Indonesian Journal of Multidisciplinary on Social and Technology
Homepage: https://journal.ilmudata.co.id/index.php/ijmst
Vol. 2 No. 1 (2024) pp: 72-76
P-ISSN: 2986-6790, e-ISSN: 2986-6782

Perbandingan Kinerja Algoritma Klasifikasi Status Mutu Air

Lidya Ningsih[1], Jajam Haerul Jaman[2], Naufal Ibnu Salam[3], Muhammad Haikal[4]
[1]Rekayasa Perangkat Lunak, Fakultas Informatika, Telkom University
[2]Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa
[3,4]Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor
[1]telulidyaningsih@telkomuniversity.ac.id, [2]jajam.haeruljaman@staff.unsika.ac.id, [3]naufalibnusalam@apps.ipb.ac.id, [4]Muhammadhaikal@gmail.com

Abstrak

Klasifikasi mutu air adalah salah satu teknik dalam melakukan penilaian terhadap air sebagai objek penelitian. Penelitian ini dilakukan dengan tujuan agar dapat memberikan pengetahuan terhadap mutu atau kualitas air, sehingga dapat menjadi solusi terbaik yang dapat dilakukan terhadap air tersebut sebelum dikonsumsi. Penelitian ini menggunakan dua skenario dengan beberapa teknik klasifikasi, diantaranya adalah Algoritme Naive Bayes, KNN, Multiclass classification, MLP, SVM, dan Random Forest. Berdasarkan hasil peneltiian yang dilakukan dengan beberapa algoritma klasifikasi tersebut, didapatkan hasil akurasi terbaik menggunakan Algoritme Random Forest dengan persentase akurasi sebesar 99,5% pada skenario pertama dan 99,7% pada skenario kedua . Sedangkan tingkat akurasi terendah ditemukan pada Algoritme Naive Bayes dengan persentase akurasi sebesar 22,3% pada skenario pertama dan 21,9% pada skenario kedua. Hal ini disebabkan karena dataset mutu air yang diperoleh tidak seimbang atau tidak terdistribusi normal (Gaussian). Selain itu, algoritme Naive Bayes memiliki kinerja baik dalam pekerjaan klasifikasi dengan data teks.

Kata kunci: *K-Nearest Neighbour, Multiclass classification, Multi Layer Perceptron, Naïve Bayes, Random Forest, Status mutu air, Support Vector Machine.*

1. Pendahuluan

Status mutu air merupakan tingkat kondisi mutu air yang menunjukkan kondisi cemar atau kondisi baik pada sumber air dalam waktu tertentu dengan membandingkan dengan baku mutu air yang ditetapkan [1]. Peraturan Pemerintah Republik Indonesia Nomor 82 Tahun 2001 Pasal 1, mutu air adalah kondisi kualitas air yang diukur dan atau diuji berdasarkan parameter-parameter dan metode tertentu berdasarkan peraturan perundang-undangan yang berlaku. Kelas air adalah peringkat kualitas air yang dinilai masih layak untuk dimanfaatkan bagi peruntukan tertentu. Adapun kriteria mutu air adalah tolak ukur mutu air untuk setiap kelas air.

Mutu air merupakan salah satu kriteria paling penting dalam perencanaan sumber daya yang berperan penting dalam menentukan kebijakan dan sumber daya air [2]. klasifikasi dalam proses kualitas mutu air sebelum dikonsumsi atau digunakan untuk berbagai kebutuhan adalah hal yang sangat diperlukan. Hal tersebut dapat mengurangi terjadinya berbagai penyakit yang ditimbulkan apabila konsumsi air dengan mutu buruk atau tercemar seperti hepatitis A, demam tifoid, dan disentri [3].

Penentuan mutu air dapat menggunakan beberapa parameter untuk mendeskripsikan mutu air dalam informasi yang terbatas, diantaranya *Total Dissolved Solid* (TDS), *Dissolved Oxygen* (DO), derajat keasaman air (pH), *Chemical Oxygen Demand* (COD), dan *Biological Oxygen Demant* (BOD). Dalam menentukan mutu air dilakukan pengambilan sampel air yang dihitung status mutu air menggunakan metode *Storet.* Metode ini diterapkan di Indonesia berdasarkan Keputusan Menteri LH Nomor 115 Tahun 2003 tentang Pedoman Penentuan Status Mutu Air [4]. Telah dilakukan beberapa penelitian sebelumnya dalam membandingkan akurasi algoritma dalam klasifikasi mutu air. Penelitian [2] membandingkan algoritma *Support Vector Machine (SVM), Probabilistic Neural Networks* dan *K-Nearest Neighbors (KNN)* dalam klasifikasi kualitas mutu air dengan hasil algoritma *Support Vector Machine (SVM)* unggul dibandingkan dengan kedua algoritma lainnya karena tidak ada data yang *error* pada saat tahap kalibrasi dan memiliki jumlah *error* paling rendah pada saat tahap validasi. Selain itu, pada penelitian [2], algoritma KNN memiliki jumlah *error* terbanyak dan performa paling buruk pada saat tahap kalibrasi maupun validasi.

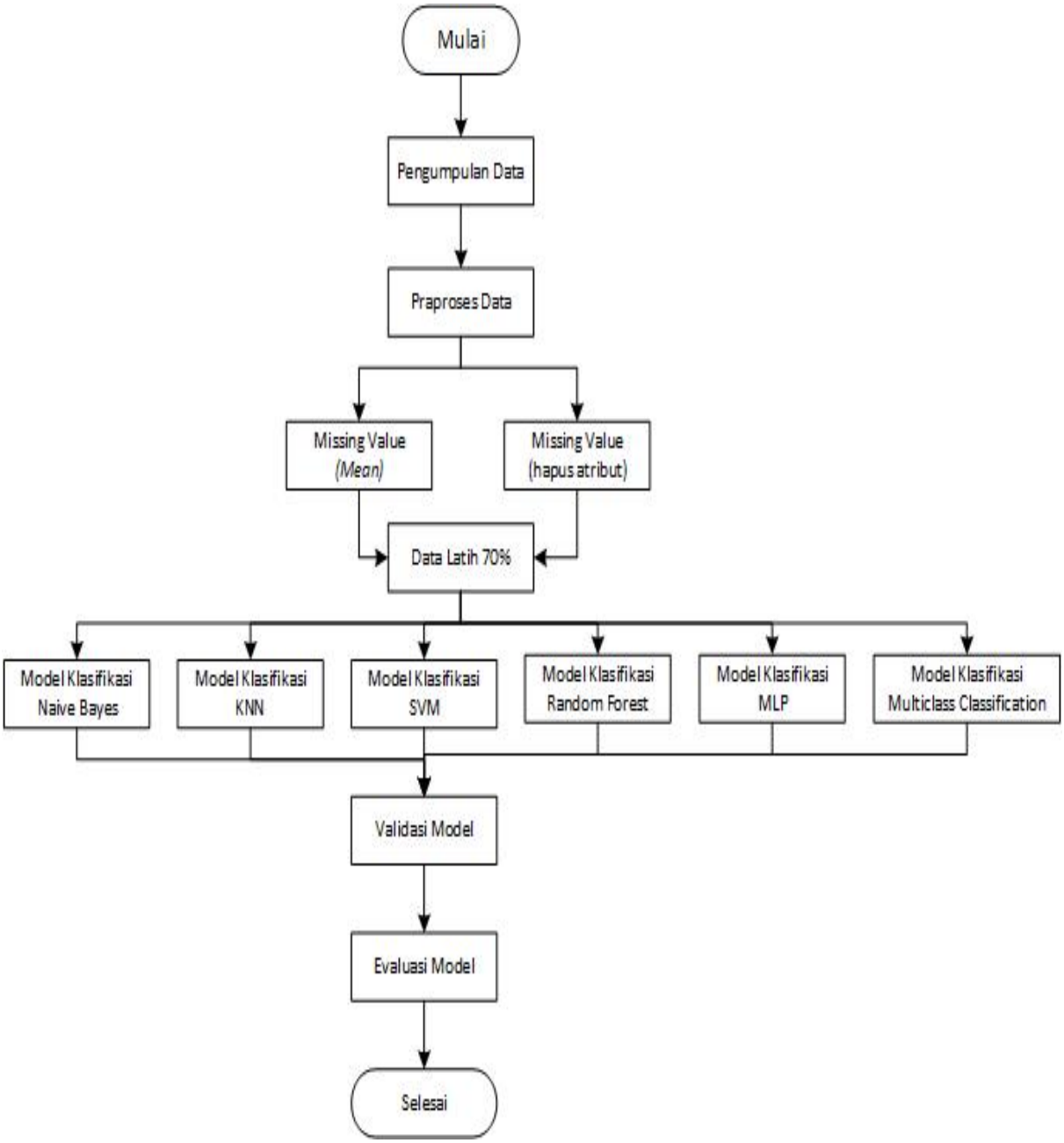
Penelitian [3] melakukan perbandingan algoritma *Support Vector Machine (SVM)* dan *K-Nearest Neighbors (KNN)* dalam klasifikasi kualitas mutu air yang akan dievaluasi dengan menggunakan kurva ROC, akurasi, dan *confusion matrix* pada faktor tingkat salinitas air dalam hal konduktivitas listrik. Hasil dari penelitian [3] menunjukkan bahwa algoritma SVM memberikan hasil klasifikasi lebih

baik dibandingkan dengan algoritma KNN dengan tingkat akurasi klasifikasi 96.6% dan *error* pada klasifikasi sebesar 3.4%.

Penelitian ini berfokus untuk mengetahui klasifikasi status air berdasarkan parameter yang dipantau dengan melakukan perbandingan beberapa algoritma klasifikasi yaitu *Naive Bayes*, *K-Nearest Neighbour* (KNN), *Support Vector Machine* (SVM), *Multiclass Classification*, *Multilayer Perceptron* dan *Random Forest* untuk melihat kinerja lagoritma terhadap data yang dilakukan penangan mengganti *missing value* dengan *mean* ddan penanganan menghapus *missing value.*

2. Metode Penelitian

Tahapan penelitian ini dilakukan untuk mendapatkan dan mengetahui informasi-informasi yang dibutuhkan dalam proses penelitian sesuai pada Gambar 1. Mencari informasi yang diperlukan dalam penelitian dapat dilakukan dengan mencari beberapa sumber yang berkaitan dengan penelitian tersebut seperti jurnal, artikel, buku dan internet untuk mencari informasi terhadap model yang diusulkan, data yang digunakan serta tahapan-tahapan yang akan dilakukan. Berikut merupakan tahapan penelitian yang dilakukan.



Gambar 1 Flowchart Klasifikasi Status Mutu Air

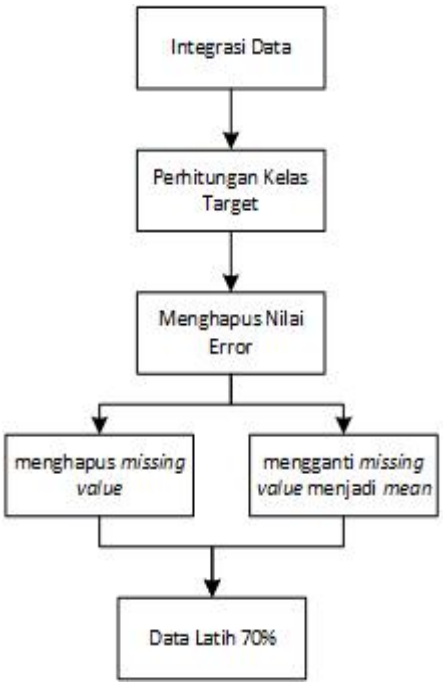
DOI: https://doi.org/10.31004/ijmst.v2i1.298

2.1. Pengumpulan data

Penelitian ini menggunakan hasil pengukuran kualitas air sungai tahun 2020 dari 23 lokasi pengukuran (ONLIMO) di 14 sungai di Indonesia. Beberapa parameter pencemar dibahas dalam penelitian ini, diantaranya total cairan terlarut (TDS), oksigen terlarut (DO), tingkat keasaman (pH), permintaan oksigen kimiawi (COD), dan permintaan oksigen biologi (BOD).

Sebelum dilakukan proses klasifikasi kualitas mutu air, terlebih dahulu dilakukan *praproses* data sesuai pada Gambar 2 sebagai berikut:

1. Tahap awal penelitian adalah integrasi data, yang menggabungkan data pemantauan kualitas air sungai dari berbagai sumber. Pekerjaan ini menghasilkan data tabular pemantauan kualitas air sungai di Indonesia.
2. Perhitungan kelas mutu air dilakukan untuk memberi label kelas kepada semua data yang digunakan. Metode Storet dan baku mutu kelas 2 digunakan untuk menentukan kelas ini [4]. Penelitian ini menghasilkan dataset dengan label kelas yang sesuai dengan kelas mutu air.
3. Menghapus nilai *error,* proses ini dilakukan dengan cara menghapus nilai parameter error. Misalnya parameter pH > 14.
4. Menangani *Missing value* menggunakan beberapa jenis penanganan, yakni mengganti *missing value* dari data *numeric* dengan nilai *mean* dan menghapus *record* dengan data tidak lengkap.
5. Proses pembagian data *training* dan data *testing* dengan persentase pembagian data adalah 70% data *training* dan 30% data *testing*



Gambar 2 Praproses Data

2.2. Naive Bayes

Naïve Bayes atau *Naïve Bayes Classifier* berasal dari *Bayes Theorem* (teorema Bayes) yang ditemukan oleh Thomas Bayes pada tahun 1770. *Naïve Bayes* merupakan salah satu teknik klasifikasi dengan metode probabilitas dan statistika untuk memprediksi peluang di masa depan berdasarkan pengalaman sebelumnya sehingga dikenal dengan istilah *teorema Bayes.* Teorema tersebut dikombinasikan dengan *Naïve* yang diasumsikan sebagai kondisi antar atribut saling bebas [5].

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai *output.* Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan *Naive Bayes* adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (*Training* Data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. *Naive Bayes* sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan [5].

2.3. K-Nearest Neighbour

K-nearest neighbour merupakan pendekatan yang dilakukan untuk proses penyelesaian kasus dengan menghitung pembobotan pada sejumlah fitur yang telah ditentukan. Pendekatan pada KNN dilakukan dengan menghitung kedekatan antara kasus yang baru dengan kasus yang lama. Tujuan algoritma KNN adalah mengklasifikasikan objek baru berdasarkan atribut dan training sample [6]

DOI: https://doi.org/10.31004/ijmst.v2i1.298

K-nearest neighbour atau dikatakan sebagai lazy learner karena berbasis pembelajaran. Metode KNN bekerja dengan cara menunda proses pemodelan data pelatihan sampai data tersebut dibutuhkan untuk mengklasifikasikan sampel data uji, sementara sampel data lain dijelaskan oleh atribut-atribut numerik pada n-dimensi dan disimpan dalam ruang n-dimensi.

K-nearest neighbour bekerja dengan mencari sampel K pelatihan yang paling dekat dengan sampel data uji. Pencarian nilai tetangga K dilakukan berdasarkan euclidean distance. Berikut ini merupakan persamaan euclidean distance yang digunakan dalam penelitian [7]

2.4. Support Vector Machine

Support Vector Machine (SVM) pertama kali dikenalkan pada tahun 1992 oleh Boser, Guyon, dan Vapnik di COLT-92. SVM adalah metode klasifikasi dan regresi yang menggunakan teori machine learning untuk memaksimalkan akurasi dari prediksi dan secara otomatis menghindari over-fit pada data [8]. SVM bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah class pada *input space* [9]. Fungsi pemisah *(hyperplane)* yang optimal untuk memisahkan observasi yang memiliki nilai variabel target yang berbeda. *Hyperplane* ini dapat berupa line pada *two dimension* dan dapat berupa *flat plane* pada *multiple dimension*.

2.5. Random Forest

Random forest (RF) merupakan pengembangan dari metode *classification and regression tree* (CART) dengan menerapkan metode *bootstrap aggregating* (*bagging)* dan *random feature selection. Random forest* merupakan metode yang dapat meningkatkan hasil akurasi karena dalam membangkitkan simpul anak untuk setiap *node* dilakukan secara acak [10]. *Random Forest* merupakan salah satu metode yang digunakan untuk klasifikasi dengan membangun pohon keputusan yang terdiri dari *root node*, internal *node*, dan *leaf node* dengan mengambil atribut dan data secara acak [11].

2.6. Multiclass classification

Multiclass classification adalah metode klasifikasi untuk memprediksi satu atau lebih kelas dari setiap contoh dan setiap sampel hanya bisa diberi label satu kelas. Contohnya ialah klasifikasi menggunakan ekstraksi fitur dari berbagai citra buah dimana setiap citra bisa menunjukkan buah jeruk, apel, ataupun pir. Setiap citra dikategorikan sebagai satu sampel dan diberi satu label dari tiga kemungkinan kelas. Pada algoritma *multiclass classification* melakukan klasifikasi setiap citra buah hanya pada satu label, misalkan citra buah A tidak bisa masuk ke dalam kelas buah apel dan buah pir. Dalam klasifikasi, data disajikan ke dalam sejumlah contoh pelatihan yang dibagi menjadi K kelas terpisah, dan kemudian membangun model *machine learning* untuk memprediksi kelas mana yang dimiliki oleh beberapa data yang sebelumnya tidak diketahui (yaitu jenis buah dari contoh sebelumnya). Dalam melihat dataset pelatihan, model mempelajari pola khusus untuk setiap kelas dan menggunakan pola tersebut untuk memprediksi keanggotaan data selanjutnya.

2.7. Multi Layer Perceptron

Multi Layer Perceptron merupakan metode jaringan syaraf tiruan berasitektur feed-forward atau biasa dikatakan sebagai perceptron multi lapis adalah metode artificial neural network (ANN) yang memiliki arsitektur jaringan yang terdiri dari minimal 3 layer yakni layer input, layer hidden dan layer output. Metode MLP bertujuan untuk mendapatkan vector bobot yang paling fit dengan data latih, metode MLP menggunakan nilai error untuk memperbaiki nilai bobot [12].

Algoritma MLP merupakan metode supervised learning (suatu metode machine learning yang menarik kesimpulan dari data-data yang telah diberi label, berupa pasangan input dan output) [13].

3. Hasil dan Pembahasan

Penelitian dilakukan dengan dua jenis skenario. Skenario pertama dilakukan penanganan data dengan menghapus *missing value* dan skenario kedua dengan mengganti *missing value* menjadi *mean*. Skenario pada penelitian ini selanjutnya dilakukan proses klasifikasi menggunakan beberapa algoritma klasifikasi terhadap kualitas mutu air.

Hasil penelitian yang dilakukan dengan penanganan menghapus data *missing value* mendapatkan hasil seperti pada Tabel 1.

Tabel 1 Hasil Pengujian Skenario 1

Skenario 1	
Algoritma	**Tingkat Akurasi**
SVM	92,7%
MLP	77%
Multiclass Classification	87,7%
Naive Bayes	21,9%
KNN	96,4%
Random Forest	99,7%

Berdasarkan hasil perbandingan dari enam algoritma klasifikas dengan menghapus data *missing value* yaitu SVM 92,7%, MLP 77%, *multiclass classification* 87,7%, *naïve bayes* 21.9%, KNN 96,4%, dan Random Forest 99,7%.

Hasil penelitian yang dilakukan pada skenario kedua dengan penanganan data mengganti *missing value* menjadi *mean.* Hal ini dilakukan untuk melihat pengaruh penanganan missing value terhadap hasil penelitian. Skenario kedua mendapatkan hasil persentase akurasi untuk setiap algoritma. Hal ini dapat dilihat pada Tabel 2.

DOI: https://doi.org/10.31004/ijmst.v2i1.298

Tabel 2 Hasil Pengujian Skenario 2

Skenario 1	
Algoritma	**Tingkat Akurasi**
SVM	92%
MLP	76,8%
Multiclass Classification	88,2%
Naive Bayes	22,3%
KNN	95,4%
Random Forest	99,5%

Berdasarkan hasil perbandingan dari enam algoritma klasifikasi mendapatkan hasil SVM 92%, MLP 76,8%, *multiclass classification* 88,2%, *Naïve bayer* 22,3%, KNN 95,4%, dan Random forest 99,5%

4. Kesimpulan

Klasifikasi mutu air adalah salah satu teknik dalam melakukan penilaian terhadap air yang menjadi objeknya. Ini dilakukan dengan tujuan agar dapat memberikan pengetahuan terhadap mutu/kualitas air, sehingga dapat menjadi solusi terbaik yang dapat dilakukan terhadap air tersebut. Penelitian ini menggunakan 6 algoritma klasifikasi untuk mengklasifikasikan status mutu air menggunakan 2 skenario. Skenario pertama melakukan penanganan dengan mengganti menghapus data *missing value* untuk setiap algoritma klasifikasi yang digunakan. Sementara skenario kedua yakni dengan *missing value* menjadi *mean* dan proses tersebut juga dilakukan pada setiap algoritma. Penelitian yang dilakukan menggunakan data 4957 *record* dengan 19 atribut. Output yang dihasilkan dari penelitian ini terdiri dari 4 label, yakni baik, cemar ringan, cemar sedang dan cemar berat. Berdasarkan hasil perbandingan dari 6 algoritma yang dilakukan, algoritma *random forest* mendapatkan hasil terbaik pada kedua skenario yang diterapkan yakni dengan hasil persentase 99.5% dan 99.7%. sementara algoritme Naive Bayes memperoleh tingkat akurasi sangat rendah (22.1%). Hal tersebut dapat disebabkan kelas mutu air tidak seimbang atau data tidak terdistribusi normal (Gaussian). Selain itu, algoritme Naive Bayes memiliki kinerja baik dalam pekerjaan klasifikasi dengan data teks.

Reference

[1] N. Lusiana, B. R. Widiatmono, and H. Luthfiyana, "Beban Pencemaran BOD dan Karakteristik Oksigen Terlarut di Sungai Brantas Kota Malang," *J. Ilmu Lingkung.*, vol. 18, no. 2, pp. 354–366, 2020, doi: 10.14710/jil.18.2.354-366.

[2] F. Modaresi and S. Araghinejad, "A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification," *Water Resour. Manag.*, vol. 28, no. 12, pp. 4095–4111, 2014, doi: 10.1007/s11269-014-0730-z.

[3] R. Prakash, V. P. Tharun, and S. Renuga Devi, "A Comparative Study of Various Classification Techniques to Determine Water Quality," *Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018*, no. Icicct, pp. 1501–1506, 2018, doi: 10.1109/ICICCT.2018.8473168.

[4] M. L. Hidup, "Keputusan menteri lingkungan hidup nomor 115 tahun 2003 tentang pedoman Penentuan status mutu air," 2003.

[5] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Syst.*, vol. 192, p. 105361, Mar. 2020, doi: 10.1016/j.knosys.2019.105361.

[6] M. Kholil, Kusrini, and Henderi, "Penerapan Metode K Nearest Neighbord Dalam Proses Seleksi Penerima Beasiswa," *Semin. Nas. Sist. Inf. dan Teknol. Inf. 2018*, pp. 13–18, 2018.

[7] A. A. Syafitri Hidayatul AA, Yuita Arum S, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 9, pp. 2546–2554, 2018.

[8] V. Jakkula, "Tutorial on Support Vector Machine (SVM)," *Sch. EECS, Washingt. State Univ.*, pp. 1–13, 2011, [Online]. Available: http://www.ccs.neu.edu/course/cs5100f11/resources/jakkula.pdf.

[9] S. anto Nugroho and arif budi Witarto, "Support vector machine teori dan aplikasinya dalam bioinformatika," 2003.

[10] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *CogITo Smart J.*, vol. 6, no. 2, p. 167, 2020, doi: 10.31154/cogito.v6i2.270.167-178.

[11] V. W. Siburian and I. E. Mulyana, "Prediksi Harga Ponsel Menggunakan Metode Random Forest," *Pros. Annu. Res. Semin.*, vol. 4, no. 1, pp. 144–147, 2018.

[12] i nyoman tri anindia Putra, ni komang ayu Sinariyani, N. Maharani, and ketut sepdayana Kartini, "decision support system for determining the type of workout using fuzzy analythical hierarchy process (F-AHP) method in STIKI GYM," 2021.

[13] I. G. R. M. Putra, M. W. A. Kesiman, G. A. Pradnyana, and I. M. D. Maysanjaya, "Identifikasi Citra Ukiran Ornamen Tradisional Bali Dengan Metode Multilayer Perceptron," *SINTECH (Science Inf. Technol. J.*, vol. 4, no. 1 SE-, pp. 29–39, 2021, [Online]. Available: https://ejournal.stiki-indonesia.ac.id/index.php/sintechjournal/article/view/552.

DOI: https://doi.org/10.31004/ijmst.v2i1.298