for analyzing and signaling surgical site infections and it was quickly became clear that these procedures were carried out manually, with the health professional analyzing large amounts of data, which makes the work more susceptible to errors and very costly in terms of time. It is in this context that the main objective of this project arises, which is to develop data mining models that are able to predict in advance interventional patients with a high probability of contracting infection from the pre and post-surgical data of each patient. By predicting a patient's risk of infection, healthcare professionals can act more quickly and appropriately to implement infection prevention and control measures. The application of these measures leads to providing better healthcare, avoiding additional complications and reducing the inherent costs.

3.2. Data Understanding

The first step in data understanding was to understand the criteria on which the professional relies to flag infections. Therefore, the criteria indicated by the professional were: the occurrence of readmission, emergency registration, antibiotic prescription and bacteriological collection, each within 30 days of surgery. Only in orthopedic surgeries this period is extended to 90 days after surgery. Thus, datasets from a portuguese hospital were provided for the period between 2018 and 2022, with data from surgeries, hospitalizations, emergencies, antibiotics, clinical analyzes and information about surgeries that did or did not result in infection. After this transfer of datasets, a first contact was made with them, where the meanings of each column were understood, as well as the null and duplicate values and those that would or would not be used in the project.

3.3. Data Preparation

In the data preparation phase, we started by eliminating irrelevant columns, treating null values and eliminating duplicate records, then creating columns with relevant information for the preparation of the dataset. It was decided to approach the problem in a logic of weeks, respecting the periods defined in the signaling criteria, so that, each week, the probability of a patient contracting an infection were updated. Thus, each orthopedic surgery had 12 records, one for each week, covering the 90 days, and each surgery of other types had 4 records, also one for each week to cover the 30 days. In this way, each record contained only the occurrences counted to the week in question. To try to obtain the best possible results, it was decided to create two approaches:

- Approach 1, where for surgeries that resulted in an infection, all records indicate that this surgery resulted in an infection, regardless of the week in which it was detected;
- Approach 2, where for surgeries that resulted in infection, it is only indicated that it existed in the week in which it occurred and in the following weeks.

In each of the approaches created, the final dataset had 40512 records, of which 1640 and 1183 had infection in approaches 1 and 2, respectively. To assess the impact of the clinical analysis data on the models, two scenarios were created in each of the approaches:

- Scenario 1, with all variables;
- Scenario 2, with all variables except those related to clinical analysis.

3.4. Modelling

In each scenario and approach, after selecting the predictor attributes and class, the predictors were scaled so that all values were on the same scale. Then, with such a large imbalance between records classified as having an infection and as not having an infection, it was necessary to use techniques to balance the number of records in each class so that the precision of the models was not negatively affected. Three different techniques were used for this purpose: SMOTE, Oversampling and Undersampling. Before applying the mentioned techniques, the dataset was randomly divided into training (70%) and testing (30%) sets, and then the balancing techniques were only applied to the training data, preserving test data for evaluating models. Then, seven algorithms were tested with each of the balancing techniques in each scenario of each approach, to analyze the first results obtained. After the tests, involving the algorithms Naïve Bayes, Decision Tree, Neural Networks, Support Vector Machines, Random Forest, Logistic Regression and K-Nearest Neighbors, it was decided to continue with only those that showed an accuracy of more than 95%. Thus, the algorithms considered in each scenario of each approach were reduced to Decision Tree, Neural Networks, Random Forest and K-Nearest Neighbors, each with SMOTE and Oversampling techniques. The remaining algorithms tested with these two techniques, as well as all algorithms tested with the Undersampling technique showed results below 95%. That said, a technique was used to provide better results to the models, in this case the grid search technique, which seeks to find the best hyperparameters of each one.

3.5. Evaluation

After carrying out the necessary analysis to try to obtain the best possible parameters, the stratified cross-validation technique was used to assess the performance of the models in a more robust way. For this approach, the original dataset was used and balancing techniques were also applied, ensuring that the proportions of the different classes in each division of the dataset remained as close as possible to the overall proportions, giving greater confidence in the results. To this end, 10 iterations, each with a value of k equal to 10, were carried out for each algorithm in each scenario considered. The results presented correspond to the average of the results obtained in the 10 iterations of each algorithm in each approach and scenario, giving them even greater consistency. To perform a comprehensive evaluation of the models, several metrics were used, each addressing a specific aspect of the algorithm's performance. Classification accuracy measures the proportion of correct predictions in relation to the total number of predictions made. Precision, in turn, focuses on the proportion of true positives to total predicted positives, emphasizing the ability to avoid false positives. Recall, or sensitivity, assesses the true positive rate, measuring the proportion of correctly identified positive instances to the total actual positive instances, emphasizing detection ability. The Kappa coefficient, on the other hand, considers the agreement between the classifications made by the model and the actual classifications, taking chance into account. In addition, the area under the ROC curve (ROC AUC) is used to assess the overall performance of the model at different sensitivity levels. It plots the curve that relates the true positive rate to the false positive rate, allowing to observe how the model behaves at different decision thresholds. It should be noted that for all these metrics, the closer the value obtained is to 1, the better the performance of the model with respect to the metric in question. The results of each algorithm, in each scenario of approaches 1 and 2, are presented in Tables 1 and 2, respectively. These tables summarize the results of the 10 iterations, highlighting concisely the performance of each algorithm in each specific scenario.

Table 1 – Results obtained in approach 1

Scenario	Algorithm	Balancing Technique	Accuracy	Precision	Recall	Kappa	ROC AUC
1	Decision Tree	SMOTE	97,96%	86,52%	87,79%	74,22%	87,79%
1	Decision Tree	Oversampling	98,71%	92,45%	90,71%	83,08%	90,71%
2	Decision Tree	SMOTE	98,15%	87,74%	89,03%	76,69%	89,03%
2	Decision Tree	Oversampling	98,80%	93,40%	90,88%	84,15%	90,88%
1	Neural Networks	SMOTE	97,38%	81,34%	91,35%	71,10%	91,35%
1	Neural Networks	Oversampling	98,14%	86,97%	90,36%	77,10%	90,36%
2	Neural Networks	SMOTE	97,44%	81,47%	93,08%	72,46%	93,08%
2	Neural Networks	Oversampling	98,56%	89,42%	93,23%	82,41%	93,23%
1	Random Forest	SMOTE	99,14%	96,66%	91,92%	88,27%	91,92%
1	Random Forest	Oversampling	99,25%	98,34%	91,84%	89,64%	91,84%
2	Random Forest	SMOTE	99,05%	96,73%	90,73%	86,98%	90,73%
2	Random Forest	Oversampling	99,31%	98,45%	92,47%	90,48%	92,47%
1	KNN	SMOTE	98,19%	86,82%	91,86%	78,29%	91,86%
1	KNN	Oversampling	98,39%	90,00%	89,15%	79,09%	89,15%
2	KNN	SMOTE	98,04%	85,62%	91,83%	76,86%	91,83%
2	KNN	Oversampling	98,26%	89,81%	87,10%	76,76%	87,10%

Table 2 - Results obtained in approach 2

Scenario	Algorithm	Balancing Technique	Accuracy	Precision	Recall	Kappa	ROC AUC
1	Decision Tree	SMOTE	98,89%	89,48%	91,54%	80,90%	91,54%
1	Decision Tree	Oversampling	99,18%	93,61%	91,63%	85,14%	91,63%
2	Decision Tree	SMOTE	99,08%	91,62%	92,46%	84,02%	92,46%
2	Decision Tree	Oversampling	99,22%	93,90%	92,23%	86,05%	92,23%
1	Neural Networks	SMOTE	98,73%	86,77%	93,34%	79,46%	93,34%
1	Neural Networks	Oversampling	98,82%	88,49%	91,80%	80,09%	91,80%
2	Neural Networks	SMOTE	99,00%	89,20%	94,87%	83,59%	94,87%
2	Neural Networks	Oversampling	99,16%	91,34%	94,67%	85,82%	94,67%
1	Random Forest	SMOTE	99,43%	95,91%	93,83%	89,65%	93,83%
1	Random Forest	Oversampling	99,47%	97,42%	93,02%	90,18%	93,02%
2	Random Forest	SMOTE	99,41%	96,54%	92,89%	89,23%	92,89%
2	Random Forest	Oversampling	99,54%	97,38%	94,32%	91,56%	94,32%
1	KNN	SMOTE	98,67%	86,31%	92,94%	78,60%	92,94%
1	KNN	Oversampling	98,82%	89,38%	90,03%	79,34%	90,03%
2	KNN	SMOTE	99,04%	90,20%	93,95%	83,93%	93,95%
2	KNN	Oversampling	98,87%	90,59%	89,23%	79,74%	89,23%

4. Discussion

As can be seen in Table 1 and Table 2, the accuracy results obtained were very positive, which was to be expected as they were selected because their initial results for this metric were above 95%. In exploring different approaches, scenarios, algorithms, and balancing techniques, and with the purpose of evaluating the models in a more comprehensive way, other evaluation metrics were used that also revealed highly satisfactory results. In both approaches tested, a trend could be observed with respect to the balancing techniques, with the Oversampling technique outperforming the SMOTE technique and showing a more positive impact on the models. From a more general view of the results, it was possible to see that the differences between scenarios 1 and 2 were insignificant, with cases in which scenario 1 was superior and vice versa, which shows that the clinical analysis variables did not have a relevant impact on the models. Still looking at overall results, the most effective choice, regardless of approach, turned out to be the Random Forest algorithm combined with the Oversampling technique. Despite the subtle differences between the scenarios and approaches with this algorithm and balancing technique, approach 2 and scenario 2, stood out as the most promising, with accuracy of 99,54%, precision of 97,38%, recall of 94,32%, Kappa coefficient of 91,56% and area under the ROC curve of 94,32%. These results highlight the capacity and consistency of the models, showing their potential to support medical decision making in the area of surgical site infections.

5. Contributions and Future Work

The realization of this project to predict nosocomial infections associated with surgeries makes it possible to automate the identification of patients at considerable risk of contracting hospital infections. Through a robust model, it is possible to detect these cases early, streamlining the clinical process and improving the healthcare offered. The use of the model also makes it possible to reduce the time spent for detection compared to traditional manual detection, improving the accuracy of predictions and reducing the possibility of human errors. By identifying patients at risk earlier, healthcare professionals can intervene in a more targeted way, reducing infections and complications, which brings significant benefits to healthcare facilities. Looking at future work to be done, some promising points can be identified with the potential to further improve the approach already realized. One of these points is the inclusion of new variables that may be relevant and could improve the accuracy of the predictions. Another point is to evaluate the models in a real context, which would offer deeper performance insights. Finally, it would be very interesting to extend the model to treatments, that is, not only to predict patients at risk of infection, but also to recommend the best treatment for each case.

Acknowledgements

This research was funded by Fundação para a Ciência e Tecnologia, within the Project Scope: UIDB/00319/2020.

References

[1] S. Tat and S. O. Samuel, “NOSOCOMIAL INFECTIONS AND THE CHALLENGES OF CONTROL IN DEVELOPING COUNTRIES,” 2010.

[2] J. A. Al-Tawfiq and P. A. Tambyah, “Healthcare associated infections (HAI) perspectives,” *J Infect Public Health*, vol. 7, no. 4, pp. 339–344, 2014, doi: 10.1016/j.jiph.2014.04.003.

[3] M. Haque *et al.*, “Strategies to Prevent Healthcare-Associated Infections: A Narrative Overview,” *Risk Manag Healthc Policy*, vol. 13, p. 1765, 2020, doi: 10.2147/RMHP.S269315.

[4] Ecdc, “Economic evaluations of interventions to prevent healthcare-associated infections Literature review www.ecdc.europa.eu,” 2017, doi: 10.2900/4617.

[5] C. Suetens, T. Kärki, and P. Diamantis, “Point prevalence survey of healthcare-associated infections and antimicrobial use in European acute care hospitals,” 2023, doi: 10.2900/474205.

[6] D. Tomar and S. Agarwal, “A survey on data mining approaches for healthcare,” *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013, doi: 10.14257/ijbsbt.2013.5.5.25.

[7] Á. Rocha, A. M. Correia, S. C. Luís, and P. Reis, “Advances in Intelligent Systems and Computing 354 New Contributions in Information Systems and Technologies Volume 2,” 2015, Accessed: Aug. 11, 2023. [Online]. Available: http://www.springer.com/series/11156

[8] D. A. da Silva, C. S. ten Caten, R. P. dos Santos, F. S. Fogliatto, and J. Hsuan, “Predicting the occurrence of surgical site infections using text mining and machine learning,” *PLoS One*, vol. 14, no. 12, p. e0226272, Dec. 2019, doi: 10.1371/JOURNAL.PONE.0226272.

[9] M. Gerontini, M. Vazirgiannis, and A. C. Vatopoulos, “Predictions in Antibiotics Resistance and nosocomial infections monitoring,” 2011, Accessed: Aug. 11, 2023. [Online]. Available: www.mednet.gr/whonet.

[10] R. M. S. Laureano, N. Caetano, and P. Cortez, “Previsão de tempos de internamento num hospital português: Aplicação da metodologia CRISP-DM,” *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, no. 13, pp. 83–98, Jun. 2014, doi: 10.4304/risti.13.83-98.