LENGTH OF STAY (LOS) PREDICTION OF TYPE 2 DIABETES MELLITUS USING CLASSIFICATION AND REGRESSION TREE (CART)

by Iik Bhakti Wiyata Kediri Perpustakaan 1

Submission date: 15-Jun-2024 09:20AM (UTC+0700)
Submission ID: 2212783030
File name: ilovepdf_merged_1.pdf (571.02K)
Word count: 3464
Character count: 17213

LENGTH OF STAY (LOS) PREDICTION OF TYPE 2 DIABETES MELLITUS USING CLASSIFICATION AND REGRESSION TREE (CART)

Eva Firdayanti Bisono[1], Jerhi Wahyu Fernanda[2], Ratna Frenty Nurkhalim[3], Krisnita Dwi Jayanti[4]

[1,3,4] *Institut Ilmu Kesehatan Bhakti Wiyata Kediri*

[2] *Insitut Agama Islam Negeri Kediri*

[1]eva.firdayanti@iik.ac.id; [2]fernanda.jerhi@iainkediri.ac.id; [3]ratna.nurkhalim@iik.ac.id; [4]krisnita.jayanti@iik.ac.id

Abstract

The prediction of LOS in type 2 patients and the influencing factors can be used as a basis for managing comorbidities and the risk of complications in patients. Predictions can be made using machine learning methods such as Classification and Regression Tree (CART). This study aims to analyse the factors that influence the LOS of type 2 DM patients. The research data was obtained from the Hospital Information System in the period 2019 to 2021 and obtained data for 541 type 2 DM patients. The study variables consisted of the dependent variable, namely LOS of DM patients type 2 and the independent variables consisted of gender, age, complications, comorbidities and urban status of type 2 DM patients. The average LOS of type 2 DM patients was 3.39 days with a median of 3 days. The results of the analysis using CART with 10-fold cross validation concluded that the morbidity variable was the variable that most dominantly influenced the LOS of type 2 DM patients. Accuracy, precision, recall, and F1 scores were respectively 0.704, 0.814, and 0.755.

Keywords: Classification and Regression Tree, Length of Stay, Prediction, Type 2 Diabetes Mellitus

I. INTRODUCTION

In the digitalization of the health services era, providing a good service in a hospital is highly prioritized because it improves the quality of health services. A patient certainly hopes for fast and appropriate services. Shorter waiting times for the patients will increase patient satisfaction. Therefore, skilled professionals are needed. One indicator used for the standard of professionalism of health services and information is indicators in health statistics such as Length of Stay (LOS)[1].

Length of Stay (LOS) is defined as the total number of days a patient has been in the hospital from being registered as an inpatient until completion of care or discharge from the hospital[2]. LOS is an indicator in health services, which can have an impact on hospital financing and patient satisfaction[3]. LOS predictions can also be used as a basis for planning and managing hospital resources to make them more effective and efficient[4]. Analysis of patient LOS factors in hospitals will have an impact on inpatient management more efficiently[5]. In some cases of chronic diseases such as Type 2 Diabetes Mellitus, LOS has a significant influence on the real costs of patient care[6].

Type 2 Diabetes Mellitus is a chronic disease whose prevalence rate always increases every year. Type 2 Diabetes Mellitus in the International Classification of Diseases (ICD) 9 is coded with code E11. Indonesian Basic Health Research results in 2018 showed that the prevalence of Diabetes Mellitus was 2.0%. This prevalence value has increased compared to 2013 with a prevalence of 1.5%[7]. This condition will indirectly increase the burden of care costs, especially for BPJS health because this disease needs large costs[8]. Steps that can be taken are to carry out accurate patient identification. To be sure, we must find out some factors that can change LOS with type DM. This can make early warning so that patients get more intensive care so that they can accelerate their recovery and improve other aspects that have an impact on more efficient costs. LOS can also be used to detect the risk of complications and comorbidities in patients with type 2 DM [9].

CART is a machine learning method with supervised learning. CART can be used to predict LOS of Type 2 DM patients by generating a rule based on existing data or information to clarify the relationship between input and output variables [10]. CART also visualizes in the form of a tree diagram that provides an overview of input variables starting with the most important variables used to predict an output. The CART study conducted by Williams, D., et. al used this method for the classification of Parkinson's Disease. Fernanda, J., W. also used CART to analyze risk factors for hypertension[11], [12]. The research results from both studies give a representation that the CART method is very flexible to use in classification because it is easy to visualize so it can be easily understood to explain the predictor variables. According to Eskandari, M., et, al, we can use the decision tree method to detect what factors can determine LOS in patients suffering from type 2 DM. Therefore, this study aims to determine the early LOS in patients who experience symptoms of type 2 DM according to some of the factors above.[13].

II. METHOD

The research data is secondary data. It obtained from one of the Private Hospital in Kediri. Data imported from the Hospital Information Management System (HIMS) starting from 2019 to 2021. This research variables consist of dependent (target) variable namely the Length of Stay (LOS) variable for type 2 DM patients. The independent variable consist of 5 variables namely gender, age, complication status, comorbidities, and patient urban status.

TABLE I
DESCRIPTION OF RESEARCH VARIABLES CATEGORIES

Variables	Categories
Length of Stay (LOS)	<=median > median
Gender	Male Female
Age	<=*45* years > *45* years
Complications	Yes No
Comorbidities	Yes No
Urban Status	Yes No

The research data obtained from HIMS is raw data. The data consists of 33 variables in HIMS and needs preprocessing data for carrying data as in Table 1 above. Preprocessing data is very important things because not all imported variables are needed for analysis and there are several variables such as complication status, comorbidity, and urban status are variables from preprocessing another variable. According to Sun, W., et. al. is very necessary to obtain accurate data analysis results[14]. Hassler, A. P., et. al. also stated that data preprocessing is an important stage in the data analysis process in healthcare[15].

The flow of data analysis in this research is divided into two steps, Step 1 is data preprocessing and Step 2 is modelling using the Classification and Regression Tree (CART) method. The preprocessing steps consist of several processes to obtain the variables used in CART modelling. The flow of data analysis can be seen in Fig. 1.

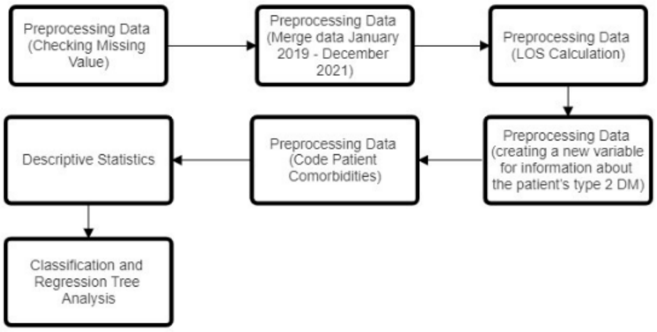
Preprocessing Data (Checking Missing Value)
Preprocessing Data (Merge data January 2019 - December 2021)
Preprocessing Data (LOS Calculation)
Preprocessing Data (creating a new variable for information about the patient's type 2 DM)
Preprocessing Data (Code Patient Comorbidities)
Descriptive Statistics
Classification and Regression Tree Analysis

Fig. 1 Flowchart of Data Analysis

A comprehensive explanation of these steps is as follows:

1. Preprocessing Data

Combining data is the first process that must be done. The data originating from HIMS in the form of Excel and separated every month from January 2019 to December 2021 period. The steps in this stage are as follows:

a. Initial data processing begins by checking missing values. Missing values are checked based on the primary diagnosis variable. At this stage ensure that the patient's primary diagnosis data is filled in and there are no blanks. After ensuring data does not have missing values, the next process is to check the type of variables in the study, especially the variables of patient entry date, patient discharge date, and patient date of birth. There are three variables that are still in the text form so the transformation process to become a date type is carried out. The date variable type for the patient's discharge date and admission date is very necessary for calculating patient LOS.

b. The next stage is to merge data from January 2019 to December 2021 into one dataset. The data has become a dataset, and the research variables are selected before. The research variables are date admissions, the patient's birth date, gender, primary diagnosis, secondary diagnoses (dtd2 and dtd3), the name of the sub-district, and the name of the patient district. After that, a filter is carried out, especially on dtd1, to filter type 2 DM patients only. Since 2019 to 2021, there were 541 patients suffering from type 2 DM. One of the ways that must be done to find out which patients have this is the diagnosis of E11

c. The LOS variable is used to calculate the patient's admission and discharge dates.. After that, the LOS variable was coded into two categories, namely <=median and >median.

d. The next step is creating a new variable for information about the patient's type 2 DM, whether only type 2 DM without complications or with complications. E11.9 is Type 2 DM without complications, while type 2 DM with codes E11.1 to E11.8 is DM with complications. Codes E11.1 to E11.8 are used to diagnose patients with type 2 DM. Patients who have these codes have complications (1), but if the code is E11.9 then the patient has secondary diagnoses and does not have complications.

e. To find out which patients have comorbidities by looking at the comorbidity variable and patients who have comorbidities or no comorbidities are coded 0.. The urban variable is a variable obtained from processing the patient's area of origin. Patients come from Kediri district and Pare sub-district, so these patients are categorized as urban, apart from that, patients are categorized as non-urban.

2. Descriptive Statistics

This stage of analysis was carried out to see the description of the variables in the research. Histograms are used to explain LOS data for type 2 DM patients during the period 2019 to 2021. Frequency Distribution Tables are also used to explain the variables of gender, age, complications, morbidity, and urban status

3. Classification and Regression Tree (CART) Analysis

Classification and Regression Tree (CART) is a statistical method used in modelling cases in health, industry, and other fields. CART is a representative of the relationship between the dependent variable (target) and the independent variable which is visualized through a tree diagram, where this tree diagram makes it easy to interpret the results of data analysis [16]. CART consists of nodes or branches and leaves. The top node or root is the variable that has the greatest level of influence. After the node will appear a leaf which represents the class/category of the target variable. In the process of forming this node, a split process is carried out using the Gini Index algorithm. The Gini index algorithm is as follows[17].

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

m is the number of classes. In the CART method, classes are divided into two (binary).
p_i is the probability of sample I to be included in class group *m*.

In this research, the CART model was validated using the cross-validation (CV) method. CV is functioned to test the model's ability to predict new data that is not visible and is not used in model construction. CV is a method that can be used to obtain optimal machine learning models[18]. This research used the 10-fold cross-validation technique with this following stages below:

a. Data divided into training data and testing data with proportion number 80% and 20 %

b. The training data is 80% (432 data), randomized into 10 new datasets that will be used in CART modeling with each dataset also divided into 800% training data and 20% testing data.

c. CART modelling was carried out on each dataset starting from dataset 1 to 10 and the error level of each model was recorded

d. Take the average score of 10 CART models from 10 datasets to get a CART model.

4. CART Model Accuracy

After obtaining CART model, prediction is done using the testing data. The prediction results of testing data are presented using the Confusion Matrix Table and based on this table, it is used to measure the goodness of the CART Model [19]. The measures used are accuracy, recall, precision, and F1 Scores [20]. These measurements are also conducted by research [21] to see the performance or accuracy of the machine learning model.

III. RESULT AND DISCUSSION

Type 2 DM patient's data from 2019 to 2021 is 541 patients. The concentration and distribution of data on the Length of Stay (LOS) of type 2 DM patients is presented in the histogram in Figure 2. The histogram provides information on the LOS of many types 2 DM patients which is 3-4 days. There are also LOS for type 2 DM patients who have a LOS more than 10 days.

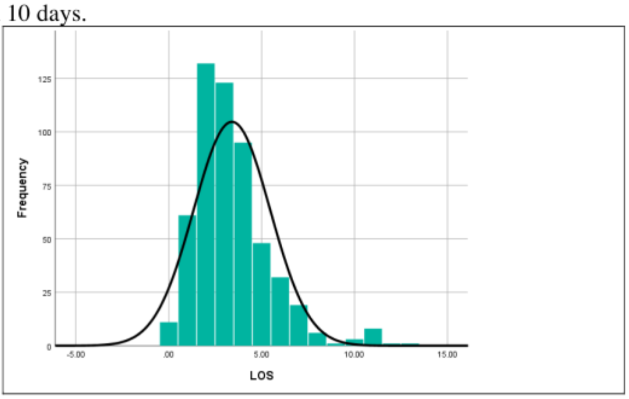

Fig. 2 Patients Type 2 DM LOS Histogram

Descriptive Statistics of LOS for Type 2 DM patients can be seen in Table 2. Patients' LOS average for type 2 DM is 3.39 days with median value of 3 days. The standard deviation of LOS Type 2 DM patients is 2.06. The median value will be used as the basis for categorizing LOS patients which used in the analysis using CART method. In the analysis using CART, LOS data will be categorized into <=3 days and >3 days.

TABLE II

DESCRIPTIVE STATISTICS PATIENTS LOS DM TYPE 2

Variable	Mean	Median	Standard Deviation
Length of Stay (LOS)	3.39	3	2.06

Respondent characteristics based on independent variables such as gender, age, comorbidities, complications, and urban area are presented in Table 3 which explains the frequency distribution of each variable. The majority case of Type 2 DM patients in this research where female with percentage is 62.9% and 37.2% were male.

TABLE III

FREQUENCY DISTRIBUTION OF PREDICTOR VARIABLES

Variable	Category	Frequency	Percentage
Gender	Female	340	62.8
	Male	201	37.2
Age	<=45	49	9.1
	>45	492	90.9
Comorbidities	Yes	121	22.4
	No	420	77.6
Complications	Yes	211	39
	No	330	61
Urban Status	Yes	105	19.4
	No	436	80.6

The majority, of type 2 DM patients are over 45 years old with a 90.05 percent. In this research, were also found patients suffering from type 2 DM aged <= 45 years with a percentage of 9.1%. There were 121 patients (22.4%) who had comorbidities. Of 541 type 2 DM patients, 211 patients (39%) have complications. Patients are categorized into urban and non-urban based on the area where they live. There were 105 patients with type 2 DM who came from urban areas (19.4%) and 436 patients who did not come from urban areas (80.6%).

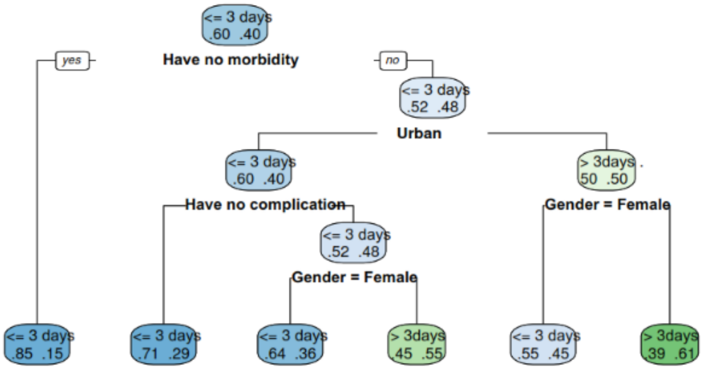

Fig. 3 Classification and Regression Tree (CART) Structure of Type 2 DM Patients

Fig. 3 is the CART Structure produced to predict LOS for type 2 DM patients. The morbidity variable is the first node in the CART structure. This condition indicates that the patient's morbidity variable is the most influential variable on the LOS of Type 2 DM patients. Patients who do not have comorbidities will have a great chance of having a LOS of less than or equal to 3 days. If the patient has comorbidities, then to predict LOS for type 2 DM patients, variables such as urban area, complications, and gender need to be taken into account.

Patients who do not have comorbidities, come from urban areas, and do not have complications experienced, have a high chance of having a LOS <= 3 days. Meanwhile, patients who do not have comorbidities, are not from urban areas, and are female will have a chance of having a LOS of <= 3 days, and if they are male, they will have a chance of having a LOS > 3 days.

TABLE III
CONFUSION MATRIX PREDICTING TESTING DATA

Prediction	Actual	
	<=3 days	**> 3 days**
< 3 days	57	24
> 3 days	13	14

The confusion matrix table in table III explains that there are 24 data which in the real data are LOS <= 3 days, predicted using the CART method to be > 3 days. Meanwhile, there were 13 actual data on patient LOS > 3 days which were predicted by the CART model to be <= 3 days. The confusion matrix table is the basis for calculating accuracy and F1 score values from the CART model.

TABLE IV
CART MODEL PERFORMANCE

Parameter	**F1 Score**
Accuracy	0.657
Precision	0.704
Recall	0.814
F1 Score	0.755

The accuracy value in this study was 0.6574 or 65.74% due to the complex characteristics of the data with various variables. Meanwhile, the precision, recall and F1 Score values are 0.704, 0.814 and 0.755.

IV. CONCLUSION

The results of the analysis using the CART method provide the conclusion that the variables morbidity, urbanity, complications, and gender are variables that can be used to predict LOS for type 2 DM patients. The morbidity variable is the variable with the highest level of importance compared to other variables. The CART model has an accuracy of 0.6574 (65.74%), and an F1 Score of 0.76.

ACKNOWLEDGEMENT

We thank to Ministry of Research and Higher Education, RS HVA Kediri, IIK Bhakti Wiyata Kediri, and IAIN Kediri who have help this research and funders for the research carried out or to people who are truly instrumental in carrying out the research.

REFERENCES

[1] R. Rosita and A. R. Tanastasya, “Penetapan mutu rumah sakit berdasarkan indikator rawat inap,” *J. Kesehat. Kusuma Husada*, pp. 166–178, 2019.

[2] M. J. Bowie, *Essentials of health information management: Principles and practices*. Cengage Learning, 2022.

[3] J. Chrusciel, F. Girardon, L. Roquette, D. Laplanche, A. Duclos, and S. Sanchez, “The prediction of hospital length of stay using unstructured data,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 351, 2021.

[4] K. Stone, R. Zwiggelaar, P. Jones, and N. Mac Parthaláin, “A systematic review of the prediction of hospital length of stay: Towards a unified framework,” *PLOS Digit. Heal.*, vol. 1, no. 4, p. e0000017, 2022.

[5] H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, and S. Yoo, “Analysis of length of hospital stay using electronic health records: A statistical and data mining approach,” *PLoS One*, vol. 13, no. 4, p. e0195901, 2018.

[6] I. Oktadiana, “Perbandingan Biaya Riil Pada Pasien Diabetes Mellitus Tipe 2 Dengan Tarif INA-CBG’S Di Rumah Sakit Umum Daerah,” *J. Farm. Tinctura*, vol. 2, no. 2, pp. 42–51, 2021.

[7] I. Irwansyah and I. S. Kasim, “Indentifikasi keterkaitan lifestyle dengan risiko diabetes melitus,” *J. Ilm. Kesehat. Sandi Husada*, vol. 10, no. 1, pp. 62–69, 2021.

[8] A. E. Pratiwi and H. Sukmawati, “ANALISIS BIAYA RATA-RATA PASIEN RAWAT INAP DENGAN PENYAKIT DIABETES MELLITUS TYPE II (STUDI DI JEMBRANA DAN GIANYAR),” *WICAKSANA J. Lingkung. dan Pembang.*, vol. 3, no. 2, pp. 21–29, 2019.

[9] B. Katipoglu, M. I. Naharci, and E. S. Yurdakul, “Risk factors predicting hospital length of stay in older patients with type 2 diabetes with Covid-19,” *J. Diabetes Metab. Disord.*, vol. 21, no. 2, pp. 1443–1449, 2022.

[10] M. Ozcan and S. Peker, “A classification and regression tree algorithm for heart disease modeling and prediction,” *Healthc. Anal.*, vol. 3, p. 100130, 2023.

[11] D. Williams *et al.*, “Visualization of Decision Tree State for the Classification of Parkinson’s Disease,” *J. Biomed. Eng. Med. Imaging*, vol. 3, no. 3, pp. 25–41, 2016.

[12] J. W. Fernanda, G. Anuraga, and M. A. Fahmi, “Risk factor analysis of hypertension with logistic regression and Classification and Regression Tree (CART),” in *Journal of Physics: Conference Series*, 2019, vol. 1217, no. 1, p. 12109.

[13] M. Eskandari, A. H. Alizadeh Bahmani, H. A. Mardani-Fard, I. Karimzadeh, N. Omidifar, and P. Peymani, “Evaluation of factors that influenced the length of hospital stay using data mining techniques,” *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–11, 2022.

[14] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, “Data processing and text mining technologies on electronic medical records: a review,” *J. Healthc. Eng.*, vol. 2018, 2018.

[15] A. P. Hassler, E. Menasalvas, F. J. García-García, L. Rodríguez-Mañas, and A. Holzinger, “Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome,” *BMC Med. Inform. Decis. Mak.*, vol. 19, pp. 1–17, 2019.

[16] C. Machuca, M. V Vettore, M. Krasuska, S. R. Baker, and P. G. Robinson, “Using classification and regression tree modelling to investigate response shift patterns in dentine hypersensitivity,” *BMC Med. Res. Methodol.*, vol. 17, pp. 1–11, 2017.

[17] T. Daniya, M. Geetha, and K. S. Kumar, “Classification and regression trees with gini index,” *Adv. Math. Sci. J.*, vol. 9, no. 10, pp. 8237–8247, 2020.

[18] I. K. Nti, O. Nyarko-Boateng, and J. Aning, “Performance of machine learning algorithms with different K values in K-fold cross-validation,” *J. Inf. Technol. Comput. Sci*, vol. 6, pp. 61–71, 2021.

[19] E. S. Kresnawati, Y. Resti, B. Suprihatin, M. R. Kurniawan, and W. A. Amanda, “Coronary Artery Disease Prediction Using Decision Trees and Multinomial NaÃ¯ ve Bayes with k-Fold Cross Validation,” *Inomatika*, vol. 3, no. 2, pp. 172–187, 2021.

[20] R. M. AlZoman and M. J. F. Alenazi, “A comparative study of traffic classification techniques for smart city networks,” *Sensors*, vol. 21, no. 14, p. 4677, 2021.

[21] A. PINAR, C. Çolak, and E. Gültürk, “Evaluation of Performance Metrics in Heart Disease by Machine Learning Techniques,” *J. Cogn. Syst.*, vol. 8, no. 1, pp. 11–15.

LENGTH OF STAY (LOS) PREDICTION OF TYPE 2 DIABETES MELLITUS USING CLASSIFICATION AND REGRESSION TREE (CART)

ORIGINALITY REPORT

9%	4%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Nurul Laily Qomariyah, Thinni Nurul Rochmah, Rizka Rosa DM. "Cost Of Illnes As A Basic For Advocacy Efforts to Prevent An Increase In The Prevalence of Type 2 Diabetes Mellitus", Jurnal Aisyah : Jurnal Ilmu Kesehatan, 2023 Publication	1%
2	Submitted to Universitas Amikom Student Paper	1%
3	kupdf.net Internet Source	1%
4	Eva Royandriani Rustamaji, Siti Zaetun, I Wayan Getas, Rohmi Rohmi, Musayadah Musayadah. "Differences Albumin Value in Type II Diabetes Mellitus Patient and Type II Diabetes Mellitus With Nefropathi Diabetic Coinfection in Sumbawa Hospital", Jurnal Analis Medika Biosains (JAMBS), 2023 Publication	1%

5	Iw Campbell. "Epidemiology and Clinical Presentation of Type 2 Diabetes", Value in Health, 2000 Publication	1 %
6	Submitted to London School of Economics and Political Science Student Paper	1 %
7	Muhammad Hilman Fadly, Monika Evelin Johan. "Web-Based Heart Disease Prediction by Comparison and Implementation of SVM, AdaBoost, and Hybrid SVM-AdaBoost Algorithms", 2023 7th International Conference on New Media Studies (CONMEDIA), 2023 Publication	1 %
8	www.fda.gov Internet Source	<1 %
9	Aagam Shah, Joshua A. Schiller, Isiah Ramos, James Serrano, Darren K. Adams, Sameh Tawfick, Elif Ertekin. "Automated image segmentation of scanning electron microscopy images of graphene using U-Net Neural Network", Materials Today Communications, 2023 Publication	<1 %
10	Submitted to University of Lancaster Student Paper	<1 %

11	mecs-press.org Internet Source	<1 %
12	www.mdpi.com Internet Source	<1 %
13	Siyang Xue, Xinke Shen, Dan Zhang, Zhenhua Sang, Qiting Long, Sen Song, Jian Wu. "Unveiling Frequency-Specific Microstate Correlates of Anxiety and Depression Symptoms", Cold Spring Harbor Laboratory, 2024 Publication	<1 %
14	bmcmedinformdecismak.biomedcentral.com Internet Source	<1 %
15	Cherubin Mugisha, Incheon Paik. "Comparison of Neural Language Modeling Pipelines for Outcome Prediction From Unstructured Medical Text Notes", IEEE Access, 2022 Publication	<1 %
16	Randy L. Maddalena, Thomas E. McKone, Michael D. Sohn. "Standardized Approach for Developing Probabilistic Exposure Factor Distributions", Risk Analysis, 2004 Publication	<1 %
17	Sayma Alam Suha, Tahsina Farah Sanam. "A Machine Learning Approach for Predicting Patient's Length of Hospital Stay with	<1 %

Random Forest Regression", 2022 IEEE Region 10 Symposium (TENSYMP), 2022
Publication

18 pure.rug.nl
Internet Source
<1 %

19 www.grafiati.com
Internet Source
<1 %

20 Mohammed Gollapalli, Latifa Alabdullatif, Farah Alsuwayeh, Moodhi Aljouali, Alhanoof Alhunief, Zaina Batook. "Text Mining on Hospital Stay Durations and Management of Sickle Cell Disease Patients", 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), 2022
Publication
<1 %

21 Sumanta Das, Jack Christopher, Armando Apan, Malini Roy Choudhury, Scott Chapman, Neal W. Menzies, Yash P. Dang. "Evaluation of water status of wheat genotypes to aid prediction of yield on sodic soils using UAV-thermal imaging and machine learning", Agricultural and Forest Meteorology, 2021
Publication
<1 %

Exclude quotes	On	Exclude matches	Off

Exclude bibliography On