JACKA
Journal of Advanced Computer Knowledge and Algorithms
Vol. 1, No. 2, April, 2024, pp. 42-46
Journal homepage: http://ojs.unimal.ac.id/jacka

K-NN with Purity Algorithm to Enhance the Classification of the Air Quality Dataset

Sujacka Retno[1*], Novia Hasdyna[2], Balqis Yafis[3]

[1] Universitas Malikussaleh, Indonesia
[2] Universitas Islam Kebangsaan Indonesia, Indonesia
[3] National Yang Ming Chiao Tung University, Taiwan

[*]Corresponding Author Email: sujacka@unimal.ac.id

Received: 30 March 2024
Revised: 31 March 2024
Accepted: 31 March 2024
Available online: 1 April 2024

Kata Kunci:
K-NN, Purity, Reduksi, Atribut

ABSTRAK

Banyaknya atribut pada dataset yang berukuran besar dapat menyebabkan penurunan tingkat akurasi pengklasifikasian. Reduksi atribut dapat menjadi solusi untuk meningkatkan performa pengklasifikasian khususnya pada algoritma K-NN. Penelitian ini membahas tentang hasil klasifikasi dari K-NN dengan reduksi atribut menggunakan Purity. Berdasarkan hasil pengujian yang telah dilakukan pada Air Quality Dataset, tingkat akurasi yang diperoleh setelah pereduksian atribut adalah sebesar 70,71%, sedangkan tingkat akurasi yang diperoleh sebelum reduksi atribut adalah sebesar 56,44%, peningkatan akurasi yang diperoleh dari pengujian dataset ini adalah sebesar 14,27%. Metode Purity yang diusulkan untuk reduksi atribut dapat meningkatkan tingkat akurasi proses klasifikasi K-NN.

Keywords:
K-NN, Purity, Reduction, Attributes

ABSTRACT

The large number of attributes in a large dataset can cause a decrease in the level of classification accuracy. Attribute reduction can be a solution to improve classification performance, especially in the K-NN algorithm. This research discusses the classification results of K-NN with attribute reduction using Purity. Based on the results of testing carried out on the Air Quality Dataset, the level of accuracy obtained after attribute reduction was 70.71%, while the level of accuracy obtained before attribute reduction was 56.44%, the increase in accuracy obtained from testing this dataset was equal to 14.27%. The proposed Purity method for attribute reduction can increase the accuracy level of the K-NN classification process.

1. INTRODUCTION

Klasifikasi bertujuan untuk menentukan sebuah kelas dari suatu objek dari sejumlah data yang tiap kelasnya tidak diketahui. Klasifikasi merupakan sebuah pola/langkah yang dilakukan dengan skema tersusun. Adapun beberapa metode yang sering digunakan untuk melakukan proses klasifikasi antara lain Decision Tree, K-Nearest Neighbor, Naïve Bayes, Neural Network, dan Support Vector Machine (Sahu et al., 2015).

Algoritma-algoritma klasifikasi sering memiliki masalah terhadap pengklasifikasian data dengan dimensi (atribut) yang tinggi, yaitu menurunnya tingkat akurasi pengklasifikasiannya. Menurut Prasetyo, (2014), salah satu solusi untuk meningkatkan akurasi dan performansi suatu algoritma klasifikasi adalah dengan mereduksi dimensi (atribut). Merujuk pada penelitian yang dilakukan oleh berbagai peneliti yang meneliti tentang reduksi atribut, seperti penelitian yang dilakukan oleh Anisah et al., (2016) yang menggunakan F-Score untuk mereduksi atribut dalam mengklasifikasi kanker payudara, Munzir et al., (2018) menggunakan PCA+GA dalam mereduksi atribut untuk mengklasifikasi microarray. Reduksi atribut mampu membuang fitur-fitur yang tidak relevan, meminimalisasi noise, serta mengurangi curse of dimensionality. Reduksi atribut juga dapat meminimalisasi jumlah waktu dan memori yang dibutuhkan dalam proses pengklasifikasian sebuah algoritma.

Penelitian tentang reduksi atribut telah banyak dikembangkan oleh berbagai peneliti, diantaranya pada penelitian yang diteliti oleh Nasution, (2019) yang menggunakan pendekatan Principal Component Analysis (PCA) sebagai metode seleksi fitur pada dataset yang berasal dari UCI Machine Learning Repository dengan pendekatan yang diusulkan memperoleh total varians yang diperoleh dari 9 variabel skrining sebesar 99%. Penelitian yang dilakukan oleh Wahyuni, (2016) menggunakan pendekatan F-Score untuk pengklasifikasian kanker payudara. Prasetyo, (2014) menggunakan metode DRC untuk mereduksi atribut pada pengklasifikasian menggunakan SVM. Adapun penelitian yang diteliti oleh Retno et al., (2019) menerapkan metode Purity untuk menentukan centroid awal yang akan digunakan untuk proses clustering menggunakan algoritma K-Means. Adapun hasil evaluasi DBI (Davies Bouldin-Index) yang

diperoleh dari penelitian tersebut adalah 1.0357 atau 2.58% lebih baik terhadap K-Means secara konvensional.

Purity digunakan untuk mengukur kemurnian dari keseluruhan atribut yang ada pada setiap data dan mampu menunjukkan atribut-atribut yang kurang relevan sehingga atribut tersebut dapat direduksi dan kemudian akan dianalisis performansinya terhadap dataset yang telah direduksi tersebut. Berdasarkan hal tersebut, maka penelitian ini akan mencoba menganalisa metode reduksi atribut dengan menerapkan metode Purity dalam menggunakan algoritma klasifikasi K-Nearest Neighbor.

Penelitian ini diajukan dengan tujuan peningkatan akurasi metode KNN melalui atribut yang direduksi yang dilakukan dengan metode Purity, hal ini diharapkan mampu mengatasi kelemahan pada algoritma KNN dan meningkatkan akurasi yang dihasilkan dalam proses pengklasifikasian yang dilakukan pada berbagai dataset. Alat ukur performansi yang digunakan dalam penelitian ini adalah *Confusion Matrix*.

2. RESEARCH METHODS

Dataset lain yang digunakan sebagai pembanding pada klasifikasi ini adalah Air Quality Dataset yang berjumlah 9357 data dengan 13 atribut. Air Quality Dataset ini diperoleh dari UCI Machine Learning Repository.

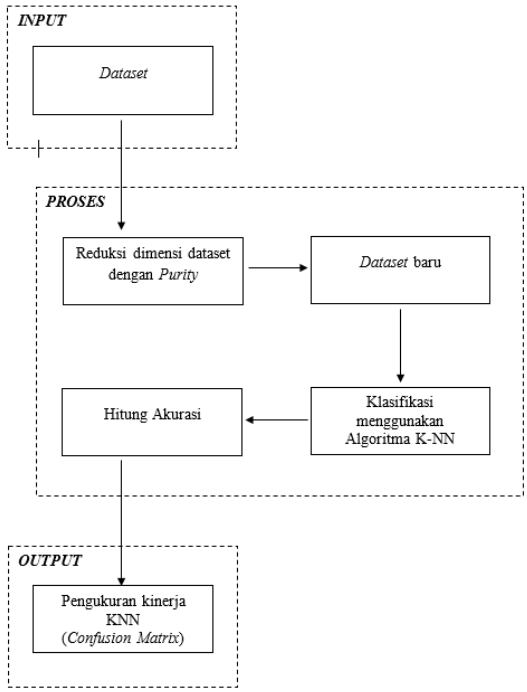
2.1 Alur Penelitian

Adapun alur penelitian dalam penelitian ini dibuat agar langkah-langkah yang diambil penulis dalam perancangan ini tidak melenceng dari pokok pembahasan dan lebih mudah dipahami, maka urutan langkah-langkah akan dibuat secara sistematis sehingga dapat dijadikan pedoman yang jelas dan mudah untuk meyelesaikan permasalahan yang ada. Urutan langkah-langkah yang akan dibuat pada penelitian ini dapat dilihat sebagai berikut.

Gambar 1. Alur penelitian

Berdasarkan gambar 1, langkah pertama yang akan dilakukan adalah pemilihan dataset, didalam penelitian ini, penulis menggunakan Air Quality Dataset dari UCI Machine Learning. Langkah selanjutnya adalah melakukan proses pengklasifikasian terhadap metode Purity+KNN yang kemudian akan dibandingkan dengan KNN secara konvensional. Langkah terakhir adalah melakukan proses analisa hasil dengan menggunakan Confusion Matix dari setiap metode dan dilakukan analisa pengujiannya.

Adapun framework dari alur penelitian yang dilakukan dalam penelitian ini mereduksi atribut dengan menggunakan Purity untuk peningkatan performansi algoritma klasifikasi K-NN adalah sebagai berikut:



Gambar 2. Framework penelitian

Pada gambar 2, langkah yang dilakukan pada proses input adalah dengan menginput dataset yang akan diklasifikasi nantinya, adapun dataset yang digunakan dalam penelitian ini adalah Air Quality Dataset yang diperoleh dari UCI Machine Learning. Pada bagian proses, dataset tersebut akan dihitung nilai Purity-nya untuk menemukan atribut yang kurang relevan sehingga akan direduksi atributnya sehingga menghasilkan dataset yang baru. Dataset yang baru ini akan diklasifikasi dengan menggunakan algoritma K-NN dan dihitung hasil akurasi pengklasifikasiannya menggunakan Confusion Matrix.

3. RESULT AND DISCUSSION

Air Quality Dataset berjumlah 9357 data dengan 13 atribut. Data ini akan diukur tingkat akurasi klasifikasinya setelah direduksi atribut menggunakan Purity K-NN. Pengujian dataset ini dilakukan untuk menguji model baru dari pendekatan K-NN dengan Purity dalam mengklasifikasi dataset yang berjumlah besar. Air Quality Dataset diperoleh dari UCI Machine Learning. Adapun atribut Air Quality Dataset dapat dijabarkan dibawah ini:

1. CO(GT) : Konsentrasi rata-rata CO dalam waktu per jam satuan mg/m^3
2. PT08.S1(CO) : Sensor timah oksida rata-rata dalam waktu per jam
3. NMHC(GT) : Konsentrasi rata-rata Non Metanic HydroCarbons satuan mikrog/m^3
4. C6H6(GT) : Konsentrasi rata-rata Benzene per jam satuan mikrog/m^3
5. PT08.S2(NMHC) : Sensor Titania rata-rata dalam waktu per jam
6. NOx(GT) : Konsentrasi rata-rata NO2 per jam satuan mikrog/m^3
7. PT08.S3(NOx) : Konsentrasi rata-rata NOx per jam satuan ppb
8. NO2(GT) : Sensor NO2 rata-rata per jam
9. PT08.S4(NO2) : Sensor Tungsten Oksida rata-rata per jam

10. PT08.S5(O3) : Sensor Indium Oksida rata-rata per jam
11. T : Suhu satuan Celcius
12. RH : Kelembaban Relatif dalam %
13. AH : Kelembaban Mutlak

Seluruh atribut yang ada pada *Air Quality Dataset* yang berjumlah 13 atribut akan dihitung nilai *Purity*-nya dengan menggunakan formula :

$$Purity(j) = \frac{1}{N_j} max(n_{ij})$$

Adapun perhitungan untuk masing-masing atribut pada Air Quality Dataset adalah sebagai berikut :

1. Nilai *Purity* untuk Atribut CO(GT)

$$Purity\ (\text{CO(GT)}) = \frac{1}{(16520,2)}(11,9) = 0.0007$$

2. Nilai *Purity* untuk Atribut PT08.S1(CO)

$$Purity\ (\text{PT08.S1}) = \frac{1}{(9887473,3)}(2039,8) = 0.00020$$

3. Nilai Purity untuk Atribut NMHC(GT)

$$Purity\ (\text{NMHC}) = \frac{1}{(199994)}(1189) = 0.00594$$

4. Nilai Purity untuk Atribut C6H6(GT)

$$Purity\ (\text{C6H6}) = \frac{1}{(90656,2)}(63,7) = 0.00070$$

5. Nilai Purity untuk Atribut PT08.S2(NMHC)

$$Purity\ (\text{PT08.S2}) = \frac{1}{(8442811,6)}(2214) = 0.00026$$

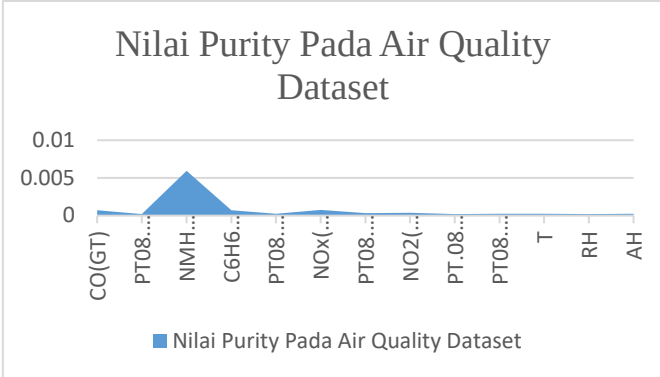
Proses perhitungan tersebut dilanjutkan hingga attibut ke 13. Adapun untuk hasil perhitungan nilai Purity secara keseluruhan dapat dilihat pada tabel 1 berikut ini:

Tabel 1. Nilai Purity pada Air Quality Dataset

No.	Nama Atribut	Nilai Purity
1	CO(GT)	0,00072
2	PT08.S1(CO)	0,00020
3	NMHC(GT)	0,00594
4	C6H6(GT)	0,00070
5	PT08.S2(NMHC)	0,00026
6	NOx(GT)	0,00077
7	PT08.S3(NOx)	0,00035
8	NO2(GT)	0,00039
9	PT.08.S4(NO2)	0,00021
10	PT08.S5(O3)	0,00027
11	T	0,00027
12	RH	0,00020
13	AH	0,00024

Berdasarkan tabel diatas terlihat bahwa atribut NMHC(GT) memiliki nilai Purity terbesar sebesar 0,00594. Atribut NOx(GT) memiliki nilai Purity terbesar kedua dengan nilai sebesar 0,00077, sedangkan nilai Purity terbesar ketiga adalah pada atribut CO(GT) dengan nilai sebesar 0,00072. Adapun data dari ketiga nilai tersebut akan direduksi atributnya pada dataset dataset tersebut dan selanjutnya akan dilakukan proses klasifikasi dengan algoritma K-Nearest Neighbor dan akan dianalisis pengaruhnya terhadap kinerja algoritma K-Nearest Neighbor secara konvensional.

Berdasarkan tabel 1, nilai Purity pada Air Quality Dataset disajikan dalam bentuk grafik pada gambar 3 berikut:



Gambar 3. Framework penelitian

```
dataset=pd.read_csv(r'D:\File\Riset\code\Python\airqualit
y_reducted.csv')
df=dataset.head(1080)
print(dataset.shape)
print(df)
x=dataset.drop('CLASS', 1)
y=dataset['CLASS']
x_train, x_test, y_train,
y_test=train_test_split(x,y,test_size=0.20,random_state=1
00)
print(x_train.shape)
print(x_test.shape)
knn=KNeighborsClassifier(n_neighbors=5,
metric='euclidean')
knn.fit(x_train, y_train)
y_pred = knn.predict(x_test)
print(y_pred)#class prediksi
print(y_test)#class actual
warnings.filterwarnings('ignore')
print('HasilAkurasi KNN adalah',
metrics.accuracy_score(y_test,y_pred), '%')
print('==========Confusion
Matrix==========')
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test,y_pred))
```

Gambar 4. Prosedur Klasifikasi Purity + KNN di Python

Berdasarkan gambar 4 yang merupakan prosedur pengklasifikasian KNN terhadap data yang telah direduksi atribut dengan menggunakan metode Purity yang dilakukan untuk mengklasifikasi data berdasarkan kelas-kelas yang terdapat pada Air Quality Dataset.

Berikut ini adalah hasil rincian dari hasil prediksi model klasifikasi Purity + KNN menggunakan data testing Air Quality Dataset yang tertera pada tabel 2.

Tabel 2. Hasil Klasifikasi pada Air Quality Dataset

No	PT08.S1 (CO)	C6H6 (GT)	PT08.S2 (NMHC)	PT08.S3 (NOx)	NO2 (GT)	PT08.S4 (NO2)	PT08.S5 (O3)	T	RH	AH	KELAS
1	1360	11,9	1046	1056	113	1692	1268	13,6	48,9	0,7578	IDEAL
2	1292	9,4	955	1174	92	1559	972	13,3	47,7	0,7255	IDEAL

3	1402	9,0	939	1140	114	1555	1074	11,9	54,0	0,7502	IDEAL
4	1376	9,2	948	1092	122	1584	1203	11,0	60,0	0,7867	IDEAL
5	1272	6,5	836	1205	116	1490	1110	11,2	59,6	0,7888	IDEAL
6	1197	4,7	750	1337	96	1393	949	11,2	59,2	0,7848	IDEAL
7	1185	3,6	690	1462	77	1333	733	11,3	56,8	0,7603	IDEAL
8	1136	3,3	672	1453	76	1333	730	10,7	60,0	0,7702	IDEAL
9	1094	2,3	609	1579	60	1276	620	10,7	59,7	0,7648	IDEAL
10	1010	1,7	561	1705	-200	1235	501	10,3	60,2	0,7517	IDEAL
11	1011	1,3	527	1818	34	1197	445	10,1	60,5	0,7465	IDEAL
12	1066	1,1	512	1918	28	1182	422	11,0	56,2	0,7366	IDEAL
13	1052	1,6	553	1738	48	1221	472	10,5	58,1	0,7353	IDEAL
14	1144	3,2	667	1490	82	1339	730	10,2	59,6	0,7417	IDEAL
15	1333	8,0	900	1136	112	1517	1102	10,8	57,4	0,7408	IDEAL
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9357	1071	11,9	1047	654	168	1129	816	28,5	13,1	0,5028	KERING

Pada tabel 2 diatas adalah hasil pengujian dari dataset Air Quality yang telah diklasifikasi dengan metode Purity + KNN. Adapun untuk nilai Akurasi dan tingkat error dalam pengujian tersebut dihitung sebagai berikut:

1. $Akurasi = \frac{479+132+712}{479+132+712+101+13+89+114+153+78} = \frac{1323}{1871} = 0{,}7071 = 70{,}71\%$
2. $Error = \frac{101+13+89+114+153+78}{479+132+712+101+13+89+114+153+78} = \frac{548}{1871} = 0{,}2928 = 29{,}28\%$

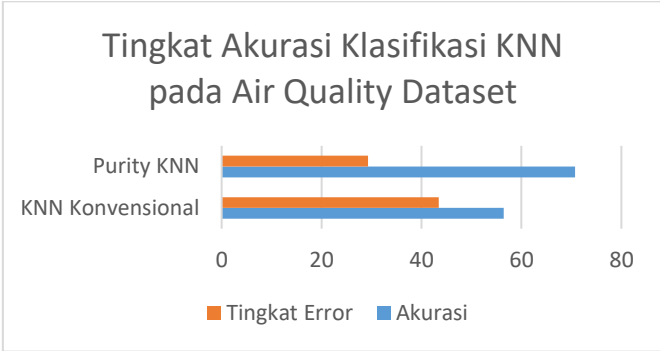
Berikut ini adalah prosedur pengklasifikasian KNN konvensional terhadap Air Quality Dataset:

```
dataset=pd.read_csv(r'D:\File\Riset\code\Python\airqualit
y.csv')
df=dataset.head(1080)
print(dataset.shape)
print(df)
x=dataset.drop('CLASS', 1)
y=dataset['CLASS']
x_train, x_test, y_train,
y_test=train_test_split(x,y,test_size=0.20,random_state=1
00)
print(x_train.shape)
print(x_test.shape)
knn=KNeighborsClassifier(n_neighbors=5,
metric='euclidean')
knn.fit(x_train, y_train)
y_pred = knn.predict(x_test)
print(y_pred)#class prediksi
print(y_test)#class actual
warnings.filterwarnings('ignore')
print('HasilAkurasi KNN adalah',
metrics.accuracy_score(y_test,y_pred), '%')
print('==========Confusion
Matrix==========')
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test,y_pred))
```

Gambar 5. Prosedur Klasifikasi KNN di Python

Hasil pengujian dari dataset Air Quality yang telah diklasifikasi dengan metode KNN konvensional dihitung sebagai berikut:

1. $Akurasi = \frac{513+166+377}{513+166+377+253+107+121+84+146+104} = \frac{1056}{1871} = 0{,}5644 = 56{,}44\%$
2. $Error = \frac{253+107+121+84+146+104}{513+166+377+253+107+121+84+146+104} = \frac{815}{1871} = 0{,}4355 = 43{,}55\%$



Gambar 6. Grafik Persentase akurasi Air Quality Dataset

Pada gambar 6 memperlihatkan perbandingan tingkat akurasi dan Error-rate klasifikasi KNN Konvensional terhadap Purity + KNN pada Air Quality dataset dimana nilai akurasi dari hasil pengujian Air Quality Dataset mencapai selisih yang signifikan yaitu sebesar 14,27%. Adapun Error-rate pada Air Quality Dataset, selisih Error-rate yang terdapat dari dua buah pengujian yang berbeda tersebut adalah sebesar 14,27%.

4. CONCLUSION

Penelitian yang dilakukan yaitu mereduksi atribut dengan menggunakan Purity untuk proses klasifikasi menggunakan K-NN menghasilkan tingkat akurasi yang lebih tinggi dibandingkan dengan pengklasifikasian K-NN secara konvensional. Pada pengujian dengan Air Quality dengan mereduksi 3 attribut yang memiliki nilai Purity terendah, Purity+K-NN berhasil meningkatkan tingkat akurasi pengklasifikasian dengan selisih sebesar 14,27% pada Air Quality Dataset dibandingkan dengan pengklasifikasian K-NN secara konvensional.

Berdasarkan hasil pengujian dari model yang diusulkan tersebut, dapat disimpulkan bahwa tingkat Error-rate yang dihasilkan oleh Purity+KNN lebih rendah dibandingkan dengan KNN secara konvensional, dimana hasil Error-rate yang diperoleh pada Air Quality Dataset dalam pengklasifikasian Purity+KNN adalah sebesar 29,28%, sedangkan secara konvensional menghasilkan Error-rate sebesar 43,55%.

REFERENCES

[1] Anisah, S., Honggowibowo, A. S., & Pujiastuti, A. (2016). Klasifikasi Teks Menggunakan Chi Square Feature Selection Untuk Menentukan Komik Berdasarkan Periode, Materi Dan Fisik dengan Algoritma Naivebayes. Compiler, 5(2), 59–66. https://doi.org/10.28989/compiler.v5i2.171

[2] Arifin, M. (2015). Ig-Knn Untuk Prediksi Customer Churn Telekomunikasi. Simetris : Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer, 6(1), 1. https://doi.org/10.24176/simet.v6i1.230

[3] Bertini, J. R., Zhao, L., Motta, R., & Lopes, A. D. A. (2011). A nonparametric classification method based on

K-associated graphs. Information Sciences, 181(24), 5435–5456. https://doi.org/10.1016/j.ins.2011.07.043
[4] Fawcett, T. (2006). An Introduction to ROC Analysis. Pattern Recognition Letters, 27(8), 861-874. https:/doi.org/10.1016/j.patrec.2005.10.010
[5] Forestier, G., Wemmert, C., & Gançarski, P. (2010). Background knowledge integration in clustering using purity indexes. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6291 LNAI, 28–38. https://doi.org/10.1007/978-3-642-15280-1_6
[6] Gorunescu, F. 2011. Data Mining Concepts, Models and Techniques. Verlah Berlon Heidelberg: Spinger.
[7] Harikumar, S., & Surya, P. V. (2015). K-Medoid Clustering for Heterogeneous DataSets. Procedia Computer Science, 70, 226–237. https://doi.org/10.1016/j.procs.2015.10.077
[8] Manning, C. D., Raghavan, P., Schutze, H., Manning, C. D., Raghavan, P., & Schutze, H. (2012). Flat clustering. Introduction to Information Retrieval, c, 321–345. https://doi.org/10.1017/cbo9780511809071.017
[9] Munzir, A. F. H., Adiwijaya & Aditsania, A. (2018). Analisis Reduksi Dimensi Pada Klasifikasi Microarray Menggunakan Mbp Powell Beale. E-Jurnal Matematika, 7(1), 17. https://doi.org/10.24843/mtk.2018.v07.i01.p179
[10] Nasution, M. Z. (2019). Penerapan Principal Component Analysis (PCA) Dalam Penentuan Faktor Dominan Yang Mempengaruhi Prestasi Belajar Siswa (Studi Kasus : SMK Raksana 2 Medan). Jurnal Teknologi Informasi, 3(1), 41. https://doi.org/10.36294/jurti.v3i1.686
[11] Park, S., & Park, N. W. (2019). Effects of class purity of training data on crop classification using 2D-CNn. 40th Asian Conference on Remote Sensing, ACRS 2019, 1–5.
[12] Park, S., & Park, N. W. (2020). Effects of class purity of training patch on classification performance of crop classification with convolutional neural network. Applied Sciences (Switzerland), 10(11). https://doi.org/10.3390/app10113773
[13] Patil, S. S., & Sonavane, S. P. (2017). Improved classification of large imbalanced data sets using rationalized technique: Updated Class Purity Maximization Over_Sampling Technique (UCPMOT). Journal of Big Data, 4(1), 1–32. https://doi.org/10.1186/s40537-017-0108-1
[14] Prasetyo, E (2014). Reduksi Dimensi Set Data dengan DRC pada Metode Klasifikasi SVM dengan Upaya Penambahan Komponen Ketiga. Prosiding SNATIF. 293–300.
[15] Retno, S., Nababan, E. B., & Efendi, S (2019). Initial Centroid of K-Means Algorithm using Purity to Enhance the Clustering Results. International Journal of Trend in Research and Development (IJTRD), 6(3), 348–351.
[16] Sahu, M., Nagwani, N. K., Verma, S., & Shirke, S. (2015). Performance Evaluation of Different Classifier for Eye State Prediction Using EEG Signal. International Journal of Knowledge Engineering-IACSIT, 1(2), 141–145. https://doi.org/10.7763/ijke.2015.v1.24
[17] Sripada, S. C. (2011). Comparison of Purity and Entropy of K-Means Clustering and Fuzzy C Means Clustering. Indian Journal of Computer Science and Engineering, 2(3), 343–346. http://www.ijcse.com/docs/IJCSE11-02-03-105.pdf
[18] Wahyuni, E. S. (2016). Penerapan Metode Seleksi Fitur Untuk Meningkatkan Hasil Diagnosis Kanker Payudara. Simetris : Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer, 7(1), 283. https://doi.org/10.24176/simet.v7i1.516.