Comparative analysis of Naïve Bayes, K Nearest Neighbor and C.45 method in weather forecast

Comparative analysis of Naïve Bayes, K Nearest Neighbor and C.45 method in weather forecast

Y Findawati[1,2,*], I R Indra Astutik[1,2], A S Fitroni[1,2], I Indrawati[1,2] and N Yuniasih[1,2]

[1] Department of Informatics, Universitas Muhammadiyah Sidoarjo, East Java, Indonesia
[2] Program Studi Sekolah Dasar, Universitas Kanjuruhan Malang, East Java, Indonesia

*yulianfindawati@umsida.ac.id

Abstract. Weather forecast in an area is unpredictable. This is due to the fact that human factors cannot predict it. The weather forecast is by applying data mining using the algorithm Naive Bayes, K-nearest Neighbor (K-NN), and C.45. Bayesian Classification is a statistical classification method that is useful for the process of determining the probability of a class membership. KNN Algorithm is a classification algorithm based on the similarity between one data and another data. C4.5 algorithms is an easy-to-use classification method interpreted. The best level of accuracy between the three algorithms can be determined by comparison. Comparison of algorithm aims to get the algorithm that is considered accurate, precision, recall and f-measure to make a prediction of a problem. the results of the comparison of the k-Nearest Neighbor, Naïve Bayes and C4.5 classification algorithms used in weather prediction case studies stating that the KNN classification algorithm is a classification algorithm that has the highest accuracy with k = 7 and fold = 5 in predicting the weather compared to Naïve Bayes classification algorithm with fold = 3 and C.45 which reached 71.58% followed by C.45 with fold = 20 having an accuracy of 69.83%. and finally Naïve Bayes 68.77%.

1. Introduction

The development of information technology at this time has an impact on everyday life that requires more accurate information. This is because information is an important element for society at this time and in the future. In making a decision from the data usually only rely on operational data, but this is not enough. Then a data analysis is needed to explore important patterns or information from large amounts of data, called data mining. Data mining is most important analysis step of knowledge discovery in database (KDD) process. The main goal of data mining is to extract the useful information from huge raw data and converting it to an understandable form for its effective and efficient use. In common, data mining tasks can be divided into two categories: descriptive and predictive classification techniques [1]. Accurate meteorological information in the world of aviation is very much needed, because it concerns flight safety. One of the influencing factors for aviation safety is the weather. Many failures from flights are caused by bad weather effects. In addition to flight, most human activities also depend on weather conditions, such as in agriculture, fisheries, transportation, plantations.

This weather condition is influenced by several factors including, air temperature, air pressure, wind direction, wind speed, air humidity, light intensity, and so on. Recent research on weather forecasts is increasing. The value of high accuracy in these predictions is a very influential thing for all human

activities, including in the world of aviation which is very dependent on weather conditions. Analysis of data mining for weather forecasting has been done. Data mining methods with the Naïve Bayes, K-Nearest Neighbour and C.45 algorithm are used to determine daily rain potential with higher accuracy values. K Nearest Neighbor is one method in the top 10 of the most commonly used data mining methods [2]. C.45 is an algorithm commonly used for retrieval decision. C.45 will find solutions to problems by making criteria as interconnected nodes to form like tree structures [3]. C.45 is a prediction model for a decision using a structure hierarchy or tree. Every tree has branches; branches represent an attribute must be fulfilled to go to the next branch until it ends in leaves (none branch again). Given the magnitude of the influence of weather factors on the world of aviation, therefore weather information is needed both when taking off, landing or during a flight. The Meteorology, Climatology and Geophysics Agency, which serves as a weather observer, has conducted observational weather predictions for weather forecasts. This paper focuses on a survey of various classification techniques that are most commonly used in data mining. The comparative study between different algorithms (K-NN classifier, Naïve Bayes and C.45) is used to show the strength and accuracy of each classification algorithm in term of performance efficiency. The best level of accuracy between the three algorithms can be determined by comparison. Comparison of algorithm aims to get the algorithm that is considered accurate, precision, recall and f-measure to make a prediction of a problem

Comparative Study between Naïve Bayes, K-nearest Neighbour and C.45 is implemented in different fields as follows to predict divorces [4] that have Result of comparison of Naive Bayes and K-Nearest Neighbor algorithm that Naive Bayes algorithm yield 72.5% accuracy and K-Nearest Neigbor algorithm yield 57.5% accuracy., identification of eye diseases [5], Comparison of C.45, Naïve Bayes and K-Nearest Neighbor Algorithms for Critical Land Prediction in Pemalang Regency [6] that have result C4.5 has the highest accuracy of 77.75% followed by Naïve Bayes 77.49% and finally k-NN has accuracy of 73.91%., in Searching Alternative Design in an Energy Simulation Tool [7] that have results Decision Tree has the fastest classification time followed by Naïve Bayes and k-Nearest Neighbor. The differences between classification time of Decision Tree and Naïve Bayes also between Naïve Bayes and k-NN are about an order of magnitude. Based on precision, Recall, Fmeasure, Accuracy, and AUC, the performance of Naïve Bayes is the best., Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques [8],[9], student performance [10]. From these problems, the authors conducted research with the title Comparative Analysis of Naïve Bayes, K-Nearest Neighbor (KNN) and C.45 Algorithms in Weather Forecast at Juanda Airport Surabaya. Based on the background described above, the problem can be formulated as follows How is the comparison of Naïve Bayes, K-Nearest Neighbor (KNN) and C.45 Algorithms in Weather Forecast at Juanda Airport Surabaya.

2. Research methods

The research aimed to find out the comparison and evaluation of the percentage results of the accuracy of the predictions of the Naive Bayes K-Nearest Neighbor (KNN) and C.45 Algorithms in the Weather Forecast at Juanda Airport, Surabaya. Data collection does not only take existing data, but must be able to describe existing data and contribute knowledge about existing data. The data must be able to provide explanations, information, and relationships. The data used in this study is daily data from BMKG in January 2015 to November 2018. The data has a record of 1422 with 8 attributes namely, Minimal Temperature, Maximum Temperature, Average Temperature, Humidity, Radiation, Wind Speed, Wind Direction, and Rain Intensity. Initial processing of data or pre-processing can be done to reduce irrelevant data and data with missing attributes. Data processing can convert excessive values or values that are too diverse to facilitate the formation of the model. From BMKG weather data Pre-processing data is carried out on the attributes of rain intensity by turning it into not rain, light rain, moderate rain, heavy rain, and very heavy rain.

3. Results and discussion

3.1. Naive Bayes manual calculation

This manual calculation is by calculating the data that has been obtained using Naive Bayes and K-Nearest Neighbor (KNN) data mining algorithms using the 20 weather dataset as follows:

Table 1. Sample Dataset Weather Conditions at Juanda Airport Surabaya.

No	Min temp	Max Temp	Avg Temp	Humidity	Radiation	Wind Velocity	Wind direction	Rain intensity
1	24,8	33,4	27,1	83	6,5	3	270	Not Rainy
2	23,6	32,5	27,5	84	1	4	270	Heavy Rain
3	24,8	30,1	27,1	84	2,1	3	315	Light Rain
4	24,3	32,6	27,3	82	6,5	3	270	Not Rain
5	24,2	32,4	27,5	85	6,5	2	0	Heavy Rain
6	24,5	32,7	28,3	79	1	2	315	Moderate Rain
7	25,1	34,2	29,8	75	5	2	90	Not Rain
8	26,6	32,8	29,1	81	7,8	2	270	Not Rain
9	26,2	33,2	29,6	74	8,3	2	90	Not Rain
10	25,3	33	29,4	80	7,5	1	90	Heavy rain
11	27,2	34	29,4	79	7,5	2	0	Not Rain
12	26,4	34,1	30,1	73	7	3	90	Not Rain
13	26,3	34,6	30,6	71	7	2	90	Not Rain
14	25,4	33,6	28,4	83	1,2	3	0	Not Rain
15	26,2	33,2	28,1	85	6,3	2	0	Not Rain
16	25,9	30,9	28,1	84	3,4	2	0	Heavy Rain
17	25,9	33	28,5	81	4,5	3	270	Not rain
18	25,4	32	28,7	79	6,3	4	270	Moderate rain
19	25	32	28	83	2	4	315	Not Rain
20	25,5	33,5	28,1	80	3,5	3	0	Not Rain
21	24,2	30,4	26,5	89	5,2	2	0	?

Calculation of Naive Bayes of weather condition dataset at Juanda Airport Surabaya is continuous data, so the first step to do the calculation is:

- Calculation of Mean Values and Standard Deviations is $\mu = 25{,}762$
- Calculation of Standard Deviation: s = 0,82415
- From the example calculation, the calculation of each class / label is obtained as follows. Function
- The next step after determining the mean and standard deviation is to calculate the predictions of the Naive Bayes algorithm with the Gaus Density function using the formula: $( ) = \frac{1}{\sqrt{2}}^{-\frac{(-)^2}{2^2}}$

  Min Temp (Not Rain | 24,2)= 0,0730
  Min temp (Light Rain | 24,2)= 0
  Min temp (Moderate Rain | 24,2)= 0,2498
  Min temp (Heavy Rain | 24,2)= 0
  Min temp (Very Heavy Rain | 24,2)= 0
  In the same way calculate the Prediction of Naive Bayes Algorithm with Gaus Density on the criteria for Maximum Temperature, Average Temperature, Humidity, Radiation, Wind Speed and Wind Direction
- Calculation of Likelihood Value. Likelihood Not Rain P(X | Not Rain)= 7,84365 x ⟦10⟧^(-12). By means of the same calculation, calculate light rain Likelihood, Moderate rain Likelihood, heavy rain Likelihood and very heavy rain.

The probability value can determine the rainfall intensity category, if the probability value is close to 1 or is equal to the value of the calculation results it is known that the testing data above is included in the rain category that has a probability value of 1.

3.2. K-Nearest Neighbor (KNN) manual calculation

This method K-Nearest Neighbor (KNN) determines the label / class seen from the data that has the closest distance to the specified k value. Suppose it is determined to be k = 5. The result can be seen in table 2.

Table 2. Results of K-NN calculation.

No	Min temp	Max Temp	Avg temp	Humidity	Radiation	Wind Velocity	Wind direction	Rain intensity	distance
1	24,8	33,4	27,1	83	6,5	3	270	Not Rain	270,0896333
2	23,6	32,5	27,5	84	1	4	270	Very Heavy Rain	270,0970381
3	24,8	30,1	27,1	84	2,1	3	315	Light Rain	315,0578042
4	24,3	32,6	27,3	82	6,5	3	270	Not Rain	270,1058681
5	24,2	32,4	27,5	85	6,5	2	0	Very heavy Rain	4,763402146
6	24,5	32,7	28,3	79	1	2	315	Moderate Rain	315,200349
7	25,1	34,2	29,8	75	5	2	90	Not Rain	91,22598314
8	26,6	32,8	29,1	81	7,8	2	270	Not Rain	270,1648386
9	26,2	33,2	29,6	74	8,3	2	90	Not Rain	91,41148724
10	25,3	33	29,4	80	7,5	1	90	Heavy Rain	90,5741133
11	27,2	34	29,4	79	7,5	2	0	Not Rain	11,64731729
12	26,4	34,1	30,1	73	7	3	90	Not Rain	91,60638624
13	26,3	34,6	30,6	71	7	2	90	Not Rain	92,01141234
14	25,4	33,6	28,4	83	1,2	3	0	Not Rain	8,263776376
15	26,2	33,2	28,1	85	6,3	2	0	Not Rain	5,622277119
16	25,9	30,9	28,1	84	3,4	2	0	Very Heavy Rain	5,825804665
17	25,9	33	28,5	81	4,5	3	270	Not Rain	270,1465158
18	25,4	32	28,7	79	6,3	4	270	Moderate Rain	270,2111212
19	25	32	28	83	2	4	315	Not Rain	315,0883844
20	25,5	33,5	28,1	80	3,5	3	0	Not Rain	9,937303457

Calculation of the distance is obtained from

√((24,8-24)^2+(33,4-32,1)^2+(27,1-27,8)^2+(83-83)^2+(6,5-3,5)^2+(3-1)^2+(270-0)^2) = 270.0292947

The results of distance calculation have been obtained, then the data is sorted from the smallest to the largest. From the sorting of the distance values, they are grouped based on the k value that has been set, which is 5 taken from the smallest distance value. Based on the grouping, it shows that the highest class / label is "no rain" so it can be predicted that the testing data above has a "no rain" class / label.

3.3. Test result

The test is carried out evaluating the value of accuracy, precision, recall value and f-measure. After obtaining the value of accuracy, precision and recall value, compared to seeing which accuracy, precision, recall value and f-measure are the highest. The test uses k fold cross-validation with fold value 3,5,10,15, and 20. Whereas in the K-nearest neighbor method, the value k is 3,5, and 7. The test is done using Weka software and the number of data sets is 1422. The test results can be seen in the table 3:

Table 3. Test result.

Algorithm	Fold	Accuration	Precision	Recall	F-Measure
Naïve Bayes	3	68,7764	0,674	0,688	0,671
	5	68,6357	0,674	0,686	0,67
	10	68,5654	0,668	0,686	0,666
	15	68,4248	0,664	0,684	0,664
	20	68,4951	0,666	0,685	0,664

Table 3. Cont.

KNN k=3	3	69,6906	0,628	0,697	0,651
	5	68,0731	0,603	0,681	0,634
	10	68,2841	0,61	0,683	0,639
	15	68,9873	0,615	0,69	0,643
	20	68,8467	0,611	0,688	0,641
KNN k=5	3	70,1828	0,626	0,702	0,654
	5	70,0422	0,628	0,7	0,654
	10	69,54	0,618	0,695	0,648
	15	69,7609	0,62	0,698	0,649
	20	69,9015	0,621	0,699	0,651
KNN k=7	3	70,4641	0,615	0,705	0,648
	5	71,5893	0,641	0,716	0,662
	10	71,3783	0,639	0,714	0,662
	15	71,4487	0,642	0,714	0,663
	20	70,8861	0,629	0,709	0,656
C.45	3	68,42	0,638	0,684	0,658
	5	68,9873	0,647	0,69	0,666
	10	69,0577	0,654	0,691	0,67
	15	69,4093	0,657	0,694	0,673
	20	69,8312	0,656	0,698	0,674

Based on the test results in table 3.13 the accuracy is above 70%, namely in the K-Nearest Neighbor method with k = 5 and k = 7. While the highest accuracy was obtained on the KNN method with k = 7 with testing fold = 5, the accuracy was 71.5893%. While the highest value of precision in the naïve Bayes method with testing uses fold = 3 and fold = 5 with the results of the precision value of 0.674. While the highest recall value is in the method on the KNN method with k = 7 with testing fold = 5, the recall result is 0.716. While the highest f-measure value is the method in the method of C.45 by testing fold = 20, the f-measure is 0.674. this is different from the research of Naïve Bayes and K-Nearest Neighbor Algorithms for Critical Land Prediction in Pemalang Regency [6] that have result C4.5 has the highest accuracy of 77.75% followed by Naïve Bayes 77.49% and finally k-NN has accuracy of 73.91%. this result is caused by a different input. in Searching Alternative Design in an Energy Simulation Tool [7] that have results Decision Tree has the fastest classification time followed by Naïve Bayes and k-Nearest Neighbor. Algorithm comparisons show that each algorithm has its own advantages and disadvantages depending on implementation. The algorithm depends on the constraint and the criteria as input.

4. Conclusion

The conclusions obtained in this study are from the results of the comparison of the K-Nearest Neighbor, Naïve Bayes and C4.5 classification algorithms used in weather prediction stating that the KNN classification algorithm is a classification algorithm that has the highest accuracy with k = 7 and fold = 5 in predicting the weather compared to Naïve Bayes classification algorithm with fold = 3 and C.45 which reached 71.58% followed by C.45 with fold = 20 having an accuracy of 69.83%. and finally Naïve Bayes 68.77%

Acknowledgement

We hereby thank you to Universitas Muhammadiyah Sidoarjo for supporting the publication of this research.

References

[1] Archana S and Elangovan K 2014 Survey of Classification Techniques in Data Mining *International Journal of Computer Science and Mobile Applications* **2**(2)

[2] Patankar B and Chavda V 2014 A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining *International Journal of Advanced Research in Computer Science and Software Engineering* **4**(12)

[3] Wu X and Kumar V 2009 *The Top Ten Algorithms in Data Mining* (Boca Raton: Chapman &

Hall/CRC)
[4] Mantas C J and Abean J 2014 Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data *Expert Systems with Applications*
[5] Zulfikar W B and Lukman N 2016 Perbandingan Naive Bayes Classifier dengan Nearest Neighbour untuk Identifikasi Penyakit Mata *JOIN (Jurnal Online Informatika)* **1**(2) 82-86
[6] Khotimah N and Istiawan D 2018 Perbandingan Algoritma C4.5, Naïve Bayes dan K-Nearest Neighbour untuk Prediksi Lahan Kritis di Kabupaten Pemalang *The 7th University Research Colloqium STIKES PKU Muhammadiyah Surakarta*
[7] Ashari A, Paryudi I and Tjoa A M 2013 Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool *(IJACSA) International Journal of Advanced Computer Science and Application* **4**(11)
[8] Jadhav S D and Channe H P 2013 Comparative Study of KNN, Naïve Bayes and Decision Tree Classification Techniques *International Journal of Science and Research*
[9] Mohanapriya M and Lekha 2018 Comparative study between decision tree and knn of data mining classification technique *Second National Conference on Computational Intelligence*
[10] Abu Amra I A and Maghari A Y 2017 Students performance prediction using KNN and Naïve Bayesian *Students performance prediction using KNN and Naïve Bayesian*