TIFDA JOURNAL
Technology Information and Data Analytic
Journal Website : http://jeptira.id/index.php/tifda
e-ISSN : 3064-0660

Decision Tree Regression Approach to Modeling Dengue, Tuberculosis, and Diarrhea Case Numbers

Muhammad Dzaki Zahirsyah[1*], Timor Setiyaningsih[1*]
[1] Department of Information Technology, Faculty of Engineering, Darma Persada University
[1] Jl. Taman Malaka Selatan No.8 Pd. Kelapa, Kec. Duren Sawit, Kota Jakarta Timur,
Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, 13450, Indonesia
* ningsihtiya.unsada@gmail.com

Abstract — The increasing incidence of Dengue Hemorrhagic Fever (DHF), Tuberculosis (TB), and Diarrhea in a district area highlights the urgent need for a data-driven prediction system to support public health policy. This study develops a predictive model of case numbers at the sub-district level using the Decision Tree Regression algorithm within the CRISP-DM methodology. Secondary data from 2020-2023 were utilized, including disease case records (Health Office), demographic data (BPS), and environmental data (BMKG). The system was implemented as a web-based application built with PHP and Python/Flask, enabling dataset management, model retraining, and interactive visualization of predictions, complemented by risk classification and recommended interventions. Experimental results demonstrate high predictive accuracy, with R^2 values of 0.9130 for TB, 0.8805 for DHF, and 0.8228 for Diarrhea. Overall, the proposed system serves as an objective and measurable decision-support tool, assisting the District Health Office in formulating preventive policies more rapidly and effectively.

Keywords – Decision Tree Regression, Prediction, Health

I. INTRODUCTION

Public health is a fundamental pillar of sustainable development. In today's digital era, information technology–based approaches have become increasingly vital in supporting decision-making within the health sector. Among these, data mining and machine learning are rapidly advancing, offering the ability to transform large-scale datasets into actionable insights for disease prediction, pattern detection, and strategic planning.

Communicable diseases such as Dengue Hemorrhagic Fever (DHF), Tuberculosis (TB), and Diarrhea remain serious threats across many regions in Indonesia, including Bekasi District. These diseases are closely linked to environmental factors, population density, sanitation levels, and the quality of public health services. Their impact on community productivity is substantial, often peaking during seasonal transitions such as the rainy or dry season. The rising incidence of these diseases in recent years underscores the urgent need for preventive measures and predictive health planning capable of identifying disease patterns and trends to inform data-driven decision-making. The introduction consists of the research background, current research and theoretical basis that supports the research you are doing. Make sure all are related to each other in order to get research results that lead to current scientific contributions.

Previous studies have extensively explored disease prediction using machine learning. One study demonstrated that the Decision Tree algorithm achieved the highest accuracy (83%) in predicting Dengue Hemorrhagic Fever (DHF) cases based on sociodemographic variables [1]. However, this research was limited to severity classification rather than predicting the actual number of cases. Another study employed deep learning models to forecast DHF transmission in Jakarta using climatic factors, yielding very high accuracy. Nevertheless, that study did not incorporate local demographic variables such as age distribution and gender ratio.

Building on these gaps, the present study introduces a different approach by applying Decision Tree Regression to predict the actual number of cases

per year at the sub-district level in Bekasi District [2]. This approach aligns with the ongoing digital transformation in public health, which emphasizes data-driven decision-making. According to the Ministry of Health, the national health information system encourages the integration of spatial and temporal data for disease prediction and outbreak-prone area planning. Thus, this research contributes to the development of a decision-support system that can be utilized by the Bekasi District Health Office in designing effective disease control strategies.

By leveraging Decision Tree Regression, this study aims to construct predictive models for DHF, TB, and Diarrhea cases using data from 2020–2023 across Bekasi sub-districts. Decision Tree Regression is particularly suitable because it can handle numerical variables while remaining interpretable, facilitating communication of results to local government stakeholders [3]. Compared to more complex algorithms such as Random Forest or Artificial Neural Networks (ANN) [4], Decision Trees generate simple rule-based models that are logically traceable. Moreover, they can identify the most influential variables, including population density, patient age, and gender ratio. The choice of this algorithm reflects considerations of efficiency, accuracy, and interpretability in the context of public health decision-making.

The predictive outcomes of this model are expected to provide valuable insights for local governments, enabling more precise preventive policies such as the allocation of health facilities, promotion of hygiene practices (PHBS), vector control, medicine distribution, and deployment of medical personnel. Unlike previous studies, the strength of this research lies in its localized scope, the inclusion of demographic variables, and the provision of quantitative predictive outputs rather than mere risk classification. Furthermore, this study supports the broader agenda of digitalizing health planning through sustainable, region-specific data utilization.

II. METHODOLOGY

A. Data Sources and Collection

This study employs secondary data gathered from multiple official institutions to ensure reliability and validity. The dataset encompasses 23 sub-districts in Bekasi District over the period 2020–2023, providing both temporal and spatial coverage for predictive modeling. The dependent variables consist of the monthly number of reported cases for Dengue Hemorrhagic Fever (DHF), Tuberculosis (TB), and Diarrhea, obtained from the Bekasi District Health Office. These variables serve as the primary targets for prediction.

The independent variables include demographic and environmental features, which were selected based on their established relevance to communicable disease transmission. Demographic data from the Central Bureau of Statistics (BPS), covering population size and density. Environmental data from the Meteorology, Climatology, and Geophysics Agency (BMKG), including rainfall, average temperature, and humidity levels.

The integration of these diverse datasets enables a multifactorial analysis, capturing both human and environmental determinants of disease incidence. This comprehensive data collection strategy strengthens the predictive capacity of the model and ensures that the results are contextually relevant for local health planning Maintaining Specification Integrity

B. Data Preprocessing

To ensure the dataset was suitable for predictive modeling, several preprocessing procedures were applied:

a) Data Cleaning: Missing values within the dataset were handled by replacing them with zero (0), under the assumption that unrecorded data represented no reported cases. This approach maintained dataset completeness while avoiding bias from imputation.
b) Data Splitting: The cleaned and transformed dataset was divided into two subsets: 80% training data for model construction and 20% testing data for performance evaluation. This split ensured that the model was assessed on unseen data, providing a reliable measure of generalization capability.
c) Data Transformation: The categorical variable representing sub-districts was converted into numerical format using One-Hot Encoding. This transformation expanded the single sub-district column into 23 binary columns, each corresponding to one sub-district. This allowed the algorithm to treat each sub-district as an independent feature without implying ordinal relationships.

C. Modeling Technique

The modeling technique applied in this study is Decision Tree Regression, chosen for its ability to capture non-linear relationships among variables while producing models that remain highly interpretable. The algorithm constructs a tree-like structure to predict numerical target values, enabling clear visualization of decision rules.

The tree-building process is conducted recursively, partitioning the dataset into subsets based on feature values. At each node, the optimal split is determined using the Variance Reduction (VR) criterion, which aims to minimize the Mean Squared Error (MSE) across the resulting child nodes.

To ensure disease-specific accuracy, three separate models were developed one each for Dengue

Hemorrhagic Fever (DHF), Tuberculosis (TB), and Diarrhea. This separation allows the model to adapt to the unique characteristics and influencing factors of each disease. The implementation was carried out using the Python programming language with the Scikit-learn machine learning library, which provides robust tools for model construction, training, and evaluation.

D. Evaluation Metrics

To assess the performance and accuracy of the predictive models, three standard regression evaluation metrics were employed:

a) R-squared (R^2): Measures the proportion of variance in the dependent variable explained by the model. Values range from 0 to 1, with values closer to 1 indicating stronger model performance.
b) Root Mean Squared Error (RMSE): Represents the square root of the average squared differences between actual and predicted values. RMSE provides an estimate of prediction error in the same unit as the target variable, where smaller values indicate higher accuracy.
c) Mean Absolute Error (MAE): Calculates the average of the absolute differences between actual and predicted values. MAE offers a straightforward measure of prediction error, regardless of direction (positive or negative).

III. RESULTS AND DISCUSSION

This section presents the outcomes of the system design, implementation, and testing of the disease prediction framework. The discussion begins with the system architecture design using UML, followed by the interface implementation, functional testing scenarios, and finally the quantitative evaluation of the predictive model's performance.

A. Use Case Diagram Design

The workflow and user interactions within the system were designed using Unified Modeling Language (UML). The Use Case Diagram was employed to map the core functionalities and the interactions between users (actors) and the system [9][10].

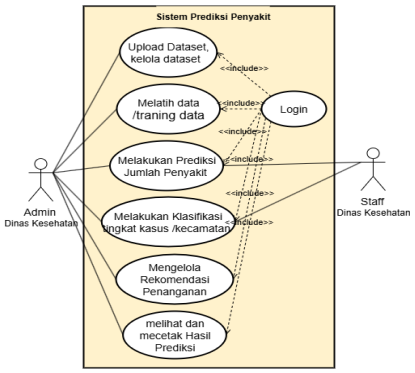

Fig.1. Use case Diagram Predicting Model System

As illustrated in Fig. 1, the system defines two primary actors, (a) Admin: Holds full access rights, including dataset management and model training. (b) Health Office Staff: Has access to perform predictions and view historical records. This design ensures a clear separation of roles and responsibilities, enabling secure and efficient system operation while supporting both administrative control and practical use in public health decision-making.

B. Implementation Web Aplication Interface

The case prediction page is accessible to both user roles: the Health Office Administrator and the Health Office Staff. Within this menu, users can generate disease case predictions by providing specific input parameters. To perform a prediction, the user must enter the year to be forecasted, select the month, and choose the relevant sub-district. This design ensures that both administrative and operational users can interact with the system effectively, supporting data-driven decision-making in public health management.

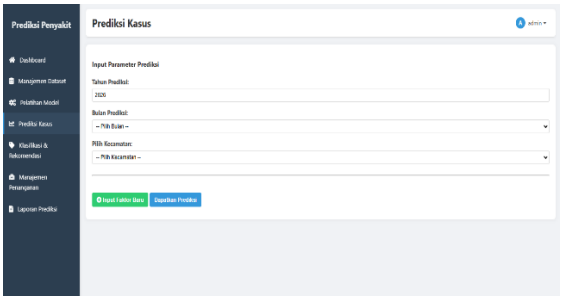

Fig.2. The Case Prediction Page

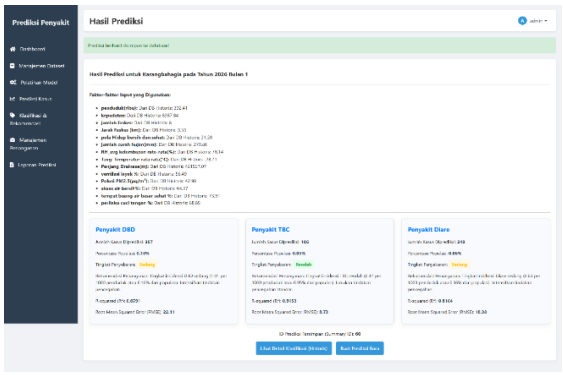

Fig.3. The Prediction Results Page

The prediction results page displays the outcomes generated by the Decision Tree Regression model. This page presents the predicted number of disease cases (DHF, TB, and Diarrhea), along with the estimated level of disease spread and corresponding recommendations for handling and preventive measures. By combining quantitative predictions with contextual risk assessment, the system provides actionable insights that can guide the Health Office in resource allocation, intervention planning, and public health education initiatives.

C. *Model Performance Evaluation and Feature Analysis*

Quantitative evaluation was conducted to measure the accuracy of the trained Decision Tree Regression models. Testing was performed on 20% of the dataset that was not exposed to the model during training, ensuring an unbiased assessment of predictive capability. The performance of the models for each disease is summarized in Table 1.

Table 1. The performance of the models Decision Tree Regression

Model	R-squared (R^2)	RMSE	MAE
DHF	0.8805	22.1106	16.7341
TB	0.9130	8.7303	6.9167
Diarrhea	0.8228	18.3816	14.6230

The regression metrics obtained are consistent with expectations, demonstrating that the Decision Tree Regression algorithm effectively captures historical data patterns for all three diseases. Based on the results presented in Table 1, it can be concluded that all models exhibit very strong performance. Tuberculosis (TB) Model, achieved the highest performance with an R^2 of 0.9130, indicating that the model explains 91.3% of the variance in TB case data. The average prediction error is relatively small, with an RMSE of 8.7303, equivalent to approximately 9 cases.

Dengue Hemorrhagic Fever (DHF) Model also demonstrated robust predictive capability, with an R^2 of 0.8805, confirming its reliability in capturing disease incidence patterns. Diarrhea Model although it recorded the lowest R^2 among the three, the value of 0.8228 is still considered highly satisfactory, showing that the model remains dependable for prediction tasks.

Beyond accuracy evaluation, feature importance analysis was performed to identify the most influential predictors in each model. The Decision Tree Regression algorithm provides a measure of how much each feature contributes to reducing prediction error. Key findings include (a) Population density emerged as a dominant predictor for all three diseases, reflecting its strong correlation with transmission risk. (b)Rainfall and humidity were highly influential for DHF, consistent with the role of climate in mosquito breeding. (c) Average temperature showed significant impact on diarrhea incidence, likely due to its effect on waterborne pathogen survival. (d) Demographic factors such as population size and gender ratio contributed moderately across models, supporting their contextual relevance.

IV. CONCLUSION

The predictive model for estimating the number of Dengue Hemorrhagic Fever (DHF), Tuberculosis (TB), and Diarrhea cases was successfully developed using the Decision Tree Regression algorithm and integrated into a hybrid web application architecture (PHP & Python/Flask). The resulting models demonstrated very high accuracy and reliability, confirming their suitability for practical implementation.

Overall, these findings confirm that the Decision Tree Regression algorithm is highly effective in modeling and predicting disease case counts, providing accurate and interpretable results across different disease categories. Compared to manual approaches, the system provides significant optimization by delivering objective, rapid, and multi-factor quantitative predictions. This functionality establishes the system as a proactive decision-support tool for the District Health Office, enabling more effective planning and preventive policy formulation.

For future research, several avenues can be explored to enhance the robustness and applicability of predictive modeling in public health. First, alternative machine learning algorithms such as Random Forests, Gradient Boosting, or Neural Networks may be investigated to compare their performance against Decision Tree Regression and potentially achieve higher predictive accuracy. Second, hybrid modeling approaches that integrate statistical techniques with machine learning could provide more comprehensive insights by capturing both temporal dynamics and complex non-linear relationships. Third, extending the dataset to cover longer temporal periods beyond 2020–2023 would allow the models to account for seasonal variations and long-term epidemiological trends, thereby improving generalizability. Finally, incorporating socio-behavioral analysis, including factors such as human mobility, education levels, and awareness campaigns could enrich the models by highlighting behavioral determinants of disease spread, offering a more holistic perspective for public health decision-making.

REFERENCES

[1] R. G. Wardhana, G. Wang, and F. Sibuea, “Penerapan Machine Learning Dalam Prediksi Tingkat Kasus Penyakit Di Indonesia,” Journal of Information System Management (JOISM), vol. 5, no. 1, 2023. USA: Abbrev. of Publisher, year, ch. x, sec. x, pp. xxx–xxx.

[2] Kusumastuti, E. Anggita, M. H. Purnomo, and E. M. Yurniano, “Prediksi Penyebaran Kasus Demam Berdarah DKI Jakarta,” Jurnal Teknik Institut Teknologi Sepuluh Nopember (ITS), vol. 12, no. 3, 2023.

[3] I. S. I. Putri, R. S. Pradini, and M. Anshori, “Decision Tree Regression Untuk Prediksi Prevalensi Stunting di Provinsi Nusa Tenggara Timur,” Jurnal Teknologi Informatika dan Komputer, vol. 10, no. 2, pp. 413–427, Sep. 2024, doi: 10.37012/jtik.v10i2.2179.

[4] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. O’Reilly Media, 2019.

[5] Y.-T. Tsan et al., “The Prediction of Influenza-like Illness and Respiratory Disease Using LSTM and ARIMA,” Journal of Environmental Research and Public Health, vol. 19, p. 1858, 2022, doi: 10.3390/ijerph.

[6] X. Zhang et al., "Predicting influenza-like illness trends based on sentinel surveillance data in China from 2011 to 2019: A modelling and comparative study," Infect Dis Model, vol. 9, no. 3, pp. 816–827, Sep. 2024, doi: 10.1016/j.idm.2024.04.010.

[7] R. Rofiani, L. Oktaviani, D. Vernanda, and T. Hendriawan, "Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4.5," Jurnal Tekno Kompak, vol. 18, no. 1, 2024.

[8] B. Khusnul Khotimah and M. S. Rochman, "Model Peramalan Jumlah Penyakit Demam Berdarah Dengan Pendekatan Metode Fuzzy Linear REGRESSION (FLR)," Jurnal Ilmiah NERO, vol. 6, no. 1, p. 2021, 2021.

[9] D. Wira, T. Putra, and R. Andriani, "Unified Modelling Language (UML) dalam Perancangan Sistem Informasi Permohonan Pembayaran Restitusi SPPD," Jurnal TEKNOIF Teknik Informatika Institut Teknologi Padang, vol. 7, no. 1, 2019.

[10] A. Hendini, "Pemodelan Uml Sistem Informasi Monitoring Penjualan Dan Stok Barang (Studi Kasus: Distro Zhezha Pontianak)," Jurnal Khatulistiwa Informatika, vol. IV, 2020.

[11] Hindrayani, K.M., Fahrudin, T.M., Aji, R.P. and Safitri, E.M., 2020, December. Indonesian stock price prediction including covid19 era using decision tree regression. In 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 344-347). IEEE.

[12] Sishi, M. and Telukdarie, A., 2021, April. The application of decision tree regression to optimize business processes. In Proceedings of the International Conference on Industrial Engineering and Operations Management (pp. 48-57).

[13] Putri, I.S.I., Pradini, R.S. and Anshori, M., 2024. Decision Tree Regression untuk Prediksi Prevalensi Stunting di Provinsi Nusa Tenggara Timur. Jurnal Teknologi Informatika dan Komputer, 10(2), pp.413-427.

[14] Ihfandi, A., 2018. Implementasi Data Mining Untuk Prediksi Daerah Rawan Penyakit Demam Berdarah Menggunakan Algoritma C4. 5 (Studi Kasus: Dinas Kesehatan Kabupaten Tangerang) (Doctoral Dissertation, Universitas Satya Negara Indonesia).

[15] Fahri, A. and Ramdhani, Y., 2022. Visualisasi Data dan Penerapan Machine Learning Menggunakan Decision Tree Untuk Keputusan Layanan Kesehatan COVID-19. J. Tekno Kompak, 17(2), pp.50-60.