Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Application of naive bayes algorithm for dominant disease classification in coastal environments

Adli Abdillah Nababan
Bisnis Digital, STMIK Pelita Nusantara, Jl. St. Iskandar Muda No.1, Medan, Indonesia

Article Info

Keywords:
Naive Bayes,
Disease Classification,
Coastal Health,
Machine Learning,
Epidemiology.

ABSTRACT

This research focuses on the implementation of the Naive Bayes algorithm to classify prevalent diseases in coastal areas. Coastal regions, characterized by unique environmental factors and limited healthcare accessibility, pose distinct challenges to public health. The primary objective of this study is to enhance the precision and understanding of disease diagnosis within these regions. By employing data analysis and machine learning techniques, the research aims to contribute significantly to the prevention, management, and treatment of diseases in coastal areas, ultimately improving the well-being of local communities. Additionally, the findings have the potential to assist governments and health institutions in formulating targeted and efficient health policies for coastal areas. A comprehensive understanding of dominant disease patterns enables data-driven decision-making, influencing the allocation of health resources, distribution of vaccines and medicines, and the design of tailored prevention programs. Overall, this research is poised to yield substantial benefits by advancing healthcare and enhancing the quality of life in coastal communities.

Corresponding Author:
Adli Abdillah Nababan
Bisnis Digital, STMIK Pelita Nusantara
Jl. St. Iskandar Muda No.1, Medan, Indonesia
adlinababan@pelitanusantara.ac.id

INTRODUCTION

The importance of focusing on public health in coastal regions is often overlooked in global health discussions. Coastal areas possess unique geographical and environmental characteristics that can significantly impact the health of their populations. For instance, they may be more vulnerable to the effects of climate change, such as rising sea levels and tropical storms, which can elevate the risk of certain diseases. Additionally, marine pollution and instability in coastal ecosystems can contribute to the spread of diseases caused by marine microorganisms.

The significance of this research is also linked to the limited access to healthcare facilities experienced by communities in coastal regions (Dahlui et al., 2020). Particularly for those residing in remote areas, gaining access to quality medical care and accurate diagnosis can pose significant challenges (Haimi, 2023). Therefore, the development of efficient disease classification methods, such as the implementation of the Naive Bayes algorithm (Meidina & Abidin, 2023) & (Harumy et al., 2022), could be a crucial step in

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

enhancing early detection and disease management in these regions. The utilization of advanced information technology and data analysis can also provide deeper insights into the environmental factors impacting the public health of coastal communities.

In an era where health data and computing are rapidly advancing, this research has the potential to fill knowledge gaps and make a significant contribution to improving the understanding and well-being of coastal communities. By combining data science and health understanding, this research can provide more accurate and timely information about dominant diseases in coastal regions, ultimately supporting more effective health planning and better decision-making for the local populations (Subrahmanya et al., 2022).

In the specific context of this research, the Naive Bayes algorithm, a classification method leveraging probability and statistics to identify the most probable category or class for each data instance (Chen et al., 2021), is applied to analyze disease data collected from coastal areas.

Several studies, such as the one by (Maliha et al., 2019), have utilized the Naive Bayes algorithm to predict cancer disease. This research, which compared three classification algorithms (Naive Bayes, K-Nearest Neighbor, and J48), highlighted factors like accuracy, error rate, sensitivity, specificity, precision, and F-score, emphasizing the significance of considering lifestyle factors, such as smoking and tobacco use, in cancer prediction.

Research conducted by Aditiya (Hermawan, 2021) developed a Naïve Bayes classification model to assess the mental health of social media users, achieving an accuracy of 92.5% categorized as "Very Good." The study suggested further exploration with a larger dataset and validation from additional experts to enhance the model's validity, along with the exploration of alternative classification algorithms for potentially improved results.

In their research, (Putri et al., 2023) concentrated on developing a data mining model for classifying Alzheimer's disease using the Naive Bayes Classifier (NBC) algorithm. The study employed hold-out data sharing methods and K-means clustering, concluding that the NBC algorithm demonstrated higher accuracy (91.89%) with K-Means clustering compared to the Hold-Out method (85.52%). The optimal number of clusters was identified as three (k = 3), with a DBI value of 0.074.

Furthermore, research conducted by (Barus et al., 2023) addressed the challenge of identifying potential heart failure patients through machine learning, specifically utilizing the Naïve Bayes technique. The study revealed promising results, with the Naïve Bayes algorithm achieving a classification accuracy of 74.58%, precision of 97.67%, sensitivity of 75%, and an AUC of 0.857, suggesting its potential as an effective early warning system for individuals at risk of heart failure.

The next research conducted by (Meidina & Abidin, 2023) explored the integration of Particle Swarm Optimization (PSO) feature selection and gain ratio techniques with the Naïve Bayes algorithm for diagnosing heart disease using the Cleveland dataset. Their research revealed that the combination of PSO and gain ratio significantly improved the accuracy of the Naïve Bayes algorithm. Without feature selection, the algorithm achieved

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

an 86.88% accuracy, whereas with PSO and gain ratio, it increased to 93.44%, reflecting a notable 6.56% improvement.

Collectively, these studies underscore the versatility and efficacy of the Naive Bayes algorithm across diverse healthcare applications, including heart disease, mental health, Alzheimer's, and cancer prediction. The success of Naive Bayes in classifying health data is attributed to its adept handling of common factors such as numerous features, class imbalance, and intricate feature interactions. Therefore, in alignment with the demonstrated effectiveness of the Naive Bayes algorithm, this research employs it for dominant disease classification in coastal regions, as its adaptability addresses the unique healthcare challenges in these areas and contributes to early detection and prevention.

METHODS

This research uses experimental research methods with a focus on the application of the Naive Bayes algorithm for the classification of dominant diseases in coastal areas. The following are the methodological steps that will be taken:

Data Collection

In the data collection, the dataset for this research is obtained from (Harumy et al., 2022) which focused on the public health issues in coastal areas, specifically the classification of the most prevalent dominant diseases among coastal communities. The research subject encompasses coastal community areas in Belawan, Serdang Bedagai, and Central Tapanuli districts in the North Sumatra Province, Indonesia.

The objective of this study is to conduct classification using the Naive Bayes Algorithm to identify the most common dominant diseases prevalent in coastal regions. The parameters considered in this research include geographical location, respondents' professions, education levels, environmental conditions, access to healthcare facilities, weather, and the types of dominant diseases experienced by the community.

The chosen research locations specifically cover three districts in the North Sumatra Province, namely Belawan, Serdang Bedagai, and Central Tapanuli. The dataset used in this research comprises 100 responses collected from these coastal areas. Following are the details of the dataset used:

Table 1. Details of the dataset

No.	Region	Work	Education	Environment	Health Access	Weather	Disease
1	Tapteng	Fisherman	Elementary School	Good	Near	Hot	Diarrhea
2	Sergai	Businessman	Junior High School	Not good	Currently	Hot	Hypertension
3	Belawan	Housewife	Senior High School	Good	Far	Combination	Diarrhea
4	Tapteng	Fisherman	Bachelor	Good	Currently	Hot	Dengue Fever
5	Sergai	Private Sector Employee	Elementary School	Good	Far	Combination	Hypertension
6	Belawan	Fisherman	Junior High School	Good	Near	Combination	Diarrhea

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

No.	Region	Work	Education	Environment	Health Access	Weather	Disease
7	Sergai	Businessman	Senior High School	Not good	Currently	Hot	Diarrhea
8	Belawan	Housewife	Bachelor	Good	Near	Combination	Dengue Fever
9	Tapteng	Fisherman	Senior High School	Good	Currently	Combination	Hypertension
10	Sergai	Private Sector Employee	Bachelor	Good	Far	Hot	Hypertension
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	Tapteng	Fisherman	Senior High School	Good	Far	Combination	Diarrhea

By utilizing data from a prior study, this research will involve further analysis using the Naive Bayes Algorithm to classify dominant diseases in coastal communities. The anticipated outcomes of this research include gaining deeper insights into the patterns and trends of dominant diseases in coastal regions, thereby aiding in the formulation of more effective health policies for the local communities.

Data Processing

In the framework of the research Naive Bayes Algorithm for Dominant Disease Classification in Coastal Regions, the data processing process is a critical stage that forms the basis of accurate analysis. Data collection was carried out from coastal areas, involving information about the dominant types of disease, environmental factors, and the health profile of the local community. This stage involves data cleaning, where missing values or outliers are identified and addressed.

Dataset Splitting

The division of the dataset into training and testing sets is a crucial step in assessing the model's performance. This process involves allocating a certain proportion of the data for training the model, while the remaining portion is used to test the performance of the model. Experiments are conducted with various splitting ratios, namely 50:50, 60:40, 70:30, and 80:20.

In the 50:50 split, half of the data is used for training the model, while the other half is employed for testing. This approach provides a balanced distribution between training and testing. With the 60:40, 70:30, and 80:20 splits, the proportion of data allocated for training increases, while the portion used for testing decreases. This type of division allows the model to train more on the training data but may increase the risk of overfitting if the training data is too large.

Variations in the dataset division offer insights into how well the model can generalize to new data and provide more accurate information about the model's performance in classifying dominant diseases in coastal regions. Evaluating the model's performance at different splitting ratios helps choose the optimal configuration to address classification challenges in the context of coastal public health (Nguyen et al., 2021).

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Naive Bayes Model Training

Naive Bayes is categorized as a "probabilistic classifier," representing a straightforward technique that combines algorithms with shared principles. In this approach, each attribute is classified independently of other featured values (Dada et al., 2019). The method operates under the assumption that the influence of an attribute value on a class is not contingent on other attribute values (Afdhaluzzikri et al., 2022). Notably, Naive Bayes is recognized for its simplicity and is particularly beneficial for handling extensive datasets. The computation involves determining the posterior probability, $P\ (c_k|x)$, derived from $P\ (c_k)$, P (x), and $P\ (x|c_k)$.

$$P\ (c_k|x) = \frac{P\ (X|c_k)P(c_k)}{P(x)}$$

where:

$P\ (c_k|x)$: Posterior Probability, which is the probability of class c_k given data *x.*

$P\ (X|c_k)$: Likelihood, representing the probability of observing data *x* given class c_k.

$P\ (c_k)$: Prior Probability, the probability of class c_k before observing the data.

$P\ (x)$: Evidence or Probability of Data, the overall probability of the data.

This formula is an extension of the binary classification Naive Bayes formula to accommodate multiple classes. The classification decision is typically made by selecting the class c_k that maximizes the posterior probability $P\ (c_k|x)$.

Model Evaluation

Evaluating the effectiveness of a classification model requires the use of specific metrics. These metrics are derived from a confusion matrix, which is a tabulation that cross-references the predicted class with the actual observations in the response variable. The confusion matrix components represent various classification outcomes, including True Positive (TP) for accurately classified positive instances, False Positive (FP) for negative instances misclassified as positive, True Negative (TN) for correctly classified negative instances, and False Negative (FN) for positive instances misclassified as negative. By analyzing these components, we can assess and measure the performance of the classification model (A. A. Nababan et al., 2022).

In gauging the effectiveness of the classification models employed in this research, the primary metric considered is Accuracy. It serves as a benchmark to evaluate the classifier's performance, with higher accuracy indicating superior model performance (A. H. Nababan et al., 2021), (Syaliman, 2021). Accuracy, as an assessment method, relies on the count of correctly predicted outcomes in the classification. The accuracy value is computed using the following equation:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictons}$$

This metric provides a comprehensive measure of the model's ability to make correct classifications, offering valuable insights into its overall performance.

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

RESULTS AND DISCUSSION

In this section, we conduct experiments to evaluate the performance of the Naive Bayes Algorithm model on a personal computer. The computer used has an Intel Core i5 processor, 4 GB RAM, and runs on the Windows 10 platform. The model is implemented using the Hypertext Preprocessor (PHP) programming language. We present the results and engage in a comprehensive discussion on the applicability of the developed classification model. Our methodology involves simulations on datasets related to the Classification of Dominant Diseases in Coastal Areas with various proportions of split datasets, assessing their performance.

Dataset

In this stage, data collection is conducted to gather information that will be used in the application. The dataset includes information about various factors relevant to disease classification in coastal environments, such as region, work, education level, environmental conditions, access to healthcare facilities, and weather. This dataset serves as a crucial foundation in the application development process.

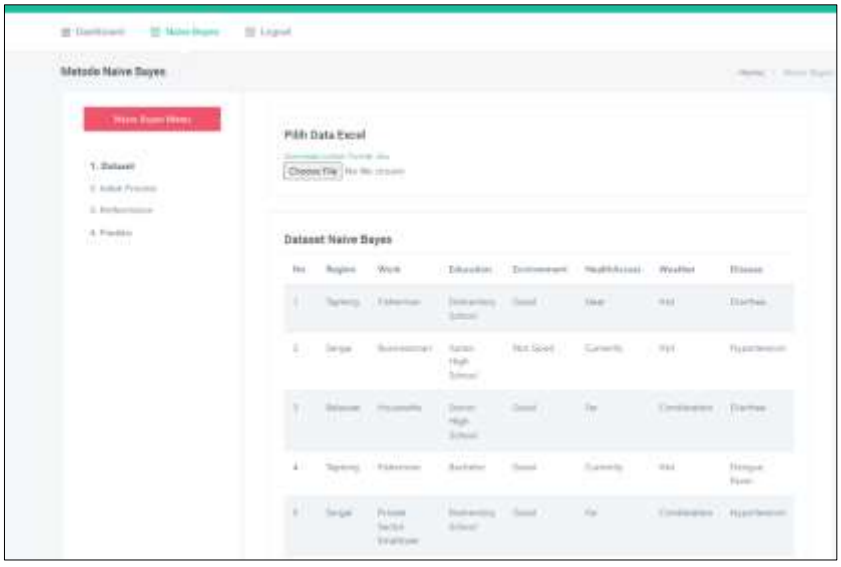


Figure 1. Import Dataset

Initial Process

The initial process includes processing and preparing data before being entered into the classification model. This includes determining supporting attributes and target labels. This step is important to ensure the quality and accuracy of the data used in the classification process.

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

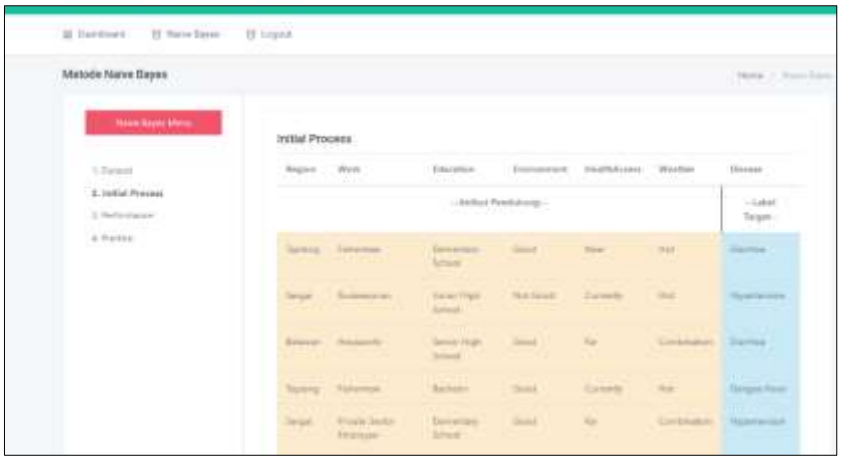


Figure 2. Initial Process

Performance

In this stage, the classification model is trained using the Naive Bayes algorithm with the pre-processed dataset. The performance of the model is evaluated by splitting the dataset into training and testing subsets. Accuracy evaluation metrics are used to measure the model's performance in classifying diseases based on the information provided.

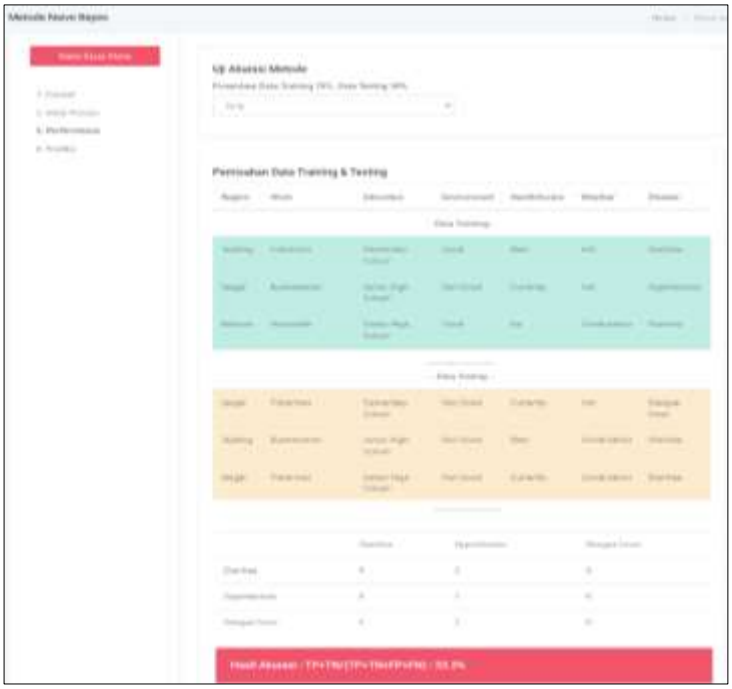


Figure 3. Test Method Accuracy

Prediction

After the classification model is trained and evaluated, the application will receive new input that includes information about the region, work, education, environment, health

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

access, and weather. The model will utilize this information to make predictions about the dominant diseases that individuals or communities in the coastal environment may experience. These predictions can be used to aid in early identification, management, and prevention of diseases in coastal regions.

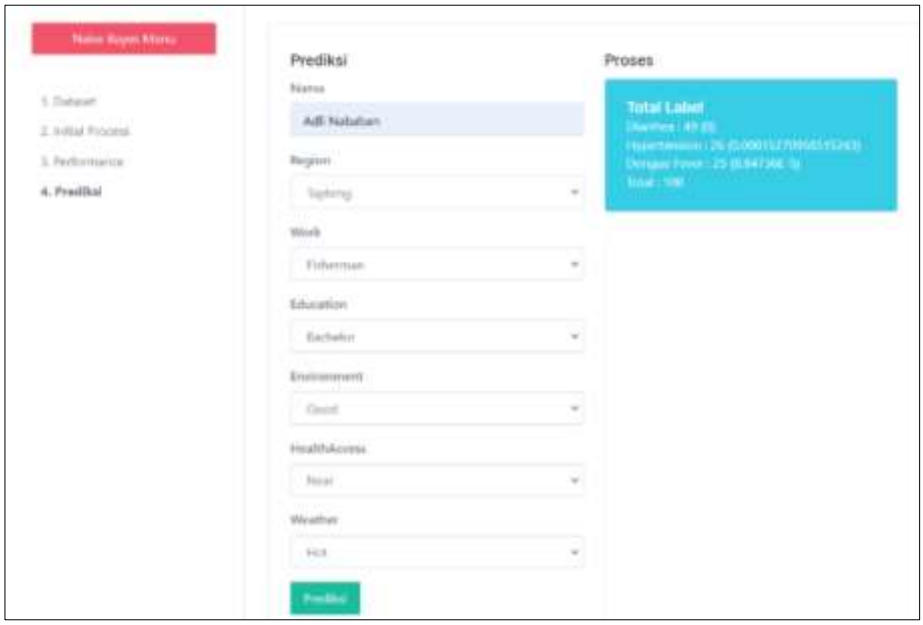
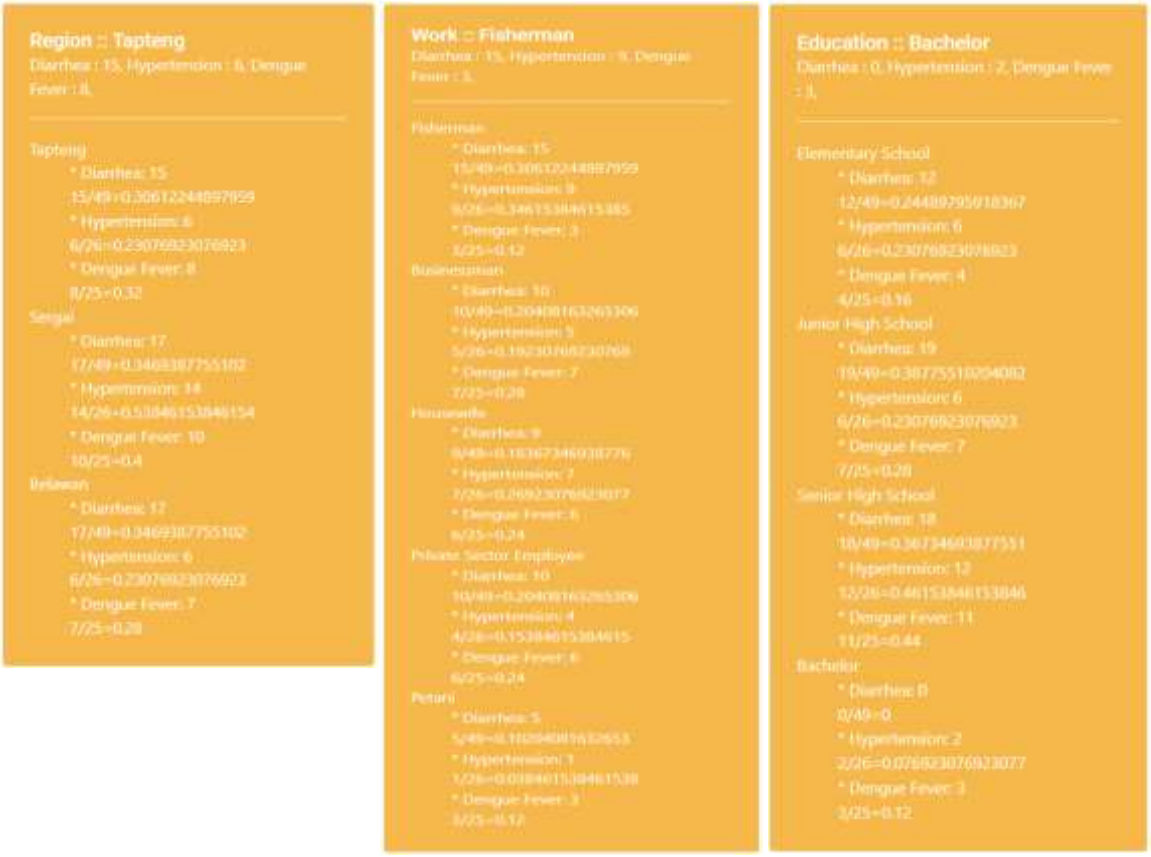


Figure 4. Application Prediction Results



Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

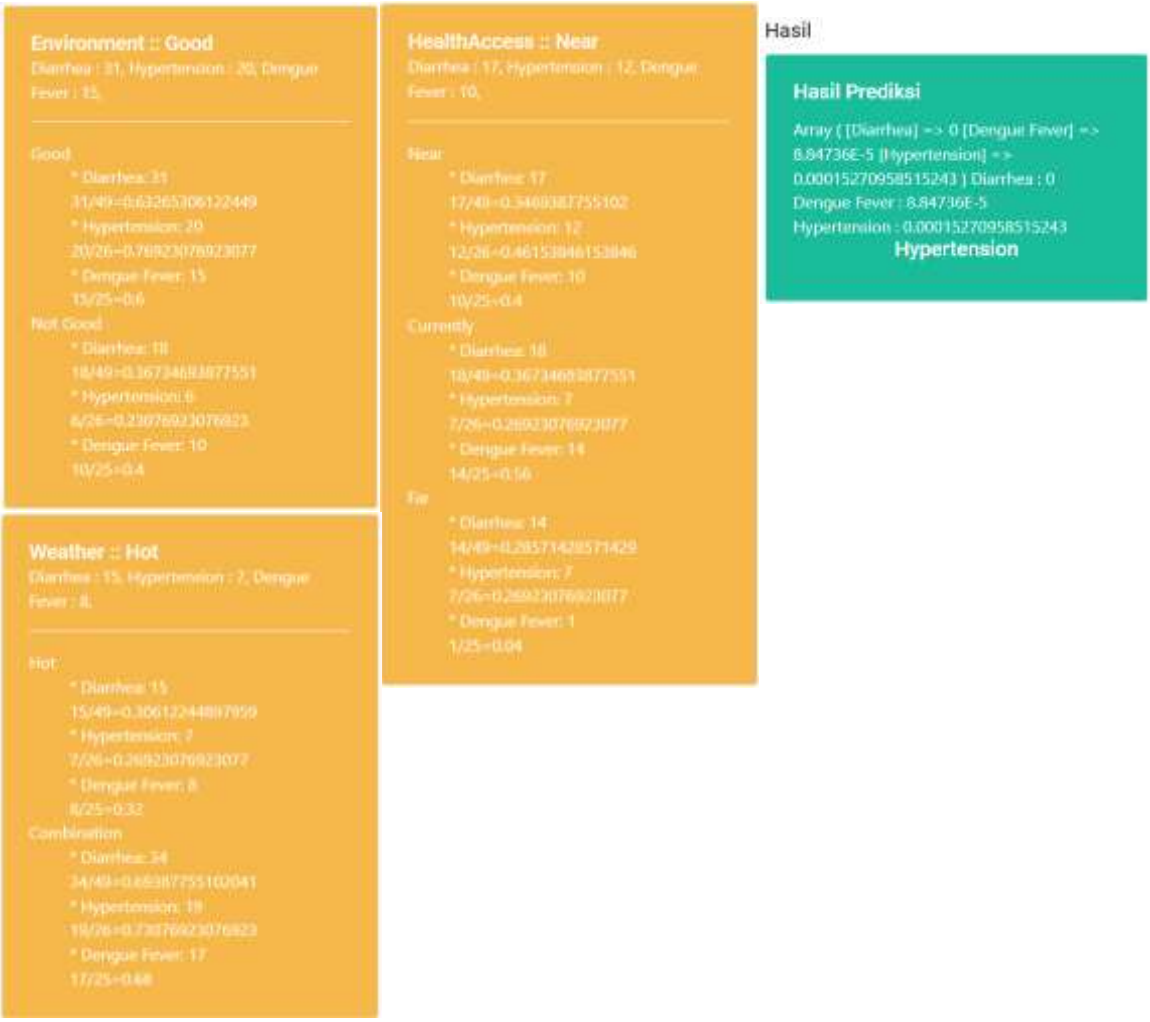


Figure 5. Probability calculations for each attribute

Figure 5 is the application prediction results obtained through probability calculations using the Naive Bayes algorithm for each attribute, including region, work, education level, environmental conditions, access to healthcare facilities, and weather. The Naive Bayes algorithm calculates the probability of each attribute given the class (i.e., the probability of observing a certain attribute value given the presence of a particular disease). By multiplying these probabilities together for each attribute, the algorithm determines the overall probability of a specific disease given the input attributes.

These prediction results provide valuable insights into the potential diseases that individuals or communities in coastal environments may experience, helping in early detection, management, and prevention efforts. As for the outcomes obtained from the training and testing datasets with a 50:50 ratio, we divided the dataset evenly into two portions, assigning 50% of the data for training purposes and allocating the remaining 50% for testing. Here are the results derived from the confusion matrix:

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Table 2. Split Ratio 50:50

		Actual		
		Diarrhea	Hypertension	Dengue Fever
Predicted	Diarrhea	13	7	9
	Hypertension	5	3	1
	Dengue Fever	7	5	0

For the subsequent splitting ratio of 60:40, we distributed the dataset in a manner where 60% of the data was assigned for training, while the remaining 40% was designated for testing. Following this division, we evaluated the model's performance and obtained the corresponding results from the confusion matrix.

Table 3. Split Ratio 60:40

		Actual		
		Diarrhea	Hypertension	Dengue Fever
Predicted	Diarrhea	15	2	6
	Hypertension	6	2	0
	Dengue Fever	6	2	1

Continuing with the splitting ratio of 70:30, we partitioned the dataset such that 70% of the data was utilized for training purposes, while the remaining 30% was reserved for testing. Subsequently, we assessed the model's performance and obtained the respective outcomes from the confusion matrix.

Table 4. Split Ratio 70:30

		Actual		
		Diarrhea	Hypertension	Dengue Fever
Predicted	Diarrhea	9	3	4
	Hypertension	6	1	0
	Dengue Fever	5	2	0

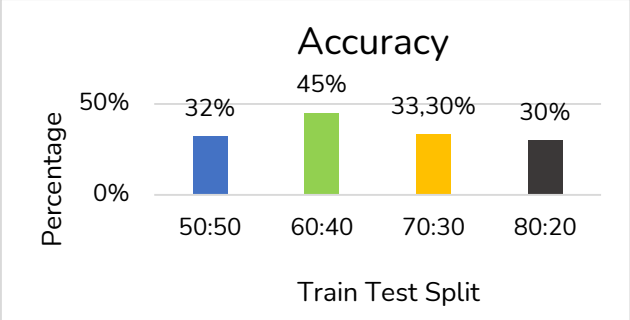
Lastly, for the 80:20 splitting ratio, we segregated the dataset into training and testing sets, allocating 80% of the data for training and 20% for testing. Following this division, we evaluated the model's performance and recorded the outcomes obtained from the confusion matrix.

Table 5. Split Ratio 80:20

		Actual		
		Diarrhea	Hypertension	Dengue Fever
Predicted	Diarrhea	6	1	2
	Hypertension	5	0	0
	Dengue Fever	4	2	0

Jurnal Info Sains : Informatika dan Sains
Volume 14 , Number 01, 2024, DOI 10.54209/infosains.v14i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Table 6. The accuracy results



CONCLUSION

The accuracy results from different dataset-splitting ratios shed light on the classification model's performance. Initially, with a 50:50 split, the model achieved 32% accuracy, indicating limited performance due to a small training dataset. Increasing the training data to 60% in the 60:40 split notably improved accuracy to 45%, showcasing better generalization. However, with a 70:30 split, accuracy dropped to 33.3%, suggesting potential overfitting. Similarly, in the 80:20 split, accuracy decreased to 30%, reinforcing overfitting concerns. Finding an optimal splitting ratio is crucial for balancing model performance and generalization. Applying the Naive Bayes Algorithm for disease classification in coastal environments using PHP involves inputting factors such as region, work, education, environment, health access, and weather. PHP's flexibility and ease of implementation make it suitable for processing diverse health-related datasets in coastal regions. Through this approach, valuable insights can aid in identifying and managing prevalent diseases in coastal areas.

REFERENCE

Afdhaluzzikri, A., Mawengkang, H., & Sitompul, O. S. (2022). Perfomance analysis of Naive Bayes method with data weighting. *SinkrOn*, *7*(3), 817–821. https://doi.org/10.33395/sinkron.v7i3.11516

Barus, O. P., Lauwren, K., Pangaribuan, J. J., & Romindo. (2023). Implementation of the Naive Bayes Algorithm to Predict the Safety of Heart Failure Patients. *Conference Series*, *4*(1), 172–177. https://doi.org/10.34306/conferenceseries.v4i1.651

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *Eurasip Journal on Advances in Signal Processing*, *2021*(1). https://doi.org/10.1186/s13634-021-00742-6

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, *5*(6). https://doi.org/10.1016/j.heliyon.2019.e01802

Dahlui, M., Azzeri, A., Zain, M. A., Mohd Noor, M. I., Jaafar, H., Then, A. Y. H., Suhaimi, J., Kari, F., Creencia, L. A., Madarcos, J. R., Jose, E., Fleming, L. E., White, M. P., Morrissey, K., Fadzil, K. S., & Goh, H. C. (2020). Health status, healthcare utilisation, and quality of life among the coastal communities in Sabah: Protocol of a population-based survey.

Medicine (United States), *99*(37). https://doi.org/10.1097/MD.0000000000022067

Haimi, M. (2023). The tragic paradoxical effect of telemedicine on healthcare disparities- a time for redemption: a narrative review. *BMC Medical Informatics and Decision Making*, *23*(1), 1–10. https://doi.org/10.1186/s12911-023-02194-4

Harumy, T. H. F., Manik, F. Y., & Altaha. (2022). Comparison of Artificial Neural Network Classification Methods for Diseases That Are Dominately Suffered By Coastal Communities. *Journal of Theoretical and Applied Information Technology*, *100*(18), 5335–5345.

Hermawan, A. (2021). Implementation of Naïve Bayes Algorithm for Classification of Mental Health of Social Media Users. *Bit-Tech*, *4*(2), 61–70. https://doi.org/10.32877/bt.v4i2.282

Maliha, S. K., Ema, R. R., Ghosh, S. K., Ahmed, H., Mollick, M. R. J., & Islam, T. (2019). Cancer Disease Prediction Using Naive Bayes,K-Nearest Neighbor and J48 algorithm. *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*, 1–7. https://doi.org/10.1109/ICCCNT45670.2019.8944686

Meidina, A., & Abidin, Z. (2023). Diagnosis of Heart Disease Using Optimized Naïve Bayes Algorithm with Particle Swarm Optimization and Gain Ratio. *Recursive Journal of Informatics*, *1*(2), 47–54. https://doi.org/10.15294/rji.v1i2.67278

Nababan, A. A., Sutarman, Zarlis, M., & Nababan, E. B. (2022). Air Quality Prediction Based on Air Pollution Emissions in the City Environment Using XGBoost with SMOTE. *2022 IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM)*, 1–6. https://doi.org/10.1109/ICOSNIKOM56551.2022.10034887

Nababan, A. H., Mahendra, R., & Budi, I. (2021). Twitter stance detection towards Job Creation Bill. *Procedia Computer Science*, *197*(2021), 76–81. https://doi.org/10.1016/j.procs.2021.12.120

Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Van Le, H., Tran, V. Q., Prakash, I., & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, *2021*. https://doi.org/10.1155/2021/4832864

Putri, W., Hastari, D., & Faizah, K. U. (2023). *Implementation of Naïve Bayes Classifier f or Classifying Alzheimer ' s Disease Using the K-Means Clustering Data Sharing Technique*. *1*(July), 47–54.

Subrahmanya, S. V. G., Shetty, D. K., Patil, V., Hameed, B. M. Z., Paul, R., Smriti, K., Naik, N., & Somani, B. K. (2022). The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal of Medical Science*, *191*(4), 1473–1483. https://doi.org/10.1007/s11845-021-02730-z

Syaliman, K. U. (2021). *ENHANCE THE ACCURACY OF K-NEAREST NEIGHBOR (K-NN) FOR UNBALANCED CLASS DATA USING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) AND GAIN RATIO (GR)*. *10*(1), 188–195.