JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 37, 1025-1038 (2021)
DOI: 10.6688/JISE.202109_37(5).0003

A Comparative Study of Machine Learning Models for Predicting Length of Stay in Hospitals

RACHDA NAILA MEKHALDI[1], PATRICE CAULIER[1], SONDES CHAABANE[1], ABDELAHAD CHRAIBI[2] AND SYLVAIN PIECHOWIAK[1]
[1]*Laboratory of Industrial and Human Automation Control*
Mechanical Engineering and Computer Science
Polytechnic University of Hauts-de-France
Valenciennes CEDEX 9, 59313 France
E-mail: {rachdanaila.mekhaldi; patrice.caulier; sondes.chaabane; sylvain.piechowiak}@uphf.fr
[2]*Alicante Company*
Seclin, 59113 France
E-mail: abdelahad.chraibi@alicante.fr

There has been a growing interest in recent years in correctly predicting the Length of Stay (LoS) in a hospital setting. Estimating the LoS on patient' admission helps hospitals in planning, controlling costs and, providing better services. In this paper, we consider predicting the LoS as a regression problem for which we implement and compare different Machine Learning (ML) algorithms. Multiple Linear Regression (MLR), Support Vector Machines (SVM), Random Forests (RF), and Gradient Boosting model (GBM) are implemented using an open-source dataset. The methodological process involves a preprocessing step combining data transformation, data standardization, and categorical data encoding. Moreover, the Synthetic Minority Over Sampling Technique for Regression (SMOTER) is applied to handle unbalanced data. Then, ML algorithms are employed, with a hyperparameter tuning phase to obtain optimal coefficients. Finally, Mean Absolute Error (MAE), R-squared (R^2), and Adjusted R-squared (Adjusted R^2) metrics are selected to evaluate the model with parameters.

Keywords: length of stay in hospitals, data preprocessing, machine learning, unbalanced data, parameters tuning

1. INTRODUCTION

The growth in the population worldwide in recent years, results in health-care institutions lacking resources. Healthcare institutions, academics, and companies in different areas have pooled their efforts to optimize hospital resources while maintaining the quality of services. Patient's Length of Stay (LoS) is an important indicator for assessing health-care services. Hence, interest in predicting the LoS in hospitals is grown. The LoS is defined as the interval between patient admission and his discharge from the hospital [1]. Predicting LoS contributes to the organization and scheduling of care activities by estimating the date of the patient's discharge and thus can be used to predict patient inflows. This helps to reduce the patient waiting times and the workload of the health-care professionals. The main purpose of predicting the LoS is to optimize the use of resources

Received September 17, 2020; revised November 17, 2020; accepted December 22, 2020.
Communicated by Maria José Sousa.

and manage budget constraints in hospitals. In all public and private French hospitals, a special program called the "Information Systems Medicalization Program" (ISMP) (In French PMSI) is implemented to calculate the correct amount of funds to be allocated to each hospital department to reduce the disparity in resources between health institutions [2]. As mentioned in [3], predicting the LoS aims to force hospitals to comply with budget constraints and to facilitate reimbursement.

In the context of predicting LoS, an essential step is to study the factors impacting this indicator in various medical units. According to [5], LoS is considered as a complex variable that can comprise the clinical and social contexts of patients and their care. This complexity is a characteristic of medical data, as medical datasets contain heterogeneous information (numerical, categorical, textual, images, *etc*.) that comes from multiple sources. Health databases for hospital stays are often large and contain incomplete and biased data. Furthermore, confidentiality must be ensured for all medical data and access must be restricted [7]. A large amount of medical data is stored in healthcare institutions, therefore, Data Mining (DM) methods are widely employed to determine factors influencing the LoS and Machine Learning (ML) models to predict it [16].

This paper aims to propose a solution for predicting the Length of Stay based on Artificial Intelligence methods including DM and ML methods. First, a review of the literature is presented to define a representation characterizing the LoS. This is a crucial step as it provides the input for the prediction algorithm. A classification process starting with data preprocessing, followed by machine learning and model evaluation is developed. SMOTE for Regression is used to address the problem of unbalanced data. Several machine learning models for regression are compared: Random Forests (RF), Gradient Boosting Model (GBM), Support Vector Machines (SVM), and Multiple Linear Regression (MLR). Finally, the results are given and discussed. Also, the conclusion and prospects are presented.

2. LITERATURE REVIEW

It is essential to understand what are the factors influencing the LoS because they are the key inputs of ML models. In [8], M. Rigal shows that Los is strongly linked to medical unit. Indeed, LoS models in an emergency or walk-in unit differ from those in other departments. In emergency departments, the patient's stay is measured in hours whereas in the cardiac department, for example, it is calculated in days. In this section, we present researches on factors impacting the LoS in most cited medical units and ML algorithms used to predict it.

In [9], R. J. Lafaro *et al.* use specific variables for predicting LoS in Cardiac Intensive Care Units (CICU) such as O2 delivery levels, hematocrit levels, serum creatinine, and blood gas analysis. These variables are selected using the ALM method (Asset Liability Management). An Artificial Neural Network (ANN) model is then applied to predict the LoS. Moreover, in a cardiac unit, authors in [10] implement and compare different algorithms: Decision Trees (DT), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) to predict the LoS of patients suffering from Coronary Artery Disease. The results show that SVM is the best fit. Once more these algorithms use variables related to the unit specialty such as: diastolic blood pressure, hemoglobin, cholesterol,

and associated comorbidities. Tsai *et al.* [11] explore factors affecting the LoS for patients with one of the three primary diseases: coronary atherosclerosis, heart failure, and acute myocardial infarction in a cardiovascular unit. The study includes the patient's address, diagnosis, comorbidities, and method of reimbursement. Multiple Linear Regression (MLR) and ANN are used for prediction. ANN produced the best results. For all the studies in the cardiac unit, the patient's age and gender are considered as basic variables impacting the LoS.

Several studies have focused on predicting the LoS in Intensive Care Units (ICU).In the study conducted by [13], the Medical Information Mart for Intensive Care database (MIMIC) is used to predict long LoS [34]. This database includes patient' admission, transfer, and discharge, as well as, medical examinations, laboratory test results, and diagnostic and demographic variables. The prediction accuracy of ANN is approximately 80% and outperformed linear models. The researchers in [14] identify demographic variables, patient medical history (cardiac, renal, and pulmonary diseases), creatinine levels, heartbeat, *etc*. as a part of relevant variables. They then compare ANN with an Adaptive Neuro fuzzy System for predicting LoS. The Adaptive Neuro-fuzzy System performed better as it considers the expert's knowledge.

In surgery units, predicting the LoS is effective in the pre-operative or post-operation phase. The survey in [15] shows that some factors differ between an urgent operation and a non-urgent operation. These factors are demographics, medical history, vital signs, laboratory tests, and caregiver' notes. The final subset of features is selected according to healthcare experts. Numerous machine learning algorithms have been tested including DT, Random Forests (RF), and SVM. The RF outperformed all the algorithms tested. Furthermore, DT, Naive Bayes (NB), and K-Nearest Neighbors are applied in [16] for predicting the LoS in a general surgery department. The overall best accuracy is obtained with DT and is 88.9%. In this study, the patient's information is gathered and type and number of operations, transfer conditions, number of visits, number of tests, and the number of previous stays are considered in the prediction model. In addition to SVM and DT, NB is used to predict LoS in an emergency department. The results show that using a subset of features improved model performance [17].

From this past research, we conclude that it is crucial to start by investigating the factors impacting the LoS in a hospital setting in order to predict it. We notice that factors impacting the LoS depends on the type of medical department to which the patient is admitted. In the majority of studies, the authors focus on predicting LoS in cardiac units, Intensive Care Units, and surgery units, as these departments request more funding from the hospital and use more resources. Besides, patient' information (demographic and medical) is commonly integrated into LoS models. Furthermore, we assume that ML algorithms, especially supervised learning techniques, can serve as valuable reference tools for predicting LoS. Artificial Neural Networks, Decision Trees, Multiple Linear Regression, Support Vector Machines, and Random Forests are the most cited algorithms.

In our study, we compare different algorithms. First, we test linear models such as MLR and SVM. Second, we test ensemble models based on DT such as RF and GBM. The first objective of our study is, to define a common representation of LoS in several medical units cited. The second objective is, to address problems present in the dataset and try to improve the LoS model prediction performance presented in [18] by applying more sophisticated methods in the preprocessing step.

3. DATASET AND MACHINE LEARNING PROCESS

In this section, the used dataset is described, and our methodological process is detailed.

3.1 Dataset Description

In our study, the real dataset is still being prepared due to administrative procedures regarding confidentiality, so, we chose to use the open-source Microsoft dataset for predicting the length of stay [20]. This dataset offers similar variables describing the LoS than those found in the literature. It includes patient' demographic information such as gender, medical history (renal disease, pneumonia, depression, *etc*.), vital signs (Body Mass Index or BMI, pulse, *etc*.), and laboratory data (hematocrit, creatinine, *etc*.). It is also a heterogeneous open-source dataset that provides categorical data (*e.g.* gender) and numerical data (*e.g.* glucose).

The dataset included 100,000 observations and 28 variables. Before any processing of the dataset, we remove the variables eid, date of entry, date of exit, and BMI. The sequential variable eid which is a sequential number does not provide any extra information. The entry and the exit dates are omitted as long as we had the time spent in hospital (LoS). Concerning the BMI, according to the study carried out by [19] the BMI is not a measure "health" or a physiological state indicating the presence (or absence) of a disorder. It is simply a measure of human size. We analyzed the correlation between BMI and LoS to check its consistency with the results of the statistical analysis. The value of the Pearson coefficient is 0.0001 with a p-value of 0.96 which shows a very low correlation.

3.2 Machine Learning Process

Our ML process includes four main steps: data analysis, data preprocessing, Machine Learning models, and model evaluation.

3.2.1 Data analysis

Univariate statistical analysis is employed. First, we distinguish numerical variables from the categorical one to adapt the data preprocessing to the data type. Then, we analyze the distribution, presence of outliers, and missing values in the variables used in this study. This defines preprocessing techniques to be used in the next step.

In the dataset used, there are no missing values in all the variables. Concerning data distribution, only the two variables *neutrophils* and *bloodureanitro* representing the "average value of neutrophils" and the "average value of blood urea" respectively are having a **non-normal** distribution. Also, all variables contain outliers. So, a standardization technique is therefore applied as will be described below, to address this issue.

3.2.2 Data preprocessing

Data preprocessing is considered as the main step in any ML classification process. It refers to the transformations applied to data before feeding it into our ML algorithms. The goal is to convert the raw data into a clean dataset. We had to apply a logarithmic

transformation and standardization to deal with the presence of outliers in the numerical used variables used. Data encoding is applied to categorical variables to convert them to numerical format, as ML algorithms require numerical data.

Data transformation

Data transformation is required when the data distribution is **non-normal**. Firstly, histograms are plotted to assess the distribution of each variable. As an example, the following figures show the histograms and the distribution of the variable **neutrophils** before (Fig. 1) and after (Fig. 2) transformation.

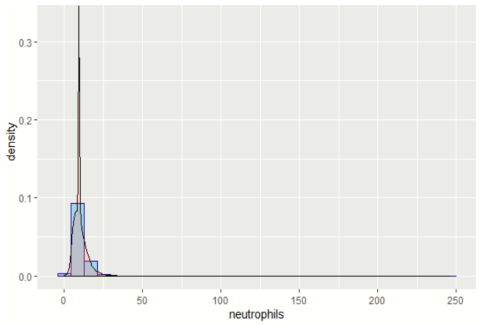


Fig. 1. Histogram before transformation.

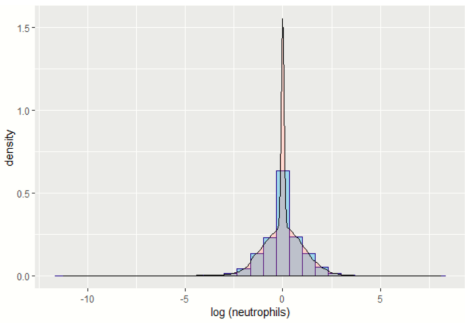


Fig. 2. Histogram after transformation.

The distribution of this variable is non-normal, a transformation using the logarithmic function is used. This decreases the variance of a variable by transforming its distribution to normal, reducing the effect of outliers or eliminating any outliers. In the dataset used, only two variables are not normally distributed with a high number of outliers.

Data standardization

Data Standardization is applied when the input features (continuous variables) are normally distributed with different means and standard deviations. For this purpose, we chose the **Z-score** method. The **Z-score** method used is as follows, where X_i represents an observation, μ is the average of the variable X, and σ is its standard deviation:

$$Z_i = \frac{X_i - \mu}{\sigma}$$

This transformation converts all the variables into the same scale and harmonizes the data structure [32].

Categorical data encoding

For encoding categorical data, the One-Hot encoding method is the most widespread approach and works very well. In this approach, we simply create additional features based on number of unique values in the categorical variable. Each value is transformed to new column with binary value indicating ether the modality is checked or not. This

method is used to ignore the natural ordered relationship between integer values [22, 21]. A categorical variable containing 4 values is splited into 4 different variables, for example. The table below illustrates this method for the variable "*rcount*", which represents the number of readmission over the previous 180 days.

Table 1. Application of the One-Hot-Encoding approach to the variable *rcount*.

rcount	0	1	2	3	4	5+
rcount 0: 0 readmissions	1	0	0	0	0	0
rcount 1: 1 readmissions	0	1	0	0	0	0
rcount 2: 2 readmissions	0	0	1	0	0	0
rcount 3: 3 readmissions	0	0	0	1	0	0
rcount 4: 4 readmissions	0	0	0	0	1	0
rcount 5+: 5 or more than 5 readmissions	0	0	0	0	0	1

Synthetic Minority Over-Sampling Technique for Regression (SMOTER)

The prediction of rare extreme values of a continuous variable is very relevant in various real-world fields. In our dataset, the target value is continuous and contained outliers or extraordinary values resulting in an unbalanced dataset that making it hard to predict the minority present values. To check the LoS distribution in our dataset, we plotted its histogram and boxplot. The histogram shows the unbalanced frequencies of the LoS variables and the boxplot highlights the presence of outliers (see Figs. 3 and 4 below). From these figures, we noticed that the minority values belong to the interval [11:17]. In fact, such values represent outliers detected by the Boxplot. In addition, there is a considerable gap between the Los frequencies (for example 2 and 8).

To address this issue, we applied the Synthetic Minority Over-Sampling Technique for Regression (SMOTER). It is a variant of Synthetic Minority Over-Sampling adapted to regression problems that addresses classification problems with unbalanced class [24]. This technique uses sampling approaches to change the distribution in a training dataset to reduce unbalance, which would otherwise result in a bias towards solutions that are not within the user's objectives. SMOTE addresses classification problems with unbalanced class distribution.

For regression tasks, few works address this type of problem. The basic feature of this method is the possibility of under sampling the dominant values and/or oversampling minority values [23]. This technique is particularly useful when the target values of interest for making predictions within a given dataset are known by a domain expert. The method is explained and implemented in [33]. We define manually a specific region or a rare values in the target. In our study, for each minority value which belongs to [11, 12, 13, 14, 15, 16, 17], a binary parameter is set to 0 or 1, where 1 if the value is relevant and 0 if not.

From Fig. 5, we can see that after applying SMOTER on the minority values of the LoS variable, the number of observations is bigger. The new dataset is then used as an input of ML algorithms. The results of ML algorithms with and without applying the SMOTER are also compared.

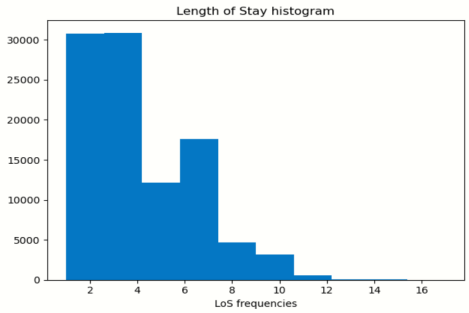


Fig. 3. Histogram of the LoS variable.

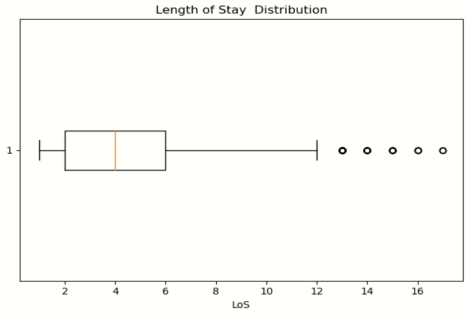


Fig. 4. Boxplot of the LoS variable.

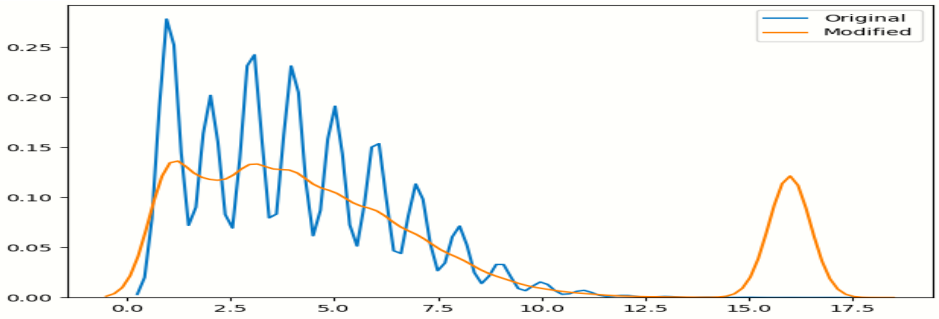


Fig. 5. SMOTER method for the variable LoS.

3.2.3 Learning methods for regression

The next step is to develop learning regression algorithms. Multiple Linear Regression (MLR) [29], Support Vector Machines (SVM) [28], Random Forests (RF) [25], and Gradient Boosting Models (GBM) [26] are compared and approved to obtain the best model. These algorithms are highlighted in much research for predicting LoS. Indeed, the study in [18] proved that RF outperforms the GBM in predicting LoS using the same dataset including all variables and without any variable selection method. The objective of this study is, therefore, to apply more preprocessing methods and then to compare supervised techniques based on linear regression (MLR and SVM) which are sensitive to outliers, with those based on DT such as ensemble methods (RF and GBM).

As it is a supervised approach, we split the initial dataset into a training dataset (70%) and a test dataset (30%). All the algorithms are applied to two different datasets: the first one applying SMOTER and the second one without applying this method. To further improve our training model, parameter tuning is conducted using the Random Search for hyperparameter optimization. James Bergstra and Yoshua Bengio proposed the idea of a Random Search for hyperparameter optimization in [30]. The method is based on defining a sampling distribution for each hyperparameter. For each algorithm, we chose to tune the most important parameters. In the case of MLR, the fit intercept parameter indicating whether to calculate the intercept for the model is used. For SVM, the kernel and its degree are adjusted. For RF, the number of trees in the forest is used in addition to

the maximum depth of the trees. Finally, when using GBM, we tuned the maximum tree depth and the learning rate, which corresponds to how quickly the error is corrected from each tree to the next [31]. All the chosen parameters are interrelated.

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the experimental results and model evaluation are presented. In any ML process, the last step is model evaluation. This step tests a model with different parameters and saves the best model. Python programming language is used in the implementation phase and a PC with an Intel(R) Core(TM) i7-8650U processor with 8 GB of RAM. Regression metrics such as *MAE*, R^2, and *Adjusted* R^2 are used to evaluate the performance of the system. MAE represents the absolute difference between the predicted and observed values and its robustness to outliers in the dataset. R^2 and *Adjusted* R^2 highlight how well the target variable explains the variability in the attributes.

The results of all the metrics used in the model evaluation (MAE, R^2, Adjusted R^2) are presented in Table 2 for both datasets used (with and without SMOTER). Also, the time taken to run the experiments is given in seconds. A discussion of results is then presented.

Table 2. Machine learning model evaluation.

Datasets	***Dataset without SMOTER***				***Dataset with SMOTER***			
Metrics	***MAE***	***R^2***	***Adj R^2***	***Time***	***MAE***	***R^2***	***Adj R^2***	***Time***
Multiple Linear Regression	0.88	0.76	0.76	3.09	0.97	0.71	0.71	3.75
Support Vector Machines	0.54	0.89	0.89	**2683.03**	0.54	0.89	0.89	**2283.29**
Random Forests	0.7	0.85	0.85	896.43	0.72	0.84	0.84	697.35
Gradient Boosting Model	**0.44**	**0.94**	**0.94**	**250.37**	**0.45**	**0.93**	**0.93**	**171.37**

Analyzing the results from Table 2 regarding the performance of the models, GBM outperformed all the other with the lowest value of MAE equal to 0.44 and an R^2 and an adjusted R^2 close to 1 with a value of 0.94. We consider these results to be good. As both GBM and RF are ensemble methods based on DT and apply respectively boosting and bagging algorithms, we noticed that the boosting method is more suitable to the dataset used than the bagging one. As boosting does more to reduce bias which is highly present in our dataset than variance which is corrected in the preprocessing step. Even the fact that in GBM the trees are built sequentially in opposite to RF where they are built in parallel, the time execution required by GBM is less than the one required by the RF. In fact, for the RF, a large number of trees may make the algorithm slow for real time prediction.

Comparing the SVM algorithm which is supporting linear and non linear regression and the MLR algorithm, SVM performed slightly better than the MLR (MAE = 0.54 against 0.88 and R^2 and adjusted R^2 = 0.89 against 0.76). As MLR is simple to be implemented and attends to minimize the error between prediction and target, unlike the SVM that makes sure the errors do not exceed a threshold, the time execution for SVM is broadly larger than in MLR. This is due to the number of parameter tuned for MLR which is less than the one tuned for SVM. Another remark is that the SVM surpassed the RF as the output of the RF is simply the mean of all trees output generated by the algorithm.

SVM always takes a longer time execution as it requires a lot of time for training the model.

Concerning the results of ML models applied on dataset without SMOTER, the results are not improved and are very close comparing the evaluation metrics used. The behavior of all algorithms is not changed with the application of SMOTER method. SMOTER method, thus, is not enough to overcome the problem of unbalanced data. As mentioned in [33] this technique is most useful when the target values are known by a domain expert. In our case, these values are just obtained based on outliers values.

To further analyze the output of the best trained model, we plot the histogram representing the percentage of deviations between the predicted values and the target values. A perfect prediction is with a deviation equal to 0. Almost half of the predicted values are correct. Also, 43% of deviations values are equal to 1 day error. In terms of hospital management, this gap can be acceptable. Thus, the developed tool can attend 92% of correct predicted values. Only 3% of difference values are greater or equal to 3 days. In most cases, this percentage corresponds to the target belonging to minority values.

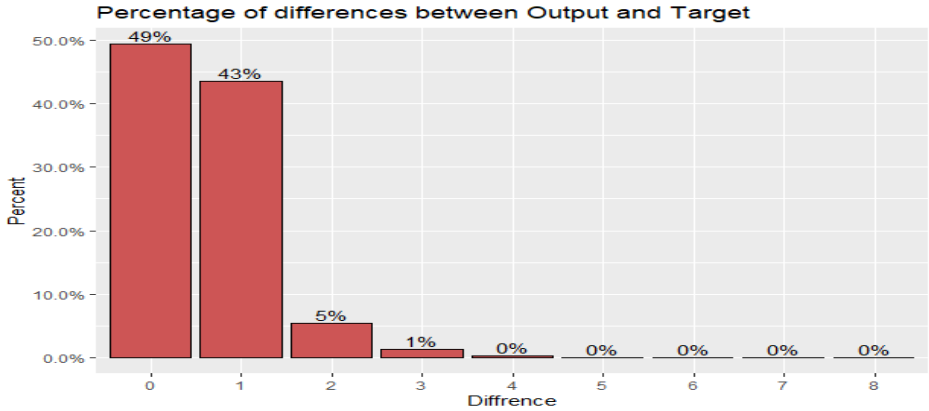


Fig. 6. Difference between output and target values.

From the experiments, GBM outperformed MLR, SVM, and RF in all the metrics cited above. Furthermore, the execution time for the GBM is shorter than for the other algorithms. For a real case study, if LoS needs to be predicted in real time, GBM is a good algorithm regarding to the variables used in this study. We had to find a good compromise between the two constraints in our design which are a minimum error in predictions and less time execution. The best models have been saved and will be trained on a real dataset later. All implemented ML methods proved their efficiency in past studies [10, 11, 15, 18] and in this study. Regarding the factors used to predict the LoS, previous research suggest to employ variables related to a specific unit. In our study, we used common variables to several units (cardiac, ICU and surgery). There are the classical demographic information such as gender, age, *etc*. In addition to variables describing biological tests results such as hematocrit, creatinine, hemoglobin, cholesterol. Also, all patient's medical history is reported (asthma, renal disease, iron deficiency, *etc*.). Moreover, we consider the comorbidity, heart pulse and psychological patient's state [9, 10, 13, 15].

Taking into account these results and to improve the performance of the models, the

role of medical experts is crucial in the preprocessing step. Indeed, they could decide if a value represents an outlier or not. We applied the SMOTER technique to address the issue of the presence of outliers in the variable of interest (LoS). For the ML phase, after tuning the parameters, it is extremely important to find the right combination parameter values to train the model. Furthermore, the objective of the study must be defined before choosing the ML method (*e.g.* real time prediction or not).

5. CONCLUSION

In this paper, we compared different Machine Learning algorithms for predicting the Length of Stay (LoS). We first investigated the factors impacting the LoS and the most common ML methods in the literature. The open source Microsoft dataset for predicting LoS is used to develop the models. The preprocessing step involves data transformation, data standardization, and categorical data encoding. The Synthetic Minority Over Sampling Technique for Regression is tested to handle the problem of unbalanced data. Several ML algorithms are explored: Multiple Linear Regression, Support Vector Machines, Random Forests, and Gradient Boosting Model. A hyperparameter tuning phase is carried out using the Random search method. The results show that the GBM outperformed other algorithms used in this study with an MAE lower than 0.44 and an R^2 and Adjusted R^2 greater than 0.94 when the SMOTER technique is not used.The implementation of SMOTER method did not improved the results as its application requires medical expert to define the rare values. We consider that the results are satisfactory with regard to the nature of the dataset we used.

One potential limitation of our study is the necessity to conduct an additional study on a real dataset to confirm the results. We had to use the Microsoft dataset as an example due to the requirements in anonymizing the data in a real case study. This dataset allows us to highlight the complexity of medical data. For better results, large real datasets can be used and a more sophisticated technique to transform the target variable can be employed.

ACKNOWLEDGMENTS

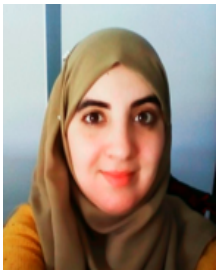
The authors would like to thank the European Regional Development Funds of Hauts-de-France Region (ERDF) and Alicante company (https://www.alicante.fr/) for the funding and their support and contributions.

REFERENCES

1. O. Khosravizadeh, S. Vatankhah, P. Bastani, R. Kalhor, S. Alirezaei, and F. Doosty, "Factors affecting length of stay in teaching hospitals of a midOPodle-income country," *Electronic Physician*, Vol. 8, 2016, pp. 3042-3047.
2. Fédération Hospitalière de France, "Le PMSI : des objectifs et une standardisation des donnees pour le service public hospitalier et les e tablissements prives de soin," 2019, pp. 1.

3. E. M. Carter and H. W. Potts, "Predicting length of stay from an electronic patient record system: A primary total knee replacement example," *BMC Medical Informatics and Decision Making*, Vol. 14, 2014, pp. 1-13.
4. L. Turgeman, J. H. May, and R. Sciulli, "Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission," *Expert Systems with Applications*, Vol. 78, 2017, pp. 376-385.
5. S. Shea, R. V. Sideli, W. Dumouchel, G. Pulver, R. R. Arons, and P. D. Clayton, "Computer-generated informational messages directed to physicians: Effect on length of hospital stay," *Journal of the American Medical Informatics Association*, Vol. 2, 1995, pp. 58-64.
6. S. Aghajani and M. Kargari, "Determining factors influencing length of stay and predicting length of stay using data mining in the general surgery department," *Hospital Practice and Research*, Vol. 1, 2016, pp. 53-58.
7. S. Haas, S. Wohlgemuth, I. Echizen, N. Sonehara, and G. Muller, "Aspects of privacy for electronic health records," *International Journal of Medical Informatics*, Vol. 80, 2011, pp. e26-e31.
8. M. Rigal, "Management des lits et duree moyenne de sejour : Exemple de recherche d Optimisation au Centre Hospitalier dvignon," *Healthcaresystems*, 2009.
9. R. J. Lafaro, S. Pothula, K. P. Kubal, *et al*., "Neural network prediction of ICU length of stay following cardiac surgery based on pre-incision variables," *PLoS One*, Vol. 10, 2015, pp. 1-19.
10. P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," *Healthcare Informatics Research*, Vol. 19, 2013, pp. 121-129.
11. P. F. Tsai, P. C. Chen, Y. Y. Chen, H. Y. Song, H. M. Lin, F. M. Lin, and Q. P. Huang, "Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network," *Journal of Healthcare Engineering*, Vol. 2016, 2016.
12. A. Almashrafi, H. Alsabti, M. Mukaddirov, B. Balan, Baskaran, and P. Aylin, "Factors associated with prolonged length of stay following cardiac surgery in a major referral hospital in Oman: A retrospective observational study," *BMJ Open*, Vol. 6, 2016.
13. T. Gentimis, A. J. Alnaser, A. Durante, K. Cook, and R. Steele, "Predicting hospital length of stay using neural networks on MIMIC III data," in *Proceedings of IEEE 15th International Conference on Dependable, Autonomic and Secure Computing*, 2017, pp. 1194-1201.
14. H. Maharlou, S. R. N. Kalhori, S. Shahbazi, and R. Ravangard, "Predicting length of stay in intensive care units after cardiac surgery: Comparison of artificial neural networks and adaptive neuro-fuzzy system," *Healthcare Informatics Research*, Vol. 24, 2018, pp. 109-117.
15. M. T. Chuang, Y. H. Hu, C. F. Tsai, C. L. Lo, and W. C. Lin, "The identification of prolonged length of stay for surgery patients," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 2016, pp. 3000-3003.
16. S. Aghajani and M. Kargari, "Determining factors influencing length of stay and predicting length of stay using data mining in the general surgery department," *Hospital Practice and Research*, Vol. 1, 2016, pp. 53-58.

17. S. Benbelkacem, F. Kadri, B. Atmani, and S. Chaabane, "Machine learning for emergency department management," *International Journal of Information Systems in the Service Sector*, Vol. 11, 2019.
18. R. N. Mekhaldi, P. Caulier, S. Chaabane, A. Chraibi, and S. Piechowiak, "Using machine learning models to predict the length of stay in a hospital setting," *Advances in Intelligent Systems and Computing*, Vol. 1159, 2020, pp. 202-211.
19. R. H. Shmerling, "How useful is the body mass index (BMI)?," *Harvard Health Publishing*, 2016, pp. 3-5.
20. Microsoft, "Predicting hospital length of stay," *Healthcaresystems*, 2017.
21. K. Potdar, S. Taher, and D. Chinmay, "A comparative study of categorical variable encoding techniques for neural network classifiers," *International Journal of Computer Applications*, Vol. 175, 2017, pp. 7-9.
22. J. Brownlee, "Why one-hot encode data in machine learning?,"*Machine Learning Mastery*, 2017, pp. 1-46.
23. B. Kovacs, F. Tinya, C. Nemeth, and P. Odor, "Unfolding the effects of different forestry treatments on microclimate in oak forests: results of a 4-yr experiment," *Journal of Artificial Intelligence Research*, Vol. 16, 2002, pp. 321-357.
24. E. Oliveira, J. Gama, Z. Vale, H. Lopes, C. Eds, R. Goebel, "SMOTE for regression," *Progress in Artificial Intelligence*, Vol. 8154, 2013, pp. 378-389.
25. A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends in Computer Graphics and Vision*, Vol. 7, 2011, pp. 81-227.
26. H. Singh, "Understanding gradient boosting machines," *Towards Data Science*, 2018, pp. 1-10.
27. R. R. Kabra and R. S. Bichkar, "Performance prediction of engineering students using decision trees," *International Journal of Computer Applications*, Vol. 36, 2011, pp. 8-12.
28. V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geology Reviews*, Vol. 71, 2015, pp. 804-818.
29. M. Orus-Lacort and C. Jouis, "Multiple linear regression analysis," *American Journal of Orthodontics and Dentofacial Orthopedics*, Vol. 149, 2018, p. 581.
30. J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, Vol. 13, 2012, pp. 281-305.
31. R. Carnevale, "Understanding gradient boosting, Part 1," *Data Stuff, LoS Classification*, 2015.
32. D. Loshin, "Data consolidation and integration," *Master Data Management*, 2009, pp. 177-199.
33. P. Branco, R. P. Ribeiro, L. Torgo, B. Krawczyk, and N. Moniz, "SMOGN: A pre-processing approach for imbalanced regression," in *Proceedings of Machine Learning Research*, Vol. 74, 2017, pp. 36-50.
34. "Overview of the MIMIC-III data," https://mimic.mit.edu/gettingstarted/overview/, 2017.



Rachda Naila Mekhaldi is currently working toward a Ph.D. in with the Polytechnic University of Hauts-de-France and the LAMIH CNRS UMR 8201, Computer Science Department (Valenciennes, France). Her research interests include the application of artificial intelligence methods particularly machine learning models in healthcare systems. She has a master's degree in machine learning for data science from Paris Descartes University (Paris, France).

Patrice Caulier is an Associate Professor at the National Institute of Science and Technology (INSA) of Hauts-de-France, a french graduate engineering school, and the LAMIH CNRS UMR 8201, Automatic Control Department. His main research interest focuses on management assistance, in presence of dysfunctions or in a crisis situation, of complex, products or services, production systems by using AI approaches. By improving the operational safety and resilience of these systems, the assistance aims to anticipate, cancel, if not limit, consequences of dysfunctions or crisis situations

Sondès Chaabane is an Engineer in Computer Science from ENSI (Tunis, Tunisia). She received her Ph.D. from INSA of Lyon, France. She is currently a Full Assistant Professor at the INSA HdF and UPHF (Valenciennes, France) and researcher in LAMIH UMR CNRS 8201 Laboratory in Automation and Control Department. Her research is based on optimization and simulation with application on automotive industry and hospital management. She has experience working with metaheuristic approaches, simulation models and artificial intelligence techniques.

Abdelahad Chraibi obtained his Ph.D. in 2015 in the field of operational research and decision support from Jean Monnet University at Saint Etienne. He is currently at the Head of Data-LAB within Alicante society to manage a team of data scientists, statisticians and developers with the objective of realization of AI solutions to assist hospitals in their daily problems.

Sylvain Piechowiak is a Full Professor in Computer Science at Polytechnic University Hauts-de-France since 2003. He is a Researcher in LAMIH UMR CNRS 8201 Laboratory in Computer Science Department. His research area concerns: constraint based reasoning (temporal, distributed, *etc*.) and uncertainty based reasoning. He uses these paradigms of reasoning to solve model-based diagnosis in transport systems or in health organizations, to support multi-agent systems in traffic simulation, to elaborate timetables in university and to improve management in hospitals.