Analisis *Matthew Correlation Coefficient* pada *K-Nearest Neighbor* dalam Klasifikasi Ikan Hias

Novia Hasdyna[1], Rozzi Kesuma dinata[2]
[1]Program Studi Teknik Informatika, Universitas Islam Kebangsaan Indonesia
[2,]Program Studi Teknik Informatika, Universitas Malikussaleh
[1]noviahasdyna@gmail.com, [2]rozzi@unimal.ac.id

ABSTRACT
K-Nearest Neighbor (K-NN) is a machine learning algorithm that functions to classify data. This study aims to measure the performance of K-NN algorithm by using Matthew Correlation Coefficient (MCC). The data that used in this study are the ornamental fish which consisting of 3 classes named Premium, Medium, and Low. The analysis results of the Matthew Correlation Coefficient on K-NN using Euclidean Distance obtained the highest MCC value in Medium class which is 0.786542. The second highest MCC value is in Premium class which is 0.567434. The lowest MCC value is in Low class which is 0.435269.
Overall, the MCC values is statistically which is 0,596415.

Keyword: *Matthew correlation coefficient, machine learning, knn, ikan hias*

1. Introduction

Saat ini perkembangan ruang lingkup analisis data *science* yang sedang berkembang pesat adalah *machine leraning* [1]. *Machine learning* atau pembelajaran mesin merupakan bagian dari artificial intelligence yang terdiri dari *supervised learning, unsupervised learning*, dan *reinforcement learning* [2]. Algoritma supervised learning antara lain adalah *support vector machine, naïve bayes, k-nearest neighbor*, dan *random forest* [3]. *Supervised learning* berfungsi untuk mengklasifikasikan data berdasarkan pola yang sudah ada [4].

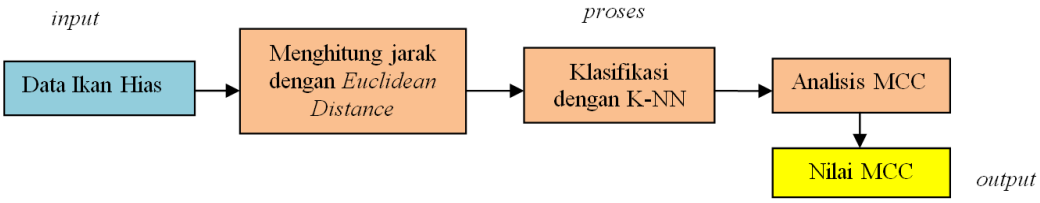
Banyak metode yang dipakai untuk mengukur kinerja algoritma *machine learning* [5]. Pada penelitian ini, untuk mengukur kinerja algoritma *K-Neirest Neighbor* digunakan *matthew correlation coefficient*. *Matthew correlation coefficien*t mengukur performansi klasifikasi data dengan range -1, 0, +1 [6]. Semakin nilai MCC mendekati +1 maka semakin baik kinerja algoritma klasifikasinya. Sebaliknya, jika nilai nya mendekati -1, maka semakin buruk kinerja algoritma klasifikasi. Pengukuran terhadap kinerja suatu sistem klasifikasi merupakan hal yang penting [7]. Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data.

Penelitian ini menggunakan algoritma k-nn untuk mengklasifikasikan data ikan hias. Adapun klasifikasi ikan hias terdiri dari 3 class, yaitu *premium*, *medium* dan *low*. Hasil yang diharapkan pada penelitian ini berupa nilai performance *Matthew Correlation Coefficient* (MCC) pada algoritma knn dalam mengklasifikasikan data ikan hias. Pada proses klasifikasi ikan hias ini akan memberikan informasi hasil pengujian data ikan hias termasuk kedalam *grade premium, medium* atau *grade low* berdasarkan konversi nilai pada *atribut* perawatan, warna, pakan, dan harga.

2. Research Method

2.1. *Framework* Penelitian

Adapun *framework* penelitian ini adalah seperti pada gambar 1.

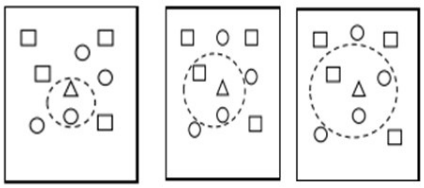


Gambar 1. *Framework* Penelitan

Berdasarkan gambar 1, skema penelitian ini terdiri dari tiga proses yaitu *input*, proses dan *output*. Setelah melakukan inputan data ikan hias, selanjutnya akan dilakukan proses perhitungan jarak dengan menggunakan Euclidean distance. Jarak terkecil akan dilakukan proses pengklasifikasian dengan *k-nearest neighbor.* Hasil klasifikasi berdasarkan data *training* dan data *testing* akan dianalisis *performance* nya dengan menggunakan *matthew correlation coefficient*. Semakin nilai mcc mendekati +1 maka semakin baik kinerja algoritma klasifikasinya. Sebaliknya, jika nilai nya mendekati -1, maka semakin buruk kinerja algoritma klasifikasi.

2.2. *K-Nearest Neighbor*

Salah satu metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumya adalah K-NN [8]. K-NN termasuk dalam golongan *supervised learning*, dimana hasil query instance yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam K-NN [9]. Berikut ilustrasi K-NN seperti pada gambar 2.



Gambar 2. Ilustrasi K-NN dengan k=1, k=2, dan k=3

K- Nearest Neighbor bekerja mencari jarak yang paling dekat antara data yang akan di evaluasi dengan k neighbor (tetangga) yang terdekat di dalam sebuah data traning [10].

Berikut urutan proses kerja algoritma *K- Nearest Neighbor* [11]:

1. Tentukan Parameter k jumlah tetangga paling dekat.
2. Hitung *Euclidean Distance* masing masing objek terhadap sampel data yang ada.

$$d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \qquad (1)$$

Keterangan:

x1 = Sampel data
x2 = Data uji atau data testing
i = Variabel data
d = Jarak
p = Dimensi data

3. Kemudian mengurutkan objek-objek tersebut kedalam kelompok yang mempunyai jarak *Euclid* kecil.
4. Mengumpulkan kategori Y (Klasifikasi *Nearest Neighbor*).

2.3 *Matthew Correlation Coefficient*

Matthew Correlation Coefficient (MCC) merupakan salah satu metode yang digunakan untuk mengukur kinerja algoritma klasifikasi. Perhitungan awal mcc dilakukan dengan menggunakan *confusion matrix*. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya.

Koefisien Korelasi Matthews (MCC) memiliki rentang -1 hingga 1 di mana -1 menunjukkan klasifikasi biner yang sepenuhnya salah sedangkan 1 menunjukkan klasifikasi biner yang sebenarnya.
Berikut adalah formula dari MCC:

$$MCC = \frac{TPx\ TN - FP\ x\ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (2)$$

Keterangan:

FN = *False Negative*
TP = *True Possitive*
TN = *True Negative*
FP = *False Possitive*
MCC = *Matthew Correlation Coefficient*

3. Result and Analysis

3.1 *Data Training*

Adapun dataset ikan hias dapat dilihat pada tabel 1.

Tabel 1. *Data Training*

Nama Ikan	Perawatan	Warna	Pakan	Harga	Grade
Fish A	5	Sangat Menarik	Binatang Kecil	Puluhan Juta	Premium
Fish B	5	Sangat Menarik	Binatang Kecil dan Pelet	Puluhan Juta	Premium
Fish C	5	Sangat Menarik	Pelet dan Tumbuhan	Jutaan	Premium
Fish D	4	Sangat Menarik	Pelet dan Tumbuhan	Jutaan	Premium
Fish E	5	Sangat Menarik	Pelet	Ratusan Ribu	Premium
Fish F	4	Menarik	Binatang Kecil dan Pelet	Puluhan Ribu	Medium
Fish G	3	Sangat Menarik	Pelet dan Tumbuhan	Puluhan Ribu	Medium
Fish H	3	Sangat Menarik	Pelet	Puluhan Ribu	Medium
Fish I	3	Menarik	Pelet	Puluhan Ribu	Medium
Fish J	3	Menarik	Tumbuhan	Puluhan Ribu	Medium
Fish K	2	Menarik	Pelet	Ribuan	Low
Fish L	2	Netral	Pelet	Ribuan	Low
Fish M	2	Kurang Menarik	Pelet	Ribuan	Low
Fish N	2	Sangat Kurang Menarik	Pelet	Ribuan	Low
Fish O	1	Kurang Menarik	Tumbuhan	Ribuan	Low

Pada tabel 1, menampilkan data training yang terdiri dari 15 data, dengan kriteria perawatan, warna, pakan, harga, dan grade ikan hias terdiri dari tiga class yaitu premium, medium dan low. Untuk data atribut dan nilainya dapat dilihat pada tabel 2.

Tabel 2. Atribut dan Nilainya

No	Nama	Jenis atribut dan Nilainya
1	Ikan Hias	Kategori (Nama Ikan)
2	Kemudahan Perawatan	Numerik
3	Warna	Kategori (Sangat Menarik, Menarik , Netral, Kurang Menarik, Sangat Kurang Menarik)
4	Pakan	Kategori (Binatang Kecil dan pelet, Pelet dan Tumbuhan, Binatang Kecil, Tumbuhan, Pelet)
5	Harga	Kategori (Puluhan Juta, Jutaan, Ratusan ribu, Puluhan ribu, Ribuan)

3.2. *Data Testing*

Data testing pada proses klasifikasi ikan hias dengan KNN ditampilkan pada tabel 3.

Tabel 3. Data Testing

Nama Ikan	Perawatan	Warna	Pakan	Harga	Grade
Blue Tang	5	4	2	3	?
Pare Tawar	4	2	5	3	?
Clown Fish	4	4	2	2	?
Gurame	3	4	1	1	?

Pada tabel 3, menampilkan data testing yang terdiri dari 4 data, yaitu ikan blue tang, pare tawar, clown fish, dan gurame. Data testing tersebut akan dilakukan proses klasifikasi dengan KNN. Selanjutnya pada tabel 4 menampilkan data konversi warna.

Tabel 4. Konversi Warna

Kriteria Warna	Nilai
Sangat Menarik	5
Menarik	4
Netral	3
Kurang Menarik	2
Sangat Kurang Menarik	1

Data konversi warna seperti pada tabel 4 diatas terdiri dari 5 kriteria warna, yaitu sangat menarik, menarik, netral, kurang menarik dan sangat kurang menarik. Adapan data konversi pakan dapat dilihat pada tabel 5.

Tabel 5. Konversi Pakan

Kriteria Pakan	Nilai
Binatang Kecil	5
Binatang Kecil dan Pelet	4
Pelet dan Tumbuhan	3
Pelet	2
Tumbuhan	1

Untuk data konversi harga dengan kriteria puluhan juta, jutaan, ratusan ribu, puluhan ribu, dan ribuan ditampilkan pada tabel 6.

Tabel 6. Konversi Harga

Kriteria Harga	Nilai
Puluhan Juta	5
Jutaan	4
Ratusan Ribu	3
Puluhan Ribu	2
Ribuan	1

Adapun konversi data training dapat dilihat pada tabel 7.

Tabel 7. Konversi Data Training

Nama Ikan	Perawatan	Warna	Pakan	Harga	Grade
Fish A	5	5	5	5	Premium
Fish B	5	5	4	5	Premium
Fish C	5	5	3	4	Premium
Fish D	4	5	3	4	Premium
Fish E	5	5	2	3	Premium
Fish F	4	4	4	2	Medium
Fish G	3	5	3	2	Medium
Fish H	3	5	2	2	Medium
Fish I	3	4	2	2	Medium
Fish J	3	4	1	2	Medium
Fish K	2	4	2	1	Low
Fish L	2	3	2	1	Low
Fish M	2	2	2	1	Low
Fish N	2	1	2	1	low
Fish O	1	1	1	1	Low

3.3 Hasil Klasifikasi K-NN

3.3.1 Menghitung parameter k

Untuk menghitung parameter k jumlah tetangga paling dekat akan ditentukan secara manual. Pada penelitian ini mengambil nilai k=3.

3.3.2 Hasil Perhitungan Jarak dengan *Euclidean Distance*

Untuk menghitung jarak tetangga terdekat dengan *Euclidean Distance* pada data testing pertama dapat dilakukan dengan cara seperti berikut:

$$\text{d(Blue Tang), d(fish A)} = \sqrt{(5-5)^2+(4-5)^2+(2-5)^2+(3-5)^2} = 3{,}74$$

$$\text{d(Blue Tang), d(fish B)} = \sqrt{(5-5)^2+(4-5)^2+(2-4)^2+(3-5)^2} = 3{,}00$$

$$\text{d(Blue Tang), d(fish C)} = \sqrt{(5-5)^2+(4-5)^2+(2-3)^2+(3-4)^2} = 1{,}73$$

$$\text{d(Blue Tang), d(fish D)} = \sqrt{(5-4)^2+(4-5)^2+(2-3)^2+(3-4)^2} = 2{,}00$$

$$\text{d(Blue Tang), d(fish E)} = \sqrt{(5-5)^2+(4-5)^2+(2-2)^2+(3-3)^2} = 1{,}00$$

Adapun hasil perhitungan jarak pada data testing ke-1 dengan menggunakan *Euclidean distance* berdasarkan data testing terhadap data training dapat dilihat pada tabel 8.

Tabel 8. Hasil Perhitungan Jarak Data Test 1 (Blue Tang)

Nama Ikan	Jarak	Grade	k=3
Fish A	3,74	Premium	
Fish B	3,00	Premium	
Fish C	1,73	Premium	K2
Fish D	2,00	Premium	K3
Fish E	1,00	Premium	K1
Fish F	2,45	Medium	
Fish G	2,65	Medium	
Fish H	2,45	Medium	
Fish I	2,24	Medium	
Fish J	2,45	Medium	
Fish K	3,61	Low	
Fish L	3,74	Low	
Fish M	4,12	Low	
Fish N	4,69	Low	
Fish O	5,48	Low	

Berdasarkan tabel 8, diperoleh hasil jarak terkecil senilai 1,00 sehingga ikan Blue Tang termasuk ke dalam klasifikasi ikan grade premium. Adapun hasil perhitungan jarak pada data testing kedua ditampilkan pada tabel 9.

Tabel 9. Hasil Perhitungan Jarak Data Test 2 (Pare Tawar)

Nama Ikan	Jarak	Grade	k=3
Fish A	3,74	Premium	K3
Fish B	3,87	Premium	
Fish C	3,87	Premium	
Fish D	3,74	Premium	K2
Fish E	4,36	Premium	
Fish F	2,45	Medium	K1
Fish G	3,87	Medium	
Fish H	4,47	Medium	
Fish I	3,87	Medium	
Fish J	4,69	Medium	

Fish K	4,58	Low
Fish L	4,24	Low
Fish M	4,12	Low
Fish N	4,24	low
Fish O	5,48	Low

Berdasarkan tabel 9, diperoleh hasil jarak terkecil senilai 2,45 sehingga ikan Pare Tawar termasuk ke dalam klasifikasi ikan grade medium. Adapun hasil perhitungan jarak pada data testing ketiga ditampilkan pada tabel 10.

Tabel 10. Hasil Perhitungan Jarak Data Test 3 (Clown Fish)

Nama Ikan	Jarak	Grade	k=3
Fish A	4,47	Premium	
Fish B	3,87	Premium	
Fish C	2,65	Premium	
Fish D	2,45	Premium	
Fish E	1,73	Premium	
Fish F	2,00	Medium	
Fish G	1,73	Medium	
Fish H	1,41	Medium	K3
Fish I	1,00	Medium	K1
Fish J	1,41	Medium	K2
Fish K	2,24	Low	
Fish L	2,45	Low	
Fish M	3,00	Low	
Fish N	3,74	low	
Fish O	4,47	Low	

Tabel 10 menampilkan hasil perhitungan jarak terkecil senilai 1,00 sehingga ikan Clown Fish termasuk ke dalam klasifikasi ikan grade medium. Adapun hasil perhitungan jarak pada data testing ketiga ditampilkan pada tabel 11.

Tabel 11. Hasil Perhitungan Jarak Data Test 4 (Gurame)

Nama Ikan	Jarak	Grade	k=3
Fish A	6,08	Premium	
Fish B	5,48	Premium	
Fish C	4,24	Premium	
Fish D	3,87	Premium	
Fish E	3,16	Premium	
Fish F	3,32	Medium	
Fish G	2,45	Medium	
Fish H	1,73	Medium	
Fish I	1,41	Medium	K3
Fish J	1,00	Medium	K1
Fish K	1,41	Low	K2
Fish L	1,73	Low	
Fish M	2,45	Low	
Fish N	3,32	low	
Fish O	3,61	Low	

Berdasarkan tabel 10, diperoleh hasil jarak terkecil senilai 1,00 sehingga ikan Gurame termasuk ke dalam klasifikasi ikan grade medium.

3.4 Hasil Analisis *Matthew Correlation Coefficient*

Untuk menganalisis MCC, proses yang pertama adalah dengan menghitung confusion matrix. Nilai dari *confusion matrix* ditampilkan pada tabel 12.

Tabel 12. Proses I Confusion Matrix

		premium	medium	low
predicted class	premium	3	0	0
	medium	0	4	1
	low	0	0	4

Proses kedua confusion matrix dapat dilihat pada tabel 13.

Tabel 13. Proses II Confusion Matrix

		TP	FP	FN	TN
predicted class	premium	3	0	0	2
	medium	4	0	0	1
	low	4	0	0	1

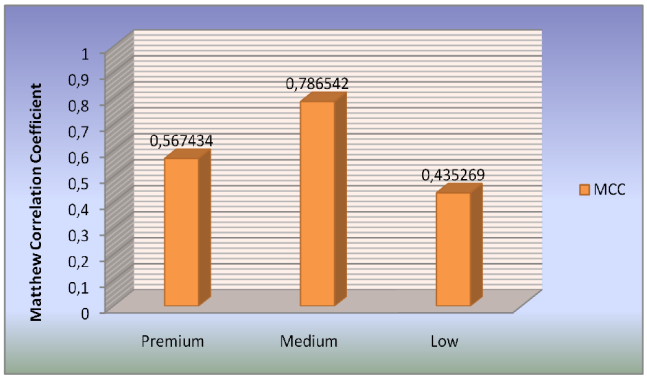
Setelah proses confusion matrix, maka akan dihitung nilai MCC dengan menggunakan persamaan 2. Adapun hasil perhitungan MCC ditampilkan pada tabel 14.

Tabel 14. Nilai Matthew Correlation Coefficient

No	Class	MCC
1	Premium	0,567434
2	Medium	0,786542
3	Low	0,435269
	Rata-rata MCC	0,596415

Berdasarkan tabel 14, dapat dilihat bahwa nilai MCC pada class medium senilai 0,786542 merupakan nilai MCC terbaik. sSedangkan nilai MCC terendah adalah pada class low dengan nilai 0,435269. Semakin mendekati angka +1 maka semakin bagus kinerja algoritma klasifikasi, sedangkan semakin mendekati -1 maka semakin buruk kinerja algoritma klasifikasi.

Adapun nilai MCC dalam bentuk grafik dapat dilihat pada gambar 3.



Gambar 3. Grafik hasil analisis *Matthew Correlation Coefficient* pada Klasifikasi Ikan Hias dengan KNN

4. Conclusion

Penelitian ini menganalisis kinerja algoritma klasifikasi KNN dalam pengklasifikasian data ikan hias. Hasil penelitian ini menunjukkan bahwa nilai *matthew correlation coefficient* pada k-nn dengan menggunakan *Euclidean distance* diperoleh nilai MCC tertinggi pada class medium sebesar 0,786542. Nilai mcc tertinggi kedua pada class premium senilai 0,567434. Nilai mcc terendah adalah pada class low sebesar 0,435269. Dengan demikian, dapat dikatakan nilai korelasi matthew yang paling baik adalah pada class medium karena class medium adalah yang paling mendekati +1. Sebaliknya nilai MCC pada class low adalah yang paling mendekati 0. Adapun rata-rata nilai MCC pada klasifikasi data ikan hias dengan KNN adalah senilai 0,596415. Oleh karena itu, kinerja algoritma klasifikasi KNN adalah baik.

Adapun saran dari penulis adalah agar pembaca dapat melakukan analisis kinerja algoritma klasifikasi dengan metode yang lain, seperti metode *k-fold cross validation*.

References

[1] M. E. Saputra, H. Mawengkang, and E. B. Nababan. "Gini Index With Local Mean Based For Determining K Value In K-Nearest Neighbor Classification." *Journal of Physics: Conference Series.* Vol. 1235. No. 1. IOP Publishing, 2019.

[2] R. K. Dinata, F. Fajriana, K. Khairunnisa,"Penerapan Algoritma Classification And Regression Trees (Cart) Pada Penerimaan Anggota Baru Unit Kegiatan Mahasiswa (UKM) Di Universitas Malikussaleh Berbasis WEB". *TECHSI-Jurnal Teknik Informatika*, 10(2), 74-81, 2018

[3] W. Wahyono, I. N. P. Trisna, S. L. Sariwening, M.Fajar, D. & Wijayanto, "Perbandingan penghitungan jarak pada k-nearest neighbour dalam klasifikasi data tekstual". *Jurnal Teknologi dan Sistem Komputer*, 8(1), 54-58, 2020.

[4] W. Wardhani, A. Khrisna, "Implementasi Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Kajen Pekalongan," *J. Transform.*, vol. 14, no. 1, pp. 30–37, 2016.

[5] L.Farokhah, "Implementasi K-Nearest Neighbor untuk Klasifikasi Bunga Dengan Ekstraksi Fitur Warna RGB". *Jurnal Teknologi Informasi dan Ilmu Komputer*, 7(6), 2020

[6] R. P. Saputri, W. S. Winahju, K. Fithriasari, "Klasifikasi Sentimen Wisatawan Candi Borobudur pada Situs TripAdvisor Menggunakan Support Vector Machine dan K-Nearest Neighbor". *Jurnal Sains dan Seni ITS*, 8(2), 349-356, 2020

[7] S. A. Naufal, A. Adiwijaya, W. Astuti "Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray". *JURIKOM (Jurnal Riset Komputer),* 7(1), 162-168, 2020.

[8] R. K. Dinata, F. Fajriana, & N. Hasdyna, "Klasifikasi Sekolah Menengah Pertama/Sederajat Wilayah Bireuen Menggunakan Algoritma K-Nearest Neighbors Berbasis Web". *Computer Engineering, Science and System Journal*, 5(1), 33-37, 2020.

[9] I. Hasimah, M. A. Mukid, H. Yasin, "Klasifikasi Calon Debitur Kredit Pemilikan Rumah (Kpr) Multiguna Take Over Menggunakan Metode K Nearest Neighbor Dengan Pembobotan Global Gini Diversity Index". *Jurnal Gaussian*, *8*(4), 407-417, 2020

[10] Y. I. Kurniawan, T. I. Barokah, "Klasifikasi Penentuan Pengajuan Kartu Kredit Menggunakan K-Nearest Neighbor". *Jurnal Ilmiah Matrik*, *22*(1), 73-82, 2020

[11] M. A. P. Arsyad. "Klasifikasi Penyakit Breast Cancer Menggunakan Naïve Bayes Dan KNN". *PhD Thesis. Universitas Muhammadiyah Malang*, 2020