Supply Chain Analysis in the Health Sector Using Gradient Boosting Regression Algorithm

Bayu Angga Wijaya[a], Nestina Halawa[b]
[a,b]Department of Science and Technology, Universitas Prima Indonesia
Corresponding Author:
[a]bayuanggawijaya@unprimdn.ac.id

ABSTRACT
Supply chain analysis in healthcare is a crucial aspect in ensuring efficient and optimized resource distribution. This study uses the Gradient Boosting Regression algorithm to predict demand in healthcare supply chains to improve the accuracy of stock planning and management trained using supply datasets from hospitals. The model evaluation results show that most of the predictions are close to the actual values, as seen from the points clustered around the reference line. Despite the slight deviations, the Mean Absolute Error (MAE) value of 157.16 indicates that the average prediction error is relatively small compared to the demand scale which ranges from 0 to 14,000. This indicates that the Gradient Boosting Regression model performs reasonably well in estimating supply chain demand in the healthcare sector. Thus, this approach has the potential to be used in more accurate decision-making, in order to improve the efficiency of distribution and availability of health resources.

Keywords : Gradient Boosting Regression, Healthcare Supply Chain, Demand Prediction, Stock Management.

INTRODUCTION

The supply chain in healthcare plays a crucial role in ensuring the availability of drugs, medical devices, and other medical resources to optimally meet the needs of patients (Elektronik, 2024). Efficiency in this supply chain is critical to avoid shortages or overstocks that can impact the quality of healthcare services (Chukwu, 2020). In some cases, inaccuracies in demand forecasting can lead to distribution disruptions, unnecessary inventory buildup, or even an urgent availability crisis. Therefore, there is a need for methods that can improve the accuracy of demand forecasting to ensure more effective distribution of resources.

In addition to challenges in stock management, healthcare supply chains also face constraints in response to fluctuating demand. Factors such as pandemics, natural disasters or population increases can cause sudden spikes in demand. If the distribution system is not able to adjust quickly, shortages can occur which can have a serious impact on health services (Villarreal et al., 2023). Therefore, the use of data-driven methods and artificial intelligence technology can help in making faster and more precise decisions.

Artificial intelligence (AI) and machine learning technologies have grown rapidly in helping various sectors, including healthcare. One method that has proven effective in forecasting is Gradient Boosting Regression, which is a decision tree-based technique to improve prediction accuracy (KANTINIT, 2023). By using this method, it is expected that a more accurate prediction model can be obtained so that it can assist decision making in supply chain management in the healthcare sector. The algorithm is able to identify patterns in historical data as well as consider influential factors in healthcare demand (Ummah, 2019).

The application of Gradient Boosting Regression algorithm in health supply chain analysis provides various advantages, such as improved prediction accuracy and efficiency in stock management (Fadmadika et al., 2024; Setyawan & Wakhidah, 2025). With this approach, supply chain managers can more easily adjust inventory quantities to changing demand levels. In addition, this model can help in reducing resource wastage by optimizing the distribution of goods based on more precisely predicted demand (Jadhav & Deshmukh, 2022). Therefore, the use of this method has the potential to improve efficiency in the overall healthcare supply chain.

In today's digital era, the use of data in decision-making has become very important, especially in the healthcare sector. By utilizing historical data and appropriate modeling techniques, healthcare institutions can improve accuracy in managing the supply and distribution of medical devices and drugs. By doing so, hospitals, clinics, and pharmacies can minimize the risk of out-of-stocks or excess inventory that could result in financial losses (Spieske et al., 2022)(Arji et al., 2023).

In addition, the role of governments and policy makers in the implementation of analytics technology in the healthcare supply chain is also very important. Regulations that support the implementation of these technologies can drive greater efficiency and improve the overall quality of healthcare. Collaboration between the government, healthcare providers, and technology companies is needed to ensure that the use of algorithms such as Gradient Boosting Regression can be applied optimally and sustainably (Apriska et al., 2024; Carolina, 2022; Rahim et al., 2021).

Considering the importance of accurate prediction aspects in the healthcare supply chain, this study aims to implement Gradient Boosting Regression as a solution in improving the efficiency of stock management and distribution. By analyzing the model's performance in predicting demand, this research is expected to provide insights for supply chain managers in data-driven decision making. Thus, healthcare services can be more responsive in facing challenges that arise in the future.

METHODS

Research Stages

This research consists of several main stages starting from data collection, data processing, data exploration, modeling, and evaluation of the built model designed to build an accurate prediction model in the health supply chain using the Gradient Boosting Regression algorithm. Each stage has an important role in ensuring the success of model analysis and implementation in improving the efficiency of medical goods distribution. The following are the stages carried out in this research:

Figure 1. Flowchart of Research Method

Literature Study

The first stage in this research is to conduct a literature study to understand the concept of supply chains in the health sector as well as demand forecasting methods that have been used before. Supply chains in the healthcare sector have unique characteristics because they are related to the availability of medicines, medical devices, and medical materials that are crucial in patient care. The literature study will include various sources such as academic journals, industry reports, and previous research that has applied Machine Learning-based forecasting methods, including Gradient Boosting Regression. In addition, comparisons will be made with other methods such as Linear Regression, Random Forest, or Neural Networks to understand the advantages and disadvantages of each technique. This study aims to provide a strong theoretical foundation for the research as well as ensure that the methods used can address specific challenges in the healthcare supply chain.

Data Collection

The dataset used in this research is taken from the Supply Chain in Healthcare data obtained from Kaggel (https://www.kaggle.com/datasets/vanpatangan/hospital-supply-chain). The data is data that has a track record of supply chain management. This data is important to ensure that the prediction model built can represent the real conditions in the field. In collecting data, it is necessary to ensure that the information obtained is of high quality, does not contain many missing values, and reflects sufficient variation for the model to learn optimally.

Data Processing and Cleaning

The data that has been collected is often not ready for immediate use in the modeling process. Therefore, the data processing and cleaning stage is a crucial step in this research. At this stage, various techniques such as deletion of invalid data, imputation of missing values, normalization of data scales, as well as transformation of variables needed to fit the format that can be processed by the Gradient Boosting Regression algorithm are performed. In addition, exploratory analysis was conducted to understand the distribution of the data, detect outliers, and identify initial patterns that could aid in the construction of the prediction model. With clean and ready-to-use data, the prediction model can work more optimally and produce more accurate estimates in predicting the demand for medical goods in the healthcare supply chain.

	Date	Item_ID	Item_Type	Item_Name	Current_Stock	Min_Required	Max_Capacity	Unit_Cost	Avg_Usage_Per_Day	Restock_Lead_Time	Vendor_ID
0	2024-10-01	105	Consumable	Ventilator	1542	264	1018	4467.55	108	17	V001
1	2024-10-02	100	Equipment	Ventilator	2487	656	3556	5832.29	55	12	V001
2	2024-10-03	103	Equipment	Surgical Mask	2371	384	5562	16062.98	470	6	V001
3	2024-10-04	103	Consumable	Surgical Mask	2038	438	1131	744.10	207	15	V002
4	2024-10-05	107	Equipment	IV Drip	2410	338	1013	15426.53	158	12	V003

	column	dtype	instances	unique	sum_null	duplicates
0	Date	object	500	500	0	0
1	Item_ID	int64	500	10	0	0
2	Item_Type	object	500	2	0	0
3	Item_Name	object	500	5	0	0
4	Current_Stock	int64	500	475	0	0
5	Min_Required	int64	500	393	0	0
6	Max_Capacity	int64	500	480	0	0
7	Unit_Cost	float64	500	500	0	0
8	Avg_Usage_Per_Day	int64	500	317	0	0
9	Restock_Lead_Time	int64	500	29	0	0
10	Vendor_ID	object	500	3	0	0

Figure 1. Warehouse data

	Patient_ID	Admission_Date	Discharge_Date	Primary_Diagnosis	Procedure_Performed	Room_Type	Bed_Days	Supplies_Used	Equipment_Used	Staff_Needed
0	P001	2024-10-06 05:30:28	2024-10-23 01:11:34	Diabetes	Appendectomy	General Ward	2	Gloves, IV	Surgical Table	2 Surgeons
1	P002	2024-10-24 11:07:58	2024-10-15 05:16:54	Fracture	Appendectomy	ICU	10	Gown, IV	MRI Machine	1 Nurse
2	P003	2024-10-22 21:43:43	2024-10-24 10:56:30	Fracture	Chest X-ray	ICU	10	Gloves, IV	X-ray Machine	1 Nurse, 1 Doctor
3	P004	2024-10-05 17:04:05	2024-10-30 14:10:01	Diabetes	Chest X-ray	ICU	11	Gloves, IV	X-ray Machine	1 Nurse
4	P005	2024-10-21 17:04:00	2024-10-08 15:53:22	Appendicitis	MRI	ICU	2	Gloves, IV	X-ray Machine	2 Surgeons

	column	dtype	instances	unique	sum_null	duplicates
0	Patient_ID	object	500	500	0	0
1	Admission_Date	object	500	500	0	0
2	Discharge_Date	object	500	500	0	0
3	Primary_Diagnosis	object	500	4	0	0
4	Procedure_Performed	object	500	4	0	0
5	Room_Type	object	500	2	0	0
6	Bed_Days	int64	500	14	0	0
7	Supplies_Used	object	500	3	0	0
8	Equipment_Used	object	500	3	0	0
9	Staff_Needed	object	500	3	0	0

Figure 2. Patient Data

	Staff_ID	Staff_Type	Shift_Date	Shift_Start_Time	Shift_End_Time	Current_Assignment	Hours_Worked	Patients_Assigned	Overtime_Hours
0	S001	Surgeon	2024-10-22 04:44:49	06:00 PM	07:00 PM	ER	8	9	1
1	S002	Nurse	2024-10-03 05:51:36	08:00 AM	06:00 PM	General Ward	9	3	0
2	S003	Technician	2024-10-15 15:11:14	08:00 AM	06:00 PM	ER	8	6	4
3	S004	Surgeon	2024-10-09 20:07:58	07:00 AM	06:00 PM	General Ward	11	6	3
4	S005	Surgeon	2024-10-12 05:01:02	07:00 AM	06:00 AM	General Ward	9	5	3

	column	dtype	instances	unique	sum_null	duplicates
0	Staff_ID	object	500	500	0	0
1	Staff_Type	object	500	3	0	0
2	Shift_Date	object	500	500	0	0
3	Shift_Start_Time	object	500	3	0	0
4	Shift_End_Time	object	500	3	0	0
5	Current_Assignment	object	500	3	0	0
6	Hours_Worked	int64	500	4	0	0
7	Patients_Assigned	int64	500	9	0	0
8	Overtime_Hours	int64	500	5	0	0

Figure 3. Staff Data

	column	dtype	instances	unique	sum_null	duplicates
0	Date	object	500	500	0	0
1	Expense_Category	object	500	3	0	0
2	Amount	float64	500	500	0	0
3	Description	object	500	3	0	0

	Date	Expense_Category	Amount	Description
0	2024-10-01	Staffing	29391.86	Surgical masks
1	2024-10-02	Supplies	47757.71	Surgical masks
2	2024-10-03	Supplies	43996.60	Ventilators
3	2024-10-04	Supplies	27908.42	Surgeons' salaries
4	2024-10-05	Equipment	39719.60	Ventilators

Figure 4. Financial Data

Data preprocessing is a crucial step in data analysis that aims to prepare raw data for effective processing and analysis. Raw data often contains inconsistencies, errors, or formats that are not suitable to be directly processed by algorithms. Through preprocessing, data can be cleaned, transformed, and integrated to produce more accurate and relevant information.

Prediction Model Building

The core stage of this research is to build a prediction model using the Gradient Boosting Regression algorithm. This algorithm was chosen because it has the ability to combine the weaknesses of decision tree-based models through a boosting approach, thus improving prediction accuracy by minimizing errors. At this stage, the model will be trained using pre-processed historical data. In addition, an exploration of the model parameters (hyperparameter tuning) is conducted to find the best combination that results in optimal performance. Some of the parameters to be tested include the number of estimators (n_estimators), the learning rate, and the maximum depth of the decision tree (max_depth). By applying this algorithm, it is expected that the model can capture patterns in medical goods demand more accurately than conventional methods.

The existing data is then grouped into several categories. Data grouping where data from “total stock” is combined with “daily stock” into maximum space, data from “estimated stock-out time” and “daily stock” are combined into “average usage per day”, “minimum stock” and “daily stock” into “minimum needs”. After getting the results from the combination of existing data, then we create a target variable, namely “estimated demand” which is taken from a combination of average usage per day data combined with restock lead time.

Model Evaluation

Once the prediction model is built, the next step is to evaluate its performance using evaluation metrics such as Mean Absolute Error (MAE). This evaluation is important to understand the extent to which the model can accurately predict the demand for medical goods. MAE is used to measure the average absolute error in prediction. If the evaluation results show that the error rate is still high, then the model can be improved by making parameter adjustments or adding new features that can improve the prediction accuracy. Careful evaluation will ensure that the model used can really help in supply chain management in healthcare.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is one of the evaluation metrics used to measure the average absolute error between model predicted and actual values. MAE is calculated by taking the average of the absolute differences between the predicted and actual values. The MAE formula is mathematically expressed as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where:

$\boldsymbol{y_i}$ is the actual value,
$\boldsymbol{\hat{y}_i}$ is the predicted value, and
$\boldsymbol{n}$ is the number of data samples.

MAE gives an idea of how far the model prediction misses the actual value in the same unit as the data used. The smaller the MAE value, the more accurate the model is in making predictions. Conversely, the larger the MAE value, the higher the prediction error produced by the model.

In the context of “Supply Chain Analysis in the Health Sector Using Gradient Boosting Regression”, MAE becomes an important metric in evaluating the performance of medical goods demand prediction models. Prediction accuracy is crucial in the healthcare supply chain as errors in forecasting can have a significant impact on the availability of medicines and medical equipment. If the prediction model has a high MAE, then errors in demand estimation can lead to stockouts or overstock. Stock shortages can hamper healthcare services as medical personnel do not have enough resources to handle patients, while overstocks can lead to a waste of resources, especially for medical products that have a limited expiration period.

Using the Gradient Boosting Regression algorithm, this study aims to build a prediction model that is able to minimize MAE so that demand estimation becomes more accurate. Thus, the supply chain can be managed more efficiently, ensuring timely distribution of medical goods and avoiding wastage or shortages that are detrimental to the health system as a whole. Therefore, in this study, the MAE value will be used as the main indicator in measuring the success of the model in helping optimize stock management in the healthcare sector.

RESULTS

Analysis Result

After we understand the relationship between the variables, we create a model by putting Estimated Demand as the y factor and Inventory Date, Admission Date, Discharge Date, Shift Date, Financial Date as the x factor and train (80%) and test (20%). After that, we train into the GBR model to get the predicted value.

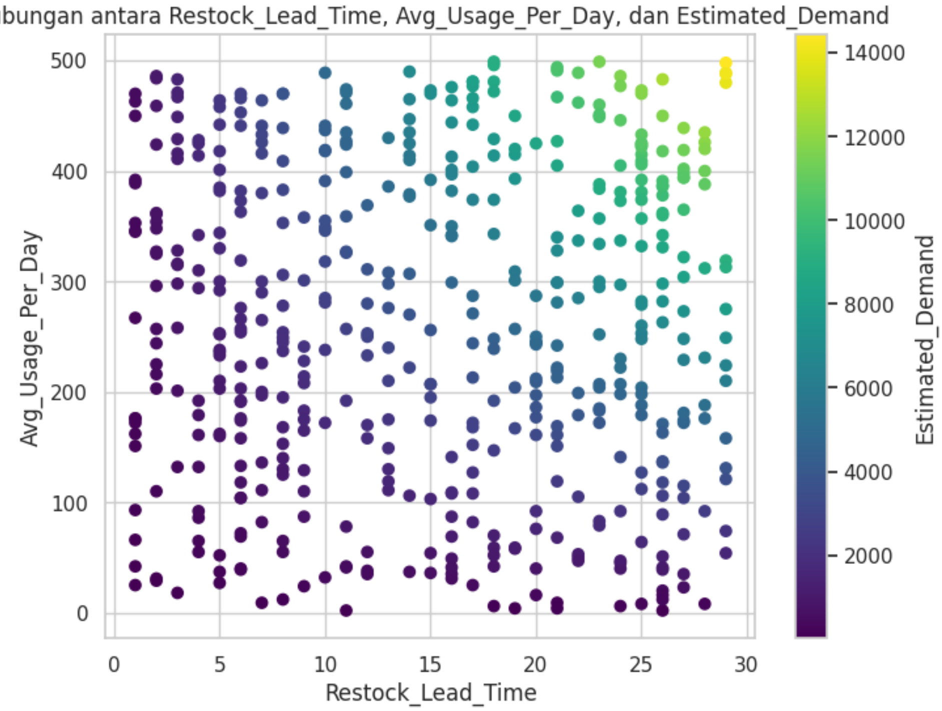

Figure 5. Correlation between data

After doing the GBR train, we look for the MAE value of the y_error and y_pred values and get a Mean Absolute Error value of 157.16. After that we make a visualization of the prediction value and the actual value, namely between Actual Demand and Predicted Demand as shown in Figure 7.

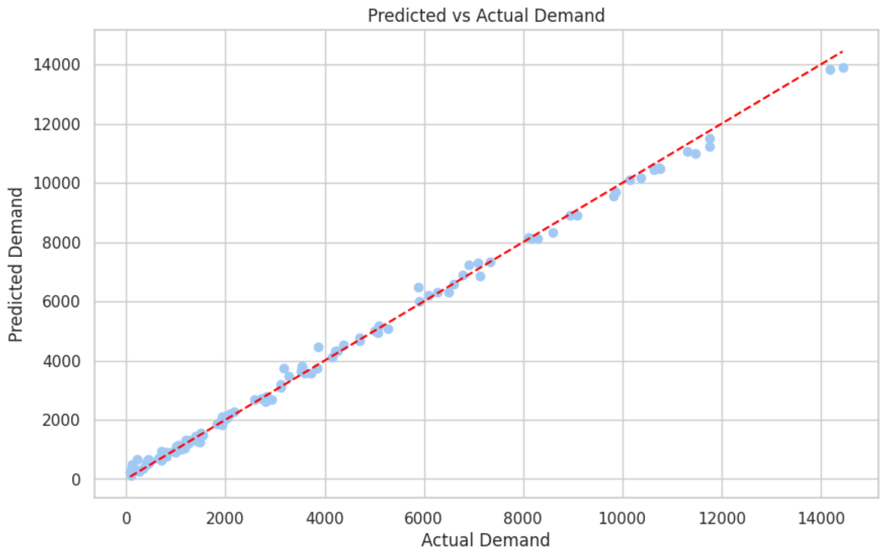

Figure 6. Predicted vs Actual Demand

DISCUSSION

In this research, a prediction model is created using the Estimated Demand variable as the target variable (y), while the independent variables (x) consist of Estimated Demand, Inventory Date, Admission Date, Discharge Date, Shift Date, and Financial Date. The modeling process begins by dividing the dataset into two parts, namely 80% of the data for training (training set) and 20% of the data for testing (test set). This division aims to ensure that the model can learn from the available historical data and test its ability to make predictions on data that has never been seen before.

After the data sharing process, the Gradient Boosting Regression (GBR) model was applied to train the model with the training data. GBR was chosen for its ability to handle complex data as well as its ability to capture non-linear patterns in the predictor variables. In the training stage, the model attempts to find patterns of relationships between the independent variables and the target variable to produce accurate predictions of future demand for medical goods.

Furthermore, after the GBR model is trained, testing is conducted using test data to measure the performance of the model. The evaluation was conducted by calculating the Mean Absolute Error (MAE), which is the average absolute difference between the actual value (Actual Demand) and the value predicted by the model (Predicted Demand). The evaluation results showed that the model produced an MAE value of 157.16, which means that the average prediction error of the model was around 150 units from the actual demand. The relatively small MAE value compared to the scale of the data indicates that the model performs quite well in predicting the demand for medical goods. When compared to the MAE value of the Random Forest algorithm which gets 147.09 and Linear Regression 150.01 as shown in Figure 10, the model built using the Gradient Boosting Regression algorithm is clearly superior.

Komparasi Model

	MAE
GB Regression	157,16
Random Forest	147,09
Linear Regression	150,01

Figure 8. GRB vs RF vs LR

To gain a deeper understanding of the model's performance, a visualization of the comparison between the actual and predicted values was conducted, as shown in Figure 7. From the visualization results, it can be seen that most of the data points are spread around the red line, which indicates that the model predictions are within a fairly accurate range. Although there are some deviations, in general the GBR model can capture the demand pattern well and provide predictions that are close to the actual values.

Based on the results of this discussion, it can be concluded that the application of Gradient Boosting Regression in health supply chain analysis provides fairly accurate results in predicting the demand for medical goods. With a relatively small MAE value and a prediction pattern that follows the actual demand trend, this model can be a useful tool in assisting supply chain managers in making strategic decisions related to the stock and distribution of medical goods. The implementation of this model is expected to improve the efficiency of healthcare supply chain management, reduce the risk of shortages or overstocks, and ensure optimal availability of medical resources.

REFEENCES

Apriska, L., Irwan, M., Suprapti, I. A. P., & Anggara, J. (2024). Pengaruh Pengeluaran Pemerintah Sektor Kesehatan, Pertumbuhan Ekonomi, Dan Kemiskinan Terhadap Indeks Pembangunan Manusia (Kabupaten/Kota Di Provinsi NTB). *Jurnal Sosial Ekonomi Dan Humaniora*, *10*(1), 1–11. https://doi.org/10.29303/jseh.v10i1.451

Arji, G., Ahmadi, H., Avazpoor, P., & Hemmat, M. (2023). Identifying resilience strategies for disruption management in the healthcare supply chain during COVID-19 by digital innovations: A systematic literature review. *Informatics in Medicine Unlocked*, *38*(February), 101199. https://doi.org/10.1016/j.imu.2023.101199

Carolina, M. (2022). Pengaruh Pengeluaran Pemerintah Di Bidang Subsidi, Pendidikan, dan Kesehatan Terhadap Kemiskinan. *Jurnal Budget : Isu Dan Masalah Keuangan Negara*, *7*(1), 165–180. https://doi.org/10.22212/jbudget.v7i1.125

Chukwu, E. (2020). *A Systematic Review of Blockchain in Healthcare : Frameworks , Prototypes , and Implementations*. *8*.

Elektronik, K. (2024). *Kaysa Ramadhani Universitas Janabadra Yogyakarta, Indonesia Email: 4*(2), 116–119.

Fadmadika, F., Handayani, H. H., Mudzakir, T. Al, Indra, J., Komputer, F. I., Buana, U., & Karawang, P. (2024). *Pengaruh smote terhadap performa algoritma random forest dan algoritma gradient boosting dalam memprediksi penyakit stroke*. *7*, 837–846. https://doi.org/10.37600/tekinkom.v7i2.1575

Jadhav, J. S., & Deshmukh, J. (2022). A review study of the blockchain-based healthcare supply chain. *Social Sciences and Humanities Open*, *6*(1), 100328. https://doi.org/10.1016/j.ssaho.2022.100328

KANTINIT. (2023). *Gradient Boosting: Pengertian, Cara Kerja dan Contoh Skripsi*. https://kantinit.com/kecerdasan-buatan/gradient-boosting-pengertian-cara-kerja-dan-contoh-skripsi/

Nabillah, I., & Ranggadara, I. (2020). Mean Absolute Percentage Error untuk Evaluasi Hasil Prediksi Komoditas Laut. *JOINS (Journal of Information System)*, *5*(2), 250–255. https://doi.org/10.33633/joins.v5i2.3900

Rahim, A., Sutanty, M., & Anggita, P. (2021). Pengaruh Pengeluaran Pemerintah Bidang Pendidikan Dan Kesehatan Terhadap Ipm Kabupaten Sumbawa Tahun 2015-2020. *Jurnal Ekonomi & Bisnis*, *9*(3), 276–284. https://doi.org/10.58406/jeb.v9i3.513

Setyawan, N. H., & Wakhidah, N. (2025). *Analisis perbandingan metode logistic regression, random forest, gradient boosting untuk prediksi diabetes*. *10*(1), 150–162.

Spieske, A., Gebhardt, M., Kopyto, M., & Birkel, H. (2022). Improving resilience of the healthcare supply chain in a pandemic: Evidence from Europe during the COVID-19 crisis. *Journal of Purchasing and Supply Management*, *28*(5), 100748. https://doi.org/10.1016/j.pursup.2022.100748

Ummah, M. S. (2019). No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title. *Sustainability (Switzerland)*, *11*(1), 1–14. http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_SISTEM_PEMB

ETUNGAN_TERPUSAT_STRATEGI_MELESTARI

Villarreal, E. R. D., Garcia-Alonso, J., Moguel, E., & Alegria, J. A. H. (2023). Blockchain for Healthcare Management Systems: A Survey on Interoperability and Security. *IEEE Access*, *11*(December 2022), 5629–5652. https://doi.org/10.1109/ACCESS.2023.3236505