PENERAPAN DATA MINING UNTUK KLASIFIKASI SPESIES IKAN DI LINGKUNGAN AKUATIK AIR TAWAR

Erfan Karyadiputra[1)], Andie[1)], Hasanuddin[3)]

[1,2,3]Fakultas Teknologi Informasi, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari Banjarmasin
Email : erfantsy@gmail.com
Email : andina777@gmail.com
Email : hasan.uniska@gmail.com

Abstrak

Kualitas air sangat mempengaruhi produksi budidaya ikan akuatik air tawar karena berdampak pada kelangsungan hidup ikan yang dapat bertahan pada kondisi air tertentu. Pembudidaya ikan terkadang masih kesulitan dalam menentukan spesies ikan yang cocok dan sempurna untuk dibudidayakan di lingkungan akuatik air tawar. Penelitian ini menerapkan perbandingan teknik data mining antara algoritma random forest, naive bayes dan k-nearest neighbor untuk mengklasifikasi spesies ikan berdasarkan karakteristik lingkungan akuatik yang berbeda seperti jenis ikan, tingkat pH air, suhu, dan kekeruhan. Adapun metode evaluasi performance akan diukur menggunakan teknik cross validation, confusion matrix dan paired t-test. Hasil penelitian menunjukkan bahwa algoritma random forest paling akurat dalam mengklasifikasi dataset spesies ikan karena memiliki nilai akurasi tertinggi yaitu sebesar 86,22% dengan nilai kappa statistik yang paling mendekati angka 1 yaitu sebesar 0,796.

***Keywords** : Air Tawar, Data Mining, Kualitas Air, Spesies Ikan*

1. PENDAHULUAN

Lingkungan akuatik (ekosistem perairan) merupakan jenis lingkungan yang sebagian besar lingkungannya didominasi oleh air [1]. Banyak faktor yang mempengaruhi ekosistem akuatik seperti pencahayaan sinar matahari, temperatur dan material garam yang terlarut. Jika pada suatu lingkungan terdapat jumlah kadar garam yang rendah maka termasuk kedalam kategori ekosistem air tawar sedangkan jika memiliki kadar garam tinggi maka termasuk kedalam ekosistem air laut. Ekosistem air tawar merupakan ekosistem perairan tawar di daratan yang dipengaruhi oleh kondisi geografis daratan disekitarnya sehingga pada perairan darat tertentu memungkinkan memiliki ciri-ciri khusus yang berbeda-beda dan khas [2].

Di lingkungan akuatik tertentu masih sulit bagi pembudidaya ikan untuk memilih jenis ikan yang cocok dibudiyakan pada proses akuakultur. Akuakultur sendiri mengacu pada peternakan hewan air atau tanaman air di lingkungan air tawar dan asin terutama untuk makanan [3]. Akuakultur juga sangat dipengaruhi oleh kualitas air, hal ini karena kualitas air yang baik mempengaruhi produksi ikan yang dibudidayakan [4].

Salah satu teknik yang dapat diterapkan untuk mengklasifikasi dataset spesies ikan adalah data mining. Data Mining adalah proses ekstraksi data menjadi informasi dan pola-pola baru yang sebelumnya belum diketahui [5]. Penelitian ini menerapkan perbandingan teknik data mining antara *random forest, naive bayes* dan *k-nearest neighbor* untuk mengklasifikasi spesies ikan berdasarkan karakteristik lingkungan akuatik yang berbeda seperti jenis ikan, tingkat pH air, suhu, dan kekeruhan.

2. METODE PENELITIAN

Penelitian ini merupakan penelitian eksperimen untuk menemukan algoritma terbaik dalam mengklasifikasi spesies ikan di lingkungan akuatik lahan rawa.

Gambar 1. Tahapan Metode Penelitian

a. Pengumpulan Data

Teknik pengumpulan data merupakan teknik atau cara-cara yang dapat digunakan untuk mengumpulkan data. Data yang diperoleh adalah dataset sekunder dari *kaggle datasets*.

Tabel 1. Data Spesies Ikan

Variabel	Kategori	Keterangan
Y	Spesies Ikan	Patin  Nila  Lele  Pentet  Papuyu 

Tabel 2. Data pH, Temperatur & Kekeruhan Air

Variabel	Kategori	Keterangan
X1	pH	pH ukuran keasaman air
X2	Temperatur	Suhu air
X3	Kekeruhan	Kekeruhan air

b. Pengolahan Data Awal

Data yang dikumpulkan kemudian dianalisa dan selanjutnya dilakukan *preprocessing* terhadap data tersebut sehingga diperoleh data yang valid dan dapat diproses ketahapan eksperiman.

c. Eksperimen dan Pengujian Model

Pada tahap ini dilakukan eksperimen dalam mengimplementasikan pengujian algoritma *random forest* (RF), *naive bayes* (NB) *dan k-nearest neighbor* (K-NN) menggunakan *cross validation*.

Random forest (RF) merupakan metode *boostrap aggregating* dengan membangkitkan sejumlah pohon dari data sampel data dimana pembuatan satu pohon pada proses *training* tidak bergantung terhadap pohon sebelumnya kemudian dalam pengambilan keputusannya diambil berdasarkan *voting* terbanyak [6].

$$Gini\ Index(D) = 1 - \sum_{i=1}^{m} P_i^2$$

Keterangan:
Pi : rasio jumlah data label kelas *i* dalam *D*
m : kelas

Algoritma *naive bayes* (NB) merupakan metode pengklasifikasian paling sederhana dengan menggunakan konsep peluang, yang mana diasumsikan bahwa setiap atribut bersifat saling lepas satu sama lain berdasarkan atribut kelas [7].

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

Keterangan:
X : data kelas belum diketahui
H : data *X*
P(H|X) : probabilitas *H* berdasarkan kondisi *X*
P(H) : probabilitas *H*
P(X|H) : probabilitas *X* berdasarkan kondisi *H*
P(X) : probabilitas dari *X*

K-nearest neighbor (K-NN) merupakan metode klasifikasi terhadap objek berdasarkan similaritas dengan label data pembelajaran yang jaraknya paling dekat dengan objek tersebut [8].

$$Similarity(T,S) = \frac{\sum_{i=n}^{n} f(T_i, S_i) * w_i}{w_i}$$

Keterangan:

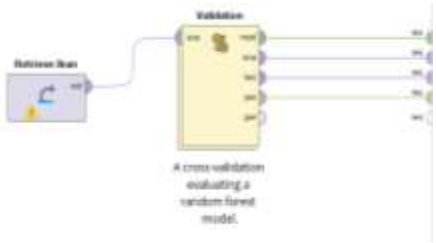
T : kasus baru
S : kasus yang ada dalam penyimpanan
n : jumlah atribut dalam setiap kasus
i : atribut individu antara 1 s/d n
f : fungsi *similarity* atribut *i* antara kasus *T* dan kasus *S*
w : bobot yang diberikan pada atribut ke-*i*

d. Evaluasi dan Validasi hasil

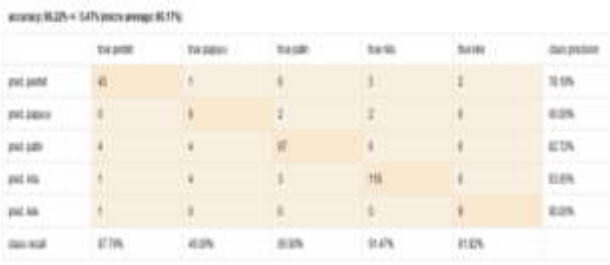
Tahap selanjutnya melakukan evaluasi dan validasi hasil pengujian *cross validation*. pada tahapan ini dilakukan evaluasi terhadap *performance* dari masing-masing algoritma menggunakan c*onfusion matrix* dan *paired t-test.*

3. HASIL DAN PEMBAHASAN

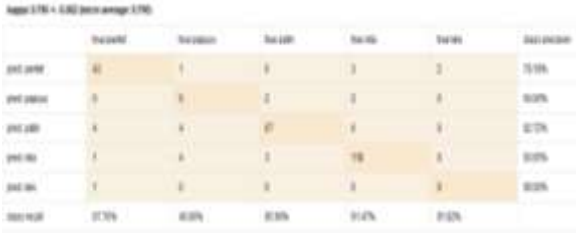
Implementasi hasil pengujian *cross validation* dari algoritma *random forest* disertai evaluasi *confusion matrix* sehingga diketahui tingkat akurasi dan nilai *kappa*.



Gambar 2. Implementasi Pengujian Algoritma



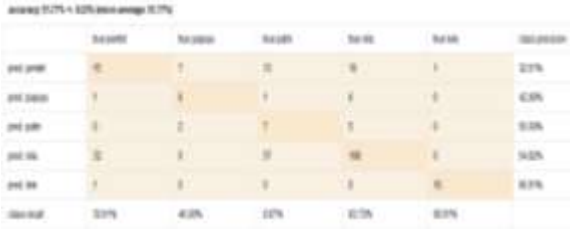
Gambar 3. Akurasi *Random Forest*



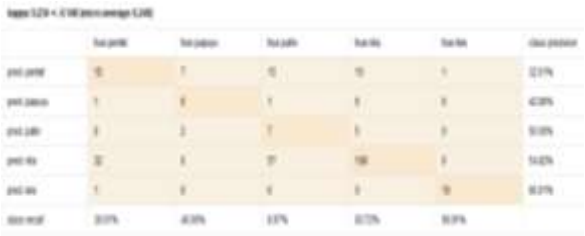
Gambar 4. Nilai *Kappa Random Forest*

Dari hasil pengujian didapatkan akurasi *random forest* sebesar 86,22% dengan nilai kappa 0,796.

Selanjutnya hasil dari pengujian *cross validation* dari algoritma *naive bayes* disertai evaluasi menggunakan *confusion matrix.*



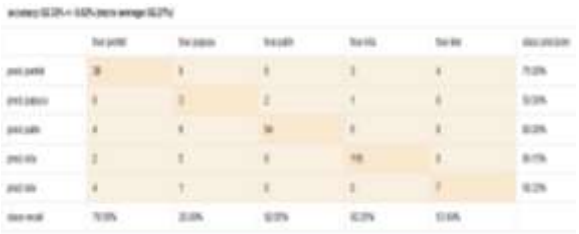
Gambar 5. Akurasi *Naive Bayes*



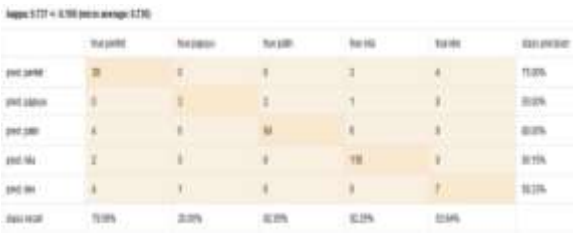
Gambar 6. Nilai *Kappa Naive Bayes*

Dari hasil pengujian didapatkan akurasi *naive bayes* sebesar 51,77% dengan nilai kappa 0,234.

Hasil pengujian *cross validation* dari algoritma *k-nearest neighbor* disertai evaluasi menggunakan *confusion matrix.*



Gambar 7. Akurasi *K-Nearest Neighbor*



Gambar 8. Nilai *Kappa K-Nearest Neighbor*

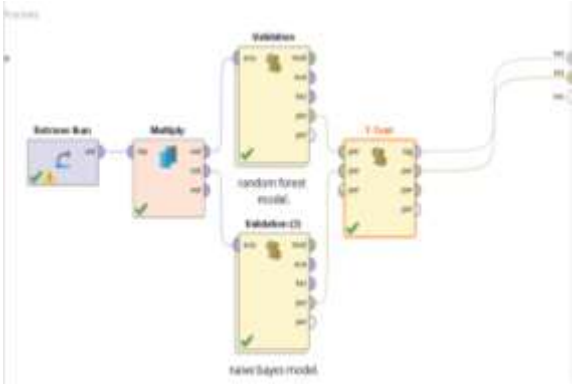
Dari hasil pengujian didapatkan akurasi *k-nearest neighbor* sebesar 82,32% dengan nilai kappa 0,737.

Tabel 3. Komparasi *Performance* Algoritma

	RF	NB	K-NN
Accuracy	86,22%	51,77%	82,32%
Kappa	0,796	0,234	0,737

Berdasarkan hasil dari perbandingan *performance* pada Tabel 3 dapat diketahui bahwa algoritma *random forest* memiliki akurasi tertinggi dan memiliki nilai *kappa* yang paling mendekati angka 1. Nilai kappa sendiri merupakan bentuk normalisasi dari nilai accuracy yang mana semakin mendekati angka 1 maka model dianggap mendekati sempurna [9].

Selanjutnya implementasi pengujian menggunakan *paired t-test. T-test* merupakan metode pengujian hipotesis menggunakan satu objek penelitian dengan dua perlakuan yang berbeda [10].



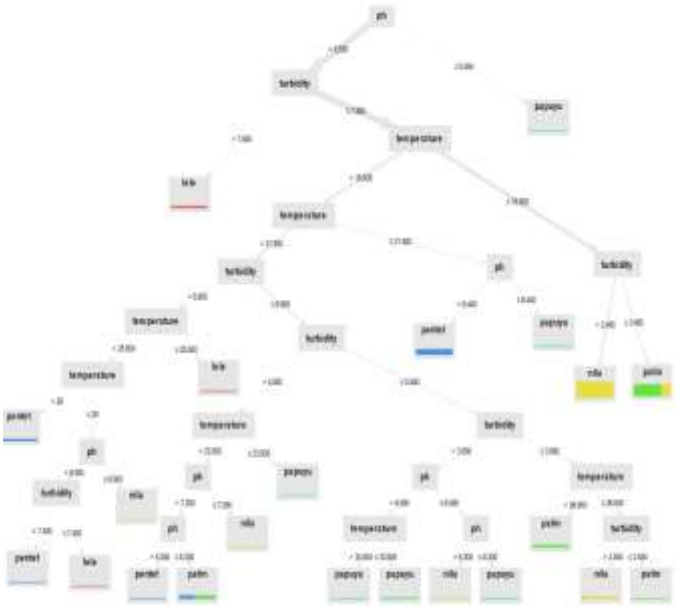
Gambar 9. Pengujian T-Test Algoritma

Hasil pengujian ini dengan membandingkan dua algoritma secara bergantian sehingga akan didapatkan hasil perbandingan seperti Tabel 4 dibawah ini :

Tabel 4. Komparasi Uji Statisti T-Test

	RF	K-NN	NB
RF		0.330	**0.000**
K-NN	**0.000**		**0.000**
NB	0.401	**0.000**	

Berdasarkan hasil dari perbandingan *performance* pada tabel 4 dapat diketahui bahwa algoritma *k-nearest neighbor* lebih dominan terhadap algoritma lainnya meskipun nilai *accuracy* tidak menunjukkan hasil yang tertinggi. Adapun algoritma *random forest* memiliki *accuracy* tertinggi daripada algoritma lainnya meskipun hanya dominan terhadap algoritma *naive bayes*.



Gambar 10. Model *Random Forest*

Adapun hasil komparasi *performance* algoritma dari semua pengujian dapat terlihat pada tabel 5 berikut :

Tabel 5. Hasil Perbandingan Seluruh Pengujian

	RF	NB	K-NN
Accuracy	86,22%	51,77%	82,32%
Kappa	0,796	0,234	0,737
T-Test	tidak dominan	tidak dominan	dominan

4. KESIMPULAN

Dari rangkaian hasil pengujian yang dilakukan, maka dapat disimpulkan bahwa algoritma *random forest* merupakan algoritma yang paling akurat karena memiliki *accuracy* yang tinggi dan nilai *kappa* yang paling mendekati angka 1. Meskipun demikian, *random forest* tidak dominan terhadap algoritma *k-nearest neighbor* namun *random forest* tetap dominan terhadap *naive bayes*. Adapun algortima *k-nearest neighbor* menjadi algoritma yang paling dominan terhadap algoritma lainya meskipun akurasinya lebih rendah daripada *random forest*.

5. REFERENSI

[1] Budianto, "Ekosistem Akuatik," 2015. https://www.pengertianilmu.com/2015/04/pengertian-ekosistem-akuatik.html (accessed Apr. 16, 2022).

[2] S. W. Utomo and S. A. Chalif, "Ekosistem Perairan," *Ekosistem Perairan*, vol. 02, no. 03. pp. 9–17, 2014.

[3] M. M. Islam, M. A. Kashem, and J. Uddin, "Fish survival prediction in an aquatic environment using random forest model," *IAES Int. J. Artif. Intell.*, vol. 10, no. 3, pp. 614–622, 2021, doi: 10.11591/ijai.v10.i3.pp614-622.

[4] L. N. Ayuniar and J. W. Hidayat, "Analisis Kualitas Fisika dan Kimia Air di Kawasan Budidaya Perikanan Kabupaten Majalengka," *J. Enviscience*, vol. 2, no. 2, pp. 68–74, 2018, doi: 10.30736/2ijev.v2iss2.67.

[5] A. H. Matondang, F. Basuki, and R. A. Nugroho, "Pengaruh Lama Perendaman Induk Betina Dalam Ekstrak Purwoceng (Pimpinela Alpina) Terhadap Maskulinisasi Ikan Guppy (Poecilia Reticulata)," *J. Aquac. Manag. Technol.*, vol. 7, no. 1, pp. 10–17, 2018.

[6] N. Widjiyati, "Implementasi Algoritme Random Forest Pada Klasifikasi Dataset Credit Approval," *J. Janitra Inform. dan Sist. Inf.*, vol. 1, no. 1, pp. 1–7, 2021, doi: 10.25008/janitra.v1i1.118.

[7] M. Rifqi, "Aplikasi Data Mining Untuk Diagnosis Penyakit Diabetes Menggunakan Algoritma C4.5 Dan Naïve Bayes Classification," 2016.

[8] F. Febrian, "Algoritma Klasifikasi Data Mining Pada Akseptasi Data Fakultatif," Sekolah Tinggi Manajemen Informatika Dan Komputer Eresha Jakarta, 2011.

[9] R. D. Syah, "Metode Decision Tree Untuk Klasifikasi Hasil Seleksi Kompetensi Dasar Pada Cpns 2019 Di Arsip Nasional Republik Indonesia," *J. Ilm. Inform. Komput.*, vol. 25, no. 2, pp. 107–114, 2020, doi: 10.35760/ik.2020.v25i2.2750.

[10] K. Hastuti, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif," in *Seminars in Neurology*, 2012, vol. 14, no. 1, doi: 10.2307/j.ctv11hppt6.3.