ARTIFICIAL INTELLIGENCE THEORY and APPLICATIONS

ISSN: 2757-9778 || **ISBN :** 978-605-69730-2-4

More information available at **aita.bakircay.edu.tr**

Prediction Length of Stay in Intensive Care Unit in the Presence of Missing Data

Zeliha ERGUL AYDIN *, Zehra KAMISLI OZTURK

Eskisehir Technical University, Dept. of Industrial Engineering, Turkey

* Corresponding Author: *Eskisehir Technical University*, *Dept. of Industrial Engineering*, Turkey
Tel.: +90 (0222) 321 35 50. E-Mail: zergul@eskisehir.edu.tr

Publication Information

Keywords :
- Length of stay prediction
- Classification algorithms
- Intensive care unit
- MIMIC-III

Category : Research Article

Received : 09.06.2021
Accepted : 25.09.2021

A B S T R A C T

Patients' length of stay (LOS) in intensive care units (ICU) is an important factor for managing limited ICU resources such as beds, staffing, medicines, and medical devices. The goal of this study predicts that the ICU length of stay of patients is more than 3 days or not with Support Vector Machine (SVM), Logistic Regression (LR), XGBoost classifiers. We retrieved the 37,600 ICU patients' demographics data and last measured vital signs in their first 12 hours of stay from the MIMIC-III database. We filled the missing patients' data with three missing data imputation methods, namely k-nearest neighbor imputation (KNN), multivariate imputation by chained equations (MICE), and Soft-Impute. Our results indicated that filling missing data with the Soft-Impute yielded the highest AUC score for all classifiers. We obtained the highest area under the curve score as 66.1% with the XGBoost classifier and Soft-Impute missing data imputation.

1. Introduction

Accurate evaluation of the physiological status of patients in intensive care units (ICU) is critical for the allocation of hospital resources such as beds, staffing, and medical equipment [1]. Therefore, patient length of stay (LOS) reflecting the physiological status is a critical parameter for the hospital resources allocation. Most of the studies [2], [3], [4] diagnosed the disease successfully with machine learning methods in the literature. Similarly, existing studies have attempted to predict the length of stay in ICU with machine learning methods. While some of the researchers focused on a specific disease [5], [6] or ICU unit [7], [8], some of them focus on whole ICU units in LOS prediction with machine learning [9], [10].

There is a large amount of missing data in the dataset used for LOS prediction like other medical data. The missing data problem can be solved by ignoring or filling it. Ignoring a large amount of missing data should not be preferred because it can cause information loss. There are methods used to fill missing data such as k-nearest neighbor imputation (KNN), multivariate imputation by chained equations (MICE), Soft-Impute, mean, median, zero, linear regression, etc. Existing studies in ICU LOS prediction generally used only the mean or median method to fill missing data. Accordingly, this study aims to predict patients' LOS in the ICU with machine learning classifications models by using different missing data imputation methods. We use Support Vector Machine (SVM), Logistic Regression (LR), XGBoost classification algorithms and

KNN, MICE, and Soft-Impute imputation methods. To predict whether the LOS of the patient is longer than three days or not, we use 37,600 ICU patients' data with 22 features drawn from the Medical Information Mart for Intensive Care (MIMIC-III) database [11] retrospectively.

2. Material and Methods

We fetch 37,600 ICU patients' data with 22 features taken from the Medical Information Mart for Intensive Care (MIMIC-III) database MIMIC-III [11] to predict whether the patient's LOS is more than three days or not. This dataset includes patients whose ages between 18 and 89, and patients who stay in Medical ICU (MICU), Surgical ICU (SICU), CCU (Coronary care unit), TSICU (Trauma/surgical ICU), or Cardiac Surgery Recovery Unit (CSRU). Besides, we consider the patients with an ICU stay of more than 12 hours and less than 10 days. The names and data types of the 22 features in the data set are given in Table 1. The first 12 features in the table are vital signs, and others are demographics features. MIMIC III is a longitudinal database, so the measures of the vital features have different timestamps. We consider the last measurement taken within the first 12 hours for vital features.

Table 1. Missing data percentage in each feature.

Feature	Data Type
Temperature	Numerical
Heart Rate	Numerical
Glucose	Numerical
Oxygen saturation	Numerical
Glasgow Coma Scale	Numerical
Diastolic blood pressure	Numerical
Mean blood pressure	Numerical
Respiratory rate	Numerical
Systolic blood pressure	Numerical
Capillary refill rate	Categoric
PH	Numerical
Fraction inspired oxygen	Numerical
Height	Numerical
Weight	Numerical
Age	Numerical
Insurance	Categoric
Ethnicity	Categoric
Admission Type	Categoric
Marital status	Categoric
Native language	Categoric
ICU unit	Categoric

Figure 1 illustrates the missing data percentage in each feature. There is an 80% rate of missing data in capillary refill rate and fraction-inspired oxygen. Only five features (admission type, gender, intensive care unit, insurance, and age) don't have any missing data.

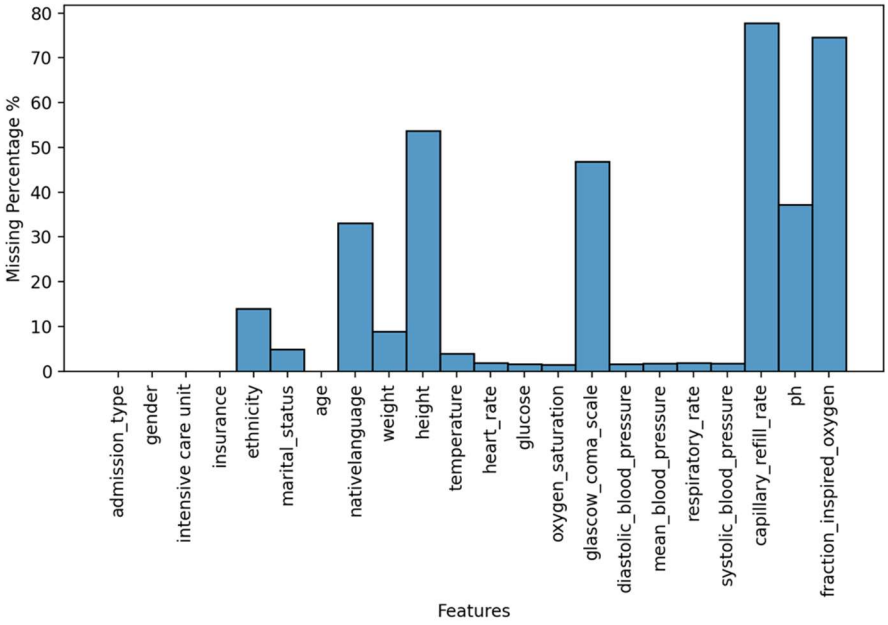

Figure 1. Missing data percentage in each feature.

Most of the machine learning models can only perform with datasets without missing data. That's why we need to fill the missing data first. KNN, MICE, Soft-Impute methods are used for filling missing records. After filling missing data, LOS prediction is performed with SVM, LR, and XGBoost classifier.

The KNN fills the missing data in a patient record by using the most similar patients' data. The MICE [12] performs the missing data filling in a patient record by using sequentially trained multivariate regression models. The Soft-Impute [13] fills the missing data with soft-thresholded singular value decomposition.

In medical predictive models, the SVM [14], LR, and XGBoost classifiers are widely used. The SVM aims to find the hyperplane parameters that distinguish the patients' LOS is longer than 3 days or not in the feature space. The LR, which is a kind of linear regression, uses the features in the patients' records as an independent variable and predicts the patients' LOS situation (LOS>3 days or LOS<3 days) as a dependent variable. XGBoost (eXtreme gradient boosting) [15] uses the gradient boosted decision trees to predict the patient LOS.

3. Results and Discussions

The dataset was split into two-part as 85% training and 15% test as same Harutyunyan et al. [9]. Firstly, we filled the missing data in the training set with the missing data imputation method and then filled the missing data in the test set with the missing data imputation method according to the training set. We trained classifiers on the filled training dataset and report area under curve (AUC) score on the filled test dataset. Hyperparameters of classifiers were tuned with a simple grid search algorithm.

The AUC scores of different LOS prediction models are given in Table 2. The best AUC score obtained by the XGBoost classifier and Soft-Impute missing data imputation was 66.1%.

Table 2. The AUC scores of prediction models

Missing Data Method	Classifier	AUC
MICE	XGBoost	0.643
MICE	SVM	0.614
MICE	LR	0.618
Softimp	XGBoost	0.661
Softimp	SVM	0.628
Softimp	LR	0.634
KNN	XGBoost	0.644
KNN	SVM	0.615
KNN	LR	0.628

In addition, Figure 2 visualizes these results. As shown in the figure, each classifier reached its own best AUC score with a soft-impute missing data imputation method. Besides, it is obviously seen that the SVM classifier has the lowest AUC score among these classifiers. The XGBoost classifier has outperformed the SVM and LR in lately medical predictive models ([16], [17], [18]). Our findings also support these previous results.

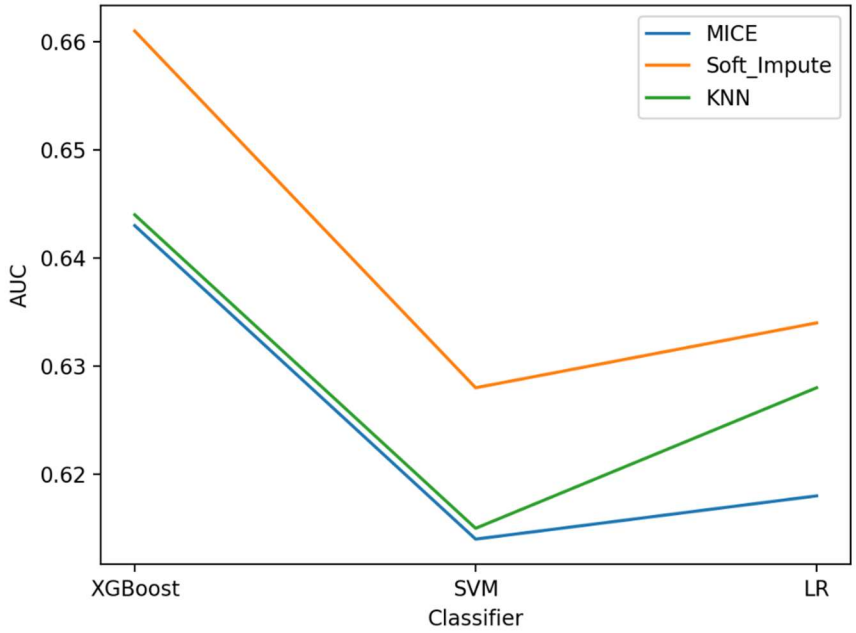

Figure 2. The AUC scores of prediction models.

When we compare these results with Wang et al. [19] who reported an AUC of 73.6 %, we conclude that this model can be improved by adding new features and using vital measurements taken at different time intervals.

4. Conclusion

In this paper, we predicted the ICU patients' LOS whether is longer than 3 days or not with machine learning algorithms. Firstly, we filled the missing data with KNN, MICE, and Soft-Impute, and then we performed SVM, LR, and XGBoost classifiers. We conclude that this result can be increased by adding

new features and measurements taken at different times. The Soft-Impute and XGBoost achieved the best results based on the AUC score. Furthermore, our findings show that filling missing data with the Soft-Impute yielded the highest AUC score for all classifiers. We plan to predict different LOS durations (5 days and 7 days) in the future, as well as develop a two-stage prediction model that includes mortality prediction.

Acknowledgement

This study is supported by Eskisehir Technical University Scientific Research Projects Committee (ESTUBAP-20DRP025).

References

[1] Ma, X., Si, Y., Wang, Z., &Wang, Y. (2020). Length of stay prediction for ICU patients using individualized single classification algorithm. *Computer Methods and Programs in Biomedicine, 186,* 105224. https://doi.org/10.1016/j.cmpb.2019.105224

[2] Olmez, E., Areta, O., & Er, O. (2021). Classification of Breast Cancer using Artificial Neural Network Algorithms. *Artificial intelligence theory and applications,1,* 57-68.

[3] Magesh, P. R., Myloth, R. D., &Tom, R. J. (2020). An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery. *Computers in Biology and Medicine, 126,* 104041. https://doi.org/10.1016/j.compbiomed.2020.104041

[4] Akgül, G., Çelik, A., Ergül Aydın, Z., & Kamışlı Öztürk, Z. (2020). Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı. Bilişim Teknolojileri Dergisi, 13 (3), 255-268. DOI: 10.17671/gazibtd.710728

[5] Daghistani, T. A., Elshawi, R., Sakr, S., Ahmed, A. M., Al-Thwayee, A., & Al-Mallah, M. H. (2019). Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *International journal of cardiology, 288,* 140–147. https://doi.org/10.1016/j.ijcard.2019.01.046

[6] Triana, A. J., Vyas, R., Shah, A.S., &Tiwari V. (2021). Predicting Length of Stay of Coronary Artery Bypass Grafting Patients Using Machine Learning. *Journal of Surgical Research, 264,*68-75. https://doi.org/10.1016/j.jss.2021.02.003.

[7] Maharlou, H., Niakan Kalhori, S. R., Shahbazi, S., & Ravangard, R. (2018). Predicting Length of Stay in Intensive Care Units after Cardiac Surgery: Comparison of Artificial Neural Networks and Adaptive Neuro-fuzzy System. *Healthcare informatics research, 24(2),* 109–117. https://doi.org/10.4258/hir.2018.24.2.109

[8] Thompson, B., Elish, K. O., & Steele, R.(2018). Machine Learning-Based Prediction of Prolonged Length of Stay in Newborns. *17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018*, 1454-1459. doi: 10.1109/ICMLA.2018.00236.

[9] Harutyunyan, H., Khachatrian, H., Kale, D.C. et al. (2019). Multitask learning and benchmarking with clinical time series data. *Sci Data* 6**,** 96 https://doi.org/10.1038/s41597-019-0103-9

[10] Gentimis, T., Alnaser, A. J., Durante, A., Cook, K. & Steele, R.(2017) Predicting Hospital Length of Stay Using Neural Networks on MIMIC III Data. *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 2017,* 1194-1201, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191.

[11] Johnson, A., Pollard, T., Shen, L. et al. (2016) MIMIC-III, a freely accessible critical care database. *Sci Data 3,* 160035. https://doi.org/10.1038/sdata.2016.35

[12] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software 45 (3)*, 1–67,doi:10.18637/jss.v045.i03.

[13] Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of MachineLearning Research, 11 (80)*,2287–2322.

[14] Cortes, C., Vapnik, V. (1995). Support-vector networks. *Mach Learn, 20,* 273–297, https://doi.org/10.1007/BF00994018

[15] Chen, T., Guestrin, C. (2016). XGBoost. *in: Proceedings of the 22ndACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM*, New York. doi:10.1145/2939672.2939785.

[16] Liu, J., Wu, J., Liu, S., Li, M., Hu, K. & Li, K. (2021) Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. *PLoS One*, *16(2).* doi: 10.1371/journal.pone.0246306

[17] Leung, W. K., Cheung ,K. S. , Li, B., et al.(2021) Applications of machine learning models in the prediction of gastric cancer risk in patients after Helicobacter pylori eradication. *Aliment Pharmacol Ther, 53 (8), 864– 872.*

[18] Pang, X., Forrest, C. B., Lê-Scherban, F., Masino, A. J. (2021) Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics. 150, 104454.* https://doi.org/10.1016/j.ijmedinf.2021.104454.

[19] Wang, S., McDermott, M.B.A., Chauhan, G., Ghassemi, M., Hughes, M.C., Naumann, T., & Ghassemi, M. (2020). Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii*.,Proceedings of the ACM Conference on Health, Inference, and Learning*, 222-235. https://doi.org/10.1145/3368555.3384469