Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 7 No. 2, August 2022, pp. 183 – 192

Visitor Prediction Decision Support System at Dieng Tourism Objects Using the K-Nearest Neighbor Method

Eko Hari Rachmawanto*[1], Christy Atika Sari[2]
Universitas Dian Nuswantoro, Semarang
E-mail : eko.hari@dsn.dinus.ac.id[1], christy.atika.sari@dsn.dinus.ac.id[2]*
**Corresponding author*

Heru Pramono[3], Wellia Shinta Sari[4]
Universitas Dian Nuswantoro, Semarang
E-mail : heru.pramono.hadi@dsn.dinus.ac.id[3], wellia.shinta@dsn.dinus.ac.id[4]

Abstract - A tourist target is anything that attracts a visitor or tourist to come to visit a place or area. Tourism goods play an important role in a country or region, becoming a source of national foreign exchange, increasing human resources, and improving the economy of surrounding communities. The problem posed in this study is how to implement a decision support system in predicting visitor numbers for Dieng tourists using the k-nearest neighbor method. The purpose of this study is to help the local government and surrounding communities to improve facilities such as restaurants, places of worship, parking lots, clean toilets so that tourists can feel safe and comfortable when visiting Dieng. Helps manage tourism targets. is what you give. These attractions using a decision support system as a process to predict visitors. The number of visitors who visited in December 2017 was 421,394, which serves as a reference for predicting the number of visitors who will visit Dieng in the following year. The predicted result is 29569.25 visitors with a parameter value of k = 8 and a minimum RMSE value of k = 1/0.

Keywords – Decision Support System, K-Nearest Neighbor Method, Root Mean Square Error

1. INTRODUCTION

Since the tourism industry can absorb labor that cannot be replaced by machines, it can be used as a source of foreign exchange and plays an important role for countries that can increase their human resources. The tourism sector also plays an important role in the economy as it can boost the economy and productivity of the surrounding areas[1].

One of Indonesia's most famous tourist destinations is the Dieng Tourism Object, west of the Mount Sindoro and Mount Sumbing complex, which averages 2000 meters above sea level and is contained in the central Wonosobo Regency and Banjarnegara Regency. Java state. Dieng retains the beautiful natural scenery and cultural dignity of the past, and has added value unique to Dieng compared to other tourist destinations. There are many attractions in Dieng such as lakes, Hindu temples, craters, Dieng Volcanic Theater, Dieng Kailasa Museum which stores relics and provides information on nature (geology, flora and fauna), Selayu River Springs. Known as Tuk Bima Lukar[1].

In order to help the Banjarnegara Regency government to better prepare all the appropriate facilities such as parking lots, toilets, concessions, places etc., this study aims to

increase the number of visitors to visit Dieng tourists in the following years. Learn how to create a predictive decision support system. Worship so that visitors feel comfortable when they visit.

Decision Support System (DSS) is a system that is able to provide problem solving skills and communication skills for problems with semi-structured and unstructured conditions where no one knows for sure how decisions should be made [2].

Decision Support System aims to provide information, guide, predict and direct information users to make better decisions. An example of implementing a Decision Support System that has been done previously is in predicting student graduation at STMIK Sinar Nusantara Surakarta using the K-Nearest Neighbor (K-NN) method.

2. RESEARCH METHOD

Several previous studies have been used as a reference in this study. Here are some related studies:

The first research conducted by Christian Gratia Nugroho et al in 2015 discussed the use of the Decision Support System which was applied to the system for selecting contraceptive methods for couples of childbearing age. This research uses the K-Nearest Neighbor Algorithm. In this study, it was found that the accuracy of the calculation of the K-NN Algorithm was 95% [3].

The second study conducted by Mustakim et al in 2016 discussed the K-Nearest Neighbor Classification Algorithm which was applied to predict student achievement predicates. The level of accuracy obtained from this study is 82% [4].

The third study conducted by Rofiq Harun et al in 2020 discussed Data Mining which was applied to determine the potential for daily rain. In this research, the K-Nearest Neighbor Algorithm is used. The results obtained accuracy of 37.3% in this study [5].

The fourth study conducted by Widi Setyoko et al in 2018 discussed the use of a decision support system to predict the potential of the credit quality of prospective debtors. The K-Nearest Neighbor method is also used in this study. The accuracy results obtained from this study were 81.8% [6].

2.1. Decision Support System

Decision Support System (DSS) is a system that is able to provide problem solving skills and communication skills for problems with semi-structured and unstructured conditions where no one knows for sure how decisions should be made [2].

Decision Support System aims to provide information, guide, predict and direct information users to make better decisions [7].

In general, Decision Support Systems are built by three major components, namely Database Management, Model Base and User Interface.

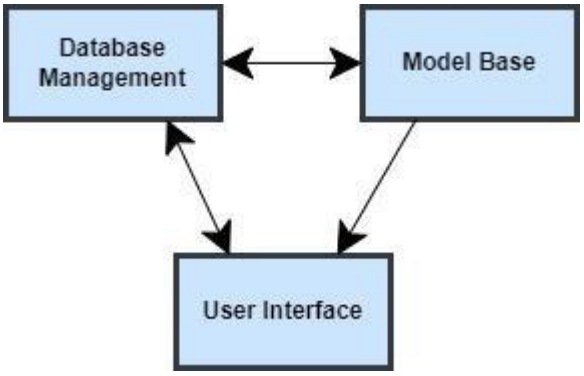

Figure 1. Decision support system components

Database Management is a data subsystem organized in a database. Data which is a decision support system can come from outside or within the environment. Decision Support System, requires data that is relevant to the problem to be solved through simulation [8][9].

Model Base is a model that represents problems in a quantitative format as a basis for simulation or decision making, including the objectives of the problem (objective), related components, existing constraints (constraints), and other related matters [8]. [9].

While the User Interface is a merger between the two previous components, namely Database Management and Model Base which are united in a third component (User Interface), after previously presented in the form of a model that is understood by computers. The User Interface displays the system output for the user and receives input from the user into the Decision Support System [8][9].

2.2. Data Mining

Data Mining is a merging process from several fields of science including Databases, Information Retrieval, Statistics, Algorithms and Machine Learning. [10]. Data mining techniques are used to examine a large database and to find new and useful patterns. Not all jobs looking for information are declared as data mining.

Data Mining is a way to find information hidden in databases and is part of the Knowledge Discovery in Databases (KDD) process to find useful information and patterns for that data [10].

2.3. Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is the evaluation process of the prediction technique used to measure the level of accuracy of a model. RMSE is the result of a function of several error characters, not one character, namely the average Error value [11].

RMSE is also used to identify any discrepancies in the model itself. The smaller the value obtained, the better the results obtained. The following is the RMSE calculation formula:

$$MSE = \sqrt{\frac{\sum(y_t - \hat{y}_t)^2}{n}} \qquad (1)$$

Information :
y_t = index actual value
$\hat{y}_t$ = index predictive value
n = number of samples or data

2.4. K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) is a method that uses a supervised algorithm, where the results of the new query instance will be classified based on the category of K-NN. The purpose of this algorithm is to classify the new object based on attributes and training sample data, the classifier also does not use any model to be matched based on memory. Then from the query point will be given a number of k objects or training points that are closest to the query point. So this classification uses a voting system of the highest value among the classification of k objects. The K-Nearest Neighbor (K-NN) method uses description classification as the predictive value of the new query instance [12].

To make a decision (class) between coming or not a class, it can be seen from the majority of decisions or closest neighbors. The neighbors will be selected based on the approach of the most. Near or far neighbors are usually calculated based on the value of Euclidean Distance, or can also use the following formula:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \qquad (2)$$

Information:
x = sample datay = test data
d = distance

$$d(a,b) = \sqrt{\sum_{i=1}^{n} \left(a_i - b_j\right)^2} \qquad (3)$$

Information:
$d(a,b)$ = distance between training data and testing data
a_i = value of x in training data
b_j = value of x on testing data
n = limit on the amount of data

Values a and b are the reference values of the training data and the new test/data values, which will be classified with a total of n. After we get the Euclidean Distance value, we can do the classification by finding the k value based on the closest training data to the testing data to be tested. After we get the value of k, then input the data into the training data class which has a value of k [12].

2.5. Proposed method

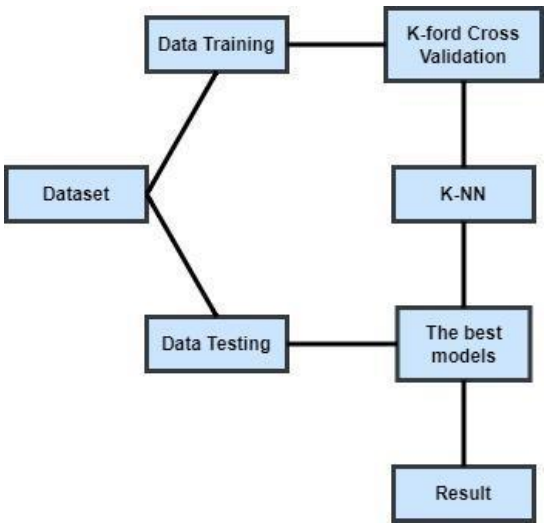

Figure 2. Proposed method process flow

Based on Figure 2, it can be explained from the working system of the proposed method that from the dataset to be obtained, the data will be divided into two, namely training data and testing data.

After dividing the data, the next step is to change the shape or format of the univariate dataset to a multivariate dataset. Training data will be used as several independent variables, such as 1 period to 5 periods with each dependent variable.

Example of 27 period data such as xt-1 : xt, 2 input variables such as xt-2, xt-1 : xt and 5 period data such as xt-5, xt-4, xt-3, xt-2, xt-1 : xt, as well as data testing.

The training data will be converted into several periods to get the right model. Then from the distribution of the data, it will proceed to the next process, namely into the K-NN method process, in this process an experiment will be carried out by changing the neighbor value or k value in each training data experiment starting from 1 to 5, this is done to get a model good so that the smallest RMSE value is obtained. After obtaining the right model, the testing data will be processed and get a predictive value.

3. RESULTS AND DISCUSSION

3.1. Data

Source The source of data or dataset in this study comes from the Banjarnegara Regency Culture and Tourism Office, namely visitor data for Dieng tourism objects from 2014-2017 in the form of monthly time series data, which is divided into local and foreign visitor data which will then be converted into univariate datasets. in 48 data records arranged in descending order.

The univariate dataset obtained will be converted into a multivariate dataset because it is still a single dataset. The following is a univariate dataset before processing :

Table 1. Data Sample

No	Year	Month	Total
1	2017	12	38909
2	2017	11	7601
3	2017	10	9711
4	2017	9	14510
5	2017	8	191826
6	2017	7	42471
7	2017	6	37347
8	2017	5	32529
9	2017	4	9218
10	2017	3	10764
11	2017	2	12252
12	2017	1	14256
…	…	…	…
…	…	…	…
…	…	…	…
37	2014	12	34383
38	2014	11	9675
39	2014	10	9449
40	2014	9	11417
41	2014	8	65273
42	2014	7	45101
43	2014	6	40370
44	2014	5	37762
45	2014	4	11193
46	2014	3	9575
47	2014	2	8693
48	2014	1	12598

The univariate dataset in the table above will be converted into multivariate data which is divided into 5 periods. The initial stages of data processing to multivariate are used to determine the output or target value. In this case the researcher uses the xt label which is the number of tourists visiting the tourist attraction. The next stage is to determine the input value (xt-i) according to the period determined by the researcher. The following is a univariate dataset after processing:

Table 2. 5-period Multivariate Dataset Sample

Data to-	Xt-5	Xt-4	Xt-3	Xt-2	Xt-1	Xt
1	42471	191826	14510	9711	7601	38909
2	37347	42471	191826	14510	9711	7601
3	32529	37347	42471	191826	14510	9711
4	9218	32529	37347	42471	191826	14510
5	10764	9218	32529	37347	42471	191826
…	…	…	…	…	…	…
…	…	…	…	…	…	…
44	0	10542	8821	9675	13021	37762
45	0	0	10542	8821	9675	13021
46	0	0	0	10542	8821	9675
47	0	0	0	0	10542	8821
48	0	0	0	0	0	10542

Based on the multivariate dataset in the table above, the number of tourists visiting in 5 periods, where xt is the output/target value and the values xt-1, xt-2, xt-3, xt-4, xt-5 are input values. Furthermore, the data will be tested using the K-Nearest Neighbor (K-NN) method by setting the value or parameter k (in this case the researcher uses k = 1 to k = 10) which will then be compared and look for the value of the Root Mean Square Error (RMSE).) best of all available data.

3.2. *Application of K-NN Metode Method*

After the dataset is processed, at this stage a calculation will be carried out using the K-Nearest Neighbor (K-NN) method to find the closest distance between the training data and the testing data, where the calculation will use the Euclidean Distance method.

The dataset used in the calculation of the K-NN method this time is a multivariate dataset with 5 periods. This dataset has 5 inputs (xt-1 to xt-5) and one output/target (xt) for a total of 48 data records.

Then from the 48 records the dataset is divided into two, namely training data and testing data, namely data from 1-35 are used as training data and data to 36-43 are used as testing data. The following is the training data before the calculation process using the K-NN method is carried out:

Table 3. Sample Data Training

Data to-	Xt-5	Xt-4	Xt-3	Xt-2	Xt-1	Xt
1	42471	191826	14510	9711	7601	38909
2	37347	42471	191826	14510	9711	7601
3	32529	37347	42471	191826	14510	9711
4	9218	32529	37347	42471	191826	14510
5	10764	9218	32529	37347	42471	191826
...	...	...	...	...	...	...
...	...	...	...	...	...	...
31	20886	11293	10878	15286	33573	46394
32	34383	20886	11293	10878	15286	33573
33	9675	34383	20886	11293	10878	15286
34	9449	9675	34383	20886	11293	10878
35	11417	9449	9675	34383	20886	11293

After the training data is grouped, then calculate the distance of the nearest neighbor between each value from the training data to the testing data using the Euclidean Distance method. The following is an example of calculating the Euclidean Distance method on training data as follows:

1. $d_1 = \sqrt{\begin{matrix}(38909-7601)^2+(7601-9711)^2+(9711-14510)^2 \\ +(14510-191826)^2+(191826-42471)^2\end{matrix}} = 233999{,}0454$

2. $d_2 = \sqrt{\begin{matrix}(7601-9711)^2+(9711-14510)^2+(14510-191826)^2 \\ +(191826-42471)^2+(42471-37347)^2\end{matrix}} = 231951{,}7574$

3. $d_3 = \sqrt{\begin{matrix}(9711-14510)^2+(14510-191826)^2+(191826-42471)^2 \\ +(42471-37347)^2+(37347-32529)^2\end{matrix}} = 231992{,}1955$

4. $d_4 = \sqrt{\begin{matrix}(14510-191826)^2+(191826-42471)^2+(42471-37347)^2 \\ +(37347-32529)^2+(32529-9443)^2\end{matrix}} = 233111{,}0274$

5. $d_5 = \sqrt{\begin{matrix}(191826-42471)^2+(42471-37347)^2+(37347-32529)^2 \\ +(32529-9443)^2+(9443-9864)^2\end{matrix}} = 151334{,}6535$

In the 6th to 35th data, the calculation process is the same as the calculations for the 1st to 5th data. The following is the result of calculating the distance to the nearest neighbor based on training data :

Table 4. Euclidean Distance Value from Training Data

Data to-	Xt-5	Xt-4	Xt-3	Xt-2	Xt-1	Xt	d
1	42471	191826	14510	9711	7601	38909	233999,0454
2	37347	42471	191826	14510	9711	7601	231951,7574
3	32529	37347	42471	191826	14510	9711	231992,1955
4	9218	32529	37347	42471	191826	14510	233111,0274
5	10764	9218	32529	37347	42471	191826	151334,6535
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
31	20886	11293	10878	15286	33573	46394	24706,6944
32	34383	20886	11293	10878	15286	33573	25064,15201
33	9675	34383	20886	11293	10878	15286	30071,29214
34	9449	9675	34383	20886	11293	10878	29747,32296
35	11417	9449	9675	34383	20886	11293	29809,46195

The next step is the process of sorting the data on the Euclidean Distance (d) value from the smallest data to the largest data based on the distance of the closest neighbors and grouped into the y category with the smallest k value.

Table 5. Euclidean distance value from training data after sorting

Data To-	Xt-5	Xt-4	Xt-3	Xt-2	Xt-1	Xt	d
19	12703	9471	8626	7640	17793	21545	11370,70613
18	9471	8626	7640	17793	21545	29811	13681,1509
30	11293	10878	15286	33573	46394	42456	23106,33989
7	14256	12252	10764	9218	32529	37347	23984,08057
6	12252	10764	9218	32529	37347	42471	24443,31158
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
1	42471	191826	14510	9711	7601	38909	233999,0454
14	21545	29811	227203	8941	6085	5250	294413,1163
15	17793	21545	29811	227203	8941	6085	294435,839
16	7640	17793	21545	29811	227203	8941	294596,9959
13	29811	227203	8941	6085	5250	34379	295735,1094

After performing the fknn search process or estimation on the application of the K-Nearest Neighbor (K-NN) method, a test will be carried out on that value to find the error value. This test is carried out using the Root Mean Square Error (RMSE) method on testing data with the following details:

Table 6. Testing Data

Data to-	Xt-5	Xt-4	Xt-3	Xt-2	Xt-1	Xt
36	65273	11417	9449	9675	34383	20886
37	45101	65273	11417	9449	9675	34383
38	40370	45101	65273	11417	9449	9675
39	37762	40370	45101	65273	11417	9449
40	11193	37762	40370	45101	65273	11417
41	9575	11193	37762	40370	45101	65273
42	8693	9575	11193	37762	40370	45101
43	12598	8693	9575	11193	37762	40370

Table 7. Calculation Results of Xt_Prediction, Error, Error2, and RMSE . values

Data to-	k	Xt_Actual	Xt_Prediction	Error	Error2	RMSE
36	k = 1	20886	20886	0	0	0

37	k = 2	34383	27635	6749	45549001	2386
38	k = 3	9675	21648	-11973	143352729	4233
39	k = 4	9449	18598	-9149	83704201	3235
40	k = 5	11417	17162	-5745	33005025	2031
41	k = 6	65273	25181	40093	1607448649	14175
42	k = 7	45101	28026	17075	291555625	6037
43	k = 8	40370	29569	10801	116661601	3819

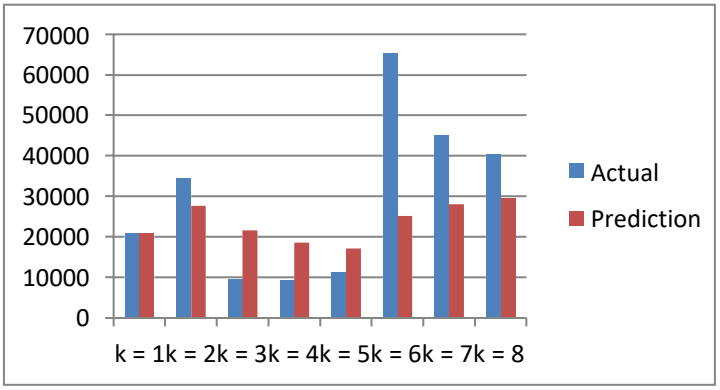

Figure 3. Prediction value search result graph

3.3. RMSE Prediction Test and Results

This system can perform the process of estimating K-NN predictions and will produce an RMSE value that will be different from the number of visitors who come by month. The following is the process of testing predictions and RMSE results:

Table 8. Prediction test results and RMSE value

Test Item Name	View RMSE predictions and results		
Purpose	Show RMSE predictions and results		
Initial Condition	Fill in the blank data		
Scenario			
1. Open the prediction menu 2. Input the amount of k 3. Then click process			
Result			
Provided data	Purpose	Observation	Conclusion
Number of visitor data, RMSE value	Displays forecast forecast and RMSE value	The data is successfully processed and displays the RMSE value	Valid

4. CONCLUSION

From the explanation that has been done, it is found that the K-Nearest Neighbor (K-NN) method can solve the problem of predicting visitors to Dieng tourism objects where the attributes are numerical data. Then in December 2017 the number of visitors who came was 421394 which will be a reference in predicting the number of visitors who come to Dieng in the following year. The prediction results are 29569.25 visitors with a parameter value of k=8 and the smallest RMSE value is at k=1 of 0. So, with the most data from the previous 3 years in August, this data can be used as a tool for local governments to prepare all the facilities in Dieng in the following year, especially in August to better prepare the facilities so that they can reach all visitors who come by adding or thoroughly tidying up facilities such as parking lots, toilets, canteens and places of worship so that visitors feel comfortable when visiting.

REFERENCES

[1] Ningsih, A. S., Waspiah, & Salsabilla, S. (2019). Indikasi Geografis atas Carica Dieng Sebagai Strategi Penguatan Ekonomi Daerah. Suara Hukum, 105-120.

[2] Ningsih, E., Dedih, & Supriyadi. (2017). SISTEM PENDUKUNG KEPUTUSAN MENENTUKAN PELUANG USAHA MAKANAN YANG TEPAT MENGGUNAKAN WEIGHTED PRODUCT (WP) BERBASIS WEB. ILKOM Jurnal Ilmiah Volume 9 Nomor 3, 244-254.

[3] Nugroho, C. G., Nugroho, D., and Fitriasih, S. H. DECISION SUPPORT SYSTEM FOR THE SELECTION OF CONTRACEPTION METHODS IN WOMEN OF RELIABLE AGE WITH K-NEAREST NEIGHBOUR (KKN) ALGORITHM. Jurnal Ilmiah SINUS. 2015; 21-29.

[4] Mustakim, & Oktaviani, G. (2016). Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa. Jurnal Sains, Teknologi dan Industri, Vol. 13, No.2, 195-202.

[5] Pelangi, H. R., Chandra, K., and Yuliyanti, L. APPLICATION OF DATA MINING TO DETERMINE THE POTENTIAL OF DAILY RAIN USING THE K NEAREST NEIGHBOR (KNN) ALGORITHM. LPPM STMIK Lombok. 2020; 3(1).

[6] Setyoko, W., Hasbi, M., & Fitriasih, S. H. (2018). SISTEM PENDUKUNG KEPUTUSAN PREDIKSI POTENSI KUALITAS KREDIT CALON DEBITUR MENGGUNAKAN METODE K-NEAREST NEIGHBOR PADA BPR KARTASURA MAKMUR DI SUKOHARJO. Jurnal Tikomsin, 8.

[7] Rusydiana, A. S. (2020). PREDIKSI PERTUMBUHAN PERBANKAN SYARIAH DI INDONESIA TAHUN 2020 DENGAN QUANTITATIVE METHODS. Jurnal Ekonomi Syariah, Vol. 4, No. 2., 75-91.

[8] Cholifah, W. N., Yulianingsih, & Sagita, S. M. (2018). PENGUJIAN BLACK BOX TESTING PADA APLIKASI ACTION DAN STRATEGY BERBASIS ANDROID DENGAN TEKNOLOGI PHONEGAP. Jurnal String Vol. 3 No.2 Desember 2018, 206-210.

[9] Munawar. (2018). Analisis Perancangan Sistem Berorientasi Objek dengan UML. Bandung: Informatika.

[10] Riszky, A. R., & Sadikin, M. (2019). Data Mining Menggunakan Algoritma Apriori untuk Rekomendasi Produk bagi Pelanggan. Jurnal Teknologi dan Sistem Komputer, Volume 7, Nomor 3, 103-108.

[11] Budiman, A. S., & Parandani, X. A. (2018). UJI AKURASI KLASIFIKASI DAN VALIDASI DATA PADA PENGGUNAAN METODE MEMBERSHIP FUNCTION DAN ALGORITMA C4.5 DALAM PENILAIAN PENERIMA BEASISWA. Jurnal SIMETRIS, Vol. 9 No. 1, 565-578.

[12] Nugraha, P. D., Al Faraby, S., & Adiwijaya. (2018). Klasifikasi Dokumen Menggunakan Metode k-Nearest Neighbor dengan Information Gain. e-Proceeding of Engineering : Vol.5, No.1, 1541.