Jurnal Komputer dan Informatika
Vol 19 No 2, oktober 2024: hlm 103-110

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

KLASIFIKASI KUALITAS *INDEX* UDARA DUNIA DENGAN METODE *ARTIFICIAL NEURAL NETWORK* DAN *SUPPORT VECTOR MACHINE KERNEL* RBF

Vincent Wijaya
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara,
Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
E-mail: Vincent.535220064@stu.untar.ac.id

ABSTRAK

Pencemaran udara telah menjadi banyak masalah besar di seluruh dunia, kualitas udara harus diperhatikan agar manusia tidak menderita gangguan fisiologis maupun kematian akibat gangguan pernafasan. Dalam penelitian ini meilihat bagaimana dua algoritma machine learning yaitu *Artificial Neural Network* (ANN) dan *Support Vector Machin*e (SVM) berfungsi untuk memberi pengetahuan mengenai algoritma machine learning untuk mengatasi pencemaran udara di dunia. Hasilnya ANN memiliki akurasi sempurna sedangkan *Support Vector Machine* (SVM) kernel RBF memiliki akurasi 0.99. Tujuan Utama dari penelitian ini adalah untuk memprediksi kualitas indeks udara dunia.

Kata kunci—Pencemaran Udara, *Support Vector Machine, Machine Learning, Artificial Neural Network,* Kualitas Indeks Udara Dunia

ABSTRACT

Air pollution has become a significant problem worldwide, and air quality must be addressed to prevent physiological disorders and respiratory-related deaths. This study examines how two machine learning algorithms, Artificial Neural Network (ANN) and Support Vector Machine (SVM), function to provide insights into machine learning algorithms for addressing air pollution worldwide. The results show that ANN achieved perfect accuracy, while Support Vector Machine (SVM) with RBF kernel attained 0.99 accuracy. The main objective of this research is to predict the air quality index worldwide.

Keywords— *Air Pollution, Support Vector Machine, Machine Learning, Artificial Neural Network, World Air Quality Index*

1. PENDAHULUAN

Udara merupakan salah satu sumber kebutuhan untuk hidup bagi manusia. Pencemaran udara telah menjadi masalah besar di banyak negara di seluruh dunia. Kualitas udara harus dipantau dan dijaga untuk kesejahteraan hidup manusia. Akibat pencemaran udara, jutaan manusia di seluruh dunia menderita gangguan fisiologis dan kematian akibat gangguan pernapasan. Pencemaran udara merupakan masalah lingkungan yang merajalela di abad ke-21. Seiring dengan cepatnya industrialisasi dan urbanisasi, pencemaran udara semakin memburuk, yang sangat mempengaruhi lingkungan hidup dan kesehatan kita.[6]

Ancaman polusi udara, terutama dari bahan-bahan berpartikel (PM), cukup serius sehingga dapat menyebabkan tingkat kematian yang lebih tinggi, seperti yang disarankan oleh Organisasi Kesehatan Dunia (WHO) [1-2]. Selain itu, jumlah kendaraan yang semakin meningkat bertanggung jawab atas peningkatan polutan seperti NO2, CO, NH3, PM2.5, dan PM10, sedangkan polutan seperti SO2, CO, O3, B (Benzene), T (Toluene), dan X (Xylene) berasal dari sumber-sumber industri. *Machine Learning* (ML) adalah bagian penting dari kecerdasan buatan (AI) yang memungkinkan suatu sistem untuk belajar dan memperbaiki diri secara otomatis dari pengalaman tanpa harus diprogram secara eksplisit [3]. Teknik-teknik yang digunakan dalam ML didasarkan pada pemeriksaan mendalam terhadap data untuk menemukan tren dan memperbarui diri sesuai dengan

temuan tersebut [4-5]. Ada enam polutan udara konvensional yang digunakan untuk mengukur kualitas udara: sulfur dioksida (SO2), nitrogen dioksida (NO2), materi partikulat dengan ukuran partikel kurang dari 10 mikron (PM10), materi partikulat dengan ukuran partikel kurang dari 2,5 mikron (PM2.5), ozon (O3), dan karbon monoksida (CO) [7-8]. Penelitian ini dibuat untuk membandingkan dua algoritma klasifikasi dari cabang ilmu *machine learning* yaitu *Support Vector Machine* (SVM) kernel *Radian Basis Function* (RBF) dan *Artificial Neural Network* (ANN).

2. METODE PENELITIAN

2.1 Klasifikasi

Klasifikasi merupakan serangkaian tes yang dilakukan setelah pelatihan, dan hasilnya akan menghasilkan model yang paling akurat untuk melakukan klasifikasi [9].

2.2 Data Collection

Dataset yang dipakai dalam penelitian ini adalah Data Kualitas Data Indeks Dunia pada tahun 2022. Terdapat 14 atribut dalam dataset yaitu: *Country, City, AQI Value, AQI Category, CO AQI Value, CO AQI Category, Ozone AQI Value, Ozone AQI Category, NO2 AQI Value, NO2 AQI Category, PM2.5 AQI Value, PM2.5 AQI Category, lat, ing*

2.3 *Machine Learning*

Tanpa diprogram secara eksplisit, sistem dapat belajar secara otomatis dan meningkatkan kemampuan mereka berdasarkan pengalaman mereka sendiri. Ini merupakan salah satu aplikasi atau bagian dari kecerdasan buatan [18].

2.4 Data *Pre-processing*

Pre-processing data adalah proses pemilihan data agar dapat digunakan secara lebih terstruktur dengan menghilangkan atribut yang tidak perlu, membuat data lebih sistematis, dan mengurangi noise [10]. Pembersihan data merupakan tahap pre-processing di mana data yang hilang dan duplikat diperiksa. Jika terdapat data yang hilang, mereka dapat diimputasi dengan menggunakan mean atau median, atau dihapus [11]. Jika terdapat data yang duplikat, mereka dapat dihapus. Langkah pertama sebelum proses klasifikasi dilakukan adalah transformasi data, di mana data kategori diubah menjadi data numerik.

2.5 Data latih dan Data uji

Proses membagi data menjadi dua bagian, yaitu data uji dan latihan, dikenal sebagai pemisahan data [12]. Model splitting data, yang digunakan untuk mempartisi dataset, mempengaruhi seberapa baik model klasifikasi berfungsi pada algoritma pembelajaran mesin. Berbagi data bertujuan untuk memastikan bahwa model yang dibangun dapat diterapkan pada data baru. Dalam penelitian ini, algoritma SVM kernel RBF dan ANN akan menggunakan data latih 80% dan data uji 20%.

2.6 *Support Vector Machine*

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang diawasi yang digunakan untuk menganalisis dan mengidentifikasi pola. Sebelum menggunakannya untuk klasifikasi SVM mengubah dokumen teks menjadi vektor [13]. SVM juga menarik garis ideal untuk menjadi pemisah kedua kelas data dengan garis batas pemisah yang paling besar [14]. Menurut [15], SVM bertujuan untuk menemukan hiperrata terbaik, yang memungkinkan klasifikasi dan prediksi yang akurat dengan memisahkan secara maksimal.Menurut [16], metode pengklasifikasian SVM dapat digunakan secara linier atau non-linier. Tujuannya adalah untuk menghitung hasil support vector, sehingga peneliti hanya perlu mengetahui fungsi kernel dan non-liniernya. Metode ini memungkinkan peneliti menemukan hyperplane terbaik dengan menemukan titik yang maksimum dan menghitung margin hyperplane.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Persamaan *Support Vector Machine*:

$$f(x) = w^{t}\Phi(x) + b$$

Keterangan :

b = Bias
$x = (x_1, x_2, \ldots, x_D)^T$ = Variabel Input
$W = (W_0, W_2, \ldots, W_D)^T$ = Parameter Bobot
$\Phi(x)$ = Fungsi Transformasi Fitur

Gambar 1. Persamaan SVM

2.6 *Artificial Neural Network*

Artificial Neural Network digunakan untuk klasifikasi dan prediksi dalam penambangan data. Pada awalnya, pembelajaran mesin digunakan untuk mencoba meniru neurofisiologi otak manusia dengan menggabungkan elemen komputasi dasar (*neuron*) ke dalam sistem yang saling berhubungan[17]. Untuk melakukan ANN, data pelatihan digunakan sebagai input untuk menemukan model yang tepat. Konstruksi algoritma ANN akan diuji secara eksperimental untuk menentukan jumlah neuron dan lapisan yang tepat untuk mencapai kinerja terbaik. Berikut ini adalah persamaan yang digunakan:

$$y = f(\sum^{n}{}_{i=1} X_i W_i)$$

Gambar 2. persamaan ANN

Hubungan Antara ketiga komponen pada persamaan diatas adalah:

a. Sinyal x berupa vektor berdimensi n (x1, x2 xn) y akan mengalami penguatan oleh synapse (w1, w2, .wn) y.
b. Akumulasi penguatan ditransformasikan oleh fungsi aktifasi f. Fungsi f ini memantauapabila akumulasi penguatan sinyal itu telah melebihi batas tertentu, maka sel neuron yang awalnya berada di kondisi "0", akan mengeluarkan sinyal"1".
c. Dengan nilai keluaran (y), sebuah neuron dapat berada dalam dua keadaan yaitu: "0" atau "1". Neuron dapat dikatakan sebagai aktif dalam ketika menghasilkan nilai keluaran "1".

2.7 *Evaluasi Klasifikasi*

Table *matrix confusion* digunakan untuk menghitung kinerja model data atau algoritma [19]. Setiap baris matrix menunjukkan kelas data aktual dan kelas prediksi, atau sebaliknya.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	True Negative (TN)
Actual Negative	False Positive (FP)	False Negative (FN)

Gambar 3. *Confusion Matrix*

Berdasarkan gambar diatas, *Confusion Matrix* memiliki 4 nilai yaitu :

a. True Positive: Seberapa banyak data yang aktual kelasnya positif, dan model memprediksi positif
b. True Negative: Seberapa banyak data yang aktual kelasnya negatif, dan model memprediksi negatif
c. False Positive: Seberapa banyak data yang aktual kelasnya negatif, namun model memprediksi positif

d. False Negative: Seberapa banyak data yang aktual kelasnya positif, namun model memprediksi negatif

Evaluasi Akurasi, Presisi, F1 Score dan Recall dapat dihitung melalui persamaan berikut Akurasi adalah total keseluruhan seberapa sering model benar mengklasifikasi.

$$\frac{TP+TN}{TP+FP+FN+TN}$$

Gambar 4. Persamaan Akurasi

Rentang Nilai	Klasifikasi Performa
90%-100%	Sangat Baik
80%-90%	Baik
70%-80%	Cukup
60%-70%	Buruk
<=60%	Sangat Buruk

Gambar 5. Klasifikasi performa berdasarkan Akurasi

Presisi adalah ketika model memprediksi positif, seberapa sering prediksi itu benar.

$$\frac{TP}{TP+FP}$$

Gambar 5. Persamaan Presisi

Recall adalah ketika kelas aktualnya positif, seberapa sering model memprediksi positif.

$$\frac{TP}{FN+TP}$$

Gambar 6. Persamaan Recall

F1-Score adalah merupakan rata-rata harmonik dari precision dan recall.

$$\frac{2\ (Recall * Precision)}{(Recall+Precision)} (6)$$

Gambar 7. Persamaan F1-Score

2.8 *Numpy*

Numpy merupakan sebuah library untuk bahasa pemrograman Python, mendukung himpunan dan matriks multidimensi yang besar dan memiliki koleksi besar fungsi matematika yang dapat digunakan pada himpunan ini [20].

2.9 data *visualisasi*

Menampilkan data yang sesuai dengan data dalam dataset.

3. HASIL DAN PEMBAHASAN

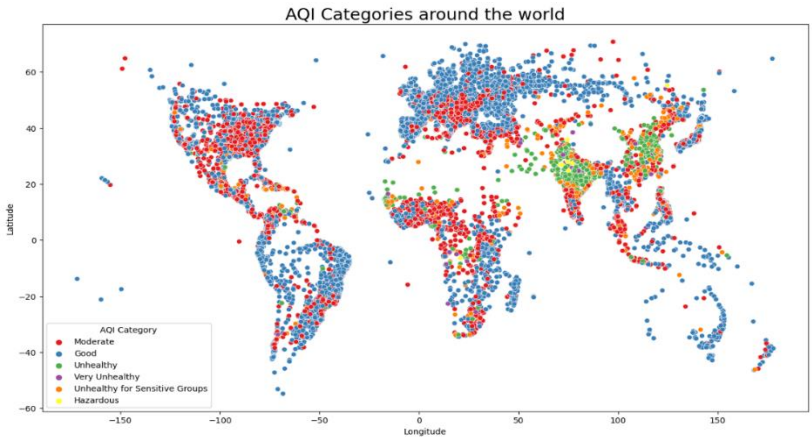
Pada Tahap ini, saya akan menunjukkan hasil eksperimen dengan dataset yang saya gunakan dalam penelitian ini. Index Kualitas Udara Dunia merupakan dataset yang saya gunakan, sumber dari dataset ini: https://www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates/data.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Dalam dataset ini terdapat total 16695 data dengan 14 fitur, dataset ini masih belum dilakukan pembersihan data. Terdapat 2 algoritma untuk melakukan klasifikasi pada dataset ini yaitu Artificial Neural Network (ANN) dan Support Vector Machine (SVM) kernel RBF dengan menggunakan bahasa pemogramman phyton, file Jupyter dan menggunakan numpy. Hasil dari penelitian ini akan membandingkan antara ke dua algoritma manakah yang memiliki nilai tertinggi untuk dataaset inI Penjelasan fitur - fitur yang ada pada dataset tersebut

1. *Country*: Negara yang dilihat AQInya
2. City: Kota yang dilihat AQInya
3. AQI Value: Penilaian indeks kualitas udara
4. AQI Categories: Kategori penilaian indeks kualitas udara
5. CO AQI Value: Penilaian terhadap gas yang tidak berwarna dan tidak berbau yang dihasilkan dari pembakaran tidak sempurna bahan bakar fosil.
6. CO AQI Categories: Kategori Penilaian terhadap gas yang tidak berwarna dan tidak berbau yang dihasilkan dari pembakaran tidak sempurna bahan bakar fosil.
7. Ozone AQI Value: Penilaian terhadap gas yang dapat terbentuk di atmosfer melalui reaksi kimia antara sinar matahari dan polutan lainnya.
8. Ozone AQI Categories: Kategori penilaian terhadap gas yang dapat terbentuk di atmosfer melalui reaksi kimia antara sinar matahari dan polutan lainnya.
9. NO2 AQI Value: Penilaian terhadap gas beracun berwarna coklat kemerahan dan berbau tajam.
10. NO2 AQI Categories: Kategori penilaian terhadap gas beracun berwarna coklat kemerahan dan berbau tajam.
11. PM2.5 AQI Value: Penilaian terhadap partikel atau tetesan kecil di udara yang memiliki lebar 2,5 mikrometer atau kurang.
12. PM2.5 AQI Categories: Kategori penilaian terhadap partikel atau tetesan kecil di udara yang memiliki lebar 2,5 mikrometer atau kurang.
13. lat dan ing: sebagai titik letak koordinat kota

Sebelum melakukan klasikasi data, saya menggunaka letak titik koordinat tersebut untuk melakukan *data visualization*, dengan lat sebagai titik x lalu ing sebagai titik y dan aqi categories untuk menentukan keadaan AQI pada setiap kota.



Gambar 8. Data visualisasi mengenai titik letak koordinatnya menggunakan lat dan ing

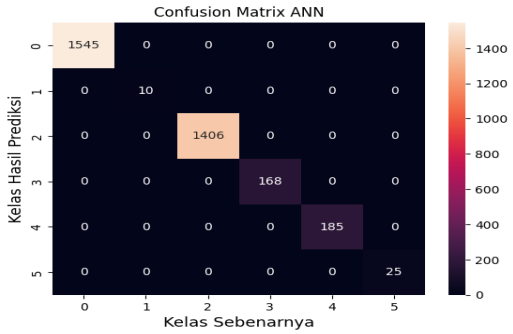
Lalu, untuk tahap *data cleansing*, saya akan menghilangkan fitur yang tidak berpengaruh dalam klasifikasi pada dataset ini. Fitur yang saya hilangkan pada dataset ini yaitu : ‘Country’, 'City','NO2 AQI Value','NO2 AQI Category','lat', 'lng' , karena fitur - fitur ini tidak terlalu berpengaruh dalam hasil klasifikasi pada dataset ini, maka fitur tersebut akan saya hilangkan dari dataset ini.

Setelah melakukan tahap *data cleansing*, saya akan melakukan tahap *data splitting*. Tahap ini saya akan membagi dataset saya menjadi dua, yaitu data latih dan data uji. Dalam eksperimen kali ini, saya akan menggunakan persentase data latih 80% dan data uji 20%, dan random state dari 1 sampai 5.

Tabel 1. Hasil klasikasi data menggunakan *Artificial Neural Network* (ANN) dengan max iter = 100

Akurasi	Presisi	Recall	f1-score	Random State
1	0.98	0.93	0.95	1
0.99	0.98	0.89	0.91	2
1	0.98	0.96	0.97	3
1	0.98	0.98	0.98	4
1	1	1	1	5
0.998	0.984	0.952	0.962	Rata - rata

Berdasarkan hasil Tabel 1. dapat dilihat bahwa hasil algoritma ANN dengan random state yang ke 5 memiliki hasil sempurna, sedangkan yang lainnya hampir sempurna. maka confusion matriks yang dimilikinya adalah



Gambar 9. Data visualisasi hasil dari confusion Matriks ANN Random state ke 5

berdasarkan gambar dari *confusion matrix* tersebut,

0 = good
1 = Hazardous
2 = Moderate
3 = Unhealthy,
4 = Unhealthy for Sensitive group
5 = Very Unhealthy

dan tidak ada kesalahan prediksi dalam data tersebut.

Tabel 2. Hasil klasifikasi data menggunakan algoritma *Support Vector Machine* (SVM) kernel RBF

Akurasi	Presisi	Recall	f1-score	Random State
0.99	0.98	0.95	0.96	1
0.99	0.97	0.95	0.96	2
0.99	0.97	0.95	0.96	3
0.99	0.99	0.95	0.96	4
0.99	0.98	0.97	0.97	5

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Berdasarkan hasil Tabel 2. dapat dilihat bahwa akurasi dari ke semua random state algoritma SVM mendapatkan hasil yang sama, namun untuk presisi, recall, dan f1-score berbeda. Nilai tertinggi diperoleh sama random state ke 4

Tabel 3. Hasil perbandingan *Evaluasi Klasifikasi* antara algoritma ANN dan SVM kernel RBF

Algoritma	Akurasi	Presisi	Recall	f1-score
ANN	0.998	0.984	0.952	0.962
SVM kernel RBF	0.99	0.978	0.954	0.962

4. KESIMPULAN

Kesimpulan yang dapat saya ambil adalah berdasarkan data hasil eksperimen nilai rata-rata yang diperoleh dari data "*AQI and Lat Long of Countries*.CSV" dengan menggunakan algoritma *Artificial Neural Network*, dan *Support Vector Machine* kernel *Radial Basis Function* dengan data latih 80%, data uji 20%, dan random state bernilai 1, 2, 3, 4, dan 5 dapat disimpulkan bahwa Algoritma *Artificial Neural Network* (ANN) memperoleh nilai tertinggi dibandingkan algoritma *Support Vector Machine* (SVM) kernel RBF, sehingga algoritma *Artificial Neural Network* memiliki kinerja terbaik. Sedangkan Algoritma *Support Vector Machine* memiliki kinerja yang lebih buruk dibandingkan algoritma *Artificial Neural Network*.

UCAPAN TERIMA KASIH

Penulis ingin mengucapkan terima kasih kepada Bu Teny Handayani sebagai dosen Machine Learning yang telah memberi bimbingan terhadap penelitian ini. Saya juga ingin berterima kasih kepada Jurnal Computatio dan IJCCS sebagai landasan bentuk format makalah ini.

DAFTAR PUSTAKA

[1] D. Fang, B. Chen, K. Hubacek, R. Ni, L. Chen et al,"Clean air for some: Unintended spillover effects of regional air pollution policies," Science Advances, vol. 5, no. 4, pp. 4707–4731, 2019.

[2] D. A. Glencross, T. R. Ho, N. Camiña, C. M. Hawrylowicz and P. E. Pfeffer, "Air pollution and its effects on the immune system," Free Radical Biology and Medicine, vol. 151, pp. 56–68, 2020.

[3] A.Y. Sun B.R. Scanlon How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions, Environ. Res. Lett. 14 (7) (Jul. 2019), 073001

[4] M. Bagheri, A. Akbari, S.A. Mirbagheri, Advanced control of membrane fouling in filtration systems using artificial intelligence and machine learning techniques: a critical review, Process Saf. Environ. Prot. 123 (Mar. 2019) 229–252

[5] F. Hassanpour, S. Sharifazari, K. Ahmadaali, S. Mohammadi, Z. Sheikhalipour, Development of the FCM-SVR hybrid model for estimating the suspended sediment load, KSCE J. Civ. Eng. 23 (6) (Jun. 2019) 2514–2523

[6] X. Li, L. Jin, and H. Kan, ''Air pollution: A global problem needs local fixes,'' Nature, vol. 570, no. 7762, pp. 437–439, Jun. 2019.

[7] Y. Ding and Y. Xue, ''A deep learning approach to writer identification using inertial sensor data of air-handwriting,'' IEICE Trans. Inf. Syst., vol. E102-D, no. 10, pp. 2059–2063, 2019

[8] M. Jia, A. Komeily, Y. Wang, and R. S. Srinivasan, ''Adopting Internet of Things for the development of smart buildings: A review of enabling technologies and applications,'' Automat. Construct., vol. 101, pp. 111–126, May 2019.

[9] L. A. Demidova, "Two-stage hybrid data classifiers based on svm and knn algorithms," symmetry, vol. 13, no. 4, p. 32, 2021.

[10] M. Nurkholifah, Jasmarizal, Y. Umar, and Rahmaddeni, "ANALISA PERFORMA ALGORITMA MACHINE LEARNING DALAM PREDIKSI PENYAKIT LIVER," Jurnal Indonesia: Manajemen Informatika dan Komunikasi, vol. 4, no. 1, pp. 164–172, Jan. 2023.

[11] N. Yolanda Paramitha et al., "Klasifikasi Penyakit Stroke Menggunakan Metode Naïve Bayes," 2023.

[12] A. Putri, C. Syaficha Hardiana, E. Novfuja, F. Try Puspa Siregar, Y Fatma, and R.Wahyuni, "Comparison of K-NN, Naive Bayes and SVM Algorithms for Final-Year Student Graduation Prediction KomparasiAlgoritma K-NN, Naive Bayes dan SVM untuk Prediksi KelulusanMahasiswa Tingkat Akhir," Institut Riset dan Publikasi Indonesia (IRPI) MALCOM Indonesian Journal of Machine Learning and Computer Science Journal Homepage, vol. 3, no. 1, pp. 20–26, 2023.

[13] D. Septhya, K. Rahayu, S. Rabbani, V. Fitria, Y. Irawan, and R. Hayami, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Implementation of Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru," vol. 3, pp. 15–19, 2023.

[14] S. S. T. E. A. R., A. N. Ade Silvia Handayani1, "KLASIFIKASI KUALITAS UDARA DENGAN METODE SUPPORT VECTORMACHINE," JIRE (Jurnal Informatika & Rekayasa Elektronika), 2020.

[15] R. abibah br Lumbantobing, "Prediksi Harga Cryptocurrency Menggunakan Algoritma Support Vector Machine," 2023.

[16] J. I. Matematika and S. Adi, "MATHunesa Tahun 2022 KOMPARASI METODE SUPPORT VECTOR MACHINE (SVM), K-NEAREST NEIGHBORS (KNN), DAN RANDOM FOREST (RF) UNTUK PREDIKSI PENYAKIT GAGAL JANTUNG Atik Wintarti".

[17] A. M. Siregar and H. H. H, "Implementasi Algoritma Neural Network untuk Mendukung Keputusan di Desa Tamanmekar," Petir, vol. 13, no. 1, pp. 21–32, 2020, doi: 10.33322 petir v13i1.768.

[18] Purba Daru Kusuma, "Machine Learning Teori, Program, Dan Studi Kasus", ISBN 9786230210839, 6230210835, 2020.

[19] Irkham Widhi Saputro, Bety Wulan Sari, "Uji Performa Algoritma Naive Bayes untuk Prediksi Masa Studi Mahasiswa", Citec Journal, Vol. 6, No. 1 ISSN: 2460-4259, Universitas AMIKOM Yogyakarta, 2019.

[20] Charles R Harris, K. Jarrod Millman, Stefan J. van der Walt, dll "Array programing with Numpy", ISSN 14764687, 2020.