Prasetiyowati and Sibaroni
Journal of Big Data (2024) 11:106
https://doi.org/10.1186/s40537-024-00958-x

Journal of Big Data

RESEARCH **Open Access**

Unlocking the potential of Naive Bayes for spatio temporal classification: a novel approach to feature expansion

Sri Suryani Prasetiyowati[1*] and Yuliant Sibaroni[1]

*Correspondence: srisuryani@telkomuniversity.ac.id

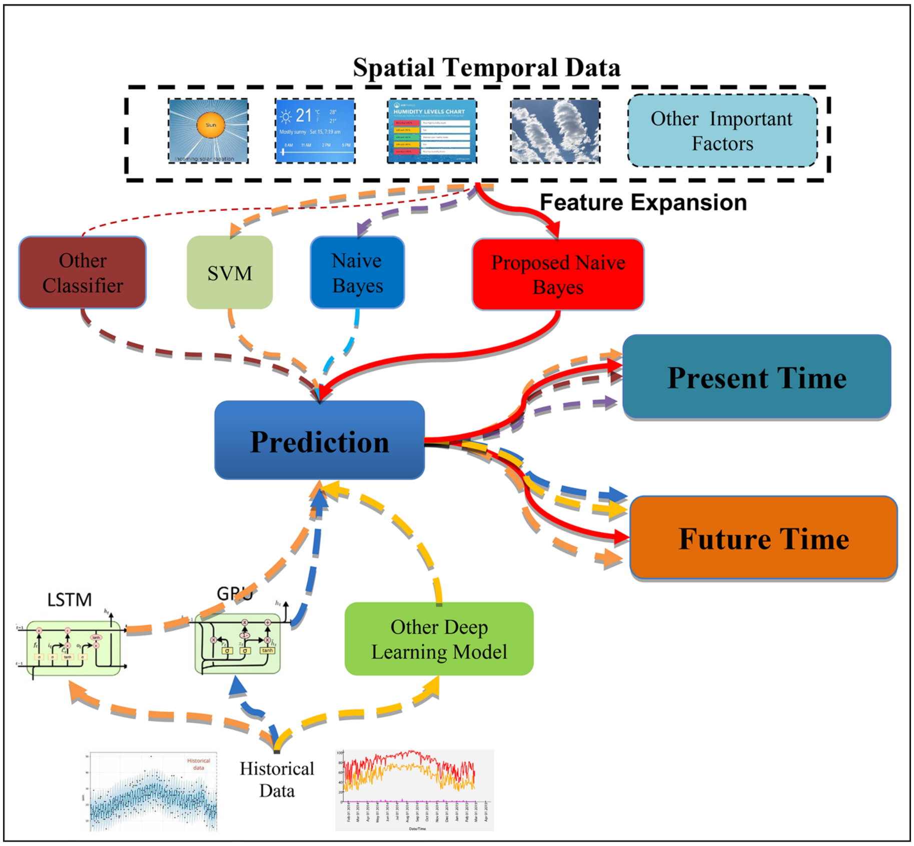
[1] School of Computing, Telkom University, Bandung, Indonesia

Abstract

Prediction processes in areas ranging from climate and disease spread to disasters and air pollution rely heavily on spatial–temporal data. Understanding and forecasting the distribution patterns of disease cases and climate change phenomena has become a focal point of researchers around the world. Machine learning models for prediction can generally be classified into 2: based on previous patterns such as LSTM and based on causal factors such as Naive Bayes and other classifiers. The main drawback of models such as Naive Bayes is that it does not have the ability to predict future trends because it only make predictionsin the present time. In this study, we propose a novel approach that makes the Naive Bayes classifier capable of predicting future classification. The process of expanding the dimension of the feature matrix based on historical data from several previous time periods is performed to obtain a long-term classification prediction model using Naive Bayes. The case studies used are the prediction of the distribution of the annual number of dengue fever cases in Bandung City and the distribution of monthly rainfall in Java Island, Indonesia. Through rigorous testing, we demonstrate the effectiveness of this Time-Based Feature Expansion approach in Naive Bayes in accurately predicting the distribution of annual dengue fever cases in 30 sub-districts in Bandung City and monthly rainfall in Java Island, Indonesia with with both accuracy and F1-score reaching more than 97%.

Keywords: Spatial-time data, Time-based feature expansion, Naïve Bayes, Classification, Prediction model

Graphical Abstract



Introduction

Prediction of future events has become an interesting topic by many researchers [1–3] because with the prediction results obtained, many parties will obtain information about future events, which is very important in preparing appropriate strategies or mitigation. Most of the prediction methods are found in statistical sciences, and usually predictions are made based on time and spatial. The predictive models are used in various fields such as health, business, climate, transportation. Predictive modeling is a statistical technique that is commonly used to predict future behavior. Predictive model development is one of the statistical techniques used to predict future behavior. The resulting solution is a data mining technique that analyzes historical and current data.

Prediction can also be implemented with a machine learning approach. In addition to prediction, machine learning can also be used to solve problems in regression, classification and clustering [4]. This method can be computationally intensive, as it involves large and complex data, so it can play an important role in solving spatial problems in various application areas, from multivariate prediction to image classification to spatial pattern detection [5–7]. But the predictions model obtained based on machine learning is limited in predicting cases at the present time, cannot be used for future predictions. This makes prediction models based on machine learning inappropriate for predicting future events. It is very interesting to develop a prediction model based on machine learning that can be used to predict future events.

Time feature -based learning that represents and analyzes the property of time elements, including mapping time series properties, such as trends, seasonal, and stationary [3, 5, 8–11] has implemented the engine learning method In various spatial time data, namely the fields of geology, epidemiology, health care, climate science, enviromental science, precision agriculture, neuroscience, social media, etc.

Several studies conducted by [2, 3, 12, 13] used machine learning methods that were applied to time spatial data, with the aim of making predictions for the future. But in all of these studies there is no entanglement and continuity between the machine learning model and the prediction process. Prediction system performance is measured based on classifier model accuracy, while the prediction process used linear regression with the independent variables is classification results, *time*, without involving its features. In these studies, the classification methods used are Artificial Neural Network, Naïve Bayes, K Nearst Network, Logistic Regression, Super Vector Machine, Decssion Tree, and Random Forest [2, 3]. Studied classification predictions on spatial-time data, using simple linear regression, with a predictive time period of 100 years and 6 days, respectively. In both studies, before predicting using regression, a classification study was carried out using machine learning and Random Forest. Classification model obtained by machine learning, using all the features in the data train, while in the predicting process of target data, the regression method is used with only involves the time feature. Predicted results with regression for the future do not contain features used in previous machine learning models. Of all the classification methods used, Naïve Bayes is the simplest classification method. Naïve Bayes is one of the most popular classification algorithms, simple but very practical. Its efficiency comes from the assumption of feature independence, although this may be met in some studies using big data [14, 15]. The Naïve Bayes classification method has the advantage of adjusting parameter freedom and is more robust. But Naïve Bayes is also still quite reliable for use in small data sets [16].

For prediction purposes, the Naïve Bayes classification method is adaptive for feature weighting and makes feature selection easier, simpler and more efficient [16–18]. Empirical results show that the Naive Bayes selection method shows high classification accuracy [15]. When compared to other feature selection approaches, Naïve Bayes obtains more competitive results regarding accuracy, sparity, and time for balanced data sets. But for the case of class-imbalanced data sets, Naive Bayes still works well, with different levels of classification for different classes still achievable [19]. Meanwhile [20] uses the Naïve Bayes Classifier with Maximal Time Series Motif to classify ECG abnormalities, with features such as record numbers and discrete sequences. The Time Series Motif Detection is proposed as Feature extraction and when combined with NB classifier it is superior with 98% accuracy, than feature selection with classifier. While in Electroencephalogram (EEG) classification research [12] and urban waterlogging classification [21], the use of feature extraction and the weighted Niave Bayes method on spatial time data has been shown to provide high accuracy. Another technique to increase the accuracy of Naive Bayes is to reduce the problem of interdependence between its features [22].

Based on previous studies, it can be concluded that the use of classifier methods such as Naive Bayes in future predictions has never been done fully based on the training model. In this study, the process of developing the feature expansion method is carried out and combined with the Naive Bayes classifier so that the obtained classifier training model can be used to predict future class events.

The main contributions of this research are as follows: (1) Produce a feature expansion method based on spatial time data that can be used to build a classification prediction model for some time in the future based on the Naive Bayes classifier, (2) Produce a prediction model for the number of annual dengue fever cases based on the Naive Bayes classifier, (3) Produce a monthly rainfall level prediction model based on the Naive Bayes classifier. The baseline used in this research is prediction model research using regression on the same data or similar data.The data used in this research are data on the number of dengue fever cases and rainfall, where both data are spatial-time data and have been used as data sets in previous studies Network [23–25]. Prediction of dengue case classes using Naive Bayes in [23] provided an accuracy of 74% while the use of a Voting-based Hybrid method for Naïve Bayes, K-Nearest Neighbor, and Artificial Neural Networks resulted in an accuracy of 90%. Meanwhile, for the rainfall dataset, there are 2 researches that have used the data. Prediction of rainfall levels using Logistic Regression [24] resulted in an accuracy of 72%, while the use of Naive Bayes and hybrid Naive Bayes-C4.5 methods [25] resulted in an accuracy of 52.98% and 64.95% respectively.

Related work

Feature expansion

In recent years, the idea of feature engineering has confirmed the outstanding performance of machine learning techniques, which can automate several applications. Feature engineering techniques such as feature extraction, feature selection and feature expansion are often applied to machine learning classification [26]. Feature extraction is the process of selecting the best subset of features from the overall feature set [27]. This approach is the creation of a new feature which is also known as feature construction. Whereas feature expansion is combining additional features from the input data, which combines the different relationships between the original features of the two objects. The goal is to extend the original vector or form a new feature, which is related to the distance from each data sample to the number of centroids found by the clustering algorithm. The use of this approach is usually applied to pattern recognition problems [28, 29], and in certain contexts, such as sentence retrieval, intrusion detection and sentiment analysis. Research [28] discusses classification with a combination of feature extraction and feature expansion, to perform a transformation from the original feature, which can increase the classification similarity score independently.

The feature expansion technique for the classification of one-dimensional time series data proposed by [30], extended features can include temporal, frequency, and statistical characteristics. The study stated that the value of classification accuracy obtained was higher than the results of conventional machine learning. Feature expansion techniques allow the classifier to consider multiple dimensions that are not feasible in low-dimensional data. Feature expansion works by taking features in the original data and doing something with or on those features, then adding additional dimensions, to see if there is an increase in the accuracy of the resulting hyperplane [31]. In feature expansion it is possible to use linear classifiers on some data by creating new features in new dimensions.

Regarding the possibility of irrelevant features appearing in the data set, feature engineering also requires substantial manual effort in designing and selecting features [32]. According to [33–35], feature selection provides an effective way, by removing irrelevant and redundant data. Feature selection is the process of selecting certain features that are considered the most influential in the classification process.

Impact of feature expansion

In some cases, [36] states that merging all features into one feature space does not guarantee optimal performance, because of dimensional problems. To overcome these problems, a variation of Bayesian approach to the multinomial probit model is used, for base expansion and kernel combinations. This model has a solid foundation in a hierarchical Bayesian framework and is capable of instructively combining available sources of information, for multinomial classification. Meanwhile [37] discusses adding features to the classification process with machine learning, to increase the accuracy or precision value of the original data classification results. The scenario used is to compare the accuracy of the addition of 2-dimensional and 4-dimensional features in the classification process. There are 5 classification methods used, namely CART, RF, Gradient Boosting, SVM, Logistics Regression, where the scenario used is with and without feature selection. The result obtained is a classification with the addition of feature dimensions can increase the value of F1-score. The other result is the classification using feature selection research apparently can not increase the value of f1-score, although in others research [33–35], feature selection can increase the value of accuracy, validity of extracted information and reduce computational costs (processing time). Adding features automatically causes additional dimensions, so it is necessary to pay attention to the balance between the problem of dimensions and the addition of new features. The simplest way to add a feature is to add the degree and logarithm function of the original feature. This process has several stages that can increase the degree of new features or the number of multipliers in new features [32]. Meanwhile, to reduce the dimensions, the main component analysis is implemented after the feature addition procedure.

The consequence of applying feature expansion to the classification process is the addition of feature dimensions. High-dimensional data analysis is a challenge in the fields of machine learning and data mining. Complex multidimensional data usually has four types of features [34] namely high-weighted features, moderately weighted features, less-weighted features, zero-weighted features. With regard to these types of features, feature engineering requires substantial manual effort in designing and selecting features [32].

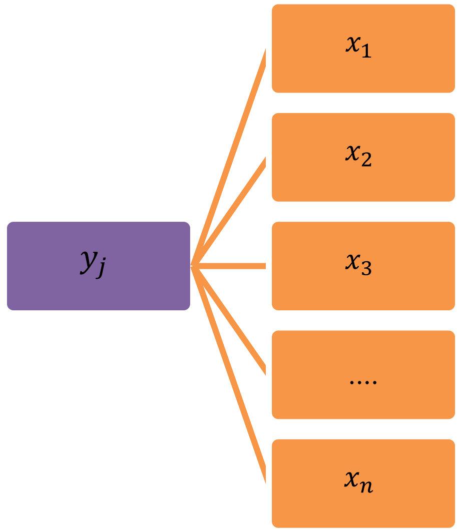


Fig. 1 Naïve Bayes Network Structure

Naive Bayes classifier

Naive Bayes learning refers to the Bayesian probabilistic model that determines the posterior class probability, namely $P(y_j|x_i)$. The simple Naïve Bayes classifier uses these probabilities to assign a class sample [37–39]. The Bayes theorem obtained is.

$$P(y_j|x_i) = \frac{P(x_i|y_j)P(y_j)}{P(x_i)} \tag{1}$$

where x_i:feature at i; y_j:class at j; $P(y_j|x_i)$:probability of even y_j given x_i has occurred; $P(x_i|y_j)$:probability of even x_i given y_j has occurred; $P(y_j)$:probability of event y_j; $P(x_i)$:probability of event x_i.

It is known that the feature set are$x_i = x_1, x_2, x_3, \ldots, x_n$, and class $y_j = y_1, y_2, y_3, \ldots, y_m$, then the relationship between class y_j and attribute x_i can be described [40], as follows:Based on the naive Bayes network structure in Fig. 1, equation (1) can be developed into

$$P(y_j|x_1, x_2, x_3, \ldots, x_n) = \frac{P(x_1, x_2, x_3, \ldots, x_n|y_j)P(y_j)}{P(x_1, x_2, x_3, \ldots, x_n)} = \frac{\prod_{i=1}^{n} P(x_i|y_j)P(y_j)}{P(x_1, x_2, x_3, \ldots, x_n)} \tag{2}$$

The denominator in Eq. (2) does not depend on the target class, but acts as a scaling factor ensuring that the posterior probabilities $P(y_j|x_i)$ are properly scaled. So that the maximum posterior rule can be used, namely assigning each instance to exactly one class, by simply calculating the value of the quantifier for each class, then selecting the class with the maximum value [38]. The resulting selected class is referred to as the Maximum A Posteriori (MAP) class with the following formula.

$$\hat{y} = arg\max_{y_j} \prod_{i=1}^{n} P(x_i|y_j)P(y_j) \tag{3}$$

Maximum A Posteriori (MAP) Estimation can also be used as an estimate of $P(y)$ and $P\left(x_i|y\right)$ [37, 39].

The proposed methods

Basically, the machine learning methods that have been used in previous research are only limited to predicting classification at this time. Likewise, the feature expansion method used in previous research was only limited to feature expansion for text data. Research on developing the Naive Bayes classification model is currently only used to predict classification at this time. Several previous studies carried out a prediction process for some time in the future using a linear regression model based on time-independent variables, by first carrying out classification using machine learning methods. In research that has been carried out previously regarding the development of classification prediction models, to date there has been no research that has developed a prediction model for classification of spatial-time data for the future directly using the time-based Naive Bayes method algorithm.

Therefore, in this research, a Naive Bayes classification prediction model was developed for the future with the scenario of expanding time-based features on spatial-time data, taking into account the stationarity of feature data over time. The proposed procedure is the development of a classification prediction model for time $t+k$, with the identification of a classification prediction model for time $t-k$, namely predicting the target class Y_t using a combination of previous time $t-k$ features. The most optimal model combination of the previous $t-k$ classification prediction models, namely the model with the best accuracy, is selected as a candidate classification prediction model for the future $t+k$ time. The data matrix development framework with extended time-based features, the Naive Bayes classification model with expanded time-based features and its architecture will be explained in the algorithm and Fig. 2.

Algorithm NB Time-Based Feature Expansion (implemeted in python)

1. Preprocessing Data
   a. Read Data
   b. Define the column that is the class column
   c. Transform the data in all columns to normal scale (0-1)
2. Create predefined time-based Data Feature Combinations
   {Create predefined time-based Data Feature Combinations
   (e.g. feature data for time-based classification 2 years or 3 years in advance or combined feature data from several years in advance)}
3. For each time-based Data Feature do:
   a. Define the features column (k columns) and class column
   b. If the data is not balanced, apply the oversampling method to balance the data
   c. for i = 1 to k do
      1) Train the NB model using features of size i
      2) Obtain the NB model with the best accuracy for for the number of features of size i
      3) Create a list containing i and the best accuracy
   d. plot graph of best accuracy vs number of features i
4. Get the best NB-based feature combination model for the next n years

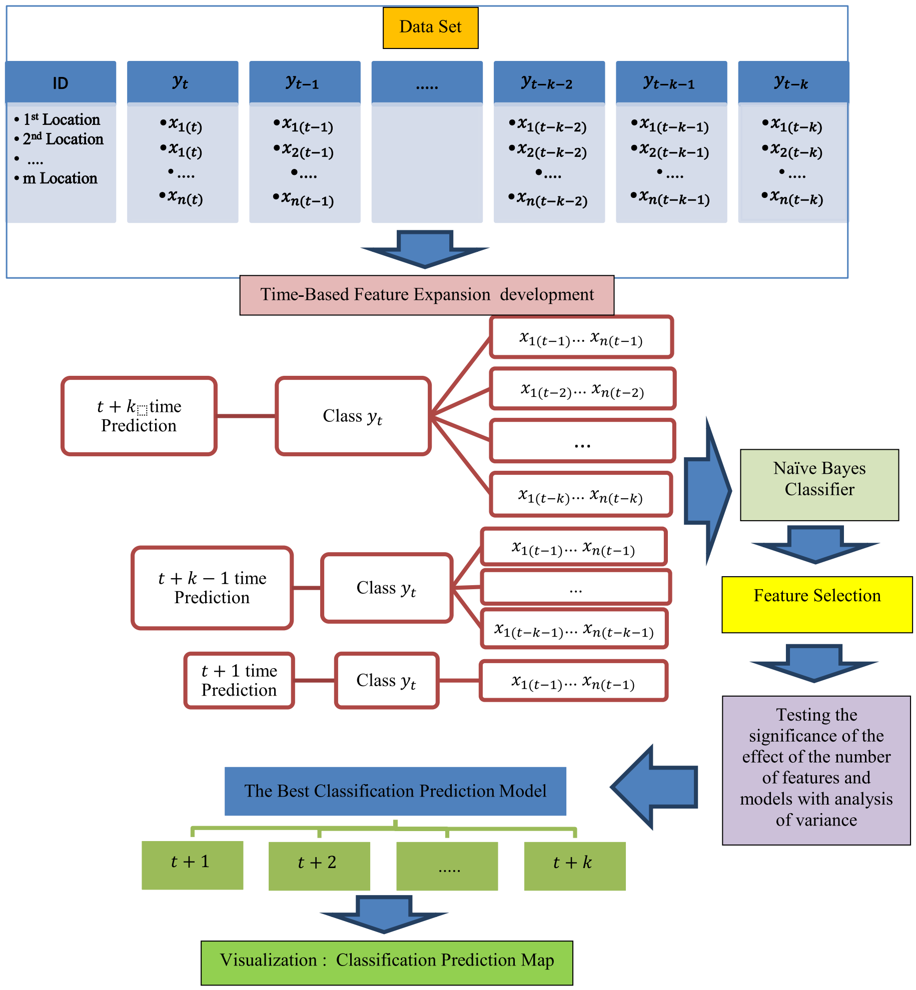


Fig. 2 Naïve Bayes-Time-Based Feature Expansion for Classification Prediction

Experiment

This section describes the development of a feature expansion model based on the previous $t-k$ time, the development of the Naive Bayes Time-Based Feature Expansion model, and experiments using Naive Bayes-Time-Based Feature Expansion to build a classification prediction model for some future time. This study used 2 data sets, namely the Dengue Hemorrhagic Fever and Rainfall which are secondary data from the Bandung City Health Office and the Meteorology, Climatology and Geophysics Agency (BMKG). These two data are used as an example of implementing the development of a time-based classification prediction model to improve the performance of Naïve Bayes classification, which so far has not been able to predict classification in the future. The use of these two data is based on the same characteristics to be tested using the proposed model. This data is research data that has been used in previous studies which meets the characteristics of time series and spatial factors, so that it can be used to build time-based classification prediction models.

Time-based feature expansion process

The feature expansion technique developed in this research makes the feature dimension (N) to be used in the Naïve Bayes model increase k times (which depends on the time characteristics of the data). This is of course very different from traditional feature selection techniques which at most only produce N-dimensional features. Feature expansion in this research is carried out to obtain a feature matrix as input to Naïve Bayes classification as a t + k classification prediction model. The feature selection technique in this study is still carried out but the selection process is on the expanded feature matrix, which is used when determining the combination of features that produce the best classification performance for the t + k prediction model.

In performing feature expansion, one must consider that the resulting output must appropriate with the input of matrix format in the classification process using machine learning methods. There are two stages carried out in this feature expansion process. The first stage is carried out to determine the shape of the feature matrix for the standard classifier model. The second stage is carried out to develop new features based on the feature matrix of the first stage. The development of the feature matrix is conducted by expanding the feature column partition based on time k from the previous target class.

Table 1 shows that to classify data in the same time, it takes n features based on the original dataset at the same time. There are k model classifiers for different times that can be obtained from this dataset. In the first scenario, all data from all time are merged and one classifier model for all time is obtained based on this dataset.

Table 2 shows that the feature matrix will have a maximum size if it is used to predict the one next period. For example, if the dataset is monthly data, then the size of the matrix will reach the maximum if it is used to predict events in the $t + k$ month later.

Naïve Bayes-Time-Based Feature Expansion

It is known that the data set consists of x_i features, namely $x_i = x_1, x_2, x_3, \ldots., x_n$ and class $y_j = y_1, y_2, y_3, \ldots., y_m$. The prediction model on machine learning will generate a numerical score for each feature, so that it can quantify the degree of class membership above in y_j class. If the dataset only consists of positive and negative classes, then the predictive model can be used as a classifier.

The model in Eq. (2) can only be used as a classification model for a moment, and cannot be used for a classification model for the future. So the contribution of this research is to develop a classification model that can be used for classification in the future, if the features are known several years before. The classification prediction model is developed by extending the features based on time.

Table 1 Feature matrix for data classification at the equal time

Features				Class
x_{1t}	x_{2t}	$\ldots$	x_{nt}	y_t
$x_{1(t-1)}$	$x_{2(t-1)}$	$\ldots$	$x_{n(t-1)}$	y_{t-1}
$x_{1(t-2)}$	$x_{2(t-2)}$	$\ldots$	$x_{n(t-2)}$	y_{t-2}
				$\ldots$
$x_{1(t-k)}$	$x_{2(t-k)}$	$\ldots$	$x_{n(t-k)}$	y_{t-k}

Table 2 Feature matrix for the future k time prediction

Features										**Class Target**	**Description**
$x_{1(t-k)}$	$\cdots$	$x_{n(t-k)}$	$\cdots$	$x_{1(t-2)}$	$\cdots$	$x_{n(t-2)}$	$x_{1(t-1)}$	$\cdots$	$x_{n(t-1)}$	y_t	$t+k$ time Prediction
			$x_{1(t-k-1)}$	$\cdots$	$x_{n(t-k-1)}$		$x_{1(t-1)}$	$\cdots$	$x_{n(t-1)}$	y_t	$t+k-1$time Prediction
				$\cdots$		$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
							$x_{1(t-1)}$	$\cdots$	$x_{n(t-1)}$	y_t	$t+1$time Prediction

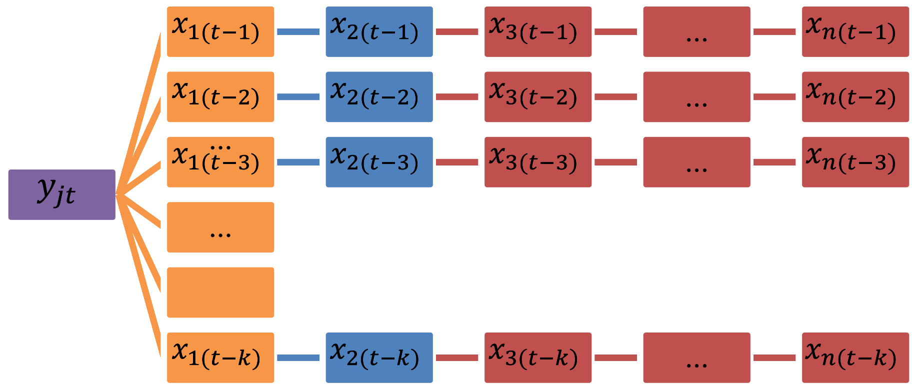


Fig. 3 Naïve Bayes-Time-Based Feature Expansion network structure

If the features and records of the data set are expanded over time, there will be an expansion of the dimensions. For example, attribute based on time $x_{i(t-k)}$, where $t = 1, 2, 3, 4, \ldots .k$, $i = 1, 2, 3, 4, \ldots n$,and $j = 1, 2, 3, 4, \ldots .m$ then $x_{i(t-1)} = x_{1(t-1)}, x_{2(t-1)}, x_{3(t-1)}, \ldots, x_{n(t-1)} : x_{i(t-2)} = x_{1(t-2)}, x_{2(t-2)}, x_{3(t-2)}, \ldots, x_{(t-2)n} : \cdots : x_{i(t-k)} = x_{1(t-k)}, x_{2(t-k)}, x_{3(t-k)}, \ldots, x_{n(t-k)}$ and class of data set is $y_{jt} = y_{1t}, y_{2t}, y_{3t}, \ldots ., y_{mt}$. It is assumed that $x_{i(t-k)}$ and y_{jt} are stasionary, then analogous to the relationship between class y_{jt} and attribute $x_{i(t-k-1)}$ can be described as in Fig. 3.

Analogous to the translation of Eq. (1) into Eq. (2), the prediction model of Naïve Bayes classification based on time is a model development obtained by expanding the features and records based on $t-k$ time before the target time, and is explained in Eq. (4).

$$\begin{aligned} & P\left(y_{jt} | x_{1(t-1)}, x_{2(t-1)}, \ldots, x_{n(t-1);}; x_{1(t-2)}, x_{2(t-2)}, \ldots, x_{n(t-2)}; \ldots; x_{1(t-k)}, x_{2(t-k)}, \ldots, x_{n(t-k)}\right) \\ & = \frac{P\left(x_{1(t-1)}, x_{2(t-1)}, \ldots, x_{n(t-1);}; x_{1(t-2)}, x_{2(t-2)}, \ldots, x_{n(t-2)}; \ldots; x_{1(t-k)}, x_{2(t-k)}, \ldots, x_{n(t-k)} | y_j\right) P\left(y_{jt}\right)}{P\left(x_{1(t-1)}, x_{2(t-1)}, \ldots, x_{n(t-1);}; x_{1(t-2)}, x_{2(t-2)}, \ldots, x_{n(t-2)}; \ldots; x_{1(t-k)}, x_{2(t-k)}, \ldots, x_{n(t-k)}\right)} \end{aligned} \tag{4}$$

By using the joint probability of the numerator of Eq. (4) and describing Eq. (4) as follows

$$\begin{aligned} & P\left(y_{jt} | x_{1(t-1)}, x_{2(t-1)}, \ldots, x_{n(t-1);}; x_{1(t-2)}, x_{2(t-2)}, \ldots, x_{n(t-2)}; \ldots; x_{1(t-k)}, x_{2(t-k)}, \ldots, x_{n(t-k)}\right) \\ & \quad = P\left(y_{jt}\right) P\left(x_{1(t-1)} | y_{jt}\right) P\left(x_{2(t-1)} | y_{jt}, x_{1(t-1)}\right) \ldots P\left(x_{n(t-k)} | y_{jt}, x_{1(t-k)}, \ldots x_{(n-1)(t-k)}\right) \\ & \\ & P\left(x_{n(t-k);} | y_{jt}, x_{1(t-1)}, x_{2(t-1)}, \ldots, x_{n(t-1);}; x_{1(t-2)}, x_{2(t-2)}, \ldots, x_{n(t-2)}; \ldots; x_{1(t-k)}, x_{2(t-k)}, \ldots, x_{(n-1)(t-k)}\right) \end{aligned} \tag{5}$$

In addition to fulfilling the stationary assumption, in Eq. (5) it must also be assumed that each feature is independent of each other. So the equation can be written in the form

$$\begin{aligned} & P\left(y_{jt} | x_{1(t-1)}, x_{2(t-1)}, \ldots, x_{n(t-1);}; x_{1(t-2)}, x_{2(t-2)}, \ldots, x_{n(t-2)}; \ldots; x_{1(t-k)}, x_{2(t-k)}, \ldots, x_{n(t-k)}\right) \\ & \quad = P\left(y_{jt}\right) \prod_{i=1}^{n} P\left(x_{i(t-k)} | y_{jt}\right), k = 1, 2, 3, \ldots t-1 \end{aligned} \tag{6}$$

Table 3 Class labeling of DHF data set

Class	Class Label	Range
Low	0	Cases < 55
Medium	1	55 ≤ Cases < 100
High	2	Cases ≥ 100

Table 4 Class labeling of rainfall data set

Class	Class Label	Range
Cloudy	1	RR < 0
Light Rain	2	0 ≤ RR < 20
Moderate Rain	3	20 ≤ RR < 50

Analogous to Eq. (3), where the denominator in (4) does not depend on the target class, but acts as a scaling factor, so that the maximum posterior rule which produces the Maximum A Posteriori (MAP) class can be expanded as follows

$$\begin{aligned}\widehat{y_t} &= arg\max_{y_{jt}} P\left(y_{jt}|x_{1(t-1)}, x_{2(t-1)}, \ldots, x_{n(t-1);}; x_{1(t-2)}, x_{2(t-2)}, \ldots\right.\\ &\quad \left., x_{n(t-2)}; \ldots; x_{1(t-k)}, x_{2(t-k)}, \ldots, x_{n(t-k)}\right)\\ &= arg\max_{y_{jt}} P\left(y_{jt}\right)\prod_{i=1}^{n} P\left(x_{i(t-k)}|y_{jt}\right), k = 1, 2, 3, \ldots, t-1\end{aligned} \tag{7}$$

If the features used are continuous, then the development of a classification prediction model based on time is also carried out on the normal distribution model or Gaussian distribution [41, 42] namely

$$P\left(X = x_{i(t-k)}|Y = y_{jt}\right) = \frac{1}{\sqrt{2\pi\sigma_{jt}}} e^{\left(-\frac{\left(x_{i(t-k)}-\mu_{jt}\right)^2}{2\sigma_{jt}^2}\right)} \tag{8}$$

where: μ: the mean of all attributes σ: standard deviation 1.

Data set

The DHF data set is a collection of data on the number of DHF cases from 30 sub-districts in Bandung City, West Java, Indonesia from 2012 to 2018. Meanwhile, the features used as factors that accelerate the spread of dengue cases are rainfall (mm), humidity (C), temperature (C), altitude (mdpl), number of people who are male, total population, number of people who do not have an education certificate, number of people with elementary school education, number of people with high school education, number of people with education high school, the number of people with undergraduate education. Table 3 contains an explanation of class index rate (IR) labeling for the number of DHF cases. IR is categorized as low if the number of cases is less than 55 per 100,000 population, moderate if the number of cases is in the range of 55 to 100 per 100,000 population, and high if the number of cases is more than 100 per 100,000 population.

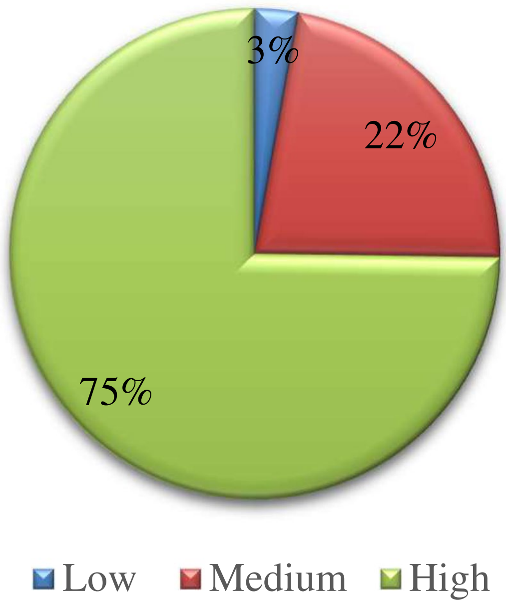


Fig. 4 The class proportion of DHF data set

The rainfall dataset in this study was obtained from the Meteorology, Climatology and Geophysics Agency at 28 location points on the island of Java, the time period from June 2021 to March 2022. The features used are the percentage of humidity (%), the length of time the sun shines (hours), wind direction (°), maximum speed, average wind speed (m/s), maximum temperature, minimum temperature, and average temperature. The data set in this study is labeled into 3 classes of rainfall (RR) that fall to the surface. Table 4 describes the distribution of data categories for class labeling.

After the labeling process is carried out, the data is transformed into time-series data, and a classification training data model will be developed to predict the class distribution of DHF and rainfall. Model development is carried out in two scenarios. The first scenario is to develop a model with feature column expansion based on the previous amount of time spent predicting the target. The second scenario is to develop a model by expanding the row of records and column of features based on the previous amount of time. The purpose of this scenario is to form a data set with new features based on time and a combination of time and records. This scenario has an effect on the formation of input data, which is made by providing a time range as input and as a boundary between training data and test data. Expansion of time-based features that are used as inputs in annual or monthly classification prediction models, based on the time range and characteristics of the data set.

The next process is data separation, which is conducted by dividing the form of time series data into several parts. It aims to form several data models before being implemented into the Naive Bayes classification method. Separation of data is done by dividing all data into several models according to the time range and characteristics of the data set. Meanwhile, feature selection in this study is used to solve multicollinearity problems, or conditions where there are several correlated variables and remove some irrelevant features. Feature selection can reduce feature dimensions, there by saving

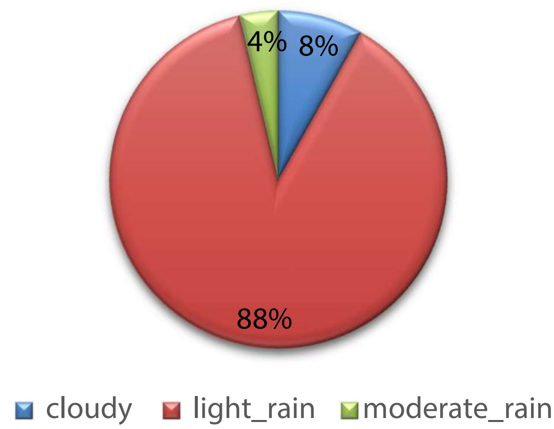


Fig. 5 The class proportion of rainfall data set

Table 5 Multiclass Classification Problem Confusion Matrix [41]

		Predicted Class			
		C_1	C_2	...	C_N
True Class	C_1	$C_{1,1}$	FP	...	$C_{1,N}$
	C_2	FN	TP	...	FN
	...	...	...	...	
	C_N	$C_{N,1}$	FP	...	$C_{N,N}$

resources for storing, processing data, and increasing the interpretability of the selected features.

The feature selection method is used to select a combination of features that affect the target column to be predicted. By selecting the relevant features, it will reduce the time complexity and can provide good accuracy for the system [39, 40].

Meanwhile, the implementation of the Naive Bayes classification method is applied to time series data with new features formed from the feature expansion process. In the data set with the new feature, each record contains the location id and the target Cases class of DHF in the DHF data set. Meanwhile, in the rainfall data, the target class is RR.

The two datasets used in this research are characterized as imbalanced data. In the DHF dataset, the "High" class has a proportion of 75%, much larger than the other two classes. In the Rainfall dataset, the "light_rain" class also has a proportion of 88%, much larger than the other 2 classes. To handle unbalanced data, an oversampling technique (SMOTE) is applied at the preprocessing stage. Figures 4 and Fig. 5 describes the class proportions of the DHF and Rainfall data sets.

Performance of Naïve Bayes-Time-Based Feature Expansion

The prediction case performed in this research is classification-based prediction. The focus of the classification is not related to the prediction result of a particular class, so the accuracy metric is actually quite appropriate in this case, which is also often used in other data mining research. However, because the data used is imbalanced, to complement the model performance measurement results, the F1-Score measurement is also added in this research. Evaluation of the performance of the feature expansion scenario in developing a classification prediction model based on $t-k$ time was previously based on the values of classification accuracy and F1-Score. This accuracy is a measure that describes the system's performance in producing correct predictions. Calculation of classification accuracy in this study uses a multiclass confusion matrix, because the number of target classes is more than two. The multiclass confusion matrix is described in Table 5, which has dimensions NxN, where N is the number of different class labels $C_1, C_2, \ldots, C_N$. Equation (9) and (10) are the formulas for calculating classification accuracy and F1-Score based on a multiclass confusion matrix [41].

$$Accuracy = \frac{\sum_{i=1}^{N} TP(C_i)}{\sum_{i=1}^{N} \sum_{j=1}^{N} C_{i,j}} x100\% \tag{9}$$

$$F1-Score = 2\frac{TPR(C_i)PPV(C_i)}{TPR(C_i)PPV(C_i)} x100\% \tag{10}$$

Meanwhile, to test the effect of the number of feature combinations and the classification prediction model of the Naive Bayes—Time-based Feature Expansion method. The experiment is carried out to see the performance of the classification prediction model for each scenario and the number of the best combinations of features. Comparisons were made using analysis of variance of two groups [42] namely based on the model and number of features. The hypotheses is defined as follows, null hypotheses (H_0) is all scenarios give the same response and alternative hypotheses (H_1) is at least there are pairs of scenarios that give different responses. Table 6 is a variance analysis table to test the significance of the influence of the number of features in the classification prediction model based on time.

The efficiency of developing a prediction model for the previous $t-k$ classification in predicting the future $t+k$ classification, is based on a comparison between the prediction accuracy using feature expansion based on time and the linear regression prediction model described in Eq. (11) [43, 44].

$$Y = \beta_0 + \beta_1 X \tag{11}$$

Vizualization of classification prediction

Results of developing a time-based classification prediction model using Naive Bayes Time-Based Feature Expansion implemented on spatial-time data. The time factor in

Table 6 Variance analysis table

Source of variation	Sum Square (SS)	Degree of free (Df)	Mean Square (MS)	F Value
Features(Between Groups (BG))	SSBG	k—1	$MSBG = \frac{SSBG}{k-1}$	$\frac{MSBG}{MSWG}$
Models(Within Groups (WG))	SSWG	k (r – 1)	$MSWG = \frac{SSWG}{k(r-1)}$	
Total (T)	SST	kr—1		

the data has been implemented in the feature expansion process to develop a classification prediction model based on the previous $t-k$ times. Spatial factors are used to visualize the prediction results of the future $t+k$ classification and are implemented on the spatial location map.

Spatial-time data is measurement data that contains location and time information [45, 46]. This spatial-time data becomes input in the estimation process. For example S_i, where $i = 1, 2, 3, \ldots n$, is a location with coordinates (x_i, y_i). Then $Y_t(S_i)$ is the prediction data for the classification class Y_t at location or coordinates S_i. Spatial data is a dependent data model, because spatial data is collected from different spatial locations which indicates a dependency between measurement data and location. Semivariogram in spatial-time analysis is a tool for measuring variability in distance, direction and time. The semivariogram model is an empirical model obtained from data, and this model is used to estimate the target class of a location. This estimation procedure is called kriging interpolation.

Several theoretical semivariogram models used to fit experimental semivariogram models are Nugget Effect, Spherical, Exponential, Gaussian and Linear. The kriging interpolation method used in this research is ordinary kriging. This method is used to visualize the classification prediction results at each location.

$$\widehat{Y}_t(S_0) = \sum_{b=1}^{n} \omega_b^{OK} Y_t(S_b) \tag{12}$$

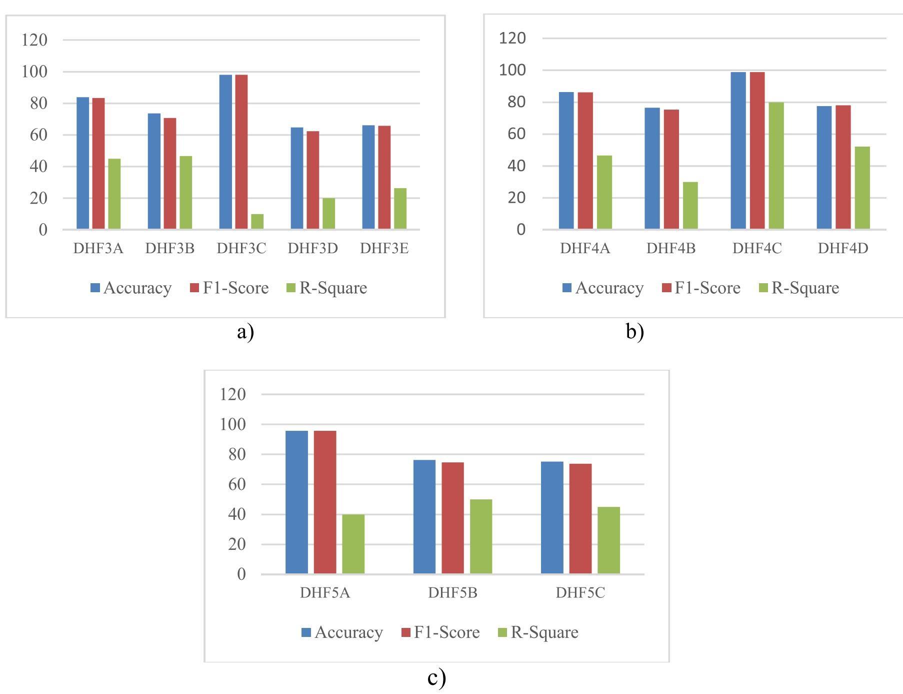


Fig. 6 Performance Comparison of accuracy and F1-Score (NBTB-FE Model) vs. R-squared (Regression) for DHF data set, **a** $t-3$, **b** $t-4$ and **c** $t-5$ model

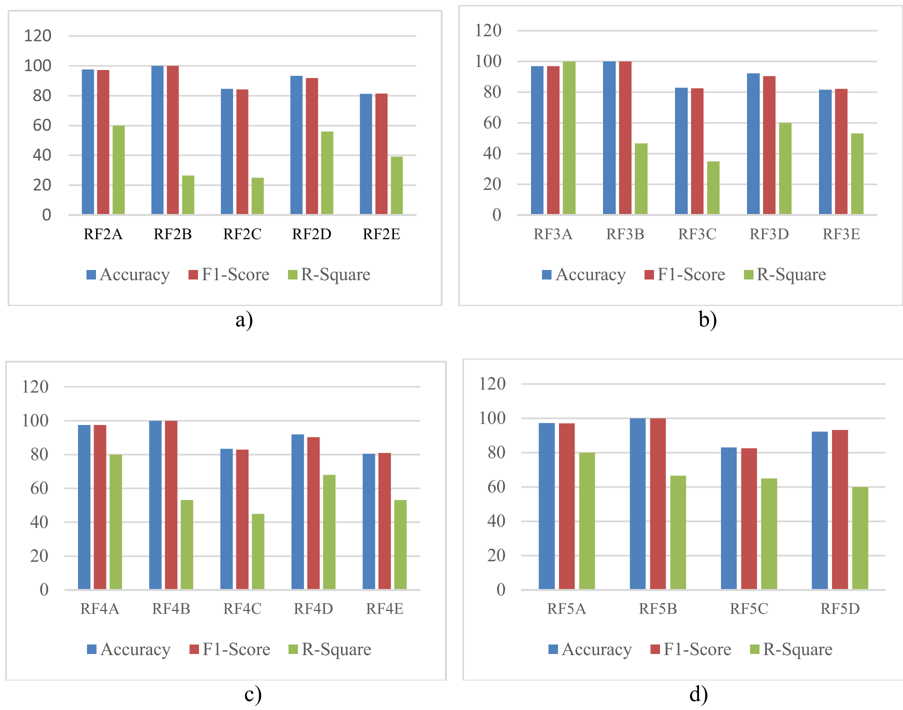


Fig. 7 Performance Comparison of accuracy and F1-Score (NBTB-FE Model) vs.R-squared (Regression) for Rainfall data set, **a** $t-2$, **b** $t-3$, **c** $t-4$, and **d** $t-5$ model

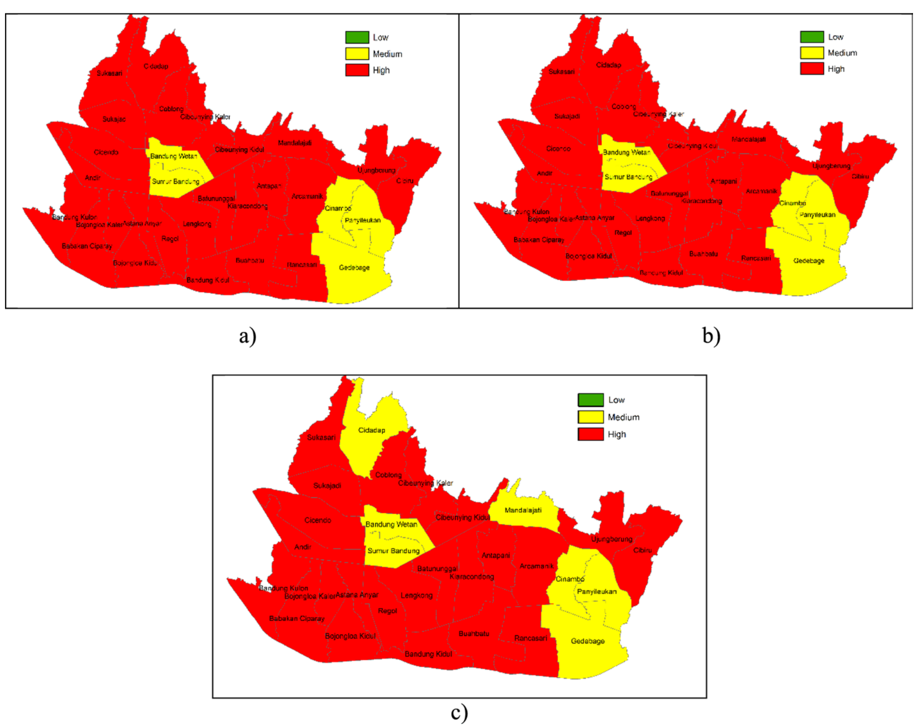


Fig. 8 DHF case classification prediction map in **a** 2021, **b** 2022, **c** 2023

Table 7 Models combination of the DHF data sets

Prediction Period	Model $t-k$	Model Name	Features Expansion	Target Prediction
Years	$t-3$	DHF3A	2015, 2016, 2017	2018
		DHF3B	2014, 2015, 2016	2017
		DHF3C	2013, 2014, 2015	2016
		DHF3D	2012, 2013, 2014	2015
		DHF3E	Combined 2012–2017	2015–2018
	$t-4$	DHF4A	2014, 2015, 2016, 2017	2018
		DHF4B	2013, 2014, 2015, 2016	2017
		DHF4C	2012, 2013, 2014, 2015	2016
		DHF4D	Cpmbined 2012–2017	2016–2018
	$t-5$	DHF5A	2013, 2014, 2015, 2016, 2017	2018
		DHF5B	2012, 2013, 2014, 2015, 2016	2017
		DHF5C	Combined 2012–2017	2017–2018

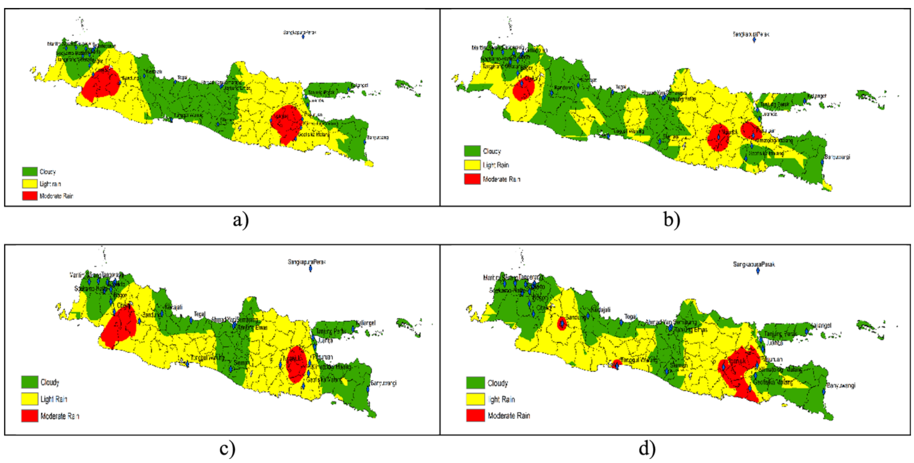


Fig. 9 Rainfall classification prediction map for May to August 2022, **a** May, **b** June, **c** July, **d** August

where $\widehat{Y}_t(S_0)$ is estimated classification class at S_0 point, ω_b^{OK} data weight value from the OK system and $Y_t(S_b)$ is sample location classification class.

Experiment result

This section describes the results of implementing the data set on the system developed in this study. The results of selection and implementation of the Naive Bayes classification method with new features resulting from feature expansion and feature selection are shown in Figs. 6 and 7. The Table 11 explains the significance of the effect of the number of features and the prediction model. classification with two-way Analysis of Variance. Meanwhile, Tables 12 and 13 shows the selected classification prediction model that is used to predict the classification for some time in the future for each data set. For spatial-time visualization of the classification prediction map shown in Figs. 8 and 9.

Table 8 Models combination of the Rainfall data sets

Prediction Period	Model $t - k$	Model Name	Features Expansion	Target Prediction
Months	$t - 2$	RF2A	01/22, 02/22	03/22
		RF2B	12/21, 01/22	02/22
		RF2C	11/21, 12/21	01/22
		RF2D	10/21, 11/21	12/21
		RF2E	Combined 10/21–02/22	12/21–03/22
	$t - 3$	RF3A	12/21, 01/22, 02/22	03/22
		RF3B	11/21, 12/21, 01/22	02/22
		RF3C	10/21, 11/21, 12/21	01/22
		RF3D	09/21, 10/21, 11/21	12/21
		RF3E	Combined 09/21–02/22	12/21–03/22
	$t - 4$	RF4A	11/21, 12/21, 01/22, 02/22	03/22
		RF4B	10/21, 11/21, 12/21, 01/22	02/22
		RF4C	09/21, 10/21, 11/21, 12/21	01/22
		RF4D	08/21, 09/21, 10/21, 11/21	12/21
		RF4E	Combined 08/21–02/22	12/21–03/22
	$t - 5$	RF5A	10/21—02/22	03/22
		RF5B	09/21—01/22	02/22
		RF5C	08/21—12/21	01/22
		RF5D	07/21—11/21	12/21

Table 9 The average performance of previous time $t - k$ prediction model using NBTF-FE and Regression

Data Set	Prediction Period	Prediction Model	NBTB-FE		Regression
			Accuracy(%)	F1-Score(%)	R-Square
	Years	$t - 3$	77.29	76.09	29.59
DHF		$t - 4$	84.80	84.59	52.22
		$t - 5$	82.39	81.42	45.00
Rainfall	Months	$t - 2$	91.39	90.97	41.34
		$t - 3$	90.74	90.36	58.95
		$t - 4$	90.66	90.32	59.86
		$t - 5$	93.11	93.22	67.90

The performance of the classification prediction model for the previous $t - k$ times

The development of the classification prediction model in this study uses the Naive Bayes classifier method, using two feature expansion scenarios. This feature expansion process will certainly increase the number of features where these additional features are not necessarily important. Furthermore, feature selection is used to maintain the balance of the addition of dimensions, which occurs during the development of the classification prediction model. The use of feature selection will optimize the number of influential feature combinations and increase the accuracy of the classification prediction model with Naive Bayes.

Development of a classification prediction model on the data set of the number of cases of DHF, expansion of features based on time in units of years. Meanwhile, in the rainfall data set, feature expansion is based on time in months. The expansion

Table 10 The best of the classification prediction model for the previous $t-k$ times using NBTF-FE

Data Set	Prediction Period	The Best Model	Accuracy (%)	F1-Score (%)
	Years	DHF3C	98.00	97.97
DHF		DHF4C	98.00	97.97
		DHF5A	89.33	89.29
Rainfall	Months	RF2C	82.44	82.44
		RF3A	98.00	97.97
		RF4C	93.55	93.54
		RF5C	95.55	95.44

Table 11 The effect significance of the NBTF-FE model variation and the evaluation metrics

Data Set	Prediction Scenario	Source of Variation	Degree of Freedom	F Value of Variation Source	P Value of Variation Source	F Critical
DHF	3-year	Models	4	454.83	**$1.44\ 10^{-5}$**	6.39
		Evaluation metrics	1	4.04	0.11	7.71
	4-year	Models	3	804.16	**$7.43\ 10^{-5}$**	9.28
		Evaluation metrics	1	0.33	0.61	10.13
	5-year	Models	2	851.36	**0.0012**	19.00
		Evaluation metrics	1	4.10	0.18	18.51
Rain Fall	2-month	Models	4	751.76	**$5.29\ 10^{-6}$**	6.39
		Evaluation metrics	1	2.48	0.19	7.71
	3-month	Models	4	340.92	**$2.56\ 10^{-5}$**	6.39
		Evaluation metrics	1	0.88	0.40	7.71
	4-month	Models	4	419.03	**$1.7\ 10^{-5}$**	6.39
		Evaluation metrics	1	0.83	0.41	7.71
	5-month	Models	3	751.76	**$9.98\ 10^{-5}$**	9.38
		Evaluation metrics	1	2.48	0.73	10.13

scenario is adjusted to the characteristics of each data set. Table 7 show models combination of the DHF data sets. The $t-k$ model in Table 7 shows the prediction model built based on the previous t years. For example, in the DHF3A model, in order to predict dengue cases in 2018, features from the previous 3 years, namely 2015, 2016 and 2017, are used. The goal is that we can have a prediction model for the next 3 years based on data from the last 3 years.

Table 8 show models combination of the Rainfall data set. The $t-k$ model in Table 8 shows the prediction model built based on the previous t months. For example, the RF2A model states the weather prediction model for the next 2 months (March 2022). The model is trained using features obtained in January and February of 2022.

Feature and record expansion scenarios are used in the training data when inputted to the Naive Bayes classifier. Simple linear regression was implemented on both data sets, used as a comparison of the prediction results with the feature expansion method in the Naive Bayes classification.

Figure 6 show the accuracy and F1-Score for $t-3$, $t-4$ and $t-5$ prediction models using NBTB-FE and regression for DHF data sets while Fig. 7 show the accuracy and

Table 12 The Feature characteristics of DHF classification prediction model for 2021, 2022, 2023

Years	Model	Features
2021	DHF3C	6
2022	DHF4C	11
2023	DHF5A	10

Table 13 The Feature characteristics of rainfall classification prediction model for May to August 2022

Months	Model	Features
May	RF2C	10
June	RF3A	10
July	RF4C	18
August	RF5C	32

F1-Score for $t-2$ until $t-5$ prediction models using NBTB-FE and regression for Rainfall data sets.

Table 9 shows the average performance of the previous t-k time classification prediction model using Naïve Bayes Time Based Feature Expansion (NBTB-FE) and Regression on both datasets. Meanwhile, Table 10 shows the performance values of the selected models by considering the influence of the number of features originating from feature expansion. Table 9 shows that in general the performance of the time-based classification prediction model using NBTB-FE outperforms all regression models.

Evaluation of the result

This chapter discuss the effect of the number of feature combinations and the classification prediction model of the Naive Bayes Time-based Feature Expansion method. The experiment is carried out to see the performance of the classification prediction model for each scenario and model performance. Comparisons were made using Analysis of Variance (ANOVA) of two groups, namely based on the model and evaluation metrics. Table 11 shows that there is a significant influence on the classification prediction model scenario with Naive Bayes Time-based Feature Expansion and the model performance values (Accuracy and F1-Score).

Tests using a 95% confidence interval show that the average of the two scenarios, both on dengue fever and rainfall data, has a significant influence. This is indicated by a P value of less than 0.05 and a large F statistical value, both based on the classification prediction model and model performance.

Features characteristics of the classification prediction model for the previous $t+k$ times

The feature selection process for each model is basically done by considering the accuracy or F1-Score value obtained and also the number of features. The number of features that is too large, even though it has the highest accuracy or F1-Score value tends to make the resulting model overfitting. Conversely, too few features will make the model

Table 14 Prediction of the DHF data set classification model for 2021, 2022, 2023

Location	2021	2022	2023
Andir	2	2	2
Cibiru	2	2	2
Sukajadi	2	2	2
Sukasari	2	2	2
Antapani	2	2	2
Arcamanik	2	2	2
Astana Anyar	2	2	2
Babakan Ciparay	2	2	2
Bandung Kidul	2	2	2
Bandung Kulon	2	2	2
Bandung Wetan	1	1	1
Batununggal	2	2	2
Bojongloa Kaler	2	2	2
Bojongloa Kidul	2	2	2
Buahbatu	2	2	2
Cibeunying Kaler	2	2	2
Cibeunying Kidul	2	2	2
Cicendo	2	2	2
Cidadap	2	2	1
Cinambo	1	1	1
Coblong	2	2	2
Gedebage	1	1	1
Kiaracondong	2	2	2
Lengkong	2	2	2
Mandalajati	2	2	1
Panyileukan	1	1	1
Rancasari	2	2	2
Regol	2	2	2
Sumur Bandung	1	1	1
Ujungberung	2	2	2

underfitting, which is when a model is too simple to capture the patterns in the data. The appearance of a feature in several models also indicates that the feature is an important feature to choose.

Tables 12 and 13 shows the number of features for the $t-k$ classification prediction model for the DBD and Rainfall dataset, which was selected by considering the characteristics of underfitting, overfitting, and the frequency of occurrence of potential features.

Tables 14 and 15 show the prediction results from the selected models in Tables 12 and 13. The prediction results in these two tables will be visualized on a classification prediction map for the distribution of the number of dengue cases and changes in rainfall.

Visualization of classification prediction

The prediction results of the classification model obtained by the Naive Bayes Time-Based Feature Extension method which are implemented on DHF and rainfall data are presented in the form of a map. The aim is to visually describe the transition of

Table 15 Prediction of the Rainfall data set classification model for May to August 2022

Location	May 2022	June 2022	July 2022	August 2022
Tangerang	2	2	2	2
Tangerang Selatan	2	2	2	2
Budiarto	2	2	2	2
Maritim Serang	2	2	2	2
Soekarno-Hatta	2	2	2	2
Sleman	2	2	2	2
Kemayoran	2	2	2	2
Tanjung Priok	2	2	2	2
Bandung	3	2	3	3
Bogor	2	2	2	2
Citeko	3	3	3	2
Kertajati	2	2	2	2
Tegal	2	2	2	2
Tunggul Wulung	2	2	3	3
Tanjung Emas	2	2	2	2
Ahmad Yani	2	2	2	2
Semarang	2	2	2	2
Sangkapura	2	2	2	2
Perak	2	2	2	2
Tanjung Perak	2	2	2	2
Kalianget	2	2	2	2
Juanda	2	2	2	2
Banyuwangi	2	2	2	2
Klimatologi Malang	3	2	3	2
Pasuruan	3	3	3	3
Nganjuk	3	3	3	3
Geofisika Malang	3	2	3	3

classification changes over time at each location. Maps are a form of visualization of prediction classification results at the spatial location of each data set, which are obtained by kriging interpolation. Meanwhile, the map development process uses ArcGis software ((https://pro.arcgis.com/).

The results of the best model predictions in Table 14, are used as the basis for making prediction maps for their classification for the next 1 to 3 years in the DHF dataset, namely 2021, 2022 and 2023 and are presented in Fig. 8. Classification of a sub-district is indicated by 3 colors, where the color is green. indicates class 0, yellow indicates class 1, and red indicates class 2. While Fig. 9 is a form of visualization of the prediction of rainfall classification in Java from May to August 2022, which is made based on Table 15.

Discussion and conclusions

The problems studied in this research are the development of time series data matrices using the feature expansion method for developing Naïve Bayes time-based (NBTB-FE) classification prediction models and also to find out the significance of the influence of the number of features of classification prediction models. The resulting Naive Bayes classification prediction model (NMTB-FE) was applied to spatio-temporal data. The selection of the DHF and Rainfall datasets for the experiment was based on their characteristics as spatio-temporal data and this data has also been implemented in previous studies [23–25] using Naive Bayes classifiers or others. The DHF and rainfall datasets used were unbalanced, so oversampling (SMOTE) was performed, and high performance was obtained in almost all models. Based on the imbalance characteristics of the two data sets, accuracy and F1-Score are used to measure the performance of the NBTB-FE model. As a base model, a regression model that uses the R-Square value as a performance measure is used. This R-Square value is then compared with the accuracy and F1-Score values of the NBTB-FE model.

Time-based feature matrix expansion is used to develop a feature expansion model for the previous time. Furthermore, the feature expansion model for time t-k with the best performance will be used for NBTB-FE for the future time t + k, at observed and unobserved locations. Currently, there is no previous research that develops Naive Bayes classification prediction models with time-based feature data input expansion like this.

In general, the performance of the NTBB-FE model implemented on the data set gives better results than the regression model as a baseline model. The level of significance of the resulting model performance varies depending on the type and quality of the data used. The implementation NBTB-FE on DHF and rainfall dataset show that the average accuracy and F1-Score of the NBTB-FE obtained is superior to the R-Square linear regression model for the previous $t-k$ time. The performance of the NBTB-FE model on the same dataset in present time prediction is also outperforms compare to the traditional Naive Bayes classifier, other classifiers and hybrids classifier [23–25].

The accuracy and F1 of feature selection results using Naïve Bayes Time-Based Feature Expansion shows very significant changes in model combinations in each time period. On the DBD data set based on the annual classification prediction model, optimal performance was obtained in the time-based feature expansion scenario with an expansion of the previous four years. Meanwhile, the optimal performance of the monthly classification prediction model on the Rainfall data set can be achieved in the time-based feature expansion scenario with an expansion of the previous three years. These characteristics are used as the basis for feature selection in the t + k prediction model. The basis for selecting optimal features in the $t+k$ prediction model is the conditions of underfitting, overfitting and the frequency of appearance of features that have the potential to improve the classification target [31].

Testing the influence of the model and the number of features in each model developed based on feature expansion in the previous $t-k$ shows that the optimal number of features greatly influences the performance of the previous $t-k$ classification prediction model. Time-based feature expansion methods can help determine the optimal number of factors and time expansion.

For the case study of the two datasets tested, the following results were obtained. The classification prediction the number of DHF cases in 2021 uses the DHF3C model, in 2022 uses the DHF4C model, while in 2023 uses the DHF5A model. Based on this, it can be seen that the combination of features from 2015, 2014, 2013 is a potential feature, because each of them was the best model, with an accuracy and F1-Score of more than 97%. However, after being expanded again with the 2016 and 2017 features, the accuracy fell by 8.67% and the F1-Score fell by 8.68%. Meanwhile, the prediction model for monthly classification of rainfall data sets, the models used are RF2C, RF3A, RF4C,and RF5C. Based on these results, it can be explained that the features of February 2022, January 2022, December 2021, are very potential, because they dominate all models.

The obtained $t+k$ classification prediction model can be visualized well in the form of classes and distribution maps to determine the pattern of vulnerability status of the increase in the number of annual dengue cases and monthly climate changes.

Based on the experimental results and discussions, it was found that the Naïve Bayes method developed with the time-based feature expansion scenario and the concept of developing a prediction model on time series, can improve its ability to predict classification for the future. The prediction results obtained using the proposed model can also outperform the prediction results using regression analysis.

The limitation of the proposed feature expansion method is related to the reduced amount of data used for training the $t+k$ -prediction model. Although the results of the feature expansion method can potentially provide more accurate results and improve the ability to predict the future, it has the consequence that building a robust $t+k$ future prediction model requires a larger data size (k times) than building a model to predict the present For future research, it is still open to the use of other classification methods and comparison with deep learning methods such as CNN, RNN or LSTM. The development of feature expansion formulas for other classifier models and model validation using various types of data can also be done as future work of this research.

Abbreviations

NB	Naïve Bayes
ANOVA	Analysis of variance
TB	Time-based
FE	Feature expansion
NBTB-FE	Naïve bayes time-based feature expansion
DHF	Dengue Hemorrhagic Fever
TP	True positive
P	Probability
F	Fisher
BG	Between groups
WG	Within groups
T	Total
SS	Sum square
SSBG	Sum square between groups
SSWG	Sum square within groups
SST	Sum square total
DF	Degree of freedom
MS	Mean square
MSBG	Mean square between groups
MSWG	Mean square within groups

Acknowledgements
We would like to thank Telkom University for providing full support for this research, so that we can complete this research.

Author contributions
The author declares full responsibility for the creation of this manuscript which includes learning the concept and design, collecting and preprocessing data sets, analyzing and interpreting results, and writing the manuscript. The author has read and approved the final manuscript. A developed concepts, formulas and wrote the main manuscript text B developed the algorithm, prepared images and finalized the text of the manuscript.

Funding
This research was not funded by a specific grant from any funding agency, whether public, commercial, or not-for-profit sectors.

Availability of data and materials

The original dataset used for this study is available in: 1. Naïve Bayes Time-Based Feature Expansion implemeted in python (https://drive.google.com/file/d/1YvkKsZ-7yVMancoSBDxKivZsuljpaz-k/view?usp=sharing). 2. Rainfall Dataset (https://docs.google.com/spreadsheets/d/1aHLg-v-CQD5lSqLt_XleAuAEMPslmZ23/edit?usp=drive_link&ouid=115398213045183747364&rtpof=true&sd=true). 3. DHF Dataset (https://docs.google.com/spreadsheets/d/1L3r-PktXYhwo2reYvp38sN9vmafhJgug/edit?usp=drive_link&ouid=115398213045183747364&rtpof=true&sd=true). 4. ArcGis Software (https://pro.arcgis.com/). Data is provided within the manuscript or supplementary information files.

Declarations

Competing interests

The authors declare no competing interests.

Received: 10 May 2024 Accepted: 13 July 2024
Published online: 05 August 2024

References

1. Robnik-Šikonja M. Explanation of prediction models with explain prediction. Inform. 2018;42(1):13–22.
2. Akhter M, Ahanger MA. Climate modelling using ANN. Int J Hydrol Sci Technol. 2019;9(3):251–65. https://doi.org/10.1504/IJHST.2019.102316.
3. Yesilkanat CM. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos Solitons Fractals. 2020. https://doi.org/10.1016/j.chaos.2020.110210.
4. Nikparvar B, Thill JC. Machine learning of spatial data. ISPRS Int J Geo-Information. 2021;10(9):1–28. https://doi.org/10.3390/ijgi10090600.
5. Ahn S, Ryu DW, Lee S. A machine learning-based approach for spatial estimation using the spatial features of coordinate information. ISPRS Int J Geo-Information. 2020. https://doi.org/10.3390/ijgi9100587.
6. Pourghasemi HR, et al. Spatial modeling, risk mapping, change detection, and outbreak trend analysis of coronavirus (COVID-19) in Iran (days between February 19 and June 14, 2020). Int J Infect Dis. 2020;98:90–108. https://doi.org/10.1016/j.ijid.2020.06.058.
7. Alkhamis MA, et al. Spatiotemporal dynamics of the COVID-19 pandemic in the State of Kuwait. Int J Infect Dis. 2020;98:153–60. https://doi.org/10.1016/j.ijid.2020.06.078.
8. Atluri G, Karpatne A, Kumar V. Spatio-temporal data mining: A survey of problems and methods. ACM Comput Surv. 2018;51(4):1–37. https://doi.org/10.1145/3161602.
9. Kolesnikov AA, Kikin PM, Portnov AM. Diseases spread prediction in tropical areas by machine learning methods ensembling and spatial analysis techniques. Int Arch Photogramm Remote Sens Spatial Inf Sci. 2019;42:221–6.
10. Mohajane M, et al. Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area. Ecol Indic. 2021;129:107869. https://doi.org/10.1016/j.ecolind.2021.107869.
11. Fouedjio F. Classification random forest with exact conditioning for spatial prediction of categorical variables. Artif Intell Geosci. 2021;2(October):82–95. https://doi.org/10.1016/j.aiig.2021.11.003.
12. MinminMiao F, et al. Discriminative spatial-frequency-temporal feature extraction and classification of motor imagery EEG: an sparse regression and Weighted Naïve Bayesian Classifier-based approach. J Neurosci Methods. 2017;278:13–24.
13. AbMunag JI, Prasadb VNK, Nickolasa S, Gangadharan GR. Representational primitives using trend based global features for time series classification. Expert Syst Appl. 2021. https://doi.org/10.1016/j.eswa.2020.114376.
14. Gao CZ, Cheng Q, He P, Susilo W, Li J. Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack". Info Sci. 2018. https://doi.org/10.1016/j.ins.2018.02.058.
15. Chen S, Webb GI, Liu L, Ma X. A novel selective naïve Bayes algorithm. Knowl Based Syst. 2020;192:105361.
16. Karabatak M. A new classifier for breast cancer detection based on Naïve Bayesian. Measurement. 2015;72:32–6.
17. Tsangaratos P, Ilia I. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: the influence of models complexity and training dataset size. CATENA. 2016;145:164–79.
18. Zhang L, Jiang L, Li C, Kong G. Two feature weighting approaches for naive Bayes text classifiers. Knowl Based Syst. 2016;100:137–44.
19. Blanquero R, Carrizosa E, Ramírez-Cobo P, Sillero-Denamiel MR. Variable selection for Naïve Bayes classification. Comput Oper Res. 2021;135: 105456. https://doi.org/10.1016/j.cor.2021.105456.
20. Padmavathi S, Ramanujam E. Naïve Bayes Classifier for ECG abnormalities using multivariate maximal time series motif. Procedia Comput Sci. 2015;47:222–8. https://doi.org/10.1016/j.procs.2015.03.201.
21. Tang X, Shu Y, Lian Y, Zhao Y, Fu Y. A spatial assessment of urban waterlogging risk based on a Weighted Naïve Bayes classifier. Sci Total Environ. 2018;15(630):264–74.
22. Viet TN, Le Minh H, Hieu LC, Anh TH. The naÏve bayes algorithm for learning data analytics. Indian J Comput Sci Eng. 2021;12(4):1038–43. https://doi.org/10.21817/indjcse/2021/v12i4/211204191.
23. Inayah FN, Prasetiyowati SS, Sibaroni Y. Classification of Dengue Hemorrhagic Fever (DHF) Spread in Bandung using Hybrid Naïve Bayes, K-Nearest Neighbor, and Artificial Neural Network Methods, Int J Inf Commun Technol 2021;7(1):10–20. https://doi.org/10.21108/ijoict.v7i1.562.
24. Gumilar A, Prasetiyowati SS, Sibaroni Y. Performance analysis of hybrid machine learning methods on imbalanced data (rainfall classification). Jurnal RESTI. 2022;6(3):481–90.
25. Sidik DD, Sen TW. Penggunaan stacking classifier Untuk Prediksi Curah Hujan. IT Soc. 2019;4(1):21–7. https://doi.org/10.33021/itfs.v4i1.1180.

26. Storcheus D, Rostamizadeh A, Kumar S. A survey of modern questions and challenges in feature extraction. 1st Int Feature Extr Mod Quest Challenges. 2015;44:1–18.
27. Guyon I. CrossRef List. Deleted. 2000, https://doi.org/10.1162/153244303322753616.
28. Yao K, Lu W, Zhang S, Xiao H, Li Y. Feature expansion and feature selection for general pattern recognition problems. ICNNSP. 2003. https://doi.org/10.1109/ICNNSP.2003.1279205.
29. Tsai C-F, Lin W-Y, Hong Z-F, Hsieh C-Y. Distance-based features in pattern classification. EURASIP J Adv Signal Process. 2011;1:2011. https://doi.org/10.1186/1687-6180-2011-62.
30. Jung D, Lee J, Park H. Feature expansion of single dimensional time series data for machine learning classification. IEEE Xplore. 2021. https://doi.org/10.1109/ICUFN49451.2021.9528690.
31. Eden J. Expand Your Horizons 2021. .
32. Kaul A, Maheshwary S, Pudi V. Autolearn—automated feature generation and selection. Proc IEEE Int Conf Data Mining ICDM. 2017. https://doi.org/10.1109/ICDM.2017.31.
33. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: a new perspective. Neurocomputing. 2018. https://doi.org/10.1016/j.neucom.2017.11.077.
34. Kumar N, Maurya V, Maurya VK. A review on machine learning (Feature Selection, Classification and Clustering) approaches of big data mining in different area of research journal of critical reviews a review on machine learning (Feature Selection, Classification and Clustering) approach. Artic J Crit Rev. 2020. https://doi.org/10.31838/jcr.07.19.322.
35. Zhao S, Wang M, Ma S, Cui Q. A feature selection method via relevant-redundant weight. Expert Syst Appl. 2022. https://doi.org/10.1016/j.eswa.2022.117923.
36. Damoulas T, Girolami MA. Combining feature spaces for classification. Pattern Recognit. 2009;42(11):2671–83. https://doi.org/10.1016/j.patcog.2009.04.002.
37. Petrusevich DA. Features addition and dimensionality reduction in classification. IOP Conf Ser Mater Sci Eng. 2020. https://doi.org/10.1088/1757-899X/919/4/042018.
38. Berrar D. Bayes' theorem and naive bayes classifier. Encycl Bioinforma Comput Biol ABC Bioinforma. 2018;1–3(September):403–12. https://doi.org/10.1016/B978-0-12-809633-8.20473-1.
39. Chakrapani HB, Chouraisa S, Saha A, Swathi JN. Predicting performance analysis of system configurations to contrast feature selection methods. Int Conf Emerg Trends Inf Technol Eng IC-ETITE. 2020. https://doi.org/10.1109/ic-ETITE47903.2020.106.
40. Le Minh T, Van Tran L, Dao SVT. A feature selection approach for fall detection using various machine learning classifiers. IEEE Access. 2021;9:115895–908. https://doi.org/10.1109/ACCESS.2021.3105581.
41. Markoulidakis I, Kopsiaftis G, Rallis I, Georgoulas I. Multi-class confusion matrix reduction method and its application on net promoter score classification problem. ACM Int Conf Proceeding Ser. 2021. https://doi.org/10.1145/3453892.3461323.
42. Sawye S. Analysis of variance : the fundamental concepts. 2017, https://doi.org/10.1179/jmt.2009.17.2.27E.
43. Hallman J. A comparative study on Linear Regression and Neural Networks for estimating order quantities of powder blends. 2019.
44. Xiao Y, Jin Z. The forecast research of linear regression forecast model in national economy. OALib. 2021;8:1–17. https://doi.org/10.4236/oalib.1107797.
45. Chowdhury AI, et al. Analyzing spatial and space-time clustering of facility-based deliveries in Bangladesh. Trop Med Health. 2019;9:1–12.
46. Cressie N, Moores MT, Moores MT. Spatial Statistis. 2021.

Publisher's Note