Prediction of Covid-19 Infection in Indonesia Using Machine Learning Methods

H Zakiyyah¹ and S Suyanto²

¹,²School of Computing, Telkom University, Bandung, Indonesia
Email: hasnazakiyyah@student.telkomuniversity.ac.id¹

Abstract. Currently, the world is experiencing a prolonged pandemic known as Covid-19. Many prediction models of Covid-19 have been developed by the governments to make the right decisions to control the outbreak. In Indonesia, there is also much research on the prediction of Covid-19 using machine learning methods, which provide the statistics to predict the total cases, the total deaths, the peak and the end of the pandemic. This paper investigates three prediction models: Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM), and Decision Tree (DT) in predicting total cases and total deaths of Covid-19 in Indonesia. First, a preprocessing is applied to change the string data to the numerical dataset using a label encoder. Second, the models are trained using the Covid-19 Indonesia Time Series All Dataset (CITSAD) with 90% and 10% train/test split. The three models are then investigated to predict new cases and new deaths. The evaluation using the CITSAD of ten provinces in Indonesia shows that DT gives the highest accuracy of 93% and provides the fastest processing time of 48.4 seconds.

1. Introduction
Based on the information in [1], the coronavirus or Covid-19 affected 214 countries and regions worldwide with 38,404,464 cases, 1,091,569 deaths, and 28,874,023 recovered. This information was taken on October 14, 2020, at 4:17 pm Indonesia time.

The Covid-19 pandemic is caused by a coronavirus that first appeared in Wuhan, China, at the end of 2019. Symptoms that often appear are fever, fatigue, and dry cough. This disease attacks all age elements in society worldwide. It is vulnerable to the elderly and people with medical records such as heart disease, diabetes, and respiratory problems [2]. On March 2, 2020, President Joko Widodo announced that the coronavirus exposed two Indonesian citizens. Both are a 64-year-old woman, and her daughter, who is 31 years old, had positive Covid-19 after contact with a Japanese citizen [3]. On August 23, 2020, according to Johns Hopkins University, Indonesia is in the 23rd position in the world confirmed 153,535 positive Covid-19 cases. The data put of 0.66 percent globally, which achieved approximately 23 million. Indonesia is in the 19th position because of the mortality rate, which is 6,680 fatalities [4].

Implementing health protocols, such as washing hands or cleaning it with alcohol, maintaining a minimum distance of two meters from people around them, avoiding crowds, and paying attention to balanced nutrition so that the body is always healthy and fit, can reduce the new cases and death rate [5]. Hence, a prediction model of Covid-19 plays an essential role in making the right decisions to control the outbreak. Many researchers in Indonesia use machine learning-based models to predict the new cases and new deaths in real-time based on the Covid-19 time series dataset. They exploit various methods, such as Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM), and Decision Tree (DT) that provide relatively high accuracies in many cases of predictions. Those methods have
advantages and disadvantages. GNB is quite simple to implement, while GNB and DT are relatively complex. SVM works for a continuous dataset, while GNB and DT can be suitable for discrete and continuous datasets. DT is traceable [6], while GNB and SVM are not. They also give different performances in terms of effectiveness (accuracy) and efficiency (cost and time). Since they are statistical data-driven methods, various datasets and parameter tuning may produce different results.

This research investigates three prediction models of GNB, SVM, and DT using the Covid-19 time series dataset in Indonesia. They are implemented using the same tools and run on the same computer to compare their performances in accuracy and processing time.

2. Literature Review

Many researchers develop various models to predict new cases and new deaths in real-time based on the Covid-19 time series dataset. They exploit two approaches of conventional machine learning (CML) and deep learning (DL) to create predictions based on the given time-series datasets.

Rosaline et al. in [7] predict the numeral, shape, and length Coverage of Covid-19, and the final term across India uses the Autoregressive Integrated Moving Average Model (ARIMA). They use water and air as benchmarks that are integrated by three regression techniques, including Support Vector Regression (SVR), Artificial Neural Network (ANN), and Linear Regression (LR). This research proves that the projection method is better than the practical model, which can be obtained previously based on predictive precision. They conclude that Covid-19 can be transmitted through water and air ecological variables so that preventive action by following health protocols is required.

In [8], Analyze and predict the age range of people affected by covid-19 using machine learning methods such as Random Forest Regressor and Random Forest Classifier, SVM, KNN + NCA, Decision Tree Classifier, Gaussian Naïve Bayesian Classifier, Multilinear Regression, Logistic Regression, and XGBoost Classifier. Furthermore, the results show that the Random Forest Regressor and Random Forest Classifier are the best methods to solve this problem.

In [9], the Network-Inference-Based Prediction Algorithm (NIPA) is proposed to estimate the Covid-19 epidemic in Hubei in the future based on community interaction in the city of Hubei. The results show that NIPA is useful for predicting epidemic outbreaks accurately. The researchers in [10] use the partial derivative regression method and nonlinear machine learning (PDR-NML) to predict the global covid-19 pandemic. The outcome shows that this method performs out the current method in the Indian population. In [11], other researchers predict the Covid-19 infection statistics by adjusting the asymptotic distribution with actual data available in China and Italy. The results suggest that it turns out that epidemiological projections should cover a wide range of uncertainties and the need to collect a dataset of maximum numbers of contaminations, counting asymptomatic patients in real-time. In [12], the researchers predict the time series of positive affected, mortality, and recovery rates of Covid-19 controlled by the average absolute error, root mean square error (RMSE), and r2_index. The best r2_value was obtained by China as much as 0.9997 for recovered cases.

In [13], researchers predict the course of the Covid-19 in Germany apply the Bateman SIZ’s innovation model and input variables based on the status quo in July 2020 by taking into account the resilience and sensitivity of the model determined by changing the input parameters for doubling-time (t) by ± 5% and ± 10%. The prediction results show that a small change, ± 5%, over time, t for the rate of increase in the number of new infections, can have a large effect, both positive and negative, throughout the pandemic. The model also estimated that the number of people infected with the virus would reach 1 million within eight years. A 5% longer T would reduce the number of infected people by 75%. While shorter doubling times could increase the number of infections over eight years to 9 million by the time the number of infected people has exceeded 100,000 by the end of 2022. This pandemic will predict to disappear by the end of 2024.

Meanwhile, a DL model predicts the increasing trend of the Covid-19 epidemic by a rolling update system depending on the data from Johns Hopkins University [14]. The result shows that Iran expected to drop in the number of active cases by 1000 by mid-November, while there will still be a more than 2,000 increase in the number of active cases in early December in Russia. Preventive action from the government is also essential to control the number of cases. Nevertheless, the DL is time-consuming in the learning process. It needs high-performance computing to learn the given dataset for
hundreds or even thousands of iterations. Moreover, it should be retrained to adapt to the new dynamic datasets. Hence, the CML approach, which is commonly more efficient than DL, is preferable to tackle those issues.

3. Proposed Model

The proposed model is illustrated in Fig. 1. The preprocessing to change the string data to the numerical dataset using a label encoder is performed. The models are then trained using a 90% and 10% train/test split of data. Next, the decision tree classifier is defined to predict new cases and new deaths. A data frame for each column of prediction, new cases, and new deaths is then created. Finally, the evaluation is carried out in terms of accuracy and processing time.

CITSAD is a collection of various open data sources, such as Covid19.go.id (pandemic data), kemendagri.go.id (demographic data), bps.go.id (demographic data) that contains a time series of Covid-19 in Indonesia from the country to province level. It also contains 37 columns and 6511 rows of data [15]. CITSAD uses some attributes with their intervals as follow:

- New Cases: [0,4823]
- New Deaths: [0,160]
- New Recovered: [0,4343]
- New Active Cases: [0,1533]
- Total Cases: [1,267000]
- Total Deaths: [0,10218]
- Total Recovered: [0,196000]

3.1. Data Preprocessing

The dataset may contain noises, missing values, and maybe in an unusable format for machine learning models [6]. In this case, the dataset has no missing values, accurable, but it has an unusable format. There are still strings in the dataset that must convert to numeric by using a label encoder to make it easier to read the province, as listed in Table 1.

| Location                  | Code |
|---------------------------|------|
| Aceh                      | 0    |
| Bali                      | 1    |
| Kepulauan Bangka Belitung | 2    |
| Bengkulu                  | 3    |
| Banten                    | 4    |
| Gorontalo                 | 5    |
| Jambi                     | 6    |
| Jawa Barat                | 7    |
| Jawa Timur                | 8    |
| DKI Jakarta               | 9    |
3.2. Data Splitting
The dataset is split into two groups: a training set and a testing set. In this research, a sample size of 10% reduces the overfitting and uses a repeatable train test split to ensure the data set is represented using a random sample from the original data set as in [16], which should be an observation characteristic from the problem empire.

3.3. Decision Tree Classifier
It is one of the machine learning known models, including a root node, branches, and leaf nodes. Each middle node indicates a test on an imputed, a branch indicates the result of a test, and the leaf node does a label of class. The highest node in the tree is the root node [17]. It also selects the best attribute used to divide the records, convert the attribute into a decision node, divide the dataset into a smaller set, finally start the tree construction, replicating this process recurrently.

A. Information Gain
*Information Gain* [18] calculates the difference between the unsplit-entropy and the average of split-entropy, which is formulated as

---

**Figure 1.** Flowchart of the proposed model.
\[ Info(D) = - \sum_{i}^n P_i \log_2 P_i, \] (1)

where \( P_i \) is the likelihood that a variable tuple in \( D \) belongs to the class \( C_i \), and \( Info(D) \) is the total mean of information that recognizes the class label in \( D \).

\[ Info_A(D) = - \sum_{j} \left[ \frac{[p_j]}{[D]} \right] \times Info(D_j), \] (2)

\[ Gain(A) = Info(D) - Info_A(D), \] (3)

where \( \frac{[p_j]}{[D]} \) is the weight of the \( j \)th partition, \( Info_A(D) \) is the expected information essential to tabulate a tuple from \( D \) according to the \( A \) splitting.

**B. Gain Ratio**

Gain Ratio grabs the affair by normalizing the information gain with Split Info, which is formulated as

\[ SplitInfo_A(D) = - \sum_{j} \left[ \frac{[p_j]}{[D]} \right] \times \log_2 \left( \frac{[p_j]}{[D]} \right), \] (4)

\[ GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}. \] (5)

where \( \frac{[p_j]}{[D]} \) is the weight of the \( j \)th partition, and \( v \) is the number of different values in attribute \( A \).

The attribute with the best gain ratio is chosen as the splitting attribute [19].

Meanwhile, the accuracy is the most crucial measure to know how well the performance of a machine learning method equals the balance of total correct forecasting \( (TP + TN + FP + FN) \) by the classifier to the entire data points. The accuracy is calculated as

\[ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \] (6)

where \( TP = \) True Positives, \( TN = \) True Negatives, \( FP = \) False Positives, and \( FN = \) False Negatives [19].

### 4. Results and Discussion

The three methods of GNB, SVM, and DT are evaluated using the CITSAD dataset. Table 2 informs that the GNB-based model produces the lowest accuracy of 70% with an execution time of 60 seconds. Meanwhile, the SVM-based model produces a higher accuracy of 80% and a relatively fast time of 51 seconds. Finally, the DT where the accuracy rate is more improved than the previous models, reaching 93% with an execution time of 48.4 seconds.

| Method | Accuracy (%) | Execution Time (second) |
|--------|--------------|--------------------------|
| GNB    | 70           | 60.0                     |
| SVM    | 80           | 51.0                     |
| DT     | 93           | 48.4                     |

Table 2. Accuracies produced by GNB, SVM, and DT.

In general, GNB can be an accurate prediction for many cases. Nevertheless, it gives low accuracy for the CITSAD dataset. This result shows that GNB cannot positively adapt to the dynamical pattern of the dataset. Meanwhile, SVM commonly performs well for various linearly separable and complex problems since the kernel trick can operate without computing the coordinate in a higher dimensional. Unfortunately, it also gives a low accuracy for the CITSAD dataset. Finally, DT gives the best performance since it can break down the complex decision-making process to be the simpler one. It is also comprehensive to force consideration of all possible outcomes of a decision and trace every path to a conclusion. Since CITSAD has many columns, DT is suitable for predicting the new cases and deaths on the Covid-19 infection in Indonesia. However, the DT accuracy can be improved by incorporating...
an optimization technique as used in [20], which increases the model accuracy for a weather prediction, or a recent method called lifelong learning [21] that performs well for dynamic churn prediction.

5. Conclusion

Three machine learning-based models: GNB, SVM, and DT, have been successfully developed to predict new cases and new deaths for the Covid-19 dataset in Indonesia. An evaluation using the CITSAD dataset shows that DT is capable of predicting the number of new cases and new deaths with an accuracy of 93%. It also the fastest method that requires an execution time of only 48.4 seconds.

6. References

[1] Worldometer, “Daily report of corona pandemics in the world,” 2020.
[2] W. H. Organization, “Media Statement : Knowing The Risks For Covid 19.” 2020.
[3] T. J. Post, “BREAKING : Jokowi announces Indonesia’s first two confirmed COVID-19 cases.” 2020.
[4] P. G. Bhwana, “Indonesia Rank 23rd in The World for Covid 19 Cases.” 2020.
[5] W. H. Organization, “Q&A: How is COVID-19 transmitted?” 2020.
[6] Iykra, “Mengenal Decision Tree dan Manfaatnya.” 2020.
[7] M. Alazab, A. Awajan, A. Mesleh, A. Abraham, V. Jatana, and S. Alhyari, “COVID-19 prediction and detection using deep learning,” Int. J. Comput. Inf. Syst. Ind. Manag. Appl., vol. 12, pp. 168–181, 2020.
[8] K. Bhanu et al., “Analysis, Prediction, and Evaluation of COVID-19 Datasets using Machine Learning Algorithms,” Int. J. Emerg. Trends Eng. Res., vol. 8, no. 5, pp. 3–8, 2020.
[9] B. Prasse, M. A. Achterberg, L. Ma, and P. Van Mieghem, “Network-inference-based prediction of the COVID-19 epidemic outbreak in the Chinese province Hubei,” Appl. Netw. Sci., vol. 5, no. 1, 2020, doi: 10.1007/s41109-020-00274-2.
[10] D. P. Kavadi, R. Patan, M. Ramachandran, and A. H. Gandomi, “Partial derivative Nonlinear Global Pandemic Machine Learning prediction of COVID 19,” Chaos, Solitons and Fractals, vol. 139, 2020, doi: 10.1016/j.chaos.2020.110056.
[11] T. Alberti and D. Faranda, “On the uncertainty of real-time predictions of epidemic growths: A COVID-19 case study for China and Italy,” Commun. Nonlinear Sci. Numer. Simul., vol. 90, 2020, doi: 10.1016/j.cnsns.2020.105372.
[12] F. Shahid, A. Zameer, and M. Muneeb, “Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM,” Chaos, Solitons and Fractals, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110212.
[13] A. D. L. B. J. W. B. G. Braun P Haffner S, “Predictions for the COVID-19 pandemic in Germany using the modified Bateman SIZ model,” Int J Clin Pharmacol Ther., vol. 58, pp. 467–474, 2020, doi: 10.5414/CP203868.
[14] P. Wang, X. Zheng, G. Ai, D. Liu, and B. Zhu, “Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran,” Chaos, Solitons and Fractals, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110214.
[15] Hendratno, “Covid 19 Indonesia Dataset.”
[16] J. Brownlee, “Train-Test Split for Evaluating Machine Learning Algorithms.” 2020.
[17] P. Banerjee, “Decision-Tree Classifier Tutorial.” 2020.
[18] A. Navlani, “Decision Tree Classification in Python.” 2020.
[19] C. Iwendi et al., “COVID-19 patient health prediction using boosted random forest algorithm,” Front. Public Heal., vol. 8, 2020, doi: 10.3389/fpubh.2020.00357.
[20] F. Ahyar and S. Suyanto, “Firefly Algorithm-based Hyperparameters Setting of DRNN for Weather Prediction,” in 2020 International Conference on Data Science and Its Applications (ICoDSA), Aug. 2020, doi: https://doi.org/10.1109/ICoDSA50139.2020.9212921.
[21] F. Sarnen, S. Suyanto, and R. Rismala, “Lifelong Learning for Dynamic Churn Prediction,” in 2020 International Conference on Data Science and Its Applications (ICoDSA), Aug. 2020, pp. 1–5, doi: 10.1109/ICoDSA50139.2020.9212924.