PRINT ISSN 1119-8362
Electronic ISSN 2659-1502

Full-text Available Online at
https://www.ajol.info/index.php/jasem
https://www.bioline.org.br/ja

J. Appl. Sci. Environ. Manage.
Vol. 27 (12) 2901-2906 December 2023

Prediction and Classification of Patients Length of Stay in a Medical Hospital in Birnin Kebbi, Kebbi State, Northwestern Nigeria

[*1]SULEIMAN, S; [2]USMAN, U; [3]BELLO, A; [4]BUNZA, MI

*[*1]Department of Statistics, Federal University Dutsin-Ma, Katsina State, Nigeria*
[2]Department of Statistics Usmanu Danfodiyo University Sokoto, Sokoto State, Nigeria
[3]Department of Computer Science, Usmanu Danfodiyo University Sokoto, Sokoto State, Nigeria
[4]Department of Statistics, Federal University Birnin Kebbi, Kebbi State, Nigeria.

**Corresponding Author Email: macostamkd@gmal.com, ssuleiman@fudutsinma.edu.ng*
Co-Authors Email: uusman07@gmail.com; bello.abdulkarim@udusok.edu.ng; bunzamuhammadibrahim@gmail.com

ABSTRACT: The objective of this paper was to predict and classify the attributes that influences patients Length of Stay (LOS) in a Medical Hospital in Birnin Kebbi, Kebbi State, Northwestern Nigeria using tree-based machine learning algorithms after data collection. When training the modes, random forest achieved an R-squared value of 0.573541 using a continuous response and classification rate of about 87% using categorical response variable. In testing the performance of the top identified modes, random forest mode had an accuracy of 72%. Linear regression model was also used in predicting patient's length of stay. Tree based models performs better than the linear regression model. The result shows that random forest outperforms decision tree and boosted tree in predicting and classifying patient LOS.

DOI: https://dx.doi.org/10.4314/jasem.v27i12.32

Cite this paper as: SULEIMAN, S; USMAN, U; BELLO, A; BUNZA, M. I. (2023). Prediction and Classification of Patients Length of Stay in a Medical Hospital in Birnin Kebbi, Kebbi State, Northwestern Nigeria. *J. Appl. Sci. Environ. Manage.* 27 (12) 2901-2906

Dates: Received: 12 November 2023; Revised: 10 December 2023; Accepted: 15 December 2023 Published: 30 December 2023

Keywords: Length of Stay, Prediction, Classification, Decision Tree, Boosted Tree, Random Forest

Patients Length of Stay (LOS) is frequently used as a performance measuring criterion by researchers in the field of hospital management (McDermott et al., 2007). The reason for LOS's popularity is attributed to its relationship with other vital hospital performance metrics. Thomas et al., (1997) studied the dependency of patient LOS on the quality of care provided by the hospital. The researchers found that the inferior quality of care was positively related to long LOS. Hassan et al., (2010) found that increase in patient LOS increases the probability of acquiring infections while in the hospital. Other studies also found that shorter than required LOS is positively related to hospital readmissions Jencks et al., (2009). Public and private health insurance providers reward hospitals for providing quality care to the patients. U.S. Centers for Medicare and Medicaid in addition to rewarding hospitals for superior care also penalizes hospitals for excess re-admissions. Therefore, hospitals aim to maximize their rewards by providing quality care to the patients and minimize re-admissions related penalties by preventing readmissions (Goantiya, 2018). Having an estimate of the number of days a patient is required to stay at the hospital can be helpful in preventing early and late discharges. Also, knowing the patient attributes that influence patient LOS can help hospitals in identifying the current good practices and areas for improvement. Numerous predictive modeling techniques, including supervised and unsupervised, can be used to predict patient's LOS. The techniques that require a training dataset containing predictor variables, with their values and their corresponding response variable values to approximate the relationship between the predictor

variables and response variables are categorized as supervised predictive modeling techniques. The techniques that don't require a training dataset containing predictor variable values and their corresponding response variable values to approximate the relationship between predictor and response variables are categorized as unsupervised predictive modeling techniques.

Supervised predictive modeling techniques were used to predict and classify patient LOS in this research. A training set is a requirement while utilizing supervised predictive modeling techniques, for this research, the training dataset was derived from patient's record of Federal Medical Centre (FMC) Birnin Kebbi. In addition, a vital component in the management of hospital resources and improved efficiency while providing adequate care is to understand the relationship of patient LOS with various medical and socio-demographic variables.

Predictive modeling techniques can also be used to identify the medical and socio-demographic variables influencing patient LOS, and some techniques can even quantify the relationship between the identified influential variables and LOS (Goantiya, 2018). Tree based modeling techniques like decision tree, boosted tree, and bootstrap forest have not been extensively utilized for the purpose of understanding patient LOS. Based on the conducted literature review, regression-based modeling techniques appear to be the most commonly used techniques in predicting patient LOS. Multiple studies suggests that tree-based techniques like decision tree, boosted tree, and bootstrap forest are less frequently used in predicting and classifying LOS.

Some studies also suggested that the performance of tree-based modeling techniques is comparable to that of regression-based techniques when applied to patient length of stay data (Goantiya, 2018).

Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can effectively generalize and thus perform tasks without explicit instructions (Koza *et al.*,1996).

Machine learning techniques can be used for prediction as well as classification. Recently many researchers have employed machine learning techniques for classification and prediction of different phenomena such as credit scoring (Suleiman et al., 2021), Diabetic status (Gulumbe et al., 2019), student academic performance (Suleiman et al., 2019) and customer accessibility (Suleiman and Usman, 2016) among others. The aim of this research work is to predict and classify patients' Length of Stay (LOS) using decision tree, boosted tree and random forest models.

This aim can be achieved by identify the modeling technique(s) that can be used by the hospital to predict or classify LOS of an incoming patient using the limited patient related information available at admittance, identify the factors or patient attributes that influence patient LOS at the hospital using decision tree, boosted trees, and random forests, test the performance of the best identified models in predicting and classifying patient LOS.

MATERIALS AND METHODS

Study Area: The proposed site for the study is Federal Medical Center (FMC), Birnin Kebbi. Birnin Kebbi is the capital city of Kebbi state in Northwest Nigeria. The state shares boundary with Sokoto state to the north, Zamfara state to the east, Niger state to the south and Benin republic to the west.

The city has one local government area (Birnin Kebbi LGA) and twenty more local government outside the capital city. The population of Bimin kebbi is approximately 380,000 people with male to female ratio of 1:1[11] and the indigenous inhabitants are mainly Hausas and Fulanis.

People of other tribes like Yorubas, Igbos, Nupes, etc also constitute significant portion of the town's residents. The people are majorly traders, farmers, civil servants and others. The hospital is a 300-bed tertiary health centre which offers primary, secondary and tertiary health services.

Data Collection: The data used for the research was secondary data and it was obtained from patients' record at Federal Medical Centre (FMC), Birnin Kebbi.

Data Analysis: The main aim of this paper was to predict and classify patients' Length of Stay (LOS) and to identify the factors or patient attributes that influence patient LOS at the hospital using data collected from patient's record in Federal Medical Centre (FMC) Birnin Kebbi for a period of six months (July to December 2022).

The dataset consists of 2,602 instances and 13 attributes. Table 1 shows the complete attributes for the Study dataset.

Table 1: Attributes of the study dataset.

No	Attribute	Description	Possible Values
1	Age	Patients age at time of admit	Min[1] Max[100]
2	Length of Stay	Calculated LOS days	Min[1] Max[212]
3	LOS Class	Three classes for the categorical LOS	A[1to5days],B[6todays], C [11 and above days]
4	State Zone	The state zone of patient's residence	4 zones (Central,North, South,others)
5	Emergency	Binary variable indicating whether the patient was admitted through the Emergency Department	Yes, No
6	Visit	The number of visits seen by the patient	1,2,3 etc.
7	Seven day readmit	Binary variable indicating whether or not the patient has been readmitted within 7 days	Yes, No
8	Thirty day readmit	Binary variable indicating whether or not the patient has been readmitted within 30 days	Yes, No
9	Insurance	Type of insurance patient used	3 types
10	Treatment team same	Binary variable indicating whether treatment team's last round was on the day of discharge	Yes, No
11	Department discharged	The department the patient was discharged from	8 Departments (Medical,Surgical,etc)
12	Same Nurse	A binary variable indicating whether or not the same nurse was the same first and last rounding provider	Yes, No
12	Type of patient	Type of patient	2 values

Decision Trees: Decision trees or Classification and Regression tree is a supervised machine learning method which works on the principle of recursive partitioning (Speybroeck, 2012). The dataset is divided into subsets by splitting the data based on one variable at a time (Loh, 2011). Decision trees can be used for classification as well as regression problems.

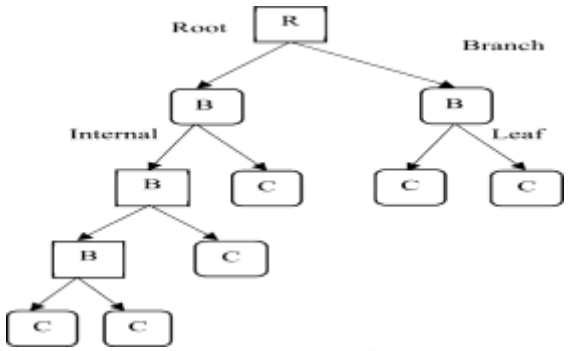

Fig.1: A typical Decision tree diagram before and after splitting, showing the root, internal and leaf node.

Regression trees: This section focused on the splitting mechanism of the decision tree when the response variable is continuous in nature. Consider a dataset R with N rows and $P+1$ columns. Out of the $P+1$ columns, P columns represent the independent variables and the remaining column is the response variable y. Let X_{ij} denote the value at the i^{th} row of the j^{th} column and, y_i be the value of the response variable for the i^{th} row, where, $i = (1,2,3,\dots,N)$ and $j = (1,2,3,\dots,P)$. The dataset R is divided into two regions R_1 and R_2 after the first split. This first split is performed at a point m on the independent variable j such that the following expression is minimized, Equation 1 is composed of two parts; the first part represents the sum of squares value of the residuals for the region R_1 and the second part represents the sum of squares value of the residuals for the region R_2. The value of y in each region is equal to the mean of actual y values in the region.

$$Min \sum_{i:Xij<m}^{n} (y_i - y)^2 + Min \sum_{i:Xij\geq m}^{n} (y_i - y)^2 \; where \; X_{ij} \in R \quad (1)$$

Classification trees: In this section, the splitting mechanism of the decision tree with categorical response variable is discussed. Suppose R_g denotes a region in the dataset R before the g^{th} split takes place, then the split will be performed at the point in R_g where the independent variable j is equal to m such that the equation 2 is minimized. Also, R_{g+1} and R_{g+2} are the two resulting regions after the split.

$$N_{R_{g+1}} * E_{R_{g+1}(j,m)} + N_{R_{g+2}} * E_{R_{g+2}(j,m)} \; (2)$$

Where,

$$E_{R_k} = Min\frac{1}{N_{R_k}}\sum_{i:X_{ij}\in R}^{n} I(y \neq y_i) \quad (3)$$

and, N_{R_k} is the number of x_{ij} in the region R_k, I is an indicator that take a value of 1 if the actual value is not equal to the classified value and 0 otherwise. The equation 3 represents the minimum value of the fraction of data points $x_{ij} \in R_k$ misclassified by a majority vote in the region R_k. Further, the resulting regions will include data points such that,

$$R_{k+1_{(j,m)}} = \{i: X_{ij} < m\} \; and \; R_{k+2_{(j,m)}} = \{i: X_{ij} > m\} \quad (4)$$

This process of splitting continues until a predefined condition is achieved. These predefined conditions can be the number of splits, minimum number of records in the data subset or region, etc. Once, a predefined condition is met, the splitting process stops, and tree-like output is produced. This output is a series of if and else statements based on the splitting point.

Boosted Trees: Boosted Tree involves boosting of the decision trees, i.e. combining the results of several decision trees to provide predictions (De'ath, G. 2007). The intention is to improve the prediction by combining results of several weak decision trees (Schapire et al., 2012). Initially, a simple tree is created using the training dataset, the predictions of this tree are then compared to the actual response values and residuals are calculated. Using these misclassifications or errors, a new tree is fitted to these residuals using all or a random sample of predictors. For continuous response variable, the scaled residual for the i^{th} observation in a leaf is calculated using the equation 5.

$$Scaled\ residual = \bar{y}' - y_i \quad (5)$$

Where $\bar{y}'$ is the mean of predicted values for the leaf and y_i is the actual response value for the i^{th} observation.

Random Forest: Random forest introduced by Breiman involves the creation of several decision trees each modeled using a random sample of the dataset and a random subset of the predictor variables for each tree split (Breiman, 2001). According to the algorithm created by Breiman, for a categorical response variable y, where y takes m discrete classes in the provided training dataset, random forest algorithm starts by creating a user defined number of categorical trees, using a random sample from the training dataset sampled with replacement and with each tree using a fixed number of random subset of predictor variables to perform splitting.

After the predefined numbers of trees are created, the random forest's classification is a result of the voting performed by all of the created classification trees. The class of the categorical response variable **y**, that receives the maximum number of votes or the class that majority of the created trees predict as their outcomes is considered as the final predicted class for any given set of predictor variable values.

Similarly, for a continuous response variable y, random forest algorithm involves creation of a user defined number of regression trees. The regression trees are created using a random sample of training dataset sampled with replacement. Each tree then uses a fixed number of randomly selected predictor variables to perform each split. After the predefined numbers of trees are created, the predictions made by each of the trees are averaged and the resulting mean value is considered as the final prediction.

RESULTS AND DISCUSSION

Using patients' attributes in table 1, several trees were created using training portion of 0.9, 0.8, 0.7, 0.6 and 0.5. The R-square values for training dataset provided by the trees created for each training portion are presented in table 2. Similarly, the Classification rate values for training dataset provided by the trees created for each training portion are presented in table 3.

Table 2: R – Square values for continuous LOS

S/N	Training Portion	R – Squared Values			
		Decision tree	Boosted tree	Random forest	Linear regression
1	0.9	0.342126	0.307609	**0.573541**	0.131042
2	0.8	0.362923	0.340953	0.566331	0.129411
3	0.7	0.340639	**0.381179**	0.567097	0.156577
4	0.6	0.341518	0.275596	0.560126	0.174497
5	0.5	**0.385744**	0.352368	0.559426	**0.192136**

Table 3: Classification rates for Categorical LOS

S/N	Training Portion	Classification Rate	
		Decision tree	Random forest
1	0.9	0.689115	0.861200
2	0.8	0.693301	0.857300
3	0.7	0.696341	0.841800
4	0.6	0.697719	**0.871200**
5	0.5	**0.705288**	0.859300

Figure 2 shows the decision tree created with the best R-Square value. From the figure, it can be observed that not all the influential variables are displayed; this is because the tree will be difficult to have a good display when all the important variables are used in creating the tree. The tree displayed only the top influential variables with a maximum depth of 4.

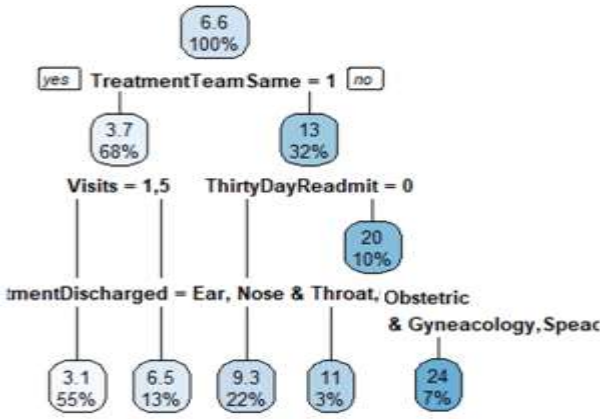

Fig. 2: A tree created with the best model using decision tree

The first split divides the training dataset into two nodes, first node includes patients with same treatment team, and the second node includes patients with different treatment team. The node containing patients with same treatment team is then split into two new nodes based on patient visit. The first node includes patients with visits of 1 or 5 and the second node includes patients with visits other than 1 and 2. The node containing patients different treatment team is then split into two new nodes based on thirty day readmit. The first node includes patients with thirty day readmit equals to zero and the second node includes patients with thirty day readmit not equals to zero. The node containing patients with thirty day readmit not equals to zero is further spitted into two new nodes based on department discharged. The first node includes patients discharged from Ear, nose and throat, Obstetric and Gyneacology and Specialty departments and the second node includes patients discharged from other departments.

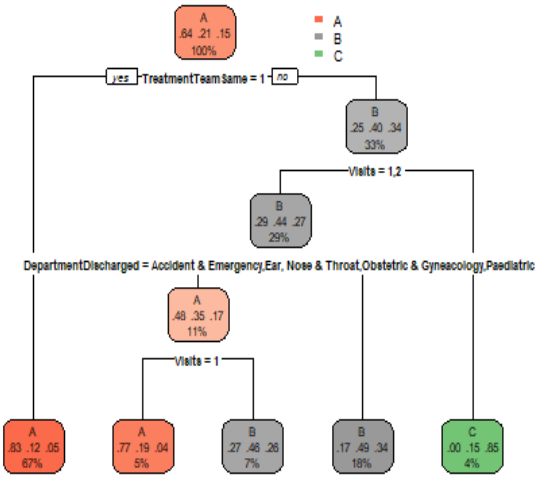

Fig. 3: A tree created with the best model using decision tree for classifying patient LOS Class.

Testing Performance of the best identified Models: In this section, the models that were identified to be the best performers in predicting patient LOS, and classifying patient LOS class. The goal is to assess the performance of each identified model on testing dataset in addition to the training dataset. To assess performance of the models on the test dataset, first, all the best performing models identified for continuous LOS were applied to the testing dataset. Later, the performance of the models that were found to be the best in classifying patient LOS were tested on the testing dataset. The R-Squared values and Classification Rates for the performance of training and testing datasets are presented in the table 4:

Table 4: Models performance on training and testing dataset.

Model	Technique	R-Square Training	R-Square Testing
When Response variable is continues LOS	Decision Tree	0.243685	**0.182250**
	Boosted Tree	0.246215	0.159154
	Random Forest	**0.355797**	0.091319
	Linear Regression	0.195711	0.192136
When Response variable is categorical LOS	Decision Tree	0.705288	0.665654
	Random Forest	**0.708800**	**0.723200**

Table 4 shows that random forests and decision tree appears to be the top performers when the objective was to predict patient LOS using the patient attributes known at the time of patient admit, as they have highest R-square values for training and testing datasets respectively than those for the other technique. Similarly, shows that random forests appears to be the top performers both when the objective was to classify patient LOS using the patient attributes known at the time of patient admit.

Table 5: Confusion Matrix for the best identified model

Prediction	Reference		
	A	B	C
A	958	71	27
B	23	222	54
C	3	32	171

Inspecting the confusion matrix above, in the testing data set, 958 patients that belong to class A are correctly classified in the LOS class A, 23 and 3 are incorrectly classified in the class B and C respectively. Also, 222 patients that belong to class B are correctly classified in the class B, 71 and 32 are incorrectly

classified in the class A and C respectively. Lastly, 171 patients that belong to LOS class C are correctly classified in the LOS class C, 27 and 54 are incorrectly classified in the LOS class A and B respectively.

Conclusion: Patients, hospitals and insurance providers will be benefited from this study, as predicted LOS and classified LOS class can help hospitals to identify patients with possible early and late discharges. Hospitals can perform additional tests to confirm whether the identified patients are fit to be discharged. This will reduce the chances of readmissions and late discharges. As a result, patient will not have to stay longer than required, hospitals can increase their throughput and, insurance providers can save money.

REFERENCES

Breiman, L (2001). Random Forests. *Mach. Learn.* 45(1), 5-32. Doi: 10.1023/A: 1010933404324

Bruser, C; Diesel, J; Zink, MD; Winter, S; Schauerte, P; Leonhardt, S (2013). Automatic Detection of Atrial Fibrillation in Cardiac Vibration Signals. *IEEE J. Biomed. Health Inform.* 17(1), 162-171.

De'ath, G (2007). Boosted Trees for Ecological Modeling and Prediction. Ecology, 88(1), 243-251.

Goantiya, R (2018) "Tree Based Modeling Techniques Applied to Hospital Length of Stay". Thesis. Rochester Institute of Technology.

Gulumbe, SU; Suleiman S; Badamasi, S; Tambuyal, AY; Usman, U (2019) Predicting Diabetes Mellitus Using Artificial Neural Network Through a Simulation Study. *Mach. Learn. Res.* 4(2): 33-38.

Hassan, M; Tuckman, HP; Patrick, RH; Kountz, DS; Kohn, JL (2010). Hospital Length of Stay and Probability of Acquiring Infection. *Intern. J. Pharm. Healthcare Mark.* 4(4), 324-338.

Hastie, T; Tibshirani, R; Friedman, J (2009). Elements of Statistical Learning: Data Mining, Inference, and Prediction (2 ed.). Springer, Stanford, California, 605-622.

Husain, W; Xin, LK; Rashid, NA; Jothi, N (2016). Predicting Generalized Anxiety Disorder Among Women Using Random Forest Approach. *Intern. Conf. Comp. Inform. Sci. (ICCOINS),* 3, 37-42. doi:10.1109/ICCOINS.2016.7783185

Jencks, S; Williams, M; Coleman, EA (2009). Rehospitalizations Among Patients in the Medicare Fee-for-Service Program. *N Engl J Med.*, 1418–1428.

Koza, JR; Bennett, FH; Andre, D; Keane, MA (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: Gero, JS; Sudweeks, F (eds) Artificial Intelligence in Design '96. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-0279-4_9

Li, ST; Chiu, NC; Kung, WC; Chen, JC (2013). Factors Affecting Length of Stay in the Pediatric Emergency Department. *Ped. & Neon.* 54(3), 179–187.

Loh, WY (2011). Classification and Regression Trees. *WIREs Data Min. Knowledge Disc.* 1(1), 14-23.

McDermott, C; Stock, G (2007). Hospital Operations and Length of Stay Performance. *Intern. J. Oper. Prod. Manage.* 27(9), 1020-1042. doi:10.1108/01443570710775847

Rezaei Hachesu, P; Ahmadi, M; Alizadeh, S; Sadoughi, F (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *HealthcInf Res.* 19(2):121–129.

Schapire, RE; Freund, Y (2012). Boosting: Foundations and Algorithms. *MIT Press.*

Speybroeck, N (2012). Classification and Regression Trees. *Intern. J. Pub. Health,* 57(1), 243–246.

Suleiman, S; Ibrahim, A; Usman, D; Isah, BY; Usman, HM (2021). g Credit Scoring Classification Performance using Self Organizing Map-Based Machine Learning Techniques, *Europ. J. Advances Eng. Technol.* 8(10), 28-35

Suleiman, S; Lawal, A; Usman, U; Gulumbe, SU; Muhammad, AB (2019). Student's Academic Performance Prediction Using Factor Analysis Based Neural Network. International Journal of Data Science and Analysis, 5(4): 61-66.10.11648/j.ijdsa.20190504.12

Suleiman, S; Usman, U (2016) prediction of Customer Accessibility of Electronic Banking logistic regression in Nigeria, *Equity J. Sci. Technol.* 4(1): 93 97

Thomas, JW; Kenneth, EG; Horvat, GG (1997). Is Patient Length of Stay related to Quality of Care? *Hosp. Health Serv. Admin.* 42(4), 489-507.