Weather Forecasting Based on Supervised Learning Using K-Nearest Neighbour Algorithm

Alvi Syahrini Utami*[1], Dian Palupi Rini[2], Endang Lestari[3]
[1,2,3]Faculty of Computer Science, University of Sriwijaya, Palembang
e-mails: *[1]alvisyahrini@ilkom.unsri.ac.id, [2]dprini@unsri.ac.id, [3]endang@ilkom.unsri.ac.id

Abstract

Weather is influenced by many natural factors causing it to change frequently at any time so that it is sometimes difficult to predict. An accurate weather prediction is needed so that people and policy-makers could anticipate this problem. Many factors that influence the weather cause difficulty in classifying the weather on a particular day. Locality Sensitive Hashing (LSH) works on training data by assigning hash values to a vectors that contain values that represent factors that affect weather and perform weather classification. Furthermore, the k-Nearest Neighbor (k-NN) algorithm will calculate the predictions of the factors that affect the weather on a certain day. Based on the tests carried out, k-NN and LSH in weather prediction has Mean Square Error (MSE) 0,301.

Keywords— *k-Nearest Neighbou r(k-NN), weather forecasting, Locality Sensitive Hasihing (LSH)*

1. BACKGROUND

Weather forecast could be thought of as a piece of knowledge about the air conditions that will be present in the future (a maximum of 5 days for moderate latitudes and 2 days for the tropics). Weather forecasting plays an important role, both for economic actors and for decision-makers, in ensuring that weather conditions in specific areas do not result in high moral or material losses.

Weather hazards, which are affected by a variety of natural causes, cause weather conditions to fluctuate, making forecasting difficult at times. Palembang's natural environment, which includes numerous rivers and swamps, has the ability to cause flooding if rain intensity is high. Correct weather forecasting is needed so that people and policymakers could set up plans ahead. The weather forecast application is expected to be an alternative solution to this problem.

The information used to help weather forecasts is normally in the form of daily data collected over a period of time, which is called a *time series*. *Time series* is a series of observations $x_i(t)$; [$i = 1, \cdots, n$; $t = 1, \cdots, m$] which is made sequential to time, where index i is the measurement performed at each time t [1]. If the $n = 1$, it is called a univariate *time series*, and if $n > 1$, it is called *multivariate time series*.

A time series could be used as data for a classification algorithm to process. One of the algorithms that can be used to classify data is *k-Nearest Neighbour* (*k-NN*). *k-NN* classifies data based on the number of *k nearest* neighbours.

Numerous research involving weather prediction and the *k-NN* system have been conducted. These studies, among others, were conducted by Barded and Patole in [2] who compared the Artificial Neural Network, *k-NN*, and the Naive Bayes Algorithm to classify and predict the weather with the best accuracy results obtained by the *k-NN* method. Another study examining the *k-NN* method for multi variate time series was conducted by [3]. The results

showed that the performance of *k-NN* was very good for classifying time-series datasets.

The researcher also used the Locality Sensitive Hashing (LSH) method as one of the methods that was supposed to provide a better solution to shorten the calculation time on *k-NN* and also used optimization theory to determine the best solution required by the device in this analysis. The *k-NN* algorithm which is hybridized with LSH is expected to provide good accuracy for predicting the weather in Palembang.

The study of [4] compared three methods for weather forecasting. These methods include Autoregressive Integrated Moving Average (ARIMA), Neural Network or neural network, and Adaptive Splines Threshold Regression (ASTAR). The three methods produce forecast values for three weather elements, namely temperature, relative humidity, and rainfall. The three methods were evaluated with correlation values and Root Mean Square Error (RMSE). The ASTAR method provided better forecasts, according to the findings. Research on weather forecasting was also carried out by [5] using methods for data mining, namely *K-medoids* and the Naive Nayes algorithm. The weather forecast is based on temperature, humidity, and wind parameters. According to the findings of this report, weather forecasting using data mining techniques produces reasonably accurate predictions.

Following that are research on the form to be used, namely *k-NN*. One such research is [6] implementing the *k-NN* method to determine the location of the closest rain post to the route of travel for the Clearroute application. Clearroute is an application designed to make it easier for people to find out the weather conditions along their planned path. Calculating accuracy using a configuration matrix and the value of $k = 3$ yields an accuracy of 73%.

Locality Sensitive Hashing (LSH) is a well-known computational tool for efficiently searching for nearest neighbours on large datasets. [7] used LSH to detect earthquakes with a data-driven science scaling case study. Meanwhile, [8] uses LSH in a data-driven predictive model that produces simulations in the form of graphs, error calculations, and computation time.

The *k-NN* and LSH methods have a reasonably high degree of precision or accuracy, according to previous studies. In this study, the *k-NN* and LSH Hybrid methods are used to forecast the weather in Palembang. Following the implementation, the prediction results were compared to the actual data to determine the method`s level of accuracy. The Mean Square Error was used to determine the degree of accuracy.

2. RESEARCH METHODOLOGY

The LSH method was used to produce hash values for each record in the training data in this analysis. Furthermore, the hash value obtained will be used to classify the test data that will be calculated using k-NN. The data used in this study were obtained from the Meteorology, Climatology and Geophysics Agency (BMKG) in Palembang. Data in the form of features that affect weather such as minimum temperature, maximum temperature, average temperature, humidity, length of exposure, wind direction, wind speed, greatest wind speed, and rainfall.

2.1 The Research Stages

The stages of the research are as follows :

1. Data collection in the form of weather data and data that did not fit the desired format were both discarded. In this analysis, data that had gone through the data filtering phase were chosen to be used in the dataset and used as training and test data.
2. The aim of data analysis is to find and analyze any problems or anomalies that might occur in the raw data collected. In such cases, preprocessing is needed to generate

normal data that can then be processed using the chosen method. At this stage, a method selection is made based on the data collected and the results of the data analysis.

3. The *k-NN* form, which has been hybridized with LSH, is used in the implementation. The design will be done in the form of a flowchart, data flow diagram, and so on at this stage. Then, in accordance with the current design, software code is developed.
4. An existing dataset was used to evaluate the software code that had been completed. There were two types of data in the dataset: training data and research data. A review of the effects of the chosen method was carried out at this point. The consistency of the system may be used to evaluate its performance. The accuracy was calculated using the Mean Square Error method (MSE).
5. The interpretation of the data and the measurement of precision was used to draw conclusions.

The research stages are available in Figure 2.1, as follows:

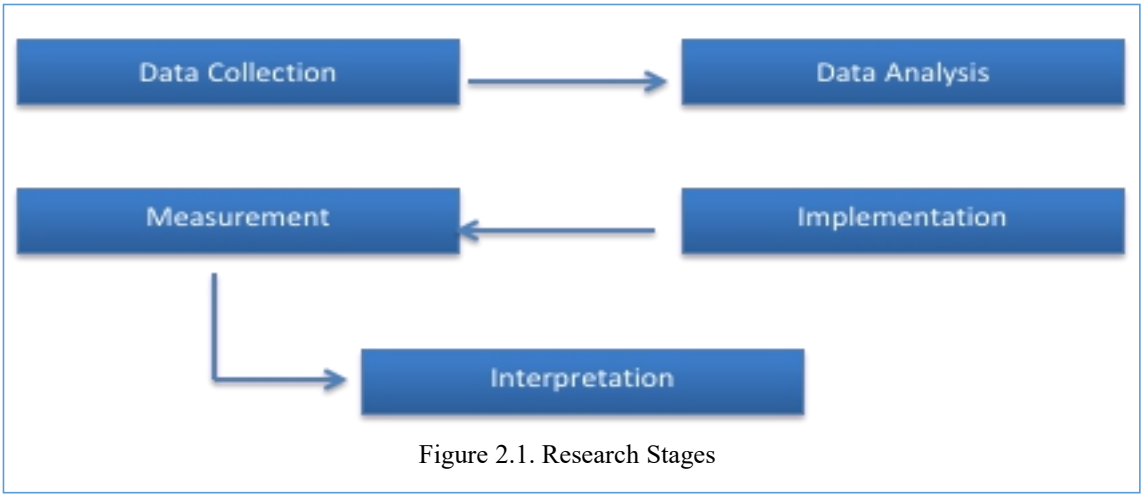


Figure 2.1. Research Stages

2. 2 Implementation Stages

An outline of the stages of implementing the k-NN and LSH methods in this study are as follows:

Inisialization. Determine the variables *n, m, A*, and *a*'s values. The row and column sizes of the matrix to be created in this case are m and n, respectively, depending on the number of features and data records. *A* is the matrix that has been formed, and *a* is a vector whose size is determined by the number of features in the dataset. The Palembang Meteorology, Climatology, and Geophysics Agency (BMKG) provided data for the method of determining variable values. A total of 1200 data records were used, which were split into training and testing data. Data preprocessing was also done at this stage.

Stage 1. A vector was formed from the large number of features; vector *a* is 9 in size in this analysis. The construction of a matrix *A* of size *m x n*, with *m* and *n* values determined by initialization values. There were 1200 data records based on the data collected, with each data having 9 attributes, resulting in a matrix measuring 9 x 1200.

The following is a vector a and a matrix A design.

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

Stage 2. The aim of data normalization is for the data to be naturally distributed. The data, for example, is normally distributed as shown in Figure 2.1.

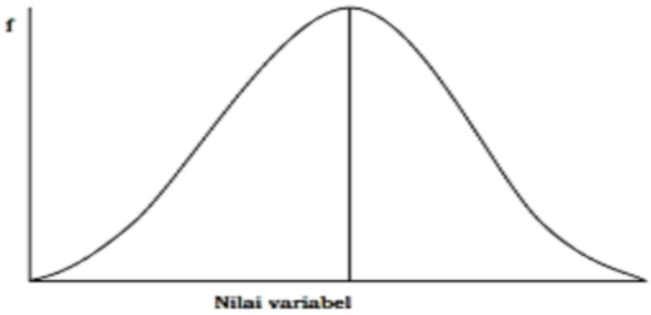


Figure 2.1. Normal Curve

Stage 3. The LSH method was used to classify the data in the matrix, where in this stage the following formula is used [9] :

$$h : \Re^d \to Z, h_{a,b}(v) = \left\lfloor \frac{a^T v + b}{r} \right\rfloor$$

Where $r > 0$ was a parameter of *hash,* $a \in R^d$ that is randomly chosen from a normal distribution, b is a scalar value of a *uniform distribution* and $\lfloor x \rfloor$ is a function of *floor,* v is variable that holds the value of a data record`s features.

Stage 4. Determining test data and compare it to training data to look for predictions of features that occur in test data. Weather prediction at $t + 1$ is performed using the current data t and the closest distance is searched with training data that has the same hash value or is +/- *1* to the hash value at *t*. For instance, if s is the closest data to t, then $s + 1$ is used to predict $t + 1$. Repeat steps 2 and 3 for normalization and classification on the test results.

Stage 5. The shortest distance was calculated using the *k-NN* method after the results of the test data classification were obtained, particularly for those with the same classification results between the training data and the test data. The class chosen is the one with the most members and is nearest to the test results. Figure 2.2 shows an example of this form.

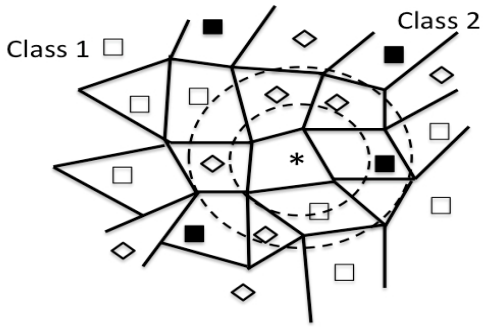


Figure 2.2. Classification *k-Nearest Neighbour* [8]

3. FINDINGS AND DISCUSSION

The data was first analyzed before being used. This research is performed to ensure that the data processed by these two methods is free of any noise or other irregularities. Blank values and anomalous values for other data are examples of noise or anomalies observed in raw data. Since the raw data obtained for the wind direction function is not in the form of numbers, such as west, southwest, or south, this feature is assigned a numerical value based on the magnitude of the wind direction angle whose zero point is measured from north and clockwise. As a result, the wind direction feature's values are in the range of 0 to 360, with a small change in wind direction causing a large change in value, affecting the prediction results.

Each data record is assigned a hash value by the LSH process, which is then used to identify the data. Figure 3.1 shows a graph of the hash value for a set of 300 data from 1200 data. The classification results of the test data calculated these hash values.

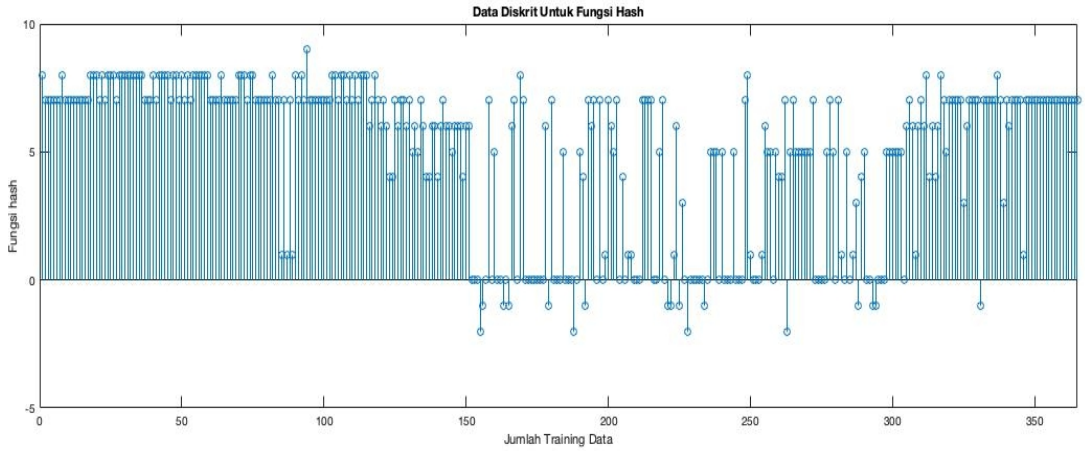
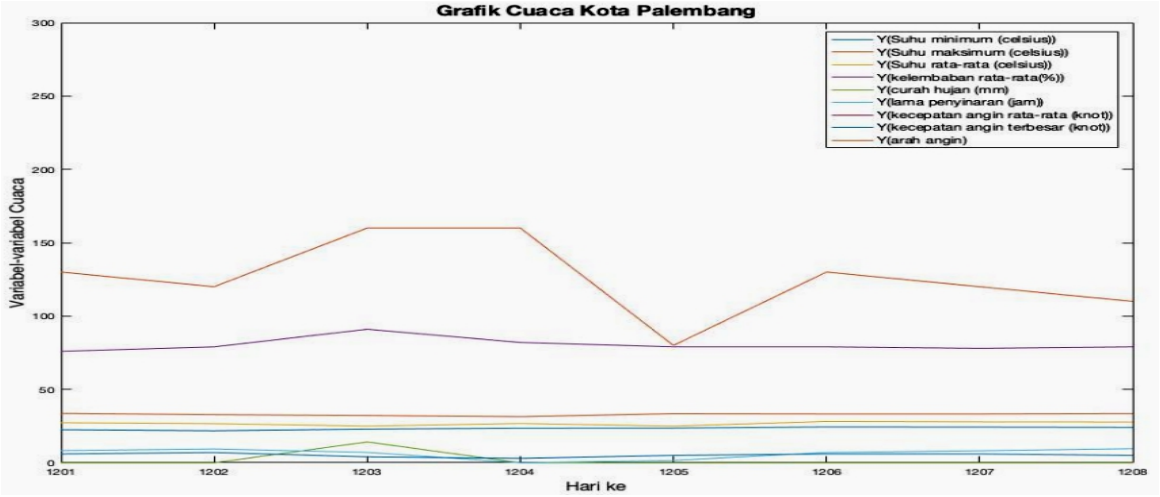


Figure 3.1 The *hash* value of 300 data

The data for the selected test measured the expected value of each weather function after obtaining the hash value for each data record. This estimate was dependent on the previous day's feature value as well as the hash value obtained during training. The test was performed on the data from 1201 to 1207. The importance of the features in each of the test data is used to make predictions. Figure 3.2 depicts the test data's real function values, while Figure 3.3 depicts the prediction outcomes.



Feature 3.2. Weather Feature

The forecast results for the features of minimum temperature, maximum temperature, average temperature, air humidity, long exposure, wind speed, greatest wind speed, and rainfall are identical to the real feature values, based on the display of the two graphs. In the case of the wind direction function, there is a major difference in the graph. This is due to the fact that the value in the wind direction feature is expressed in degrees, so even a minor shift in wind direction results in a significant change in value. To address this, the test results were also normalized, resulting in values that were usually distributed.

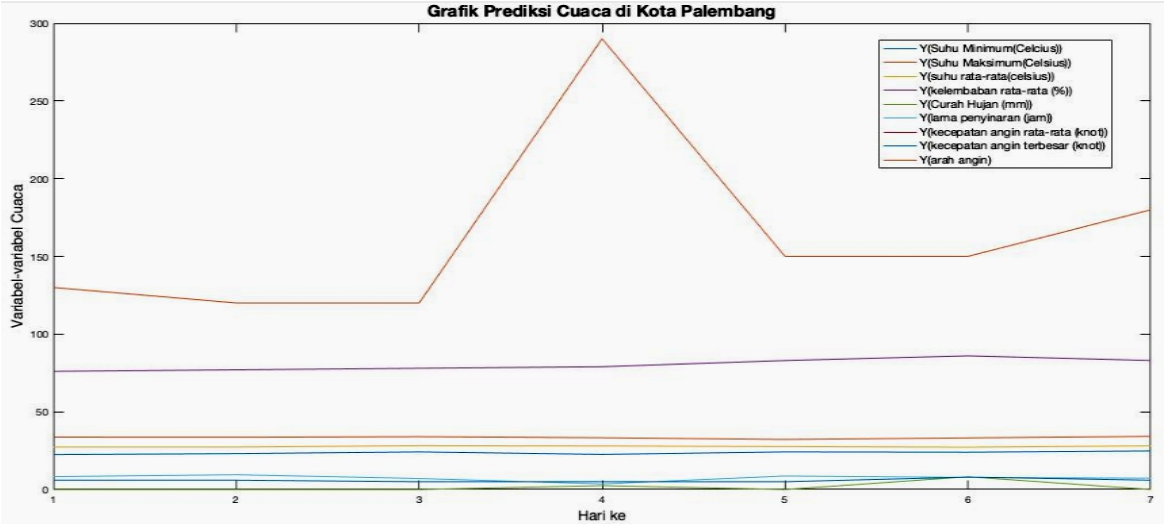


Figure 3.2. Prediction of Weather Feature

Table 3.1 shows the normalized feature values. The values in this table are used to measure accuracy using the Mean Square Error, where:

MSE : Mean Square Error
y : true data
: Predictive data
n : number of data tested

Table 3.1. Normalization of Test Data Feature

Fea-tures	Day:													
	1		2		3		4		5		6		7	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R
F1	-0.279	-0.279	-0.263	-0.298	-0.230	-0.327	-0.319	-0.247	-0.266	-0.142	-0.312	-0.243	-0.281	-0.235
F2	-0.016	-0.016	-0.0002	-0.023	0.012	-0.149	-0.203	-0.098	-0.103	0.172	-0.122	-0.036	-0.118	-0.011
F3	-0.167	-0.167	-0.158	-0.175	-0.131	-0.287	-0.260	-0.185	-0.197	-0.097	-0.246	-0.155	-0.223	-0.145
F4	0.975	0.975	1.087	1.129	1.103	0.973	0.290	0.855	0.928	1.620	0.967	1.025	0.724	1.104
F5	-0.807	-0.807	-0.842	-0.842	-0.830	-0.493	-0.536	-0.690	-0.757	-0.893	-0.639	-0.811	-0.709	-0.841
F6	-0.612	-0.612	-0.604	-0.607	-0.654	-0.629	-0.524	-0.690	-0.581	-0.842	-0.647	-0.647	-0.583	-0.637
F7	-0.666	-0.666	-0.692	-0.667	-0.706	-0.688	-0.509	-0.633	-0.656	-0.734	-0.643	-0.671	-0.605	-0.692
F8	-0.666	-0.666	-0.692	-0.667	-0.706	-0.688	-0.509	-0.633	-0.656	-0.734	-0.643	-0.671	-0.605	-0.692
F9	2.241	2.241	2.165	2.152	2.144	2.291	2.573	2.325	2.289	1.652	2.288	2.211	2.399	2.151

Explanation :

F1 : minimum temperature (celsius)
F2 : maximum temperature (celcius)
F3 : average temperature (celcius)
F4 : average humidity (%)
F5 : rainfall intensity (mm)
F6 : length of exposure (hour)
F7 : average of wind speed (knot)
F8 : greatest wind spped (knot)
F9 : wind direction (degree)
P : predictive value
R : true value

The error rate in weather prediction using LSH and *k-NN* is obtained from the MSE calculation which results in a value of 0.301.

4. CONCLUSION

Weather prediction in Palembang using a hybrid method of LSH and *k-NN* has been done and could run well. The MSE value obtained is 0.301, hence, the accuracy level of this hybrid method is around 70%.

5. SUGGESTION

In order to improve accuracy, further training and testing data or other classification methods should be used.

ACKNOWLEDGEMENT

This research was fully funded by DIPA Budgeting of Community Service Agency of Sriwijaya University Financial Year of 2020 No. SP DIPA-023.17.2.677515/2020, First Revision on 16 March 2020 based on Decision Letter of Rector No: 0684/UN9/SK.BUK.KP/2020 on 15 July 2020 and contract number: 0163.268/UN9/SB3.LPPM.PT/2020.

REFERENCES

[1] Yang, S. and Shahabi, C. (2007). An efficient k-nearest neighbour search for multivariate time series. *Information and Computation*, 205:65–98.

[2] Barde, N. and Patole, M. (2014). Classification and forecasting of waether using ann, k-nn, and naive bayes algorithms. *International Journal of Science and Research (IJSR)*, pages 1740 – 1742.

[3] Yang, S. and Shahabi, C. (2007). An efficient k-nearest neighbour search for multivariate time series. *Information and Computation*, 205:65–98.

[4] Sutikno, Bekti, R. D., Susanti, P., and Istriana (2010). Perkiraan cuaca dengan metode autoregressive integrated moving average, neural network, dan adaptive splines threshold autoregression di stasiun juanda surabaya. *Jurnal Sains Dirgantara*, 8(1):43 – 61.

[5] Biradar, P., Ansari, S., Paradkar, Y., and Lohiya, S. (2017). Weather predic- tion using data mining. *International Journal of Engineering Development and Research (IJEDR)*, 5(2):211 – 214.

[6] Rosidi, A., Ginardi, R. H., and Munif, A. (2017). Implementasi metode k- nearest neighbour untuk penentuan lokasi pos hujan terdekat dengantitik rute perjalanan pada aplikasi clearroute. *Jurnal Teknik ITS*, 6(2):A392 – A395.

[7] Rong, K., Yoon, C. E., Bergen, K. J., and Elezabi, H. (2018). Locality sensitive hashing for earthquake detection : A case study of scalling data-driven science. In *The 44th International Conference on Very Large Data Bases*, volume 44, pages 1674 – 1687, Rio de Janeiro, Brazil.

[8] Saputra, H., Malyan, A. B. J., Supani, A., and Indarto (2019). Data-driven predictive control menggunakan algoritma nearest neighbour untuk sistem yang tidak stabil. *Jurnal JUPITER*, 10(1):41 – 51.

[9] Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. (2004). Locality- sensitive hashing scheme based on p-stable distribution. Number 20, pages 253–262, Brooklyn, New York, USA. ACM.