JOURNAL OF DATA SCIENCE | Vol.2022:18
eISSN:2805-5160

Palembang Aerodrome Weather Forecast for Palembang Sultan Mahmud Badaruddin II Airport

Isti Ma'atun Nasichah [1], Tri Basuki Kurniawan [1*], Misinem [2]

[1]Program Studi Magister Teknik Informatika, Universitas Bina Darma, Palembang, Indonesia
[2]Program Studi Teknik Komputer, Fakultas Ilmu Komputer, Universitas Bina Darma, Palembang, Indonesia

[*]**Email**: tribasukikurniawan@binadarma.ac.id

Abstract

SMB II Palembang Meteorological Station is one of the weather observation points owned by BMKG in charge of carrying out weather observations, analysis, and weather forecasts at Sultan Mahmud Badaruddin II Airport Palembang. Weather information has an important role in the world of aviation, so accurate airport weather forecasts are needed. Random Forest, Naive Bayes Classifier, and Support Vector Machine methods are classification methods used to forecast rain in this study. The data used is weather parameter data from December 2012 to December 2021. Rain forecast using the SVM method produces an accuracy rate of 72%, the NBC method 66%, and the Random Forest method produces an accuracy rate of 74%. Heavy rains and very heavy rains can’t be predicted accurately using the SVM, NBC, and Random Forest methods. Based on the feature selection method, the attribute that has the most influence on rain forecasts is the average humidity.

Keywords

Weather Forecast; Random Forest, Naive Bayes Classifier; Support Vector Machine; Feature Selection

Introduction

Today's fast advancements in science and technology motivate us to carry out a variety of tasks successfully and efficiently. In order to fulfill the need for quick and accurate information, technology is crucial. Data mining, which enables the collection of massive volumes of data, is now an area of information technology development that is trending upward.

With regard to a variety of meteorological occurrences, BMKG has evolved into a resource for the public and the government. A rise in the standard of meteorological information services that are quick, precise, accurate, simple to grasp, and responsive to community requirements must coincide with the government and the public's growing interest. The most crucial element in ensuring the safety and security of air transportation against extreme weather is aviation

Submission: 10 November 2022; **Acceptance:** 12 December 2022

meteorology (Gultepe Ismail, et.al. 2019). In 23% of all flights currently in operation, weather plays a significant influence in aviation accidents (Kulesa, G. 2003).

The state of the atmosphere at a specific location and time is known as the weather (Tjasyono, Bayong, et.al, 2012). Therefore, the Sultan Mahmud Badaruddin II Palembang Meteorological Station is tasked with providing stakeholders in the Sultan Mahmud Badaruddin II Palembang Airport area with flight weather information in the form of a 24-hour airport weather forecast (aerodrome forecast), which is updated annually. 6 hours. Five categories were used in this study: "very light rain," "light rain," "moderate rain," "heavy rain," and "very heavy rain." While Random Forest, Naive Bayes Classifier, and Support Vector Machine are the three classification techniques employed, it is difficult to tell which one is more accurate. so that projections for the weather at airports may take it into account.

Methodology

In order to make airport weather forecasts, this study compared the performance of Random Forest, Naive Bayes Classifier, and Support Vector Machine. The actions done to accomplish this aim are as follows. The overall research method is depicted in the form of a flowchart in figure 1 and the research framework in figure 4.

Data source

Several meteorological variables will be employed in this study, including average air temperature, maximum air temperature, lowest air temperature, average air pressure, average dew point temperature, average air humidity, length of the day, and average daily rainfall.

Research Stages

The following are the steps of analysis used in this study:

Min-Max Normalization

The preference for qualities with a wide range of values over those with a narrow range of values. Additionally, when employing the equation below (J. Han, and M. Kamber, 2013), computations may be done without encountering numerical issues by utilizing min-max normalization in the following equation.

$$D'(i) = \frac{D(i) - min\,(D)}{max\,(D) - min\,(D)}(U - L) + L$$

(Equation 1)

D'(i) = value of the data I for the normalized attribute D.
D(i) = the data's initial value
Min (D) = is the data attribute's minimal value.
Max (D) = Maximum value of an attribute
U = upper
L = lower

Based on a comparison of input data normalizing techniques for the Support Vector Machine, the min-max normalization approach was chosen because it offers a higher degree of accuracy and performance than the zero mean and log scaling techniques (S. Ali, and A. Smith, 2010).

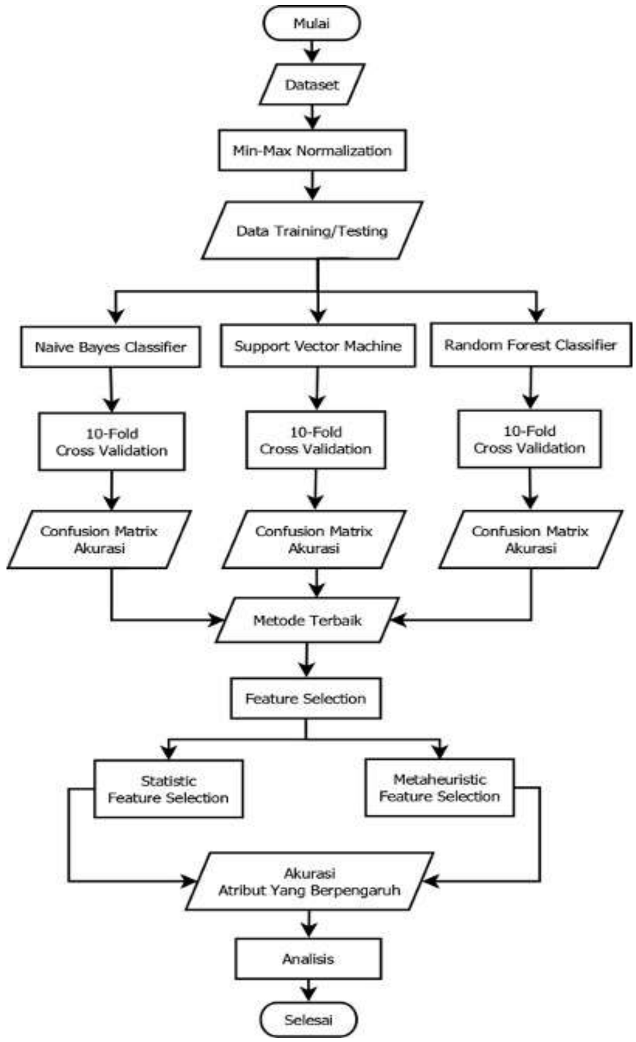

Figure 1. Research Flowchart

Data Mining

Data mining is the processing of a set of it to uncover unexpected links and to present data in a fresh form that makes it understandable (Novandya, Andhika dan Oktria, Isni, 2017.). The three methods of association rule mining, clustering, and classification are often employed in data mining (Estoatnowo, Ditya. 2016).

Finding and defining a model that can explain and discriminate between different data classes is the process of classification, which aims to utilize the data to forecast the class of an item whose status is unknown. Data collection, data cleaning to remove outliers or errors from noisy data, incomplete data (data with missing attribute values), inconsistent data (data with inconsistent attribute filling out), and data reduction to choose variables or attributes to be used in research are the stages of classification (Subhan, Ahmad; Ahmad Zainul Fanani, 2017).

Naïve Bayes Classifier

Bayesian classification, often known as the Naïve Bayes Classifier (NBC), is a classification technique that calculates the likelihood that a member of a class exists. A classification method based on straightforward probabilistic is called NBC. Similar to decision trees and neural networks, the Bayesian theorem is capable of classifying data. As a result, NBC is capable of managing irrelevant data and is effective, efficient, and dependable when handling enormous datasets. The class label is the symbol for Y, while the input vector containing the data is the symbol for X. The following is the categorization NBC equation:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X_i)}$$

(Equation 2)

For every X $= \{X_1, X_2, X_3 \ldots X_q\}$ as many as q attributes or q dimensions where:

$P(Y|X)$ = Probability of data with vector X in class Y
$P(Y)$ = Initial probability of class Y (Prior Probability)
$P(X|Y)$ = Posterior Probability
$\prod_{i=1}^{q} P(X_i|Y)$ = Independent probability of class Y in vector X
i = Observation i

This is an NBC functioning system since the value of $P(Y)$ will never change, therefore only $P(Y)\prod_{i=1}^{q} P(X_i|Y)$ is computed by choosing the probability with the highest likelihood of being utilized as the selected class in the prediction results.

Support Vector Machine

One of the categorization techniques used in data mining is the Support Vector Machine (SVM). SVM has the ability to predict both classification and regression outcomes. The benefits of SVM include its ability to accurately classify data that does not contain training data, the fact that the level of generalization is unaffected by the dimensions of the observed variables, and the fact that the computation is quicker due to the small number of observations required to create the decision function (B. Xu, el.al., 2012) (Siregar, Aplia Belina, 2017).

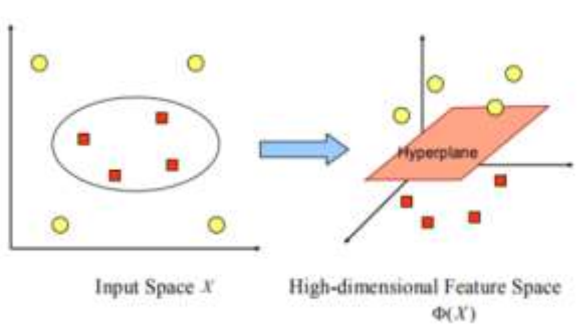

Figure 2. Hyperplane

Although SVM was originally created to operate on linear issues, it has now been developed to work on non-linear problems as well. SVM incorporates the kernel idea in a high-dimensional space where data from two classes may always be separated by a hyperplane in order to work on non-linear situations (Nugroho, et .al., 2003) (Ritonga, A. S., Purwaningsih, E. S, 2018). The separation or margin between data classes can be increased using hyperplanes. By determining the maximum point and then calculating the margin, the optimal hyperplane between the two classes may be discovered. The heart of the SVM method's procedure entails looking for the optimal hyperplane to use as a class separator (Assaffat, Luqman, 2015). The SVM concept is depicted in figure 2.

10-Fold Cross Validation

An algorithm or model's performance may be assessed statistically by dividing the data into two subsets, referred to as the training data and the testing data. This cross-validation approach is frequently employed to shorten calculation times while preserving estimate accuracy in situations when there are few samples available.

It is advised to utilize 10-fold cross validation, as illustrated in figure 3, a kind of k-fold cross validation, to choose the optimal model since it can offer a better estimate of accuracy than standard cross validation.

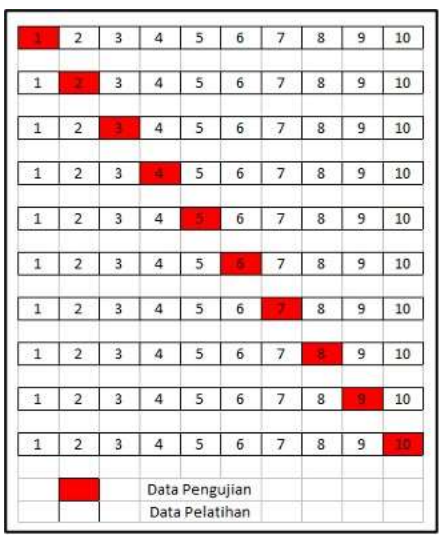

Figure 3. 10-Fold Cross Validation Scheme

Feature Selection

In order to limit the number of features that are less important and will be deleted, which is intended to enhance classification performance and reduce the computational burden, feature selection is employed when creating a new feature subset from the original feature. The purpose of the feature selection step is to locate and collect data feature information in order to exclude characteristics that are deemed less important. Several feature selection techniques that are a part of the supervised learning process are listed below and are grouped into three basic groups (Wibowo, Antoni, 2017) (Urszula, S., 2015):

a) Metode wrapper with backward and forward selection.

b) Mutual information, for example chi square, and method filter.
c) The embedded for example lasso method and the document frequency term method.

Confusion Matrix

The confusion matrix is a table that records the results of the classification (Prasetyo. 2012) as shown in table 1. Generally, the measurement of classification performance is carried out with a confusion matrix. The confusion matrix is a table that records the results of classification work. The following is the result of the confusion matrix (Novandya, Andhika dan Oktria, Isni, 2017).

Table 1. Confusion Matrix

		Prediction	
		C1	C2
Aktual	C1	F11	F12
	C2	F21	F22

Information:
F11 = the number of correct predictions is negative (True Negative).
F21 = number of false positive predictions (False Positive)
F12 = number of false negative predictions (False Negative)
F22 = Number of correct predictions are positive (True Positive)

Actual is a categorization for the status of rain that has already been assigned. The categorization of status variables produced by the program or software leads to prediction. Numerous more figures that may be utilized as classification performance values can be computed from the confusion matrix's construction (Faisal, M., and Nugrahadi, D, 2019)6. The following are these values:

a) Accuracy is the percentage of correctly predicted outcomes. The accuracy formula is:

$$Accuracy = \frac{F11 + F22}{F11 + F21 + F12 + F22}$$

(Equation 3)

b) Recall, also known as a true positive (TP), is the percentage of positive cases that are accurately detected.

$$True\ Positive = \frac{F22}{F12 + F22}$$

(Equation 4)

c) Precision (P) may be determined using the equation: Precision (P) is the percentage of valid positive case predictions.

$$P = \frac{F22}{F21 + F22}$$

(Equation 5)

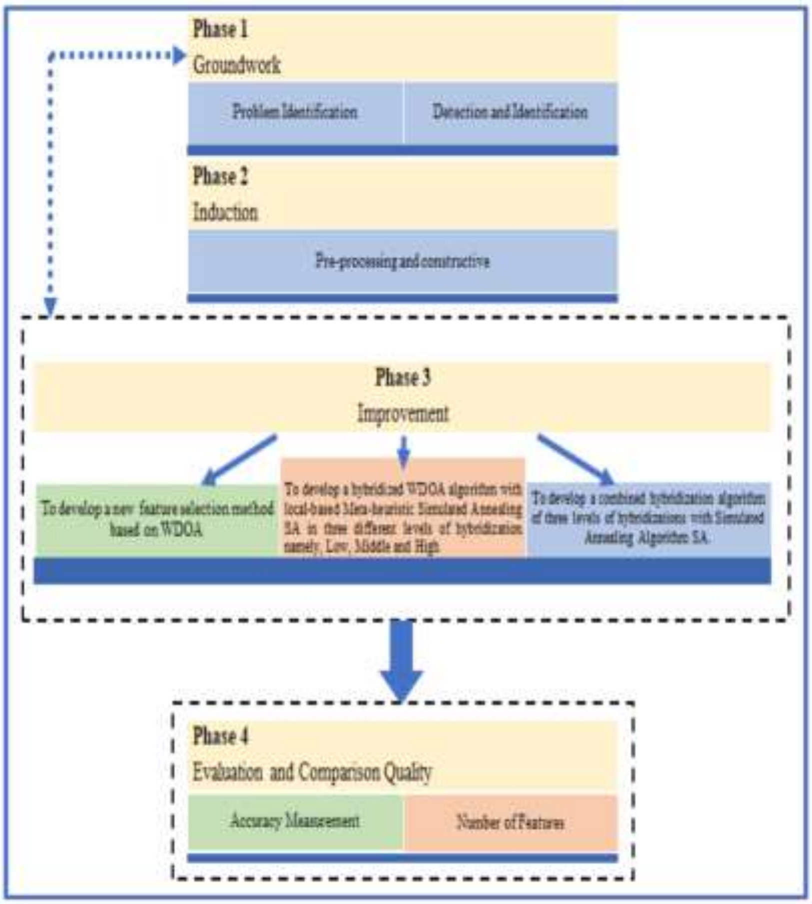

Figure 4. Research framework of Text classification

Problem Identification

The goal of this task is to identify the relevant related studies. We are mainly focusing on understanding the issues facing the development of an effective text classifier. This task was realized by reviewing the most recent text classifiers and their details to identify the existing techniques' strengths and drawbacks.

Two major text classifiers- related problems were identified in this phase. The identified problems are, first, the ambiguity performance of three classifiers with the aim of detecting the class of new documents. Second, feature reduction while maintaining high performance in FS problems.

Detection and Identification Cases

The advancement of the Internet and the increased amount of online information has significantly impacted the ability to detect and identify huge documents. The classification approach is capable of simulating human thinking, and it has been successfully applied in various areas due to its salient features in mining. However, text classification to detect and identify new events, document, or sentiments is complicated. This is because large volume of data degrades classifiers' performance due to high dimensionality of feature space.

Results and Discussion

Data Processing Results

The Palembang SMB II Meteorological Station provided the following 7 qualities of meteorological parameters for the time period of December 2012 to December 2021, which were used in this study as depicted Figure 5.

No.	Suhu Udara (°C)	Suhu Maksimum (°C)	Suhu Minimum (°C)	Kelembapan Udara (%)	Suhu Titik Embun (°C)	Tekanan Udara (mb)	Penyinaran Matahari (jam)	Curah Hujan (mm)
1.	27,0	33,6	24,3	90	24,9	1009,2	3,8	34,9
2.	25,6	32,1	23,2	92	24,0	1010,1	3,6	87,3
3.	27,0	31,4	23,2	85	24,1	1010,0	1,4	0,0
4.	26,8	29,5	23,7	87	24,3	1009,9	0,0	0,8
5.	27,7	33,1	23,8	83	24,2	1009,9	4,0	5,6
6.	26,5	32,0	23,0	88	24,1	1009,8	2,5	2,0
7.	26,0	29,4	23,0	89	23,9	1010,6	2,0	0,0
8.	25,6	32,3	23,0	89	23,5	1010,7	4,2	67,0
9.	27,7	32,8	23,4	83	24,3	1009,7	5,1	0,0
10.	27,2	32,5	23,5	85	24,2	1009,3	6,7	39,0
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
3315.	27,2	31,3	23,9	83	23,8	1010,4	5,1	0,5
3316.	27,6	31,2	24,7	83	23,4	1010,2	3,8	0,0
3317.	28,0	33,5	24,4	84	24,9	1009,8	6,3	4,2
3318.	25,3	32,4	22,8	91	23,8	1010,9	1,5	29,5

Figure 5. Dataset

Then, the rainfall data is transformed into labels for the various kinds of rain, including very light, light, medium, heavy, and very heavy.

Data Normalization

Because of the wide range of values in the data utilized in this investigation, normalization is necessary. The min-max normalization approach transforms the dataset by calculating the lowest and maximum values for each attribute with a value range of 0-1. such that the dataset's attribute ranges are evenly distributed. This can be seen at figure 6.

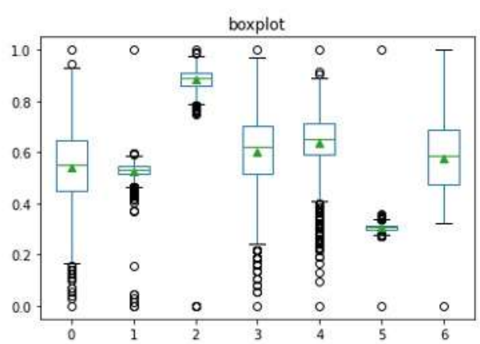

Figure 6. Boxplot Dataset Graph Following Normalization

Model Testing
We will now go through the outcomes of testing the Naive Bayes Classifier (NBC), Support Vector Machine (SVM), and Random Forest (RF) classification algorithms. The data is initially split into training and testing data before classification. The dataset is split into two sections using the train/test split technique, with the training data making up 2488 of the total data and the testing data making up 830 of the total data. Additionally, to assess the model's performance using 10-Fold Cross Validation The average value of the model's accuracy will then be derived using the test results utilizing data and a confusion matrix.

NBC Test Results
The purpose of this test is to evaluate the effectiveness of the Cross Validation and Naive Bayes Classifier (NBC) algorithms. By comparing the prediction class with the actual class, which can offer information in the form of accuracy by utilizing a confusion matrix, the outcomes of model testing are assessed. The results of this test show that the model's accuracy ranges from 0.69 to 0.62, with an average accuracy of 0.66 and a standard deviation of 0.02 after 10 tests as shown in Table 2 and Figure 7.

Table 2. Results of the NBC Accuracy Calculation

Testing	Accuray NBC
1.	0.68875502
2.	0.6626506
3.	0.65361446
4.	0.62148594
5.	0.6937751
6.	0.63654618
7.	0.68473896
8.	0.6626506
9.	0.67168675
10.	0.65261044
Average	0.662851405

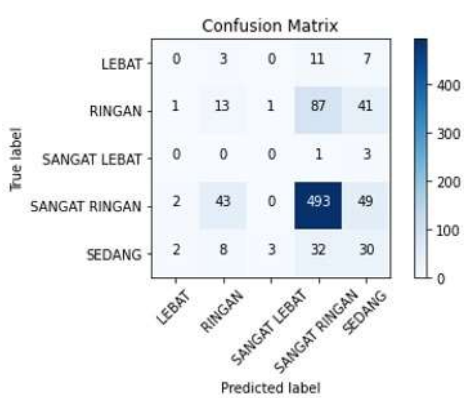

Figure 7. NBC Test Confusion Matrix

Based on the NBC test's confusion matrix, shown in Figure 5 above, it can be observed that 493 times the NBC test accurately predicted very light rain, whereas 49 times it incorrectly anticipated moderate rain and 43 times it correctly predicted light rain. Then, for light rain, it was forecasted 13 times accurately and there were 87 times when it was incorrectly projected as very light rain and 41 times when it was incorrectly predicted as moderate rain. There were 32 incorrect predictions for moderate-intensity rain, where moderate rain was forecasted as very light rain eight times and properly as light rain 30 times. Furthermore, it is impossible to anticipate heavy rain with any degree of accuracy; there have been 11 instances where heavy rain has been mistaken for very light rain and seven other occasions for moderate rain. Furthermore, it is impossible to precisely estimate rain with a very heavy intensity, and there have been three instances in which extremely heavy rain has been mistaken for moderate rain and one for very light rain.

SVM Test Results

The purpose of this test is to evaluate the effectiveness of the Cross Validation and Support Vector Machine (SVM) algorithms. After doing this test ten times, the results show that the average model accuracy value is 0.72, with a standard deviation of 0.00, with the best accuracy being 0.73 and the lowest accuracy being 0.72 as shown in Table 3 and figure 8.

Table 3. Results of the SVM Accuracy Calculation

Testing	Akurasi SVM
1.	0.72155689
2.	0.72155689
3.	0.72372372
4.	0.72289157
5.	0.72289157
6.	0.72289157
7.	0.72507553
8.	0.72727273
9.	0.72727273
10.	0.72727273
Average	0.724240593

Figure 8. SVM Test Confusion Matrix

Random Forest Test Results

The purpose of this test is to evaluate the effectiveness of the algorithms Random Forest (RF) and Cross Validation. The results of this test show that the model's accuracy ranges between 0.71 and 0.76, with an average accuracy of 0.74 and a standard deviation of 0.02 after ten tests as shown in Table 4 and Figure 9.

Table 4. Results of the RF Accuracy Calculation

Pengujian	Akurasi RF
1.	0.71257485
2.	0.7245509
3.	0.75075075
4.	0.73493976
5.	0.74698795
6.	0.72590361
7.	0.75528701
8.	0.75454545
9.	0.71515152
10.	0.73333333
Rata - rata	0.735402513

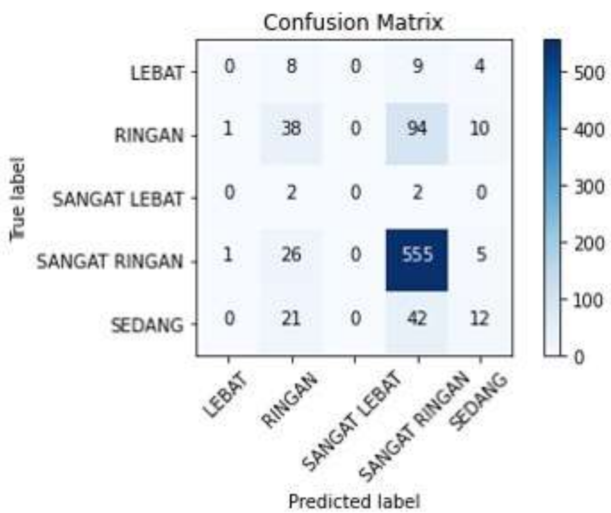

Figure 9. RF Test Confusion Matrix

According to the confusion matrix for the RF test, it can be shown that 26 times where very light rain is expected to be light rain, 555 times of the time very light rain is properly predicted. Additionally, 94 occasions where light rain was forecasted as very light rain, there were 38 instances when light rain was successfully anticipated and 38 times when it was incorrectly predicted. In cases where moderate rain was forecasted as very light rain, there were 42 prediction mistakes and 12 instances when moderate rain was accurately foreseen. Therefore, it is impossible to reliably anticipate rain with a high intensity, and there are nine times as many prediction mistakes for heavy rain than for very light rain and eight times as many for mild rain. It is also impossible to anticipate rain with a very heavy intensity precisely, and there are twice as many instances where extremely heavy rain is mistaken for light rain and twice for very light rain.

RF Parameter Optimization with Statistical Feature Selection

The Random Forest model, which has an accuracy rate of 0.74, is the best one according to the test results of the NBC, SVM, and Random Forest models with 10-Fold Cross Validation. In addition, the Random Forest model may be used to determine which characteristics (attributes) are most connected (have the strongest information connection) to the label of rain intensity. The Mutual Information approach and the F-Classification method are the two strategies employed in this feature selection statistic.

Metode Mutual Information

In order to identify the factors that have the most impact on the rain intensity label, the Mutual Information approach is now evaluated on the outcomes and functionality of the built-in model. We can see from figure 10 below.

Metode	No #	Include FS	Exclude FS	Acc	Std	Max	Min
Mutual Info	6	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Minimum' 'Kelembapan_Rata_Rata' 'Suhu_Titik_Embun_Rata_Rata' 'Tekanan_Udara_Rata_Rata'	'Penyinaran_Matahari'	0.73	0.02	0.76	0.69
[0.69461078, 0.72155689, 0.75075075, 0.7439759, 0.74096386, 0.70481928, 0.74924471, 0.75757576, 0.70606061, 0.72121212]							
	5	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Minimum' 'Kelembapan_Rata_Rata' 'Suhu_Titik_Embun_Rata_Rata'	'Tekanan_Udara_Rata_Rata' 'Penyinaran_Matahari'	0.73	0.02	0.75	0.68
[0.67964072, 0.73353293, 0.74174174, 0.72289157, 0.73192771, 0.71084337, 0.72205438, 0.75151515, 0.72424242, 0.73030303]							
	4	'Suhu_Rata_Rata' 'Suhu_Minimum' 'Kelembapan_Rata_Rata' 'Suhu_Titik_Embun_Rata_Rata'	'Suhu_Maksimum' 'Tekanan_Udara_Rata_Rata' 'Penyinaran_Matahari'	0.70	0.02	0.74	0.66
[0.66167665, 0.67964072, 0.70870871, 0.71987952, 0.68674699, 0.68072289, 0.70090634, 0.73939394, 0.70606061, 0.69393939]							
	3	'Suhu_Rata_Rata' 'Suhu_Minimum' 'Kelembapan_Rata_Rata'	'Suhu_Titik_Embun_Rata_Rata' 'Suhu_Maksimum' 'Tekanan_Udara_Rata_Rata' 'Penyinaran_Matahari'	0.68	0.03	0.74	0.64
[0.71856287, 0.72754491, 0.73273273, 0.72590361, 0.71686747, 0.71084337, 0.74018127, 0.74242424, 0.72424242, 0.74848485]							
	2	'Suhu_Rata_Rata' 'Kelembapan_Rata_Rata'	'Suhu_Minimum' 'Suhu_Titik_Embun_Rata_Rata' 'Suhu_Maksimum' 'Tekanan_Udara_Rata_Rata' 'Penyinaran_Matahari'	0.70	0.02	0.73	0.67
[0.7245509, 0.67964072, 0.67867868, 0.72590361, 0.6686747, 0.69277108, 0.72507553, 0.72424242, 0.6969697, 0.72424242]							
	1	'Kelembapan_Rata_Rata'	'Suhu_Rata_Rata' 'Suhu_Minimum' 'Suhu_Titik_Embun_Rata_Rata' 'Suhu_Maksimum' 'Tekanan_Udara_Rata_Rata' 'Penyinaran_Matahari'	0.73	0.01	0.75	0.71
[0.71856287, 0.72754491, 0.73273273, 0.72590361, 0.71686747, 0.71084337, 0.74018127, 0.74242424, 0.72424242, 0.74848485]							

Figure 10. Mutual Information + Random Forest Test Results

According to Table 6 above, the Mutual Information feature selection yielded results with an average accuracy of 0.73, or 73%, and a minimum accuracy of 0.68, or 68%. The tests with the average air humidity attribute alone, the tests using five characteristics but excluding average air pressure and sunshine, and the tests using six attributes but excluding sunlight all yielded the greatest accuracy of 73%. The test that merely employed the three parameters of average air temperature, average humidity, and minimum temperature yielded the lowest accuracy of 68%. As a result, it is clear that average air humidity has the greatest impact on the label indicating the severity of the rain, whereas sunshine has the least impact.

Metode F-Classification

In order to identify the factors that have the greatest impact on the rain intensity label, the F-Classification approach was evaluated at this stage using the model's performance and findings. We can see from figure 11 below.

Metode	No #	Include FS	Exclude FS	Acc	Std	Max	Min
f_Classif	6	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Minimum' 'Kelembapan_Rata_Rata' 'Suhu_Titik_Embun_Rata_Rata' 'Penyinaran_Matahari'	'Tekanan_Udara_Rata_Rata'	0.73	0.02	0.76	0.69
[0.69461078, 0.72155689, 0.75075075, 0.7439759, 0.74096386, 0.70481928, 0.74924471, 0.75757576, 0.70606061, 0.72121212]							
	5	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Kelembapan_Rata_Rata' 'Suhu_Titik_Embun_Rata_Rata' 'Penyinaran_Matahari'	'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.72	0.02	0.76	0.68
[0.69161677, 0.71556886, 0.72372372, 0.74096386, 0.72289157, 0.68373494, 0.70090634, 0.75757576, 0.71212121, 0.7030303]							
	4	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Kelembapan_Rata_Rata' 'Suhu_Titik_Embun_Rata_Rata'	'Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.70	0.02	0.74	0.66
[0.66167665, 0.67964072, 0.70870871, 0.71987952, 0.68674699, 0.68072289, 0.70090634, 0.73939394, 0.70606061, 0.69393939]							
	3	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Kelembapan_Rata_Rata'	'Suhu_Titik_Embun_Rata_Rata' 'Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.68	0.03	0.74	0.64
[0.64371257, 0.67964072, 0.68168168, 0.70783133, 0.66566265, 0.63855422, 0.67069486, 0.73636364, 0.69393939, 0.7030303]							
	2	'Suhu_Rata_Rata' 'Kelembapan_Rata_Rata'	'Suhu_Maksimum' 'Suhu_Titik_Embun_Rata_Rata' 'Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.70	0.02	0.73	0.67
[0.7245509, 0.67964072, 0.67867868, 0.72590361, 0.6686747, 0.69277108, 0.72507553, 0.72424242, 0.6969697, 0.72424242]							
	1	'Kelembapan_Rata_Rata'	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Titik_Embun_Rata_Rata' 'Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.73	0.01	0.75	0.71
[0.71856287, 0.72754491, 0.73273273, 0.72590361, 0.71686747, 0.71084337, 0.74018127, 0.74242424, 0.72424242, 0.74848485]							

Figure 11. F-Classification + Random Forest Test Results

Figure 10 shows that utilizing the F-Classification feature selection, the greatest average accuracy results were 0.73, or 73%, and the lowest average accuracy were 0.68, or 68%. The tests that simply utilized the average air humidity attribute and the tests that used 6 characteristics other than average air pressure had the greatest accuracy, 73%. The test that merely employed the three parameters of average air temperature, average humidity, and maximum temperature yielded the lowest accuracy of 68%. As a result, it is clear that the average air humidity has the greatest impact on the label indicating the severity of the rain, while the average air pressure has the least impact. Using the statistical feature selection approach can boost the efficacy of the system being constructed, according to the results of this test. Additionally, the improvement in classification accuracy is not necessarily impacted by the addition of data or characteristics.

Optimization of RF Parameters Using Metaheuristic Feature Selection

In order to identify the factors that had the most impact on the rain intensity label, 13 Metaheuristic Feature Selection approaches were evaluated at this stage on the performance and outcomes of the model constructed. There are several metaheuristic feature selection algorithms that are used, including Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Firefly Algorithm (FA), Differential Evolution (DE), Cuckoo Search Algorithm (CS), Bat Algorithm (BA), Flower Pollination Algorithm (FPA), Sine Cosine Algorithm (SCA), Whale Optimization Algorithm (WOA), Harris Hawk Optimization (HHO), Salp (GWO).

Using the F-Classification feature selection, the greatest average accuracy results were 0.74, or 74%, and the lowest average accuracy were 0.73, or 73%, as can be seen in table 8 above. The WOA + Random Forest approach with the usage of 7 characteristics produced the maximum accuracy of 74%. In comparison to the PSO, FA, CS, BA, SCA, HHO, and SSA techniques, the employment of the WOA method is deemed useless when comparing the accuracy outcomes of the 13 feature selection methods based on computing time. This is because the WOA approach requires more computing time, but the usage of the 7 feature selection methods yields an accuracy of 73% utilizing only 1 characteristic, namely the average humidity. Additionally, the WOA approach is unable to choose the characteristics that are significant and have the greatest impact on the label of rain intensity. Figure 12 has illustrated this explanation.

Algorithma	No #	Include FS	Exclude FS	Acc	Std	Max	Min
PSO	1	'Kelembapan_Rata_Rata'	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Titik_Embun_Rata_Rata''Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.73	0.01	0.74	0.71
[0.72649573, 0.71367521, 0.72103004, 0.72532189, 0.72103004, 0.73819742, 0.71982759, 0.73478261, 0.72608696, 0.72608696]							
GA	5	'Suhu_Maksimum' 'Suhu_Minimum' 'Kelembapan_Rata_Rata' 'Suhu_Titik_Embun_Rata_Rata' 'Tekanan_Udara_Rata_Rata'	'Suhu_Rata_Rata' 'Penyinaran_Matahari'	0.73	0.02	0.76	0.70
[0.75213675, 0.71367521, 0.70386266, 0.71244635, 0.7167382, 0.7639485, 0.72413793, 0.74782609, 0.72608696, 0.73043478]							
FA	1	'Kelembapan_Rata_Rata'	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Titik_Embun_Rata_Rata''Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.73	0.01	0.74	0.71
[0.66167665, 0.67964072, 0.70870871, 0.71987952, 0.68674699, 0.68072289, 0.70090634, 0.73939394, 0.70606061, 0.69393939]							
DE	5	'Suhu_Maksimum' 'Suhu_Minimum' 'Kelembapan_Rata_Rata' 'Tekanan_Udara_Rata_Rata' 'Penyinaran_Matahari'	'Suhu_Rata_Rata' 'Suhu_Titik_Embun_Rata_Rata'	0.73	0.02	0.75	0.69
[0.73931624, 0.71794872, 0.69098712, 0.7167382, 0.72103004, 0.74678112, 0.73275862, 0.75217391, 0.74782609, 0.72608696]							
CS	1	'Kelembapan_Rata_Rata'	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Titik_Embun_Rata_Rata''Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.73	0.01	0.74	0.71
[0.72649573, 0.71367521, 0.72103004, 0.72532189, 0.72103004, 0.73819742, 0.71982759, 0.73478261, 0.72608696, 0.72608696]							
BA	1	'Kelembapan_Rata_Rata'	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Titik_Embun_Rata_Rata' 'Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.73	0.01	0.74	0.71
[0.72649573, 0.71367521, 0.72103004, 0.72532189, 0.72103004, 0.73819742, 0.71982759, 0.73478261, 0.72608696, 0.72608696]							
FPA	5	'Suhu_Maksimum' 'Suhu_Minimum' 'Kelembapan_Rata_Rata' 'Suhu_Titik_Embun_Rata_Rata' 'Tekanan_Udara_Rata_Rata'	'Suhu_Rata_Rata' 'Penyinaran_Matahari'	0.73	0.02	0.76	0.70
[0.75213675, 0.71367521, 0.70386266, 0.71244635, 0.7167382, 0.7639485, 0.72413793, 0.74782609, 0.72608696, 0.73043478]							
SCA	1	'Kelembapan_Rata_Rata'	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Titik_Embun_Rata_Rata' 'Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.73	0.01	0.74	0.71
[0.72649573, 0.71367521, 0.72103004, 0.72532189, 0.72103004, 0.73819742, 0.71982759, 0.73478261, 0.72608696, 0.72608696]							
WOA	7	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Minimum' 'Kelembapan_Rata_Rata' 'Suhu_Titik_Embun_Rata_Rata' 'Tekanan_Udara_Rata_Rata' 'Penyinaran_Matahari'	-	0.74	0.02	0.77	0.72
[0.72649573, 0.72222222, 0.72103004, 0.7167382, 0.75965665, 0.77253219, 0.72844828, 0.74347826, 0.76086957, 0.7173913]							
HHO	1	'Kelembapan_Rata_Rata'	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Titik_Embun_Rata_Rata' 'Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.73	0.01	0.74	0.71
[0.72649573, 0.71367521, 0.72103004, 0.72532189, 0.72103004, 0.73819742, 0.71982759, 0.73478261, 0.72608696, 0.72608696]							
SSA	1	'Kelembapan_Rata_Rata'	'Suhu_Rata_Rata' 'Suhu_Maksimum' 'Suhu_Titik_Embun_Rata_Rata' 'Penyinaran_Matahari' 'Suhu_Minimum' 'Tekanan_Udara_Rata_Rata'	0.73	0.01	0.74	0.71
[0.72649573, 0.71367521, 0.72103004, 0.72532189, 0.72103004, 0.73819742, 0.71982759, 0.73478261, 0.72608696, 0.72608696]							
JA	5	'Suhu_Maksimum' 'Suhu_Minimum' 'Kelembapan_Rata_Rata' 'Tekanan_Udara_Rata_Rata' 'Penyinaran_Matahari'	'Suhu_Titik_Embun_Rata_Rata' 'Suhu_Rata_Rata'	0.73	0.02	0.75	0.69
[0.73931624, 0.71794872, 0.69098712, 0.7167382, 0.72103004, 0.74678112, 0.73275862, 0.75217391, 0.74782609, 0.72608696]							
GWO	5	'Suhu_Maksimum' 'Suhu_Minimum' 'Kelembapan_Rata_Rata' 'Tekanan_Udara_Rata_Rata'	'Suhu_Titik_Embun_Rata_Rata' 'Suhu_Rata_Rata'	0.73	0.02	0.75	0.69

Figure 12. Metaheuristic Test Results + Random Forest Test Results

Conclusion

The following conclusions can be inferred from the findings and conversations previously described:

a) The accuracy of three distinct rain-predicting techniques varies; the SVM approach has a 72% accuracy rate, the NBC method has a 66% accuracy rate, and the Random Forest method has a 74% accuracy rate.
b) The Random Forest approach outperforms the SVM and NBC methods in forecasting rain.
c) The SVM, NBC, and Random Forest algorithms cannot reliably predict severe and extremely heavy rain occurrences.
d) According to the feature selection approach, the average humidity is the trait that has the greatest impact on rain forecasts.
e) The degree of categorization accuracy does not necessarily increase when more characteristics are applied.

References

Gultepe, Ismail dkk. 2019. A Review of High Impact Weather for Aviation Meteorology. Pure and Applied Geophysics.

Kulesa, G. 2003. Weather And Aviation: How Does Weather Affect The Safety And Operations Of Airports And Aviation, And How Does Faa Work To Manage Weather-Related Effects?

Tjasyono, Bayong & Harijono, Sri W.B. 2012. Meteorologi Indonesia Volume II : Awan dan Hujan Monsun. Jakarta : Badan Meteorologi Klimatologi dan Geofisika.

J. Han, and M. Kamber. 2013. Data mining: Concepts and Techniques. Third Edition, Morgan Kaufmann Publishers, San Fransisco.

S. Ali, and A. Smith. 2010. Improved Support Vector Machine Generalization Using Normalized Input Space. AI 2010: Advances in Artificial Intelligence, vol.4304, pp.362-371,2010.

Novandya, Andhika dan Oktria, Isni. 2017. Penerapan Algoritma Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi. Jurnal Vol 6 (2) hlm 2089 – 5615.

Estoatnowo, Ditya. 2016. Klasifikasi Status Kelulusan Mahasiswa Menggunakan Metode CHAID dan Algoritma C4.5 (Studi Kasus: Mahasiswa Jurusan Statistika Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia). Skripsi. Fakultas Matematika dan Ilmu Pengetahuan Alam. Universitas Islam Indonesia. Yogyakarta.

Subhan, Ahmad. dan Ahmad Zainul Fanani. 2017. Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma Naïve Bayes. Skripsi. Fakultas Teknik Informatika. Universitas Dian Nuswantoro. Semarang.

B. Xu, J. Z. Huang, G. Williams, Q. Wang, dan Y. Ye. 2012. Classifying Very High-Dimensional Data with Random Forests Built from Small Subspaces. International Journal of Data Warehousing and Mining, Vol. 8, no. 2, hlm. 44–63.

Siregar, Aplia Belina. 2017. Pemodelan support Vector Machine Untuk Klasifikasi Curah Hujan Bulanan Di Kabupaten Indramayu. Skipsi. Statistika. Institut Pertanian Bogor.

Nugroho, Anto Satrio, Arief Budi Witarto, dan Dwi Handoko. 2003. Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika.

Ritonga, A. S., & Purwaningsih, E. S. 2018. Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas Pengelasan Smaw (Shield Metal Arc Welding). Ilmiah Edutic, 5(1), 17–25.

Assaffat, Luqman. 2015. Analisis Akurasi Support Vector Machine Dengan Fungsi Kernel Gaussian RBF Untuk Prakiraan Beban Listrik Harian Sektor Industri. Jurnal Momentum. Vol. 11 (2).

https://mti.binus.ac.id/2017/11/24/10-fold-cross-validation/, diakses tanggal 3 Januari 2022.

Urszula, S., 2015. Feature Selection by Filter, Wrapper, and Embedded Approaches. Studies in Computational Intelligence, pp. 29-44.

Prasetyo. 2012. Data Mining Konsep dan Aplikasi Menggunakan Matlab. Yogyakarta: Andi Offset.

Faisal, M., & Nugrahadi, D. 2019. Belajar Data Science Klasifikasi dengan Bahasa Pemrograman R. Banjarbaru, Kalimantan Selatan: Scripta Cendekia.