GREENATION PUBLISHER
GCIR: Greenation Computer and Information Review
E-ISSN: 9999-9999
P-ISSN: 9999-9999
+62 812-1046-7572 https://greenpub.org/GCIR greenation.info@gmail.com
DOI: https://doi.org/10.38035/gcir.v1i1

Implementation of a Machine Learning with a Python Programming Approach to Predict the Level of Disease Cases

Ridwan[1], Prio Kustanto[2]
[1]STIE Dewantara, Bogor, Indonesia, ridwans70@gmail.com
[2]Universitas Bhayangkara Jakarta Raya, Jakarta, Indonesia, pkustanto@dsn.ubharajaya.ac.id

Corresponding Author: ridwans70@gmail.com[1]

Abstract: Digital transformation invariably results in the development of data-driven insights. To counteract digital transformation in the health domain, tools that can perform predictive analysis are needed. The use of machine learning to predict the severity of a disease can aid in the prompt formulation of health care policies. Machine learning models employ several techniques in classification algorithms with a framework provided by the Python programming language. In this case, the used data include a dataset that was collected from the Indonesian Health Profile, which includes information on health-related behaviors, health-related laws, and sociodemographic trends. Decision tree modeling has very good accuracy when classifying the severity of illness.

Keyword: digital transformation, machine learning, classification, and outbreak prediction.

INTRODUCTION

The Industry 4.0 revolution has brought about a digital transformation for many sectors of society, including government (Maria Pangestika, 2020). As of 2024, the Indonesian government has set a deadline for the digital transformation. The first step in this digital transformation initiative is the data-driven citizen education and outreach process. As a result, the Indonesian Ministry of Health is actively participating in the digital transformation process by presenting the Cetak Biru Strategy for the Digital Transformation of Health 2024, which, among other things, aims to address data-driven health policy issues (Wilda Faida & Angesti, 2023). The implementation of health policies has been steadily improving over time, taking into account the conditions in each of the countries in question. As a result, the policies can be implemented successfully (Taufiqurokhman, n.d.).

As a result of the aforementioned Cetak Biru Strategi Transformasi Digital Kesehatan 2024, the Indonesian Ministry of Health must focus on both the general state of health and the specific conditions of various diseases in each of the country's regions (Politeknik et al., n.d.). In this sense, datasets produced by the Department of Health and Information Technology quantify the range of illness severity as well as the range of health care implementation, including vaccination, sanitation, and the availability of clean housing during each time. The aforementioned statistics are a component of the Profil Kesehatan Indonesia project, which comprises infographics and datasets offering a variety of facts and data on Indonesian health, illness rates, education, and demography. The Health Ministry already possesses carefully validated datasets in this area (Syihab Alfaritsi, n.d.).

In order to support digital transformation, development teams working on virtual reality, sanitation, and other health programs need to be equipped with tools that can process data and provide predictive insights. By providing predictive and automatically generated insights, machine learning can optimize the process of creating new content. Machine learning, or machine learning, is a branch of research that focuses on the principles of graph theory and computer learning in real-world scenarios. It uses learning algorithms such as discriminative and non-discriminative learning to predict and facilitate the automatic generation of answers based on large amounts of data. (Hasan & Raden, 2024). These machine learning frameworks are built upon a foundation of data preparation and modeling that can be expanded upon and implemented automatically. With the use of a mesin learning model, Pemangku Kepentingan is able to create predictive studies using a collection of Indonesian Health Profile information, which assist them in developing a policy..

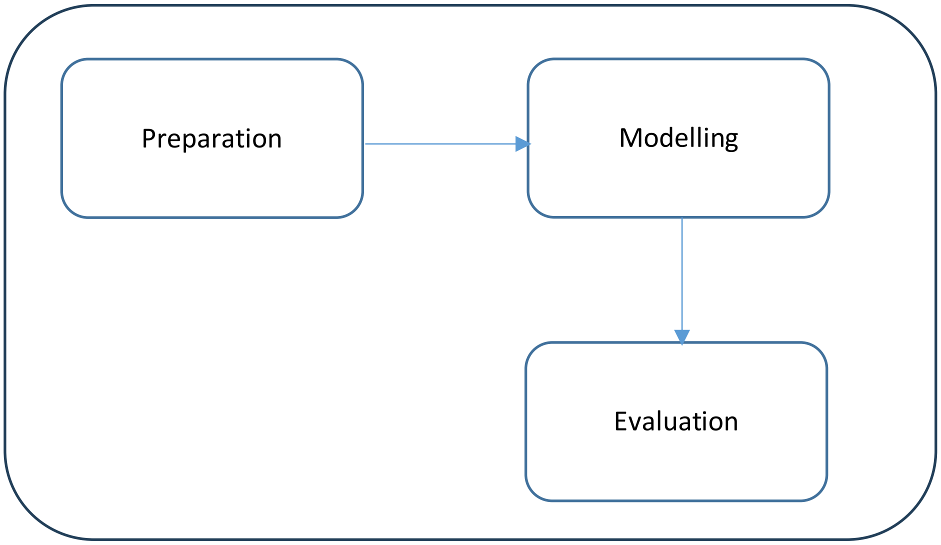
This study aims to illustrate the application of machine learning to predict the incidence of several chronic diseases in Indonesia. Numerous studies on the prediction of the peak disease severity have been conducted. In the single most comprehensive evaluation of the literature from various case studies, the application of machine learning with ARIMA modeling to predict COVID-19 illness has produced time series predictions with optimal accuracy (Salsabila et al., 2024a). Subsequently, the application of machine learning to predict the incidence rate of tuberculosis in Gambok, Malaysia using the regression linear and artificial neural network (ANN) methods based on socio-demographic and environmental parameters yielded a high margin of error [6]. Subsequently, the modeling that was done to predict the COVID-19 case from five different countries in a short period of time produced the conclusion that the MLP and ANFIS algorithms had relatively low margin error and very high efficiency compared to inferior algorithms (Novelinda et al., 2021).

The goal of this project is to apply machine learning to develop prediction models for the case rates of different diseases in every province in Indonesia, based on a variety of studies. Machine learning modeling with different classification techniques and factors including environmental circumstances and healthcare service levels. Then, the level of disease cases that serves as the target attribute for prediction is the number of cases of various infectious diseases from all provinces during specific periods, which have been normalized into five classes (1-4), where the higher the class, the higher the level of cases.

The application of classification algorithms in accordance with the requirements to forecast the proportion of cases that have been categorized into particular classes. Classification algorithms can be used to predict the categorization of various parameters, such as predicting the classification category of an individual infected with COVID-19 based on age, gender, specimen condition, and transmission category (Salsabila et al., 2024b). This research uses the Python programming language framework.

METHOD

There are three stages involved in the implementation of machine learning: data modeling, evaluation, and experimentation.

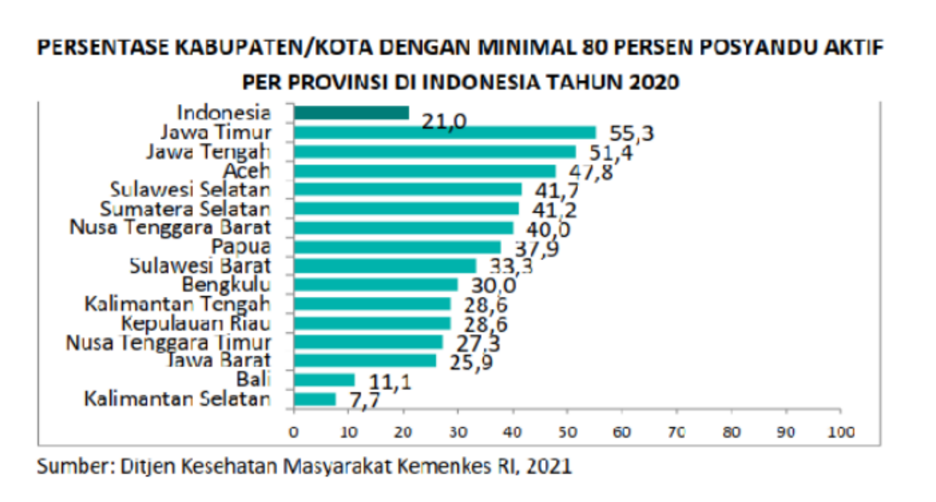


Picture 1. Research Stages

The process is carried out by analyzing the attribute or parameter that will be used in machine learning modeling. The regression from each of those parameters is constructed using the initial hypothesis (Alfaina Shidiq, 2022). The hypothesis that is developed in this study focuses on the conditions in which the incidence rate of illness is related to sociodemographic variables, environmental conditions, and the rate of health care that has been previously examined in previous studies. Subsequently, the study is conducted using data gathered from the Indonesian Health Profile from 2016 to 2021. Based on the aforementioned data collection, the variable/parameter x used in this study is as follows:

Year, Province, Disease Name, Number of Cases, Percentage of Poor Population, Human Development Index, Population Density per KM. Percentage of Districts/Cities implementing the GERMAS policy, Percentage of Food Processing Places that Meet Requirements According to Standards, Percentage of Public Places and Facilities that are Supervised According to Standards. The number of districts/cities that provide integrated services (PANDU) at each community health center by province. Percentage of Community Health Centers with Complete Basic Immunization Vaccine Availability by Province, Percentage of Districts/Cities Implementing Active Posyandu Development, The number of production facilities in the pharmaceutical field and medical equipment by province. Ratio of Community Health Centers per District, Number of Registered Primary Clinics and Main Clinics According to Ownership and Province. The ratio of health centers per sub-district, Number of Hospitals and Hospital Bed Ratio per 1000 Population by Province (Pelaksana et al., 2019).

Subsequently, the aforementioned data are extracted into a single dataset that is cleaned and formatted using data cleansing and formatting techniques to provide data that matches the required parameters. Data formatting is done by modifying the data format so that it conforms to the data characteristics needed for modeling, while data cleansing is done by removing data bars that are not accurate or sufficiently long (Ajhari, n.d.).


PERSENTASE KABUPATEN/KOTA DENGAN MINIMAL 80 PERSEN POSYANDU AKTIF PER PROVINSI DI INDONESIA TAHUN 2020

Sumber: Ditjen Kesehatan Masyarakat Kemenkes RI, 2021

Picture 2. Screenshot of Float Data Example

The following graph represents a float data example that has not yet been formatted. In this case, improper data conditions consisting of floating-point numbers cannot be resolved by pandas, necessitating data formatting and purification. This is pseudocode from the data purification process using Pandas and Numpy.

```
Import numpy as np
df.dropna()
df = df.replace(‘-‘, np.nan).astype(‘object’)
```

Next, data formatting is implemented by applying a pseudocode with the following functions to variable/parameter x:

```
Df[‘parameter/variabel x’] = Df[‘parameter/variabel x’].str.replace(‘,’, ‘.’).astype(float)
```

Under this function, numerical data that is not consistent (angka with format ',') will be normalized to a floating-point array. Subsequently, the data formatting step is carried out by normalizing the target parameter or attribute (the number of illness cases) into eight groups (1-4). The normalization process is carried out using the "KBinsDiscretizer" function from scikit-learn. The function in question will operate by applying the clustering strategy (K-Means) to automatically group the numerical data in order to group the data into four clusters. As a result, this function will change the continuous variable Jumlah Kasus Penyakit to a biner variable with a rentang of 1 through 4.

Table 1. Description of the Target Variable Normalization Function

Kbins Discretizer Parameter	**Input**
n_bins (number of cluster)	**4**
Encode (data type)	**ordinal**
Strategy (algorithm)	**k-means**

Subsequently, the study is started with a modeling exercise. The modeling in this study is done using several classification algorithms. Generally, classification model building is done by estimating the likelihood of a given sample from a single data variable (Apriliah et al., 2021). In this case, the modeling is done by utilizing the decision tree, random forest, logistic regression, support vector machine (SVM), and xgboost algorithms with the scikit-learn framework/library from Python. The final step will be comparing the results of each of the aforementioned algorithms. In essence, decision tree algorithms work by analyzing a probability structure with different attributes and features, while random forest algorithms are collections of decision tree processes that arise in parallel (Rizka & Putri, n.d.).

Table 2. Description of Algorithm Functions from the Scikit-learn Library

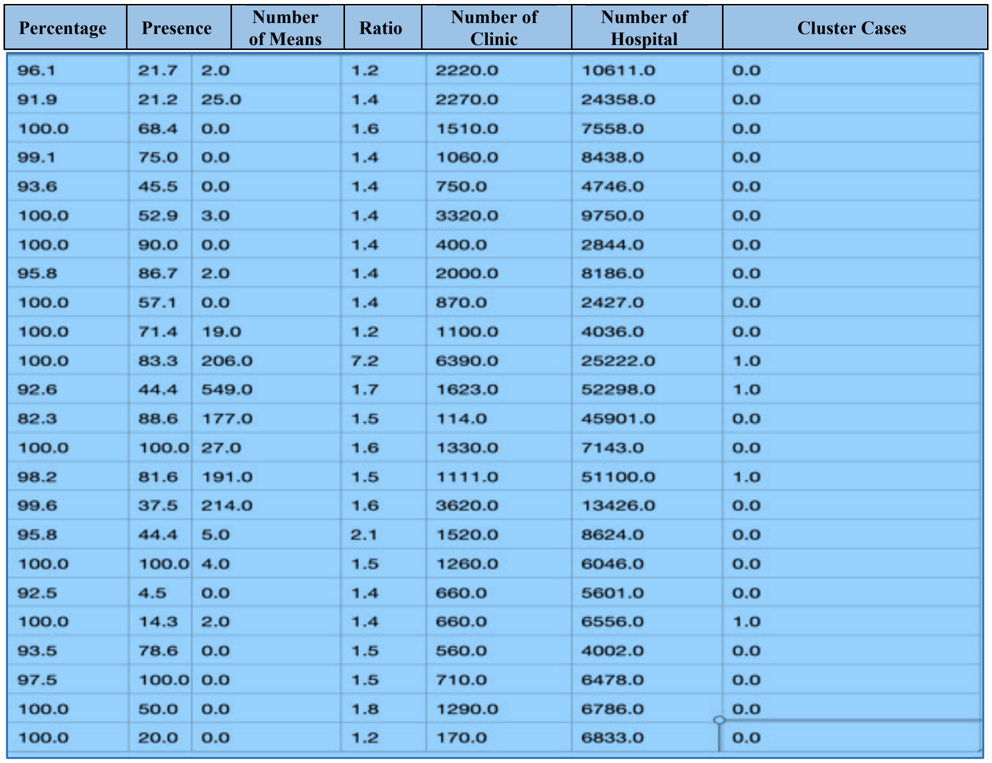
Model	Scikit-learn Function	Parameter
Logistic Regression	LogisticRegression()	**Random_state=0**
Random Forest	RandomForestClassifier ()	**n_estimators=1000, oob_score = True, n_jobs=-1 random_state =50 max_features="auto" max_leaf_nodes=30**
SVM	SVC()	**kernel=linear**
Decision Tree	DecisionTreeClassifier	
XGBoost	**XGBClassifier()**	**n_estimators = 400, learning_rate = 0.1, max_depth = 3**

After the modeling process is carried out, each model will be evaluated by comparing the accuracy of the five algorithm techniques for each type of disease. Accuracy calculations are carried out after the scikit-learn function of each algorithm is applied. Accuracy calculations are carried out by implementing the following scikit-learn formula:

sklearn.metrics.accuracy_score(*y_true,y_pred,norm alize=True*, *sample_weight=None*)

RESULTS AND DISCUSSION

The results of these four processes—implementation, modeling, evaluation, and so on will be presented in this study. Data that is suitable for modeling is produced by the data extraction, data cleaning, and data formatting stages of the process. In this case, the data cleansing process was successful in removing data points with null values. The data formatting process then succeeded in converting the numerical data type from several data bars derived from various variable x to float type. Afterwards, the data formatting process also successfully initiates the normalization of the target variable or attribute. In this case, the data variable target (number of cases) is continuously sampled from index 1 to index 4 in a few clusters/kelas, and a new variable that is representative of the sample is created (number of cases clustered).



Percentage	Presence	Number of Means	Ratio	Number of Clinic	Number of Hospital	Cluster Cases
96.1	21.7	2.0	1.2	2220.0	10611.0	0.0
91.9	21.2	25.0	1.4	2270.0	24358.0	0.0
100.0	68.4	0.0	1.6	1510.0	7558.0	0.0
99.1	75.0	0.0	1.4	1060.0	8438.0	0.0
93.6	45.5	0.0	1.4	750.0	4746.0	0.0
100.0	52.9	3.0	1.4	3320.0	9750.0	0.0
100.0	90.0	0.0	1.4	400.0	2844.0	0.0
95.8	86.7	2.0	1.4	2000.0	8186.0	0.0
100.0	57.1	0.0	1.4	870.0	2427.0	0.0
100.0	71.4	19.0	1.2	1100.0	4036.0	0.0
100.0	83.3	206.0	7.2	6390.0	25222.0	1.0
92.6	44.4	549.0	1.7	1623.0	52298.0	1.0
82.3	88.6	177.0	1.5	114.0	45901.0	0.0
100.0	100.0	27.0	1.6	1330.0	7143.0	0.0
98.2	81.6	191.0	1.5	1111.0	51100.0	1.0
99.6	37.5	214.0	1.6	3620.0	13426.0	0.0
95.8	44.4	5.0	2.1	1520.0	8624.0	0.0
100.0	100.0	4.0	1.5	1260.0	6046.0	0.0
92.5	4.5	0.0	1.4	660.0	5601.0	0.0
100.0	14.3	2.0	1.4	660.0	6556.0	1.0
93.5	78.6	0.0	1.5	560.0	4002.0	0.0
97.5	100.0	0.0	1.5	710.0	6478.0	0.0
100.0	50.0	0.0	1.8	1290.0	6786.0	0.0
100.0	20.0	0.0	1.2	170.0	6833.0	0.0

Picture 3. *Screenshot of Data Formatting Process Results*

Based on the above graph, the data formatting process converts the numerical values that exceed the koma threshold (100,0) to the titik threshold (100.0). Subsequently, the number of cases in each data variable is normalized using the clustering process carried out by the "KBinsDiscretize" function. From this process, a new variable (jumlah_kasus_clustered) appears, which is the result of normalizing the data of the number of cases in five groups with indices of 1.0, 2.0, 3.0, and 4.0. This is related to the clustering method's diskritisasi principle, which safeguards the process by which the clustering method in the diskretisasi phase measures the titik pusat from the cluster data in an accurate manner and matches each individual nilai to the titik pusat that is closest to it.

After performing the preprocessing step, the dataset is ready to be used in the modeling step. To start the modeling process, all of the functions of the Scikit-learn algorithm are implemented in order to apply the models that have been previously demonstrated in the research methodology. In the modeling process, the dataset's results from the data purification process will be combined with training and testing using the following guidelines:
X2021_train, X2021_test, y2021_train, y2021_test = train_test_split(X2021, y2021, test_size=0.4, random_state=101)

According to the above discussion, datasets that have been classified based on their condition are data training (data used to train modeling algorithms) with a composition of 65% of the total data set, and data testing (data used to evaluate algorithm performance) with a composition of 45% of the total data set.

The modeling process is carried out using a dataset that has been categorized into each of the several types of illness. When compared to other models, the tuberculosis model using decision tree and XGBoost techniques produced the lowest results.

Table 3. Tuberculosis Disease Modeling Results

Modelling	*Accuracy Score*
Logistic Regression	**0.73**
Random Forest	**0.73**
Support Vector Machine	**0.73**
Decision Tree	**0.81**
XGBoost	**0.79**

Then, modeling diarrheal disease produces the highest accuracy using decision tree and random forest techniques.

Table 4. Diarrhea Disease Modeling Results

Modelling	*Accuracy Score*
Logistic Regression	**0.4**
Random Forest	**0.90**
Support Vector Machine	**0.51**
Decision Tree	**0.93**
XGBoost	**0.87**

Modeling of dengue fever produces quite high accuracy using the Decision Tree and XGBoost techniques.

Table 5. DHF Disease Modeling Results

Modelling	Accuracy Score
Logistic Regression	**0.74**
Random Forest	**0.80**
Support Vector Machine	**0.77**
Decision Tree	**0.82**
XGBoost	**0.85**

The Decision Tree technique for Malaria also has quite high accuracy compared to other techniques.

Table 6. Malaria Disease Modeling Results

Modelling	Accuracy Score
Logistic Regression	**0.66**
Random Forest	**0.77**
Support Vector Machine	**0.66**
Decision Tree	**0.80**
XGBoost	**0.77**

For leprosy, modeling using random forest, XGBoost and decision tree techniques has an accuracy level of 0.79, which is the highest level of accuracy compared to other techniques.

Table 7. Leprosy Modeling Results

Modelling	Accuracy Score
Logistic Regression	**0.68**
Random Forest	**0.71**
Support Vector Machine	**0.68**
Decision Tree	**0.71**
XGBoost	**0.71**

In contrast to before, modeling using the decision tree technique for diphtheria does not have good accuracy. In this case, logistic regression, random forest, and support vector machine techniques have similarities in accuracy that exceeds the accuracy of decision trees.

Table 8. Diphtheria Disease Modeling Results

Modelling	Accuracy Score
Logistic Regression	**0.81**
Random Forest	**0.81**
Support Vector Machine	**0.81**
Decision Tree	**0.64**
XGBoost	**0.79**

Based on the results of every step of the modeling process with regard to any kind of illness, the evaluation process with regard to every step of the modeling process above indicates that the decision tree modeling technique used to determine the classification threshold for illness based on the patient's health status and sociodemographic characteristics of each province's residents is the most accurate model. Based on this, the committee for peer review can conduct simulations to determine the extent to which machine learning models using classification algorithms can surpass the threshold for health care quality compared to the disease severity threshold.

CONCLUSION

The results of this study indicate that the use of machine learning to the prediction of the tingkat kasus of illness using classification algorithms may be carried out by first normalizing the illness's angka and then combining modeling techniques with classification algorithms. In this study, the decision tree modeling technique produced very good results. Under this situation, the health care team can use the prediction of illness severity as a modeling tool to estimate the target of health care quality during the current period. The process of transforming analog data into digital data is initiated by the digital transformation.

For further research, the use of variables relating to health care, as well as socio-demographic aspects of society, can serve as a starting point, particularly by implementing regression and forecasting algorithms in order to predict and predict the occurrence of new diseases.

REFERENCE

Ajhari, A. A. (n.d.). *Pendeteksian Anomali Pergerakan Penerbangan Pesawat Menggunakan Data Automatic Dependent Surveillance Broadcast*. https://www.researchgate.net/publication/376270051

Apriliah, W., Kurniawan, I., Baydhowi, M., Haryati, T., Informasi Kampus Kabupaten Karawang, S., Teknik dan Informatika, F., Bina Sarana Informatika, U., Banten No, J., & Karawang Barat, K. (2021). *SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest* (Vol. 10, Issue 1). http://sistemasi.ftik.unisi.ac.id

Alfaina Shidiq, A. (2022). *Perancangan Mesin Takik Bambu dengan Pendekan User Centered Design*.

Hasan, M., & Raden, W. (2024). *Buku Digital-Ekosistem Sumber Daya M anusia dalam Tantangan Resesi Global.p df AUTHOR turnitin Ekosistem Sumber Excluded from Similarity Report*.

Maria Pangestika, m. h. s. y. h. a. n. w. m. m. h. a. j. s. y. a. h. l. s. d. d. n. z. h. j. n. t. m. p. (2020). *Smart Farming: Pertanian di Era Revolusi Industri 4.0*. Andi.

Novelinda, Fitrianah, D., & Komputasi Dan Teknologi Informasi, E.-I. (2021). *Analisa Perbandingan Algoritma CNN Dan MLP Dalam Mendeteksi Penyakit COVID-19 Pada Citra X-Ray Paru* (Vol. 3, Issue 2).

Pelaksana, T., A Littik, S. K., Ch Berek, N., Dodo, D. O., & Doke, S. (2019). *Laporan akhir pendampingan tata kelola program kesehatan di kabupaten sumba tengah provinsi nusa tenggara timur*.

Politeknik, J. K., Kemenkes, K., & Raya, P. (n.d.). *Sarjana Terapan Kebidanan*.

Rizka, S. O., & Putri, A. (n.d.). *Pemodelan Algoritma Random Forest Untuk Klasifikasi Log Access Jenis Domain Pada Pandi (Pengelola Nama Domain Internet Indonesia)*.

Salsabila, T. H., Indrawati, T. M., & Fitrie, R. A. (2024a). Meningkatkan Efisiensi Pengambilan Keputusan Publik melalui Kecerdasan Buatan. *Journal of Internet and Software Engineering*, *1*(2), 21. https://doi.org/10.47134/pjise.v1i2.2401

Salsabila, T. H., Indrawati, T. M., & Fitrie, R. A. (2024b). Meningkatkan Efisiensi Pengambilan Keputusan Publik melalui Kecerdasan Buatan. *Journal of Internet and Software Engineering*, *1*(2), 21. https://doi.org/10.47134/pjise.v1i2.2401

Syihab Alfaritsi. (n.d.). *Administrasi Publik di Era Disrupsi dan Big Data*. https://id.wikipedia.org/wiki/Inovasi_disruptif

Taufiqurokhman, D. (n.d.). *Pendelegasian tanggungjawab negara kepada presiden selaku penyelenggara pemerintahan*.

Wilda Faida, E., & Angesti, D. (2023). Readiness of Medical Recording and Health Information Education Institutions in the Digital Transformation Era of UTAUT Based. *Journal of Community Empowerment for Multidisciplinary (JCEMTY)*, *1*(2), 90–103. https://doi.org/10.53713/jcemty.v1i2.91

Primawanti, E. P., & Ali, H. (2022). Pengaruh Teknologi Informasi, Sistem Informasi Berbasis Web Dan Knowledge Management Terhadap Kinerja Karyawan (Literature Review Executive Support Sistem (Ess) for Business). *Jurnal Ekonomi Manajemen Sistem Informasi*, *3*(3), 267-285.

Hasyim, U., & Ali, H. (2022). Reuse intention models through customer satisfaction during the COVID-19 pandemic: Cashback promotion and e-service quality case study: OVO electronic money in Jakarta. *Dinasti International Journal of Digital Business Management*, *3*(3), 440-

450.

Wahono, S., & Ali, H. (2021). Peranan Data Warehouse, Software Dan Brainware Terhadap Pengambilan Keputusan (Literature Review Executive Support Sistem for Business). *Jurnal Ekonomi Manajemen Sistem Informasi*, *3*(2), 225-239.

Indarsin, T., & Ali, H. (2017). Attitude toward Using m-commerce: The analysis of perceived usefulness perceived ease of use, and perceived trust: Case study in Ikens Wholesale Trade, Jakarta–Indonesia. *Saudi Journal of Business and Management Studies*, *2*(11), 995-1007.

Faisal, F., Ali, H., & Rosadi, K. I. (2021). Sistem Pengelolaan Pendidik Dan Tenaga Kependidikan Berbasis Simdik Dalam Manajemen Pendidikan Islam. *Jurnal Ilmu Manajemen Terapan*, *3*(1), 77-85.

Nugroho, F., & Ali, H. (2022). Determinasi SIMRS: Hardware, Software Dan Brainware (Literature Review Executive Support Sistem (ESS) For Business). *Jurnal Manajemen Pendidikan Dan Ilmu Sosial*, *3*(1), 254-265.

Ashshidiqy, N., & Ali, H. (2019). Penyelarasan Teknologi Informasidengan Strategi Bisnis. *Jurnal Ekonomi Manajemen Sistem Informasi*, *1*(1), 51-59.

Djojo, A., & Ali, H. (2012). Information technology service performance and client's relationship to increase banking image and its influence on deposits customer banks loyalty (A survey of Banking in Jambi). *Archives Des Sciences*, *65*(8).

Desfiandi, A., Yusendra, M. A. E., Paramitasari, N., & Ali, H. (2019). Supply chain strategy development for business and technological institution in developing start-up based on creative economy. *Int. J. Supply Chain Manag*, *8*(6), 646-654.

Havidz, I. L. H., Aima, H. M., Ali, H., & Iqbal, M. K. (2018). Intention to adopt WeChat mobile payment innovation toward Indonesia citizenship based in China. *International Journal of Application or Innovation in Engineering & Management*, *7*(6), 105-117.