EFFECT OF NORMALIZATION OF GENRE MUSIC DATA ON CLASSIFICATION PERFORMANCE WITH RANDOM FOREST

Wahyudi[1], Mohammad Reza Faisal[2], Dwi Kartini[3],
Irwan Budiman[4], Andi Farmadi[5]
[12345]Computer Science FMIPA ULM
A. Yani St. KM 36 Banjarbaru, South Kalimantan
Email: j1f115019@mhs.ulm.ac.id[1], reza.faisal@ulm.ac.id[2], dwikartini@ulm.ac.id[3],
irwan.budiman@ulm.ac.id[4], andifarmadi@ulm.ac.id[5]

Abstract

This research is about the classification of the music genre using the Random Forest method. This test uses a dataset from GitHub or GITZAN about the music genre with 10 labels, 26 features and 1000 total data. This research is divided into two stages, namely by classifying all data without being normalized, and by using all normalized data. . In this research, Min-Max is used for data normalization method, and for accuracy calculation using Confusion Matrix method. The resulting accuracy when using all data with data that is not normalized produces an accuracy of 66.3%, while the resulting accuracy performance when using all data with normalized data results in an accuracy of 65.1%.

Keywords: Normalization, Random Forest, Min-Max, Confusion Matrix, Accuracy

1. INTRODUCTION

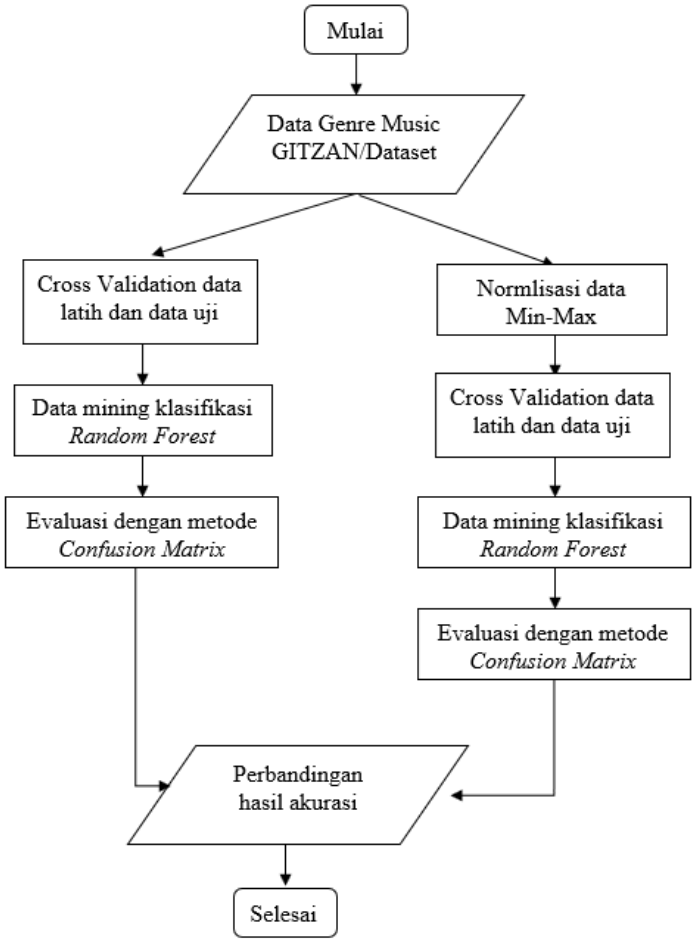
Data normalization is the process of scaling the attribute values of data so that they fall within a certain range, and classification is a process for grouping the same objects or entities and separating objects or entities that are not the same.

In a study conducted by Wei Du et al (2016) regarding the classification of music genres using the K-Nearest Neighbors method and Support Vector Machines. By using the dataset from GITZAN, where the dataset has 26 features, 10 classes with a total of 1000 data. Based on this dataset which has such a far range from each data, this research was carried out with normalization and use of the random forest classification method based on suggestions from research by Wei Du. et al.

Random Forest is also a method developed to improve decision tree methods that are prone to overfitting. Overfitting itself is something of a deficiency, which cannot recognize patterns outside of learning. Based on this, the purpose of this study is to determine the accuracy of the effect of normalization on the classification of music genres.

2. RESEARCH PROCEDURE

The work procedures that will be carried out in this study:



a. Research Dataset

Dataset from GITZAN*https://github.com/kumargauravsingh14/music-genre-classification*used for research which contains 1000 audio data, each of which is 30 seconds long and contains 10 music genres, namely blues, classical, country, disco, hip-hop, jazz, reggae, rock, metal and pop. So each genre has 100 audio data.

b. Dataset sharing

For the sharing of training data and test data on the dataset using cross validation.

c. Data Normalization

This research will also carry out the data normalization stage using the MIN-MAX method.

d. Data Mining Classification

This study uses the Random Forest method as a music genre classification.

e. Evaluation

At the evaluation stage, the method used is the Confusion Matrix. Which serves to measure the performance of the classification algorithm.

3. Results and Discussion

3.1. Result

3.1.1. Dataset

The data used in this study contained 10 music genre extraction labels with a total of 1000 audio data, each of which lasted 30 seconds. The dataset is from GITZAN or GitHuband the dataset has 26 features. The dataset can be seen in table 1.

Table 1 Dataset

C_stft	rmse	S_C	S_B	SR_off	ZCR	mfcc1	-	mfcc20	L
0.430894125103 60664	0.196221590042 11426	19,465,656,516,9 97,700	19,799,099,336,7 76,700	39,558,677,460,7 56,000	0.0974536813080 4953	- 6,777,097,956,346 ,110	-	- 0.6600370657688 395	the blues
0.338896085804 55005	0.251350194215 77454	21,414,616,556,4 43,500	21,680,155,600,4 12,000	4,627,997,015,13 5,200	0.1051512916021 6719	- 29,362,092,860,47 2,900	-	- 13,048,123,730,19 3,100	the blues
0.263016104481 94723	0.170081377029 41895	13,790,817,422,0 26,900	20,040,008,502,5 10,500	3,015,831,763,67 7,920	0.0393758012045 2786	- 20,698,758,963,23 4,400	-	- 17,835,099,945,87 2,500	the blues
0.307921016118 63837	0.131785482168 19763	145,175,414,666, 004	15,773,699,166,3 09,100	2,955,348,796,07 6,810	0.0614350026606 0371	- 17,939,544,684,23 2,500	-	- 8,638,613,351,936, 610	the blues
0.332479590883 5457	0.117412768304 34799	25,532,324,147,7 39,300	22,801,286,685,7 71,800	5,148,102,203,46 3,620	0.1468517197174 9225	- 8,515,025,014,369 ,660	-	3,752,625,581,331, 530	the blues
0.377687767406 78095	0.131890103220 93964	16,133,157,252,2 62,700	19,722,022,614,9 82,500	34,704,040,149,4 17,000	0.0582358534491 09906	- 17,749,076,092,80 9,300	-	- 5,375,209,218,828, 230	the blues
0.466435909284 9299	0.192153707146 6446	2,225,216,647,15 9,950	22,551,950,050,8 03,100	4,703,188,227,13 0,900	0.1085866449787 1517	8,635,233,803,031 ,840	-	1,238,193,206,433, 140	the blues
0.375158189190 5781	0.198281615972 51892	12,365,742,801,2 74,800	1,602,309,124,32 2,470	260,415,091,972, 233	0.0446259584220 2012	- 14,863,917,335,27 0,700	-	- 0.0487289471766 01194	the blues
0.380260210341 482	0.248262286186 21826	21,169,429,590,0 51,500	19,566,110,562,2 33,200	419,610,796,004, 257	0.1272724730311 5324	- 26,929,784,696,80 1,000	-	29,189,869,428,52 2,000	the blues
0.289932320004 5883	0.103115268051 6243	2,513,716,817,32 9,610	2,345,230,614,18 6,440	5,247,443,269,42 8,450	0.1351461971507 3528	- 11,316,777,334,11 7,400	-	36,153,485,806,44 0,000	the blues

3.1.2. Division of Research Dataset

The data obtained were 1000 data with 10 music genres, and divided into 900 training data and 100 test data using 10 cross validation. The dataset has 10 labels namelyblues, classic, country, disco, hip-hop, jazz, reggae, rock, metal and pop. To share this database, use 10Fold. The dataset sharing table can be seen in table 2.

Table 2 Distribution of training and test data with 10Fold.

No.	label	Type of data		Total
		Training data	Test data	
1	Blues	90	10	100
2	Classic	90	10	100
3	Country	90	10	100
4	Disco	90	10	100
5	Hip-Hop	90	10	100
6	Jazz	90	10	100
7	Reggae	90	10	100
8	Rock	90	10	100
9	Metal	90	10	100
10	Pop	90	10	100
Total Data		900	100	1000

3.1.3. Data normalization

Perform data normalization using the min max normalization method with a range from 0.0 to 1.0. To perform the normalization of the min max, it is done in Excel, and below is an example of manual calculation of the MIN MAX method used.

$$D'(i) = \frac{D(i) - Min\,(D)}{Max(D) - Min(D)}(U - L) + L \quad (1)$$

$$D'(i) = \frac{1000 - 300}{2000 - 300}(1{,}0 - 0{,}0) + 0{,}0$$

$$D'i(i) = \frac{700}{1700}x1$$

$$D'(i) = 0{,}4117$$

Information:

D (i): Original i data value

Min (D): Minimum value of all data i

Max (D): The maximum value of all data i

U: Maximum value desired

L: Minimum value desired

The following is a table of 3 data that has been normalized.

c_stft	rmse	S_C	S_B	SR_off	ZCR	mfcc1	-	mfcc20	L
0.0598685330861244	0.059843103893983	0.270752327691123	0.274365524919377	0.488480955434153	0.0598324013713041	0.052478155895318	-	0.0597503194278519	the blues
0.0598585641486969	0.059849077647418	0.291871321997768	0.294748711705234	0.561313101395624	0.0598332354868565	0.0566401553370082	-	0.059680451366079	the blues
0.0598503417692684	0.059840271332083	0.20925958688978	0.276976024466098	0.386618350975912	0.0598261080333474	0.0373925987985638	-	0.0578892240342063	the blues
0.0598552076800824	0.059836121576393	0.217134392247109	0.230746157892105	0.380064396956517	0.0598284983755839	0.0403824923556233	-	0.0588857583093457	the blues
0.0598578688552227	0.059834564144555	0.336490935236084	0.306897324570302	0.617671868578392	0.059837754158795	0.0505949324889417	-	0.0602284769859091	the blues
0.0598627676279183	0.059836132913170	0.23464125738983	0.273530319057892	0.435875947807547	0.059828151714736	0.0405888847400976	-	0.0592393818414474	the blues
0.0598723844054053	0.059842663096870	0.300947040089321	0.304195504693154	0.569460845556138	0.0598336077428645	0.0504646738550044	-	0.0599560122768583	the blues
0.0598624935220009	0.059843327119062	0.193817432000465	0.233448578539991	0.342008480975061	0.0598266769420892	0.043715252143276	-	0.059816560974188	the blues
0.0598630463786704	0.059848743040724	0.289214468064135	0.271840850988917	0.514513463070575	0.059835632545263	0.0569037210398494	-	0.060138143632205	the blues
0.0598532584191271	0.059833014862683	0.332209012	0.313951792	0.628436498	0.0598364857445667	0.0475589445727293	-	0.0602136016055151	the blues

3.1.4. Random Forest Classification and Evaluation

A. Classification with all data without normalization

In this research, the first test is to do classificationi Random Forest using all the data or all the features in the dataset which results in an accuracy of 66.3%. In testing using 10 music genres, namely blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. This test also uses all features, namely C-STFT, RMSE, SR-OFF, SB, SC, ZCR, MFCC1 to MFCC20. The accuracy value is obtained by using confusion matrix as accuracy calculation. The results of calculations with confusion matrix can be seen in table 4 below.

Table 4 Random Forest Confusion Matrix on all data.

Predicted	True value									
Value	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Raggae	Rock
Blues	70	0	5	1	2	1	5	1	3	8
Classical	0	93	4	1	0	7	0	0	1	0
Country	7	3	57	5	2	8	0	4	10	13
Disco	2	0	3	56	9	3	2	5	5	17
Hiphop	1	0	5	13	55	3	3	3	11	2
Jazz	5	2	12	0	1	73	0	3	2	8
Metal	7	0	0	3	9	1	84	0	1	7
Pop	0	0	3	9	9	3	0	79	6	1
Reggae	2	2	4	5	10	1	1	5	58	6
Rock	6	0	7	7	3	0	5	0	3	38

From the results of confusion matrix for all features that use 10 genres. For blues there are 70 correct numbers of data, classica l 93, country 57, disco 56, hip-hop 55, jazz 73, metal 84, pop 79, reggae 58, while rock only 38 correct data classified by the system. The calculation of the results can be seen in table 5 below.

Table 5 Accuracy Calculation Random Forest with all the data.

Accuracy Calculation	
Formula	$\frac{\text{Jumlah data yang benar}}{\text{Jumlah seluruh data}} x100$
Result	$\frac{70+93+57+56+55+73+84+79+58+38}{1000} x100$ $= 66.3\%$

B. Classification of all data with normalized data.

This second test uses all data and the data has been carried out in the normalization stage using the MIN MAX method. The results of the confusion matrix calculation using all normalized data can be seen in table 6 below.

Table 6 Random Forest Conufusion Matrix all data are normalized

Predicted Value	True value									
	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Raggae	Rock
Blues	71	0	6	1	0	4	6	0	2	8
Classical	0	94	3	1	0	7	0	0	1	0
Country	8	2	56	3	4	7	0	6	9	14
Disco	2	1	2	52	9	1	1	5	6	17
Hiphop	1	0	3	16	56	2	4	3	13	4
Jazz	4	1	13	0	0	70	0	3	3	8
Metal	8	0	0	1	8	1	86	0	1	7
Pop	0	0	3	9	8	4	0	76	4	4
Reggae	1	1	5	4	10	3	0	5	58	6
Rock	5	1	9	13	5	1	3	2	3	32

From the results of the confusion matrix for all data, the results for the blues are 71 correct amounts of data, classical 94, country 56, disco 52, hiphop 56, jazz 70, metal 86, pop 76, reggae 58, while rock is only 32 data classified with correct. And the calculations can be seen in table 7 below.

Table 7 Accuracy calculation with all data are normalized

Accuracy Calculation	
Formula	$\frac{\text{Jumlah data yang benar}}{\text{Jumlah seluruh data}} x100$
Result	$\frac{71+94+56+52+56+70+86+76+58+32}{1000} x100$ $= 65{,}1\%$

From the above calculations, an accuracy of 65.1% was obtained using all data that had been normalized. The graph of the comparison of the accuracy of all data without normalization and using all data is normalized can be seen in Figure 1 below.

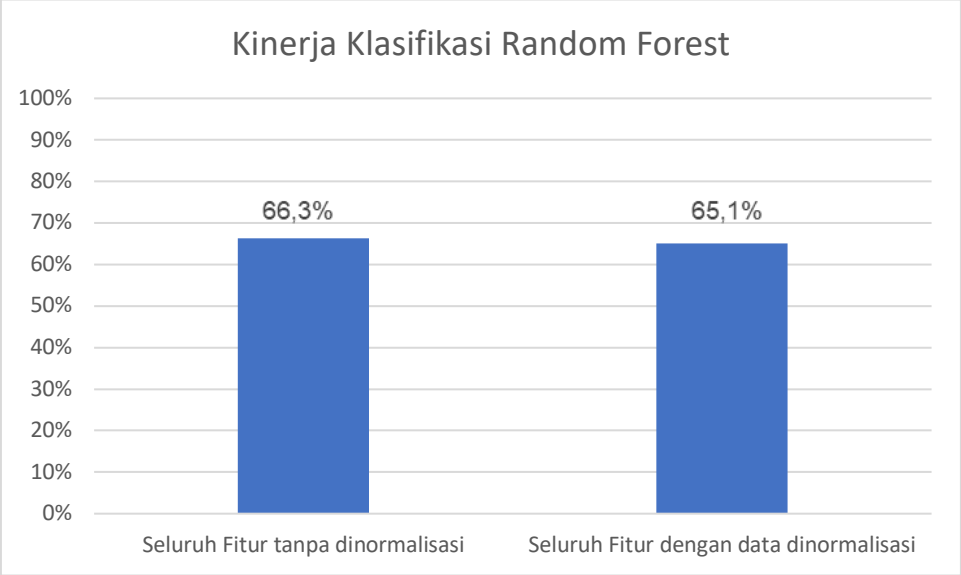


Figure 1. graphical comparison of accuracy results

3.2. DISCUSSION

In this study, the data used were labeled and extracted data with 26 features and 10 labels that had been done in previous studies. The amount of data in the GITZAN database is 1000 data genres of music that already have labels.

The label used is genre blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each label has 100 data, so the total number of data 1000. In this study, two tests were carried out, namely classifying all data without being formalized, and using all normalized data.

The features used in the test are C-STFT, RMSE, SR-OFF, SB, SC, ZCR, MFCC1 to MFCC20, while the things that must be considered in selecting a node or root are the gain value and the gain value is obtained by calculating the entrophy value first.

The classification is carried out using the random forest method and confusion matrix as a method for predicting it, the next process is to normalize the data, the normalization itself is done using the Min-Max method.

Furthermore, dividing the data into training data and test data, sharing training data and test data is done using 10 cross validations. And obtained by dividing the training data amounted to 900 and test data totaling 100 which was done using all data. Next is the classification process using the random forest method and calculating accuracy.

Evaluation or prediction is the last stage in this research, after doing the above steps, this evaluation stage using confusion matrix, obtained an accuracy of 66.3% using all data without normalization, and 65.1% with all data normalized.

4. CLOSING

4.1. Conclusion

The conclusions in this study are:

1. By using all data without normalization, the results obtained an accuracy of 66.3%.
2. And by using all the data, and the data that is done normally, the result is an accuracy of 65.1%.

From the comparison of these tests, the accuracy does not increase with normalization of the data. Even normalization can reduce the level of classification accuracy.

4.2. Suggestion

The suggestions for this research are to improve accuracy, can use other classification methods or can do the feature selection stage on the dataset.

REFERENCES

[1] Ali, S. & Smith, M. 2006. Improved Support Vector Machine Generalization Using Normalized Input Space. School of Engineering and Information Technology, Deakin Uneversity, Australia.

[2] Ardiansyah & Meilina, P. 2017. Music Genre Classification Using Support Vector Machine Method. Muhammadiyah University, Jakarta.

[3] Ashuman, G., Sheezan, M., Masood, S. & Saleem, A. 2016. Genre Classification of Songs Using Neural Network. Dapartment of Computer Engg, New Delhi.

[4] Han, J., Kamber, M. & Pei, Jian. 2012. Data Mining Concepts and Techniques. Elsevier Inc, Vol. 1, 978-381479.

[5] Wei, Du., Lin, H., Sun, J, YB & Yang, H. 2016. A New Method for Music Genre Classification. Shenyang Institute of Computing Technology, Shenyang, China.