Jurnal Masyarakat Informatika, 17(1), 2026
DOI:10.14710/jmasif.17.1.77361
ISSN: 2777-0648

# Structural Correlation Patterns in Regional COVID-19 Surveillance Data and Implications for Epidemiological Monitoring

**Herfandi Herfandi[*1)], Rafat Bin Mofidul[2)], Ijaz ahmad Khan[3)]**
*Department of Informatics, Sumbawa University of Technology, Indonesia*
*Department of Electrical Engineering, Kookmin University, South Korea*
[*] *Corresponding author: herfandi@uts.ac.id*

***Abstract***
*The Covid-19 pandemic has had a significant impact on the health sector in various regions, including Kabupaten Sumbawa. This study aims to analyze relationships among attributes in the Covid-19 dataset using the Correlation Matrix algorithm within the CRISP-DM methodology. The dataset was obtained from the official website of the Government of Kabupaten Sumbawa, comprising 10,573 records, of which 405 were cleaned after the data cleaning process. The analysis was conducted using RapidMiner 9.9 software. The findings indicate a very strong correlation between the attributes KONTAK ERAT-DISCARDE, SUSPEK-DISCARDE, and KONFIRMASI-MENINGGAL DUNIA with the increase in total Covid-19 cases. In addition, a significant negative correlation was observed between the attribute PP-MASIH KARANTINA and the number of deaths. Furthermore, an almost perfect correlation was found between PROBABLE-DISCARDE and PROBABLE-MENINGGAL. Based on these findings, it is recommended that the government prioritize monitoring cases before they are declared discarded and strengthen the quarantine system for travelers. This study provides a data-driven foundation for formulating evidence-based pandemic response policies.*

***Keywords:*** *Covid-19, correlation matrix, data mining, dataset attributes*

## 1 Introduction

Although the Covid-19 pandemic was declared globally subdued in 2023 [1] its impacts on healthcare systems, social structures, and public policy in various regions of Indonesia continue to pose challenges that require comprehensive evaluation [2]. Kabupaten Sumbawa is one of the regions significantly affected, recording a surge in cases across multiple categories, including traveler (*Pelaku Perjalanan*, PP), close contacts (*Kontak Erat*), suspected cases (*Suspek*), probable cases (Probable), and confirmed cases classified as recovered, deceased, or still active. During the pandemic, the local government implemented policies such as PPKM (Community Activity Restrictions), bans on receptions and cultural gatherings, and the optimization of Covid-19 command posts. However, the effectiveness of these measures remains to be validated through data-driven approaches [3].

Historical Covid-19 case management data from the official website of the Government of Kabupaten Sumbawa (https://covid19.sumbawakab.go.id) recorded 22,725 cases between August 4, 2020, and August 31, 2021. The dataset comprises 10,699 Pelaku Perjalanan, 6,486 Kontak Erat, 2,403 Suspek, 9 Probable, and 3,128 confirmed positive cases. Analyzing this dataset can uncover patterns of disease transmission and evaluate the effectiveness of the mitigation strategies implemented during the pandemic. One potential approach is data mining [4], [5], [6], particularly correlation analysis [7], [8], which identifies relationships between attributes in the dataset and examines the influence among

*Received: 27 August 2025; Revised: 12 February 2026; Accepted: 22 February 2026; Published: 24 February 2026*

variables. Significant correlations, both positive and negative, can serve as critical indicators for problem mapping and data-driven decision-making.

Given the varying scales and heterogeneous data types within the dataset, the Correlation Matrix algorithm is employed due to its ability to standardize attribute values through standard deviation-based normalization [9], [10]. This approach enhances analysis accuracy and minimizes potential bias.

This study aims to analyze the interrelationships among epidemiological attributes within a localized infectious disease dataset using the Correlation Matrix algorithm. Although COVID-19 cases have declined globally, data-driven surveillance and infectious disease diffusion modeling remain crucial for regional preparedness and policy evaluation. Unlike prior studies that primarily focused on mobility-driven or large-scale national/global datasets, this research emphasizes a comprehensive attribute-level correlation analysis at the regional level, providing a structured analytical foundation for evidence-based public health decision-making and future outbreak preparedness.

## 2 Literature Review

Recent studies have expanded the scope of Covid-19 correlation analysis by incorporating mobility patterns, climatic factors, and causal inference approaches. For instance, [11] employed convergent cross mapping to reveal the influence of weather-driven mobility on case numbers across Europe. Similarly, [12] demonstrated that a 10% reduction in mobility resulted in a decrease of the effective reproduction number by approximately 0.05–0.07 across 99 countries, with substantial interregional variations. In Spain, [13] reported a median $R^2$=0.85 in associating mobility trends and meteorological conditions with case growth using Principal Component Analysis (PCA) combined with time-lagged correlation analysis. Beyond correlation-focused studies, epidemic diffusion has also been widely examined through network-based perspectives, where disease spreading is modeled as a contagion process over complex contact graphs. This line of work provides a theoretical foundation for understanding how case categories may co-evolve under shared reporting and transmission dynamics in a population network. Furthermore, digital contact tracing has been analyzed using network science to assess how tracing effectiveness depends on contact structures and intervention intensity, strengthening the relevance of attribute-level monitoring for policy evaluation. In addition, social-graph contagion diffusion frameworks such as CoSoGMIR have been proposed to formalize movement–interaction–return patterns as diffusion mechanisms, which further motivates positioning this study within epidemiological diffusion modeling rather than purely descriptive correlation analysis.

While the Correlation Matrix approach has been widely adopted in pandemic-related analyses, its application to evaluating relationships among all attributes in a localized Covid-19 dataset remains limited, particularly for informing public policy. The novelty of this study lies in its comprehensive application of the Correlation Matrix algorithm to examine both positive and negative inter-attribute correlations in the Covid-19 dataset for Kabupaten Sumbawa. This approach not only serves as an exploratory data analysis method but also provides a foundation for developing more effective and evidence-based regional policy interventions.

## 3 Research Methods

This study adopts a quantitative approach utilizing the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology [15]. In accordance with the research framework in Figure 1,

the workflow is described through six CRISP-DM phases: (1) Business Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modeling, (5) Evaluation, and (6) Deployment. In implementation, the core operational activities consist of data extraction, data cleaning, data transformation, and numerical analysis [16]. These activities are explicitly mapped to the CRISP-DM phases (Data Understanding, Data Preparation, and Modeling) and described with clear inputs and outputs to improve methodological transparency and reproducibility. The dataset was obtained from the official website of the Sumbawa Regency Government (https://covid19.sumbawakab.go.id) and consisted of 10,573 records with epidemiological attributes, including Travelers (*Pelaku Perjalanan*), Close Contact (*Kontak Erat* ), Suspected Cases (*Suspek*), Probable, and Confirmed Positive cases. The analysis process employed Google Spreadsheets for automated data retrieval, Microsoft Excel for data cleaning and transformation, and RapidMiner 9.9 to perform correlation analysis using the Correlation Matrix algorithm. The primary objective is to identify and evaluate the relationships among attributes in the Covid-19 dataset of Sumbawa Regency in an objective and comprehensive manner.

## 3.1 Research Framework

The research framework was adapted from the CRISP-DM methodology to align with the dataset's characteristics, as shown in Figure 1. The Business Understanding phase focuses on problem formulation and defining data-driven objectives. Data Understanding involves data collection and reviewing the dataset structure. Data Preparation includes attribute reduction, value standardization, removal of missing or invalid entries, and merging datasets into a unified, analysis-ready format. Modeling is carried out using the Correlation Matrix algorithm to measure the strength of inter-attribute relationships. The Evaluation phase interprets correlation tables and visual graphs, both positive and negative correlations, to determine the direction and magnitude of relationships. Finally, the Deployment phase produces data-driven policy recommendations to support pandemic mitigation strategies in Sumbawa Regency.

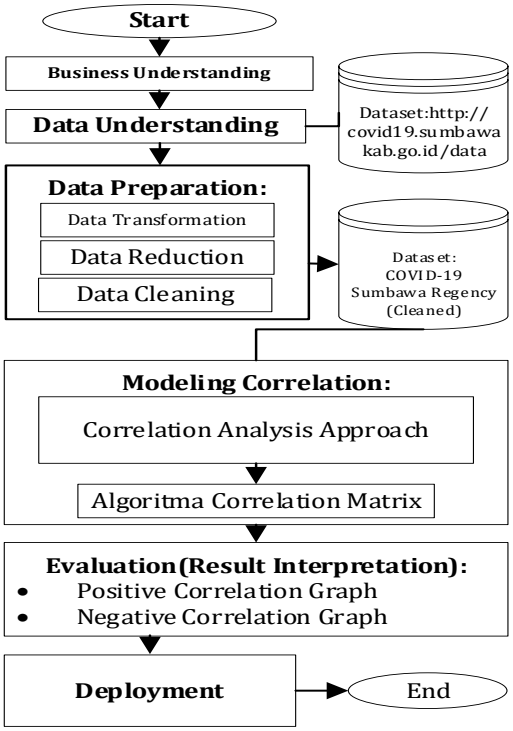



Figure 1 Research Framework

### 3.2 Research Framework

Correlation analysis is a statistical method used to measure the degree of association between attributes in a dataset. This approach is widely used across domains, including education, economics, the social sciences, and healthcare, to explore relationships between independent and dependent variables and uncover hidden patterns in data. In practice, correlation analysis produces a correlation matrix with correlation coefficients ranging from −1 to +1 for each pair of attributes. A value close to +1 indicates a strong positive correlation, a value close to −1 indicates a strong negative correlation, while a value near 0 indicates no significant relationship. The general interpretation of correlation coefficients is presented in Table 1.

Table 1 Interpretation of Correlation Coefficient

| Correlation Value (*r*) | Type | Interpretation |
|---|---|---|
| $0.80 \leq r \leq 1.00$ | Positive | Very strong correlation |
| $0.60 \leq r < 0.80$ | Positive | Strong correlation |
| $0.40 \leq r < 0.60$ | Positive | Moderate correlation |
| $0.00 \leq r < 0.4$ | Positive | No significant correlation |
| $-0.4 \leq r < 0.00$ | Negative | No significant correlation |
| $-0.6 \leq r < -0.4$ | Negative | Moderate correlation |
| $-0.8 \leq r < -0.6$ | Negative | Strong correlation |
| $-1.0 \leq r < -0.8$ | Negative | Very strong correlation |

The computation of correlations is performed using the Correlation Matrix algorithm, which first calculates the covariance matrix for all attribute pairs and then normalizes it using the standard deviation of each attribute to eliminate scale bias. Mathematically, the correlation coefficient between two attributes is expressed as:

$$\mathrm{Corr}_{a_1,a_2} = \frac{\mathrm{Cov}_{a_1,a_2}}{\sqrt{\mathrm{Var}_{a_1} \cdot \mathrm{Var}_{a_2}}} \quad (1)$$

where $Corr_{a_1,a_2}$ denotes the correlation coefficient between attributes $a_1$ dan $a_2$ and $Cov_{a_1,a_2}$ is their covariance, and $Var_{a_1}$ and $Var_{a_2}$ are the variances of each attribute, respectively. This approach is particularly effective for numerical datasets, such as the Covid-19 data from Sumbawa Regency, where attribute values vary significantly in scale. The analysis outcomes provide a more accurate quantitative representation of the direction and strength of inter-attribute relationships, serving as a solid foundation for data-driven policymaking at the regional government level.

## 4 Research Methods

This section presents the research findings and their corresponding discussion based on the CRISP-DM approach, from problem understanding and data collection and exploration to result evaluation. Data is visualized through graphs, tables, and narrative descriptions to enhance interpretability and facilitate comprehensive analysis.

### 4.1 Data Understanding

The study begins by examining the challenges faced by the Sumbawa Regency Government in mitigating the Covid-19 pandemic through various public health policies, including community

activity restrictions, mobility restrictions, and the establishment of emergency response posts. Despite these interventions, virus transmission could not be consistently suppressed and persisted over an extended period. The main focus of this research is to identify statistical relationships and influences among attributes in the Covid-19 dataset for Sumbawa Regency, which consists of numerical variables with heterogeneous value ranges that are not easily analyzed using conventional statistical methods. To address this, the Correlation Matrix algorithm is applied as a data mining technique to quantify the strength and direction of inter-attribute relationships. The results of this analysis are expected to provide an evidence-based foundation for formulating more effective, data-driven public health policies at the local level.

## 4.2 Data Understanding

Data was collected from the official website of the Sumbawa Regency Government (https://covid19.sumbawakab.go.id/data) using the =IMPORTHTML function in Google Spreadsheets. Data extraction was performed iteratively by adjusting dates, covering the period from August 4, 2020, to September 30, 2021, resulting in a total of 10,573 records representing daily Covid-19 cases across all subdistricts. The attribute structure of the dataset is outlined in Table 2.

Table 2 Interpretation of Correlation Coefficient

| Attribute | Sub-Attribute | Description |
|---|---|---|
| TANGGAL | – | Date of data recording |
| KECAMATAN | – | Name of subdistrict in Sumbawa Regency |
| TOTAL | – | Total cases per subdistrict |
| JUMLAH | – | Accumulated cases per date |
| PP | JUMLAH PP | Number of travelers |
| | DISCARDE | Travelers completing quarantine |
| | MASIH KARANTINA | Travelers still under quarantine |
| KONTAK ERAT | JUMLAH KONTAK | Number of close contacts |
| | DISCARDE | Close contacts completing observation |
| | MASIH KARANTINA | Close contacts still under quarantine |
| SUSPEK | JUMLAH SUSPEK | Number of suspected cases |
| | DISCARDE | Suspected cases cleared |
| | MENINGGAL | Suspected cases resulting in death |
| | MASIH ISOLASI | Suspected cases still under isolation |
| PROBABLE | JUMLAH PROBABLE | Number of probable cases |
| | DISCARDE | Probable cases cleared |
| | MENINGGAL | Probable cases resulting in death |
| | MASIH ISOLASI | Probable cases still under isolation |
| KONFIRMASI | JUMLAH POSITIF | Number of confirmed positive cases |
| | SEMBUH | Number of recovered cases |
| | MENINGGAL DUNIA | Number of confirmed deaths |
| | MASIH POSITIF | Number of active confirmed positive cases |

The dataset exhibits a complex structure with multiple numerical attributes varying across both subdistricts and dates. This data serves as the foundation for the subsequent stages of preparation and modeling for correlation analysis.

### 4.3 Data Preparation

The Data Preparation stage ensured the dataset was clean, relevant, and ready for modeling. In this study, Microsoft Excel was used to execute three primary sub-processes: data transformation, data reduction, and data cleaning. During data transformation, attribute names were simplified and made more interpretable. For example, the attribute "PP" with the sub-attribute "Discarde" was renamed "PP-DISCARDE," and "Suspek-Meninggal" was changed to "SUSPEK-MENINGGAL." Additionally, only total daily case counts for relevant attributes were retained to streamline the analysis. Subsequently, data reduction removed irrelevant or empty attributes such as "Tanggal," "Kecamatan," "Total kasus per kecamatan," and "PROBABLE-MASIH ISOLASI." This step reduced dataset complexity without discarding critical information. In the data cleaning stage, an additional attribute, "DATE," was introduced as a time index to record daily case information, derived from the attribute "Total Semua Kasus." The final outcome of this process was a cleaned dataset (DataSet Clean) comprising 405 records with 13 analytical attributes and 1 ID attribute in the form of a date. The structure of the cleaned dataset is provided in Table 3, and the dataset characteristics are shown in Table 4. The prepared dataset meets the required quality criteria for data mining and is ready to be modeled using the Correlation Matrix algorithm to systematically analyze the interrelationships among attributes.

Table 3 Attributes of the Cleaned Dataset

| Attribute (ID) | Description | Data Type |
|---|---|---|
| DATE | Date of case recording | Date |
| TOTAL-KASUS | Total daily cases | Integer |
| JUMLAH-PP | Total number of travelers | Integer |
| PP-DISCARDE | Travelers completing quarantine | Integer |
| PP-MASIH KARANTINA | Travelers under quarantine | Integer |
| JUMLAH KONTAK ERAT | Total close contact cases | Integer |
| KONTAK ERAT-DISCARDE | Close contacts completing observation | Integer |
| KONTAK ERAT-MASIH KARANTINA | Close contacts under quarantine | Integer |
| JUMLAH SUSPEK | Total suspected cases | Integer |
| SUSPEK-DISCARDE | Cleared suspected cases | Integer |
| SUSPEK-MENINGGAL | Deaths among suspected cases | Integer |
| SUSPEK-MASIH ISOLASI | Suspected cases under isolation | Integer |
| JUMLAH PROBABLE | Total probable cases | Integer |
| PROBABLE-DISCARDE | Cleared probable cases | Integer |
| PROBABLE-MENINGGAL | Deaths among probable cases | Integer |
| KONFIRMASI-JUMLAH POSITIF | Total confirmed positive cases | Integer |
| KONFIRMASI-SEMBUH | Total recovered cases | Integer |
| KONFIRMASI-MENINGGAL DUNIA | Total confirmed deaths | Integer |
| KONFIRMASI-MASIH POSITIF | Total active confirmed cases | Integer |

Table 4. Dataset Characteristics (Cleaned Dataset, n = 405)

| Item | Value / Description |
|---|---|
| Data source | Official website of the Sumbawa Regency Government (https://covid19.sumbawakab.go.id/data) |
| Period | 4 August 2020 – 30 September 2021 |
| Initial number of records | 10,573 |
| Number of records after cleaning | 405 |
| ID/Index | DATE (Date) |
| Number of analytical features (numeric) | 12 features (all integer) |
| Features & value ranges (min–max) | PP-DISCARDE: 8,935 – 21,398 • PP-MASIH KARANTINA: 0 – 192 • KONTAK ERAT-DISCARDE: 802 – 11,134 • KONTAK ERAT-MASIH KARANTINA: 1 – 467 • SUSPEK-DISCARDE: 587 – 3,733 • SUSPEK-MENINGGAL: 0 – 12 • SUSPEK-MASIH ISOLASI: 0 – 189 • PROBABLE-DISCARDE: 0 – 10 • PROBABLE-MENINGGAL: 0 – 8 • KONFIRMASI-SEMBUH: 56 – 3,080 • KONFIRMASI-MENINGGAL DUNIA: 2 – 186 • KONFIRMASI-MASIH POSITIF: 4 – 315 |
| Transformation/standardization | Attribute name simplification (e.g., "PP–Discarde" → "PP-DISCARDE"), removal of invalid entries, and date format harmonization (DATE). |
| Added statistical graphs | (1) Time-series trend plots for key attributes (e.g., KONFIRMASI-MENINGGAL DUNIA, KONTAK ERAT-DISCARDE, SUSPEK-DISCARDE). (2) Distribution plots (histogram/boxplot) to show dispersion and outliers of key attributes. |

## 4.4 Correlation Modeling

At this stage, the researcher used RapidMiner 9.9 to conduct a correlation analysis using the Correlation Matrix algorithm. This method was used to measure the strength and direction of relationships among numerical attributes in the dataset. The selection of RapidMiner was based on its comprehensive support for the entire data mining process, including data preparation, validation, visualization, and model optimization. The correlation modeling process is illustrated in Figure 2.

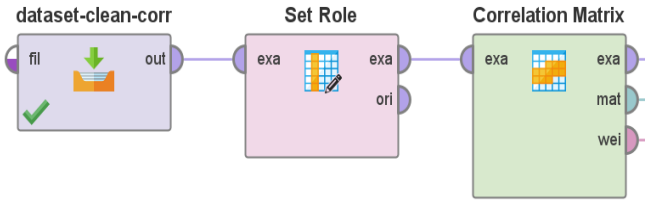



Figure 2 Correlation Modeling Process

The correlation modeling produced a correlation matrix of the COVID-19 dataset for Sumbawa Regency, showing the correlation coefficients between attributes. The complete results are presented in Figure 3. Although several attributes exhibit very strong correlations, it is important to emphasize that correlation does not imply causation. The very strong positive correlations between TOTAL-KASUS and KONTAK ERAT-DISCARDE (0.99), SUSPEK-DISCARDE (0.98), and KONFIRMASI-MENINGGAL DUNIA (0.97) indicate that the recorded increase in total cases closely aligns with increases in the major epidemiological categories captured in the dataset. This pattern may

also reflect structural dependence among variables, as the total-case indicator is naturally related to the accumulation and reporting dynamics of its constituent categories.

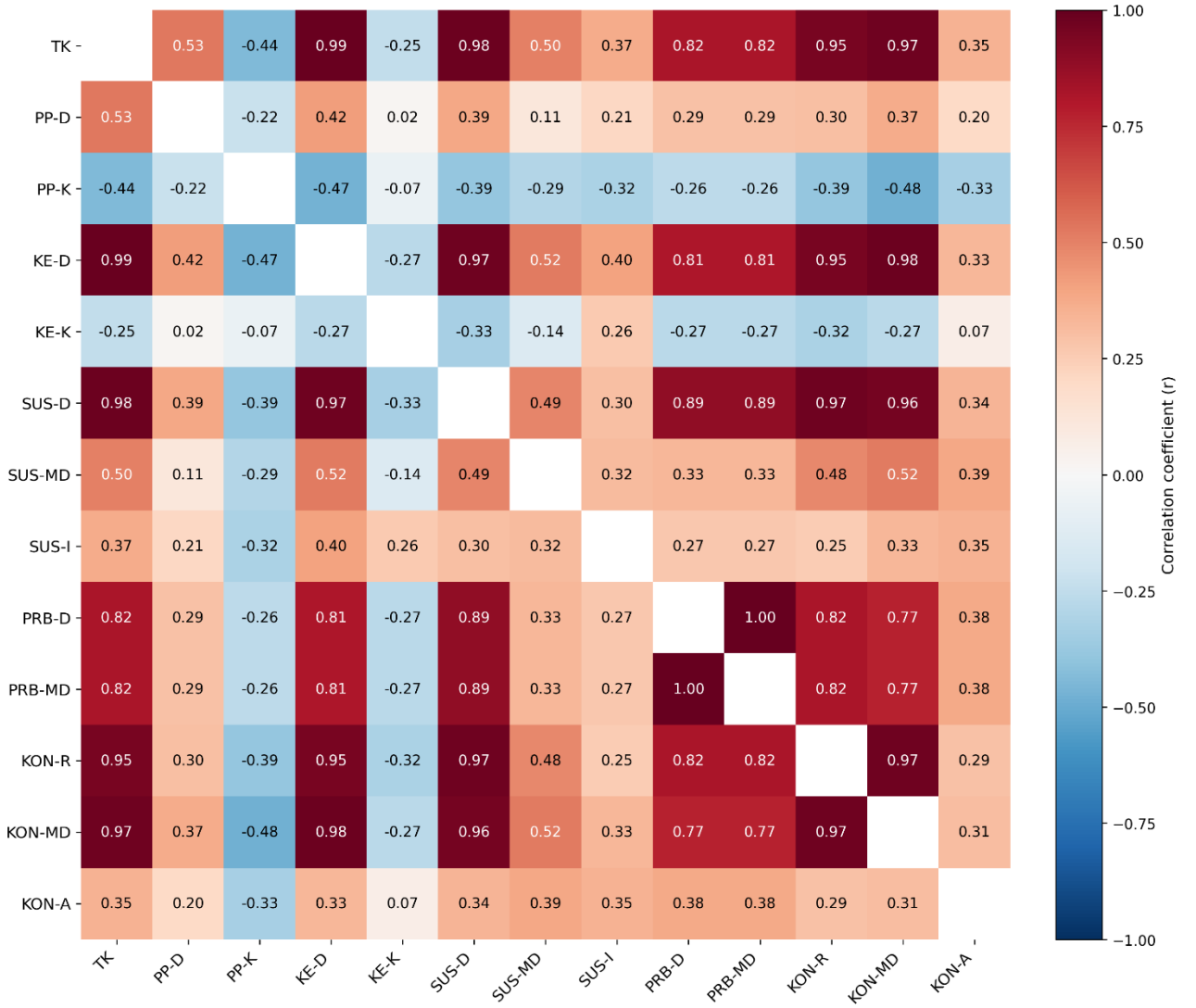



Figure 3 Correlation matrix of Attributes in the COVID-19 Dataset of Sumbawa Regency. Abbrevations: TK = TOTAL-KASUS; PP-D = PP-DISCARDE; PP-K = PP-MASIH KARANTINA; KE-D = KONTAK ERAT-DISCARDE; KE-K = KONTAK ERAT-MASIH KARANTINA; SUS-D = SUSPEK-DISCARDE; SUS-MD = SUSPEK-MENINGGAL; SUS-I = SUSPEK-MASIH ISOLASI; PRB-D = PROBABLE-DISCARDE; PRB-MD = PROBABLE-MENINGGAL; KON-R = KONFIRMASI-SEMBUH; KON-MD = KONFIRMASI-MENINGGAL DUNIA; KON-A = KONFIRMASI-MASIH POSITIF

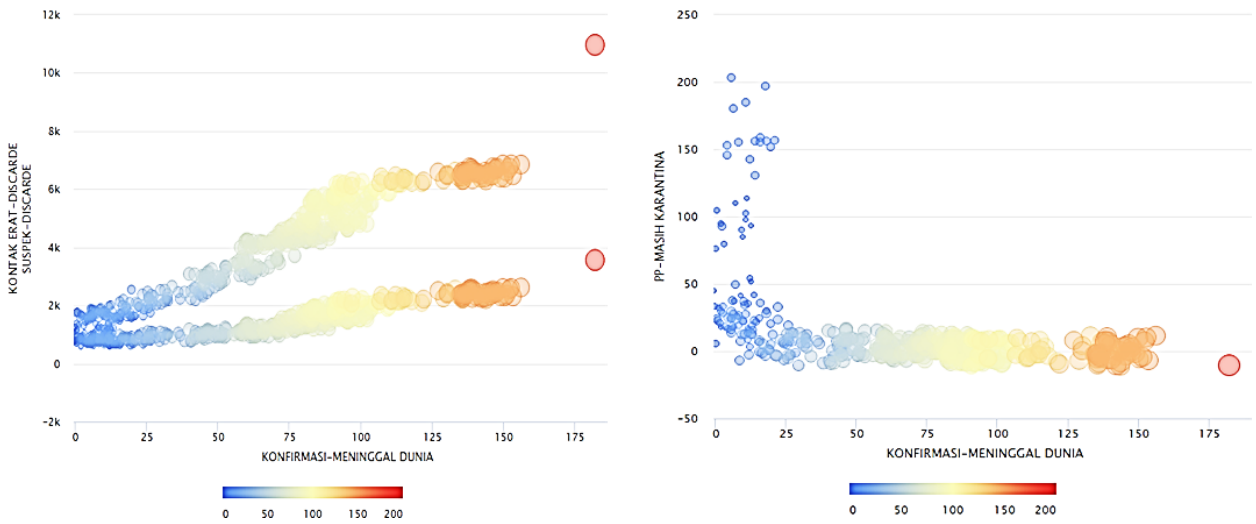
Conversely, PP-MASIH KARANTINA shows a negative correlation with KONFIRMASI-MENINGGAL DUNIA (−0.48), suggesting a statistical association in which higher numbers of travelers remaining in quarantine tend to coincide with lower confirmed mortality counts. This finding provides quantitative support for the hypothesis that quarantine may be associated with a reduced risk of severe outcomes; however, external factors, such as policy shifts, healthcare capacity, testing intensity, and temporal dynamics, were not modeled in this study. Therefore, the correlation results are interpreted as statistical associations to support monitoring priorities and policy evaluation, rather than as direct causal mechanisms.

Correlation analysis is exploratory and does not imply causality. In time-indexed epidemiological data, apparent associations may be driven by shared temporal trends, reporting

practices, or unmodeled confounders, such as policy changes, testing intensity, and healthcare capacity. Moreover, examining many attribute pairs may increase the chance of spurious findings. Therefore, the correlation matrix results should be interpreted as statistical associations for monitoring prioritization and hypothesis generation, rather than as causal evidence. Near-perfect correlations of almost 0.99 may arise from structural dependencies among indicators, such as shared accumulation and reporting mechanisms, overlapping category definitions, or partially redundant information across variables. Accordingly, these high coefficients are interpreted as strong statistical alignment and a motivation for feature redundancy checks, rather than causal evidence.

## 4.5 Evaluation

Based on the correlation modeling stage, several patterns of attribute relationships were identified. Figure 4(a) illustrates that "KONFIRMASI-MENINGGAL DUNIA" has a very strong positive correlation with "KONTAK ERAT-DISCARDE" and "SUSPEK-DISCARDE", indicating that an increase in cases declared as "discarded" tends to be followed by a rise in mortality cases. The color shift from blue to red on the graph visually supports this pattern. Conversely, Figure 4(b) shows a negative correlation between "PP-MASIH KARANTINA" and "KONFIRMASI-MENINGGAL DUNIA", suggesting that an increase in the number of quarantined travelers is associated with a decline in mortality rates, albeit with moderate correlation strength.




(a) (b)

Figure 4 Scatter plot visualization of correlation patterns involving KONFIRMASI-MENINGGAL DUNIA. (a) Strong positive correlation with KONTAK ERAT-DISCARDE and SUSPEK-DISCARDE; (b) Negative correlation with PP-MASIH KARANTINA, indicating an inverse relationship between confirmed deaths and quarantine cases.

Similarly, Tables 5 and 6 highlight the attributes most strongly correlated with "TOTAL-KASUS", with the highest correlations observed with "KONTAK ERAT-DISCARDE" (0.987), "SUSPEK-DISCARDE" (0.975), "KONFIRMASI-MENINGGAL DUNIA" (0.974), "KONFIRMASI-SEMBUH" (0.952), "PROBABLE-DISCARDE" (0.823), and "PROBABLE-MENINGGAL" (0.823). Negative correlation values are seen for "PP-MASIH KARANTINA" (-0.44), while attributes such as "SUSPEK-MASIH ISOLASI" or "KONFIRMASI-MASIH POSITIF" show no significant correlation.

Table 5. Correlation with PROBABLE-MENINGGAL and SUSPEK-MENINGGAL

| **Correlation with PROBABLE-MENINGGAL** | | |
|---|---|---|
| ***Related Attribute*** | ***Correlation Value*** | ***Interpretation*** |
| PROBABLE-DISCARDE | 0.999 | (+) Very Strong Correlation |
| SUSPEK-DISCARDE | 0.885 | (+) Very Strong Correlation |
| KONTAK ERAT-DISCARDE | 0.814 | (+) Very Strong Correlation |
| **Correlation with SUSPEK-MENINGGAL** | | |
| ***Related Attribute*** | ***Correlation Value*** | ***Interpretation*** |
| KONTAK ERAT-DISCARDE | 0.522 | (+) Some Correlation |
| SUSPEK-DISCARDE | 0.494 | (+) Some Correlation |

Table 6. Correlation between TOTAL-KASUS and Other Attributes

| **Related Attribute** | **Correlation Value** | **Interpretation** |
|---|---|---|
| KONTAK ERAT-DISCARDE | 0.987 | (+) Very Strong Correlation |
| SUSPEK-DISCARDE | 0.975 | (+) Very Strong Correlation |
| KONFIRMASI-MENINGGAL DUNIA | 0.974 | (+) Very Strong Correlation |
| KONFIRMASI-SEMBUH | 0.952 | (+) Very Strong Correlation |
| PROBABLE-DISCARDE | 0.823 | (+) Very Strong Correlation |
| PROBABLE-MENINGGAL | 0.823 | (+) Very Strong Correlation |
| PP-DISCARDE | 0.526 | (+) Some Correlation |
| SUSPEK-MENINGGAL | 0.498 | (+) Some Correlation |
| SUSPEK-MASIH ISOLASI | 0.373 | (+) No Correlation |
| KONFIRMASI-MASIH POSITIF | 0.353 | (+) No Correlation |
| KONTAK ERAT-MASIH KARANTINA | -0.25 | (–) No Correlation |
| PP-MASIH KARANTINA | -0.44 | (–) Some Correlation |

While the correlation matrix provides a clear overview of linear associations among attributes, correlation analysis remains exploratory and does not imply causation. The observed relationships may be influenced by shared reporting mechanisms (structural dependence among epidemiological indicators), unmodeled external factors (policy shifts, testing intensity, and healthcare capacity), and temporal dynamics in time-indexed data. Therefore, the results are interpreted as statistical associations to support monitoring priorities and policy evaluation rather than as direct causal mechanisms. To strengthen interpretation beyond correlation magnitude alone, future extensions may incorporate statistical significance reporting for correlation coefficients, such as p-values and/or confidence intervals, and time-lagged cross-correlation to capture delayed epidemiological effects. In addition, regression-based analysis can be used to examine multivariate interactions and reduce confounding effects when evaluating key relationships.

## 4.6 Deployment

Based on the evaluation results, several policy recommendations are proposed for the Sumbawa Regency Government. Enhanced monitoring of “KONTAK ERAT-DISCARDE” and “SUSPEK-DISCARDE” cases is crucial, as their strong correlation with mortality indicates that a discarded status does not necessarily indicate safety. Therefore, extending the monitoring period is advisable. Strengthening quarantine protocols for travelers is also essential, given the negative correlation between “PP-MASIH KARANTINA” and mortality cases, which demonstrates the effectiveness of quarantine measures; improvements in facilities, supervision, and public education on quarantine are

recommended. Additionally, preventive measures for high-risk probable cases are necessary, as the near-perfect correlation of 0.999 between "PROBABLE-DISCARDE" and "PROBABLE-MENINGGAL" suggests that probable cases pose a high fatality risk even after monitoring concludes, requiring more rigorous preventive handling. Finally, prioritizing resource allocation toward attributes most strongly influencing "TOTAL-KASUS", namely "KONTAK ERAT-DISCARDE", "SUSPEK-DISCARDE", "KONFIRMASI-MENINGGAL DUNIA", "KONFIRMASI-SEMBUH", "PROBABLE-DISCARDE", and "PROBABLE-MENINGGAL", is recommended. This data-driven approach is expected to curb the spread of COVID-19 more effectively, reduce mortality rates, and support evidence-based policymaking tailored to the local context of Sumbawa Regency.

## 5 Conclusion

Based on the analysis conducted using the Correlation Matrix algorithm, it can be concluded that there are very strong correlations between specific attributes in the COVID-19 dataset of Sumbawa Regency, particularly "KONTAK ERAT-DISCARDE," "SUSPEK-DISCARDE," and "KONFIRMASI-MENINGGAL DUNIA," which significantly contribute to the increase in total cases. Conversely, a notable negative correlation is observed between "PP-MASIH KARANTINA" and mortality cases, indicating that strict quarantine measures can effectively reduce death rates. Furthermore, "PROBABLE-DISCARDE" exhibits an almost perfect correlation with "PROBABLE-MENINGGAL," suggesting the need for a more thorough evaluation of the classification process for probable cases. Based on these findings, it is recommended that the Sumbawa Regency Government strengthen monitoring processes for discarded cases, enforce stricter quarantine measures for travelers, and prioritize resource allocation toward attributes with the greatest impact on case surges. These actions are expected to enhance data-driven policymaking in managing the pandemic and help mitigate similar impacts in future public health crises.

Future work will extend the present analysis by incorporating Spearman's rank correlation to complement Pearson correlation and strengthen robustness against monotonic or potentially non-linear relationships. In addition, time-lagged and cross-correlation analyses with a seven-day interval will be conducted to examine possible delayed epidemiological effects. A concise top-k summary of the most influential attributes, particularly those most strongly associated with TOTAL-KASUS, will also be provided, while the complete correlation matrix will be retained to ensure transparency and reproducibility.

## References

[1] Y. R. T. Hutagaol, R. P. P. Sinurat, and S. M. Shalahuddin, "Strategi Penguatan Keuangan Negara Dalam Menghadapi Ancaman Resesi Global 2023 Melalui Green Economy [Strategy to Strengthen State Finances in Facing the Threat of a Global Recession in 2023 Through a Green Economy]," *Jurnal Pajak dan Keuangan Negara (PKN)*, vol. 4, no. 1S, pp. 378–385, Dec. 2022, doi: 10.31092/jpkn.v4i1S.1911.

[2] S. Sumarno, K. Karsim, D. Dwiyanto, F. Ekobelawati, and F. Christian, "Kerja jarak jauh dan produktivitas karyawan: Mengevaluasi dampak jangka panjang dari telecommuting pasca Covid-19 [Remote work and employee productivity: Evaluating the long-term impact of telecommuting post-Covid-19]," *Journal of Management and Digital Business*, vol. 4, no. 3, pp. 775–786, Dec. 2024, doi: 10.53088/jmdb.v4i3.1265.

[3] M. H. D. R. Yahya and I. Y. N. Nanda, "Evaluasi Program Bantual Sosial Tunai (BST) Pada Masa Pandemi COVID-19 (Studi kasus; Kelurahan Langgini Kecamatan Bangkinang Kota Kabupaten Kampar) [Evaluation of the Cash Social Assistance Program During the COVID-19 Pandemic (Case Study: Langgini Village, Bangkinang District, Kampar Regency)]," *SUMUR-Jurnal Sosial Humaniora*, vol. 3, no. 1, pp. 27–34, Feb. 2025, doi: 10.58794/sumur.v3i1.1341.

[4] F. Sari, "Implementasi Data Mining Dalam Menganalisis Tingkat Kepuasan Pelanggan Menggunakan Metode Rough Set [Implementation of Data Mining in Analyzing Customer Satisfaction Levels Using the Rough Set Method]," *Jurnal Buana Informatika*, vol. 8, no. 1, Jan. 2017, doi: 10.24002/jbi.v8i1.1071.

[5] L. F. Azmi and L. Zahrotun, "Implementasi Data Mining untuk Estimasi Produksi Cabai menggunakan Metode Exponential Smoothing [Implementation of Data Mining for Chili Production Estimation using the Exponential Smoothing Method]," *Jurnal Buana Informatika*, vol. 15, no. 01, pp. 59–68, Apr. 2024, doi: 10.24002/jbi.v15i1.8333.

[6] A. A. Alya Putri and S. A. Rahmah, "Implementasi Data Mining dengan Algoritma K-Means Clustering untuk Analisis Bisnis pada Perusahaan Asuransi [Implementation of Data Mining with K-Means Clustering Algorithm for Business Analysis in Insurance Companies]," *Djtechno: Jurnal Teknologi Informasi*, vol. 5, no. 1, pp. 139–152, Apr. 2024, doi: 10.46576/djtechno.v5i1.4537.

[7] S. Ma, Y. Huang, Y. Liu, H. Liu, Y. Chen, J. Wang, and J. Xu, "Big data-driven correlation analysis based on clustering for energy-intensive manufacturing industries," *Appl Energy*, vol. 349, p. 121608, Nov. 2023, doi: 10.1016/j.apenergy.2023.121608.

[8] L. Xu, Y. Wang, L. Mo, Y. Tang, F. Wang, and C. Li, "The research progress and prospect of data mining methods on corrosion prediction of oil and gas pipelines," *Eng Fail Anal*, vol. 144, p. 106951, Feb. 2023, doi: 10.1016/j.engfailanal.2022.106951.

[9] M. Shantal, Z. Othman, and A. A. Bakar, "A Novel Approach for Data Feature Weighting Using Correlation Coefficients and Min–Max Normalization," *Symmetry (Basel)*, vol. 15, no. 12, p. 2185, Dec. 2023, doi: 10.3390/sym15122185.

[10] N. Nur, S. Situju, and S. Aulia Rachmini, "Data Mining dan Manajemen Pengetahuan [Data Mining and Knowledge Management]," 2024. [Online]. Available: https://www.researchgate.net/publication/388498331

[11] D.-R. Nichita, M. Dima, L. Boboc, and M.-G. Hâncean, "Data analysis evidence beyond correlation of a possible causal impact of weather on the COVID-19 spread, mediated by human mobility," *Sci Rep*, vol. 14, no. 1, p. 17782, Aug. 2024, doi: 10.1038/s41598-024-67918-6.

[12] N. K. Bergman and R. Fishman, "Correlations of mobility and Covid-19 transmission in global data," *PLoS One*, vol. 18, no. 7, p. e0279484, Jul. 2023, doi: 10.1371/journal.pone.0279484.

[13] D. Conesa, C. L. Roja, T. Gullon, A. T. Campo, C. Prats, E. A. Lacalle, and B. Echebarria, "A mixture of mobility and meteorological data provides a high correlation with COVID-19 growth in an infection-naive population: a study for Spanish provinces," *Front Public Health*, vol. 12, Mar. 2024, doi: 10.3389/fpubh.2024.1288531.

[14] E. Cleary *et al.*, "Comparing lagged impacts of mobility changes and environmental factors on COVID-19 waves in rural and urban India: A Bayesian spatiotemporal modelling study," *PLOS Global Public Health*, vol. 5, no. 4, p. e0003431, Apr. 2025, doi: 10.1371/journal.pgph.0003431.

[15] U. Kannengiesser and J. S. Gero, "Modelling the Design of Models: An Example Using CRISP-DM," *Proceedings of the Design Society*, vol. 3, pp. 2705–2714, Jul. 2023, doi: 10.1017/pds.2023.271.

[16] D. Pratmanto, F. F. D. Imaniawan, and V. Maarif, "Analisis Sentimen Pada Ulasan Pengguna Aplikasi Identitas Kependudukan Digital Dengan Metode Naive Bayes Dan K-Nearest [Sentiment Analysis on User Reviews of Digital Population Identity Applications Using Naive Bayes and K-Nearest Methods]," *Computatio : Journal of Computer Science and Information Systems*, vol. 7, no. 2, pp. 155–166, Dec. 2023, doi: 10.24912/computatio.v7i2.26322.