EasyChair Preprint

№ 15260

Optimizing Patient Length of Stay with Machine Learning: a Comparative Study Using Neural Networks Methods, Regression Methods, and Apriori Algorithm

Sahatas Chatnopakun, Kant Panyavanich and Maleerat Maliyaem

October 18, 2024

Optimizing Patient Length of Stay with Machine Learning: A Comparative Study Using Neural Networks Methods, Regression Methods, and Apriori Algorithm

Sahatas Chatnopakun
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
s6707011956011@email.kmutnb.ac.th

Kant Panyavanich
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
s6707011956029@email.kmutnb.ac.th

Maleerat Maliyaem
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
maleerat.m@itd.kmutnb.ac.th

***Abstract*— The duration of an inpatient stay affects hospital administration and improves hospital effectiveness in terms of controlling expenses and raising patient standards. It also assists in identifying the correlations among illnesses requiring hospitalization. For our study, we took 24,150 records from of the Open Data database pertaining to inpatient admissions in 2023. We used a number of methods, including Neural Networks, Deep Learning, Linear Regression, and Support Vector Machines, to predict the Length of Stay (LOS). We converted the data to numerical form for predictive purposes, dividing the dataset into 70% for training and 30% for testing. We assessed the model's performance using Root Mean Squared Error (RMSE) and split the forecast into four LOS categories: 0-2, 3-4, 5-7, and 8 days or more. The study also employed the Apriori algorithm to identify illness association rules that could impact LOS estimates. The results showed that identifying illness correlations is one element that might aid in enhancing the capacity to predict LOS.**

Keywords*— *Length of Stay, Neural Networks, Deep Learning, Support Vector Machines, Linear Regression, Apriori Algorithm, Association Rules, Healthcare Management.

I. INTRODUCTION

The length of hospital stay, or Length of Stay (LOS), is a crucial component in evaluating patient outcomes, cost control, and the effectiveness of healthcare. The precise forecasting of LOS has gained significance, particularly within data-driven healthcare, as innovations in machine learning and artificial intelligence have created new opportunities for predictive modeling. A multitude of studies have utilized various approaches to forecast hospital length of stay, employing both structured and unstructured data from multiple sources. A study by Zeleke et al. [1] examined the application of machine learning algorithms, including regression methods, to forecast length of stay in inpatients. This study emphasized the importance of integrating clinical and demographic data into predictive models. Yildirim and Canayaz [2] proved the efficacy of ensemble approaches in forecasting LOS, with a particular emphasis on neonatal intensive care units. Additionally, a study by Hu et al. [3] employed network analytics and machine learning methodologies to forecast length of stay in elderly patients with chronic illnesses.

This study utilized various machine learning methods, to create predictive models for LOS. The dataset, comprising numerical and categorical variables, was preprocessed and transformed for best algorithm performance. Additionally, we utilized the Apriori method of association rule mining to uncover correlations between diseases in the dataset, thereby enhancing the interpretability of the prediction outcomes. The primary objective of this study is to evaluate the efficacy of several prediction models in estimating LOS and to elucidate the critical elements that lead to prolonged hospital admissions. This paper enhances the debate on data-driven healthcare management by providing a comprehensive analysis of LOS prediction, utilizing both sophisticated machine learning methodologies and conventional statistical techniques.

II. DATA PROCESSING

A. Data Set and Summary Statistics

This study included data derived from 24,150 inpatient admission records from 2023, encompassing 5,809 cases. The data was obtained via the Open Data of the Bamrasnaradura Infectious Diseases Institute, a hospital affiliated with the Department of Disease Control of Thailand [4], with all personal patient information anonymized to ensure privacy protection. Each record comprises eight fields: Age, Gender, Treatment Field include Orthopedic (ORTHO), Surgical (SUR), Pediatric (PED), Obstetrics (OB), Medicine (MED), Gynecology (GYN), Ear, Nose, and Throat (ENT), Ophthalmology (OPH), and Urology (URO), ICD-10 Codes, Disease Name, ICD-10 Status, LOS, and Right to Treatment.

Each record denotes a patient's therapy for a singular ailment. The ICD-10 Status field is essential as it signifies whether the visit pertained to a Principal Diagnosis, whereas future records may disclose any External Causes, Comorbidities, Complications, or other considerations. The quantity of conditions addressed may considerably influence the patient's LOS.

TABLE I. PATIENT DISTRIBUTION BY LOS, GENDER, AND AGE

LOS	*Patients (%)*	Gender		Age				
		Male	*Female*	*[0,1]*	*(1,6]*	*(6,18]*	*(18,60]*	*(60,60+]*
0-2	27.51	733	865	185	258	171	628	356
3-4	33.35	789	1,148	310	227	107	820	473
5-7	21.43	523	722	168	64	46	355	612
8+	17.59	496	526	121	4	7	248	642
-	0.12	1	6	1	0	0	3	3
Percentage		43.76	56.24	13.51	9.52	5.70	35.36	35.91

Table 1 illustrates the percentage of patients categorized by LOS. It signifies that 82.29% of patients remained for a duration of 1 to 7 days. The distribution of male and female patients was relatively balanced, with a predominant majority over 19 years of age (71.27%).

B. Data Integrity

The data utilized in this investigation indicated that there were 7 absent cases in the LOS field. The absence of these values does not influence the machine learning process, akin to the results observed in the research of Mehrabani-Zeinabad et al. [5], wherein the omission of missing values in predictor variables from the dataset did not substantially alter model efficacy. This study removed records without LOS data from the analysis, as LOS could not be determined, and the low number of missing entries would not impact model development.

C. Data Selection

The disease name field provides identical information to the ICD-10 codes and was hence omitted from the machine learning process. The dataset was constrained in breadth due to privacy protection requirements that prohibit access to individual patient information. Thus, all accessible fields were utilized in the predictive modeling for this investigation, akin to the methodology employed by Desai et al. [6], who applied machine learning techniques for predictive modeling in healthcare, leveraging all available fields despite data constraints.

D. Data Transformation

This study employs a dataset that includes both numerical and textual information. All data was converted into numerical representation to enable its application in machine learning for this investigation. Textual data, encompassing ICD-10 Codes, ICD-10 Status, and Right to Treatment, was transformed into categorical fields. Each newly created field was designated a binary value of either 1 or 0. The Treatment Field served as a filter in the machine learning process, enabling predictions to be classified in two manners: utilizing data from all departments or generating predictions tailored to each department. This methodology parallels the study conducted by Samynathan [7], wherein ICD codes were converted into numerical data to enhance predictive capabilities, illustrating the efficacy of machine learning in healthcare data analysis.

III. PREDICTIVE METHODS

Forecasting LOS in healthcare is essential for optimizing resource distribution and enhancing patient management. Altair AI Studio was utilized to construct predictive models, to attain precise predictions. Furthermore, association rules generated by the Apriori algorithm were utilized to investigate correlations among different diseases in the dataset. The relationships were subsequently incorporated into the predictive models to improve the analysis of LOS outcomes.

A. Neural Network

This study employed a neural network to forecast LOS based on patient demographics and clinical characteristics. The neural network model utilized backpropagation to modify the weights of inter-neuronal connections, therefore reducing prediction errors via repetitive adjustments. Rumelhart et al. [8] assert that backpropagation is an effective method for training neural networks, facilitating the model's enhancement by minimizing the discrepancy between expected and actual results. The neural network architecture comprised an input layer, many hidden layers to capture intricate patterns, and an output layer to forecast LOS. The Rectified Linear Unit (ReLU), a frequently utilized activation function, was employed in the hidden layers to introduce non-linearity and enhance model performance [9]. The model was trained utilizing gradient descent, aiming to minimize the Root Mean Squared Error (RMSE) between predicted and actual LOS values. RMSE was used as it offers a comprehensible metric of prediction error in the identical units as the target variable [10].

B. Deep Learning

This study also used deep learning, a subclass of neural networks characterized by an increased number of layers. Since deep learning algorithms are able to automatically extract high-level features from raw data, they are particularly well-suited to handle large datasets with complex, non-linear relationships [11]. Deep learning was chosen for this study due to its large clinical dataset and ability to uncover hidden patterns that impact LOS. The deep learning approach makes it possible to identify intricate correlations between variables that traditional methods might not be able to easily identify. One technique to reduce overfitting, a common problem in deep learning when models overfit to training data, is dropout regularization [12]. Backpropagation minimizes the loss function, which is typically the root mean square error for regression problems. A lower RMSE ensures that the model improves its predictions with time.

C. Linear Regression

Linear regression, a fundamental and useful statistical tool for predictive modeling, was utilized to estimate LOS. Linear regression delineates the linear association between the dependent variable (LOS) and one or more independent factors (e.g., age, medical history, illness severity). Prior studies, like those by Draper and Smith [13], has established the effectiveness of linear regression in forecasting healthcare outcomes, especially when the interrelations among variables are presumed to be linear. This study involved constructing a linear regression model by fitting a linear equation to the data, thereby minimizing the sum of squared errors between the observed LOS and the projected values. The model is trained by minimizing the RMSE cost function.

D. Support Vector Machines

Support Vector Machines (SVM) are extensively utilized in classification and regression tasks, including healthcare forecasting. SVM creates a hyperplane in a high-dimensional space to differentiate between distinct classes or, in regression scenarios, to forecast continuous values. Cortes and Vapnik [14] developed the SVM algorithm, which has proven beneficial in diverse applications owing to its capacity to manage both linear and non-linear data. This work utilized SVM to forecast LOS, employing the Radial Basis Function (RBF) kernel to identify non-linear correlations between the characteristics and LOS. The kernel approach enables SVM to project input data into a higher-dimensional space without explicitly calculating the transformation.

E. Apriori Algorithm

The Apriori algorithm, developed by Agrawal and Srikant [15], is a prevalent method for extracting association rules from extensive datasets. It is especially adept in discerning relationships among co-occurring entities, such as diseases in this case. The system initially identifies common itemsets and

subsequently generates rules that elucidate relationships among those items. This study employed the Apriori method to identify correlations among different diseases within the patient dataset. The algorithm produced association rules according to established minimum support and confidence standards, designed to guarantee that the identified rules were both prevalent and robust.

IV. Experiments

Forecasting LOS were performed using Neural Network, Deep Learning, Linear Regression, and SVM algorithms, to evaluate and compare the performance of each model. After inputting the data, it was filtered by various factors, including LOS and department. Subsequently, the fields used for prediction were selected, and several features influencing the prediction were tested. All values were transformed into numerical format. To assess model accuracy, the dataset was randomly divided into two subsets: 70% for training and 30% for testing [16]. RMSE was used as the metric for measuring the prediction error of LOS. The results were compared with inpatient records from 9 departments. LOS prediction was categorized into four periods: 0-2 days, 3-4 days, 5-7 days, and 8 days or more. The data distribution across these periods was 27.51%, 33.35%, 21.43%, and 17.59%, respectively, with relatively equal patient counts used in the prediction process.

Furthermore, the relationship between diseases in inpatients was analyzed using the Apriori Algorithm, with relationships evaluated for each department. The confidence level for association rules was set to no lower than 90%, and the minimum support was established at 0.01 Recent studies, such as [17], highlight the utility of the Apriori algorithm in medical data analysis for discovering meaningful associations.

V. Results and Discussion

A. Prediction

The prediction of LOS produced RMSE across nine departments. However, the ENT department had no patient data for stays of 5-7 days, and there was no patient data for the OB, GYN, ENT, and OPH departments for stays of 8 days or more. Field selection testing revealed that the most influential fields in the prediction process, in order of importance, were ICD-10 Codes, Right to Treatment, Gender, Age, and ICD-10 Status.

TABLE II. RMSE by LOS Category

LOS		RMSE			
		Neural Net	***Deep Learning***	***Linear Regression***	***SVM***
All Days	Mean	4.464	4.714	4.216	4.216
	SD	3.458	3.866	3.295	3.295
0-2 Days	Mean	0.493	0.460	0.560	0.560
	SD	0.141	0.204	0.138	0.138
3-4 Days	Mean	0.491	0.500	0.588	0.588
	SD	0.063	0.051	0.155	0.155
5-7 Days	Mean	0.737	0.681	0.652	0.652
	SD	0.315	0.401	0.304	0.304
8+ Days	Mean	11.346	13.365	10.681	10.681
	SD	7.533	10.545	6.807	6.807

Table II displays the average and standard deviation of RMSE for every department. It is evident that there is no substantial difference in LOS values between departments. In comparison to stays of eight days or more, when forecast errors can reach up to fourteen days, the RMSE for LOS in the range of 0-2 days, 3-4 days, and 5-7 days does not surpass one day, indicating higher accuracy. Dividing the time periods results in a decrease in forecasting accuracy, as the overall prediction error is approximately 5 days. When the algorithms are compared, there is no observable variation in their predicted performance. On the other hand, we found that SVM outperforms other algorithms in predicting stays of eight days or more.

B. Association Rule

The association rules for diseases revealed relationships involving between 2 to 6 diseases. When analyzing disease relationships across all departments, 9 out of 15 rules were related to successful childbirth (e.g., Single Live Birth and Singleton, Born in Hospital). When the data was separated by department, the discovered disease relationships were specific to each department. For example, in the ORTHO department, most relationships involved Escherichia coli as the cause of diseases classified under other chapters (B962), primarily associated with Urinary Tract Infection, site not specified (N390). In the SUR department, conditions related to neoplasms were frequently linked to chemotherapy sessions for neoplasm treatment, among other associations.

TABLE III. The Number of Association Rules

Department	Count of HN	Count of ICD-10	Minimum Support (Rule)		
			0.01	***0.03***	***0.05***
All	5,802	1,879	15	2	0
ORTHO	430	334	15	0	0
SUR	916	559	7	0	0
PED	1,602	372	11	3	1
OB	406	150	147	21	0
MED	2,214	1,158	11	4	2
GYN	98	81	4,465	4	0
ENT	28	22	253	253	4
OPH	52	73	6,941	162	7
URO	56	106	∞	633	5

TABLE IV. Top 3 ICD-10 and Association Rules

Department	ICD-10	Association Rule
All	1. I10 2. E789 3. E119	1. O800, Z115 → Z370 2. J128, I10 → U071 3. O821 → Z370
ORTHO	1. I10 2. E789 3. M171	1. B962 → E789, N390, I10 2. B962, I10 → E789, N390 3. B962 → N390, I10
SUR	1. I10 2. E789 3. Z511	1. C20, C787 → Z511 2. Z511, C23 → C787 3. C780, C509 → Z511
PED	1. Z380 2. A099 3. A979	1. P369 → Z380 2. P221 → Z380 3. P221, P369 → Z380
OB	1. Z370 2. Z115 3. O800	1. D649 → O990 2. N736 → O342 3. D582, O800 → O990
MED	1. I10 2. E789 3. E119	1. J128 → U071 2. J128, E789 → U071 3. J128, I10 → U071
GYN	1. I10 2. Z115 3. D259	1. C56 → Z511 2. R571 → D62 3. E119 → E789
ENT	1. G473 2. I10 3. E789	1. E119 → I10 2. J310 → J351 3. J310 → J351, G473
OPH	1. I10 2. H259 3. H3602	1. Z115 → H250 2. H431 → H3602 3. H3602, Z961 → E143
URO	1. I10 2. E789 3. N40	1. N319, N179 → N136 2. A415 → N136 3. T913 → N319

Tables III and IV delineate the quantity of associated disorders and the number of association rules per department, where Confidence surpassed 90%, Lift above 3, and support were established at 0.01, 0.03, and 0.05, respectively. Through the assessment of Confidence, Lift, and Support, the most prevalent and strongly correlated diseases within each department were determined, with the top three diseases for each department emphasized.

The most commonly recognized ailment was Essential (primary) hypertension (I10), which shown significant correlations with other disorders. The number of patients influenced the association rules, with Single Live Birth (Z370) being the most strongly associated condition across all departments, particularly in the PED and OB departments. The second most commonly associated disease was Coronavirus disease (COVID-19), virus identified (U071), predominantly from the MED department.

From the experiments analyzing the relationship between diseases and LOS prediction, it was found that disease relationships are a significant factor contributing to the improved accuracy of LOS predictions. Furthermore, LOS prediction may also depend on other related factors, including patient age, presence of complications, number of concurrent diseases, and the patient's right to treatment.

VI. Conclusion

This study used inpatient data from Open Data, anonymizing all personal patient information and limiting the number of available fields to seven. Therefore, we used all variables for prediction and converted the data into a numerical format for LOS forecasting. We assessed the predictions' performance using RMSE, which revealed very minor variations in RMSE values between the algorithms. We evaluated the model by segmenting predictions into temporal intervals and found that the intervals exhibiting the highest predictive accuracy, with an error margin not exceeding one day, were 0-2 days, 3-4 days, and 5-7 days. This finding corroborates Ma et al. [18], which indicated that LOS prediction accuracy is significantly high in shorter time intervals, exhibiting low error. This study utilized the Apriori Algorithm to assess illness relationships and identify commonly co-occurring conditions to examine the relationship between diseases and inpatient LOS. The study showed that illness correlations have a big effect on how well length of stay predictions work. They should be looked at along with other important factors, such as patient age, the presence of complications, the number of comorbidities, and healthcare coverage. Comparing the length of stay prediction results from this study with other research may prove difficult due to discrepancies in data sources, the volume of data available for prediction, bed management procedures, and differing hospital rules. Precisely forecasting the length of stay improves hospital administration efficiency, particularly in terms of bed allocation and patient care within each department. Current hospital systems could enhance operational efficiency by incorporating this research's LOS prediction algorithm and disease association analysis [19]. This study found the link between LOS prediction and illness association rules in order to find important factors that affect LOS predictions. The goal was to make a good LOS prediction model for hospital systems. Enhancing the patient dataset and integrating illness association rules into the predictive model could significantly augment the precision of length of stay forecasts.

References

[1] A. J. Zeleke, P. Palumbo, P. Tubertini, R. Miglio, and L. Chiari, "Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis," Front. Artif. Intell., vol. 6, Jul. 2023, Art. no. 1179226.

[2] A. E. Yildirim and M. Canayaz, "Machine learning-based prediction of length of stay (LoS) in the neonatal intensive care unit using ensemble methods," Neural Comput. Appl., vol. 36, 2024, doi: 10.1007/s00521-024-09831-7.

[3] Z. Hu, H. Qiu, L. Wang, and M. Shen, "Network analytics and machine learning for predicting length of stay in elderly patients with chronic diseases at point of admission," BMC Med. Inform. Decis. Mak., vol. 22, no. 1, Mar. 2022, Art. no. 62, doi: 10.1186/s12911-022-01802-z.

[4] Department of Disease Control of Thailand, August 20, 2024, “Inpatient registration,” Bamrasnaradura Infectious Diseases Institute, a hospital under the Department of Disease Control of Thailand. [Online]. Available: https://opendata.ddc.moph.go.th/sv/dataset/https-ddcopendata-ddc-moph-go-th-f-datasetview-413-1

[5] K. Mehrabani-Zeinabad, M. Doostfatemeh, and S. M. T. Ayatollahi, "An efficient and effective model to handle missing data in classification," Biomed Res. Int., vol. 2020, Art. no. 8810143, 2020, doi: 10.1155/2020/8810143.

[6] R. J. Desai, S. V. Wang, M. Vaduganathan, T. Evers, and S. Schneeweiss, "Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes," JAMA Netw. Open, vol. 3, no. 1, Jan. 2020, Art. no. e1918962, doi: 10.1001/jamanetworkopen.2019.18962.

[7] S. K. Samynathan, "ICD-10 Code Prediction using Machine Learning," M.S. thesis, Dept. Comput., Nat. Col. Ireland, Dublin, Ireland, 2022.

[8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[9] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in Proc. 14th Int. Conf. Artificial Intell. Statist., Fort Lauderdale, FL, USA, 2011, pp. 315-323.

[10] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," Climate Res., vol. 30, no. 1, pp. 79–82, 2005.

[11] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, pp. 1929-1958, Jun. 2014.

[13] N. R. Draper and H. Smith, Applied Regression Analysis, 3rd ed. New York, NY, USA: Wiley, 1998.

[14] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995.

[15] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Data Bases (VLDB), Santiago, Chile, 1994, pp. 487-499.

[16] R. N. Mekhaldi, P. Caulier, S. Chaabane, A. Chraibi, and S. Piechowiak, "A comparative study of machine learning models for predicting length of stay in hospitals," J. Inf. Sci. Eng., vol. 37, no. 5, pp. 1025-1038, Sep. 2021, doi: 10.6688/JISE.202109_37(5).0003.

[17] H. Ma, J. Ding, M. Liu, and Y. Liu, "Connections between various disorders: Combination pattern mining using Apriori algorithm based on diagnosis information from electronic medical records," Biomed. Res. Int., vol. 2022, May 2022, Art. no. 2199317, doi: 10.1155/2022/2199317.

[18] F. Ma, L. Yu, L. Ye, D. D. Yao, and W. Zhuang, "Length-of-stay prediction for pediatric patients with respiratory diseases using decision tree methods," IEEE J. Biomed. Health Inform., vol. 24, no. 9, pp. 2651-2662, Sep. 2020, doi: 10.1109/JBHI.2020.2973285.

[19] R. Ippoliti, G. Falavigna, C. Zanelli et al., "Neural networks and hospital length of stay: an application to support healthcare management with national benchmarks and thresholds," Cost Eff. Resour. Alloc., vol. 19, Dec. 2021, Art. no. 67, doi: 10.1186/s12962-021-00322-3