Jurnal Infortech
Volume 6 No. 1 Juni 2024
E-ISSN: 2715-8160

Perbandingan Algoritma *Machine Learning* pada Klasifikasi Penyakit Jantung

Musriatun Napiah[1*], Sujiliani Heristian[2]

[1]Ilmu Komputer, Universitas Bina Sarana Informatika
e-mail: musriatun.mph@bsi.ac.id

[2]Teknologi Komputer, Universitas Bina Sarana Informatika
e-mail: sujiliani.she@bsi.ac.id

Diterima	Direvisi	Disetujui
01-04-2024	06-05-2024	06-05-2024

Abstrak - Jantung adalah organ utama yang memompa darah keseluruh tubuh bergerak melalui sistem peredaran darah. Menurut WHO, penyakit jantung koroner (PJK) telah menjadi masalah kesehatan yang meningkat pesat, menyebabkan 6,7 juta kematian pada tahun 2017. Banyak alternatif atau cara yang digunakan untuk mencegah dan mendeteksi penyakit jantung, kekurangan pengetahuan sering kali membuat penderita terlambat memeriksa diri ke dokter. Tujuan dari penelitian ini adalah untuk mendeteksi penyakit jantung sejak dini akibat gangguan kadiovaskular sehingga dapat memungkinkan untuk pencegahan yang lebih efektif dan pengelolaan yang baik terhadap kondisi kardiovaskular, dengan menggunakan dataset dari *http://archive.ics.uci.edu/ml/* sebanyak 1026 pasien penyakit jantung. Metode yang digunakan pada penlitian ini adalah *Machine* Learning dengan algoritma *logistic regression, naive bayes,* dan *k-nearest neighbour* (KNN) . Hasil yang didapatan dari penelitian ini yang tertinggi adalah dengan menggunakan metode *k-nearest neighbour* (KNN) yaitu akurasi sebesar 91%, sedangkan dengan algoritma *logistic regression* akurasinya sebesar 87%, dan *naive bayes* akurasinya sebesar 83%.

Kata Kunci: Jantung, *logistic regression, naive bayes,* dan *k-nearest neighbour*

Abstract *- The heart is the main organ that pumps blood throughout the body through blood vessels. WHO notes that coronary heart disease (CHD) is a health problem that is increasing rapidly, causing 6.7 million deaths in 2017. There are many alternatives or methods used to prevent and detect heart disease, lack of knowledge often makes it too late for sufferers to check themselves. doctor. The aim of this research is to detect heart disease early due to cardiovascular disorders so that it can allow for more effective prevention and good management of cardiovascular conditions, using a dataset from http://archive.ics.uci.edu/ml/ of 1026 heart disease patients. The method used in this research is machine learning with logistic regression, naive Bayes, and k-nearest neighbor (KNN) algorithms. The highest results obtained from this research were using the k-nearest neighbor (KNN) method, namely an accuracy of 91%, while with the logistic regression algorithm the accuracy was 87%, and naive Bayes the accuracy was 83%.*

Keyword : Heart, Logistic Regression, Naive Bayes, and K-Nearst Neighbor

PENDAHULUAN

Jantung adalah salah satu organ yang memegang peranan penting dalam sistem peredaran darah manusia. Penyakit jantung merupakan salah satu penyakit yang paling mematikan di dunia, dan istilah untuk semua jenis gangguan yang mempengaruhi jantung. Penyakit jantung tidak sama dengan penyakit kardiovaskular mengacu pada gangguan pembuluh darah dan jantung, sedangkan penyakit jantung mengacu hanya pada bagian hati (Annisa et al., 2019).

Di Indonesia, penyakit jantung merupakan kondisi yang paling umum terjadi pada wanita dewasa, sesuai dengan informasi tentang kasus penyakit di negara ini melalui Data Indonesia menurut provinsinya, penyakit jantung merupakan masalah serius di Indonesia, dengan jumlah kasus yang tinggi dan angka kematian yang meningkat. Fakta bahwa biaya klaim terbesar dalam program Jaminan Kesehatan Nasional (JKN) disebabkan oleh penyakit jantung menunjukan dampak yang signifikan dari kondisi ini terhadap sistem kesehatan dan ekonomi negara. Angka 15,5 juta kasus penyakit

jantung di Indonesia pada tahun 2022 menunjukkan besarnya beban penyakit ini terhadap populasiada (DataIndonesia.id, 2023). Dari informasi tersebut, terlihat bahwa banyak individu yang belum menganggap serius faktor penyebab penyakit ini. Setelah menjalani pemeriksaan kesehatan, dokter menemukan bahwa penyakit ini telah mencapai tingkat stadium yang lanjut (Journal et al., 2019).

Kurangnya akses informasi/media dalam mencari tahu tentang penyakit jantung sehingga menyebabkan keterlambatan untuk pemeriksaan awal ke dokter, merupakan penyebab banyaknya angka kematian semakin tinggi terhadap penyakit ini. Dengan demikian diagnosa sejak dini sangat penting untuk dilakukan (Utomo, 2020), Sehingga diperlukan sistem klasifikasi untuk mendeteksi sejak dini tentang penyakit jantung, agar dapat memberkan informasi tentang penyakit jantung yang di derita oleh seseorang.

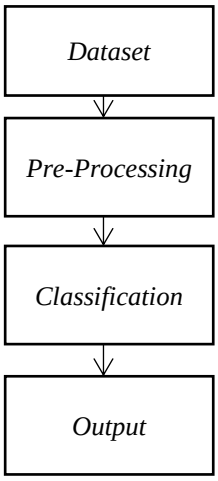
Machine Learning (ML) merupakan salah satu penerapan dari Kecerdasan Buatan (AI) yang berfokus pada pengembangan sistem yang mampu belajar secara mandiri tanpa perlu diprogram berulang kali. ML memerlukan data (biasanya disebut sebagai data pelatihan) sebagai bagian dari proses pembelajaran sebelum menghasilkan output. Menggunakan algoritma *Machine Learning* untuk mendiagnosis penyakit kanker diharapkan dapat menghasilkan prediksi yang akurat dan memungkinkan peningkatan dalam deteksi dini serta perawatan yang lebih efektif bagi pasien (Informatika & Informasi, 2020).

Naïve Bayes merupakan salah satu metode yang dapat digunakan untuk mengklasifikasikan data. Bayesian classification merupakan pengklasifikasian statistik yang dapat digunakan untuk memprediski probabilitas keanggotaan suatu class (Damuri et al., 2021). KNN merupakan algoritma yang efektif untuk data yang besar, tahan terhadap data pelatihan yang noise, dan memiliki performa yang baik (Putry et al., 2022), sedangkan *Logistic Regression* merupakan klasifikasi linier yang menangani masalah klasifikasi
multi kelas dan telah terbukti menghasilkan klasifikasi yang powerful (Prasetyo et al., 2021).

Penelitian ini akan membandingkan algoritma Naive Bayes, Regresi Logistik, dan KNN dalam klasifikasi penyakit jantung. Diharapkan penelitian ini akan memberikan acuan dan referensi serta memajukan pengetahuan untuk penelitian mendatang.

METODE PENELITIAN

Penelitian yang sudah dilakukan merupakan hasil eksperimen yang bertujuan untuk membandingkan hasil dari algoritma yang digunakan pada data penyakit jantung. Metode yang diusulkan sebagai berikut pada gambar.



Sumber : peneliti (2024)

Gambar 1. Metode Penelitian

1. Dataset

Dataset merupakan data mentah berupa tabel yang akan kita olah dengan menggunakan algoritma yang diusulkan. Penelitian ini menggunakan data pasien penyakit jantung diambil dari sumber yang terdapat di http://archive.ics.uci.edu/ml/, dengan total 1026 rekaman dan 76 atribut.

2. Pre – Processing

Pre-processing adalah proses pengolahan data atau citra asli sebelum data atau citra tersebut diolah oleh algoritma yang diusulkan pada metode peneltian (Cnn et al., 2020). Preprocessing data adalah tahap di mana data yang akan digunakan untuk pelatihan disiapkan terlebih dahulu. Dalam penelitian ini, preprocessing data dilakukan dengan proses untuk mengatasi data yang kosong atau hilang, yang melibatkan metode seperti mencari rata-rata atribut untuk kelas yang serupa (Rizki et al., 2020).

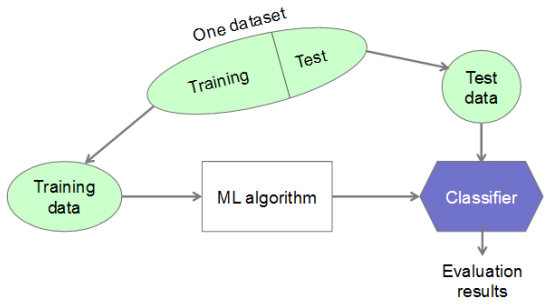
Langkah-langkah pada tahap *pre-processing* meliputi pembagian data menjadi dua bagian, sebagai berikut:

1. Train Data

Data training merupakan himpunan data yang telah diberi label atau kelas tertentu, yang digunakan oleh mesin untuk mengenali fitur-fiturnya serta memahami pola yang ada. Dengan memahami pola ini, mesin dapat membuat model atau pola data yang digunakan untuk tujuan tertentu, seperti klasifikasi atau prediksi (Musu et al., 2021). Data pelatihan (train data) sebanyak 20% untuk mengevaluasi algoritma.

2. Testing Data

Data testing adalah kumpulan data yang juga memiliki label atau kelas, yang digunakan untuk menguji keakuratan pola atau model yang telah dibuat dalam mengklasifikasikan data. Saat melakukan pengujian model, atribut label dari data testing disembunyikan selama proses klasifikasi, dan kemudian digunakan untuk membandingkan hasil klasifikasi dengan label sebenarnya. Hal ini memungkinkan untuk menilai seberapa baik model tersebut dalam melakukan klasifikasi dengan akurasi yang sesuai dengan data sebenarnya (Musu et al., 2021). Data pengujian (testing data) sebanyak 80% untuk menentukan metode algoritma yang paling sesuai dalam penelitian.



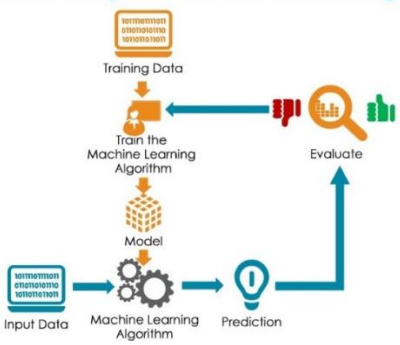
Sumber : (Caballe, 2019)

Gambar 2. Data Train dan Data Testing

3. *Machine Learning*

Machine learning adalah penerapan komputer dan algoritma matematika yang menggunakan pembelajaran dari data untuk menghasilkan prediksi pada periode yang akan datang (Roihan et al., 2020).

Machine Learning berhubungan erat dengan komputasi variabel, fokusnya adalah pada membuat prediksi menggunakan komputer. Studi tentang optimasi matematika menyediakan metode, teori, dan aplikasi di berbagai bidang pembelajaran mesin. Data mining adalah disiplin studi dalam pembelajaran mesin, yang menitikberatkan pada eksplorasi dan analisis data melalui pembelajaran tanpa pengawasan (Pratama, 2020).

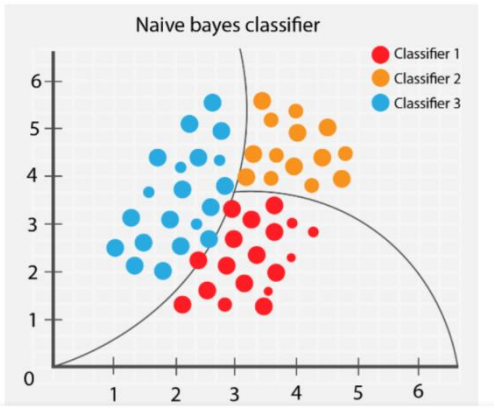


Sumber : (dqlab.id, 2021)

Gambar 3. Cara Kerja *Machine Learning*

Klasifikasi adalah proses yang didasarkan pada kelas dan *variable dependen* (Prasetio & Ripandi, 2019). Klasifikasi adalah metode pembelajaran yang disupervisi (*supervised learning*) yang memerlukan data pelatihan yang sudah diberi label untuk menghasilkan aturan yang membagi atau mengklasifikasikan data uji ke dalam kelompok atau kelas yang telah ditentukan sebelumnya (Bimo et al., 2020). Model klasifikasi yang populer adalah *Decission Tree, Naïve Bayes, Neural Network, Genetic Algorithm, dan Support Vector Machine* (Prasetio & Ripandi, 2019).

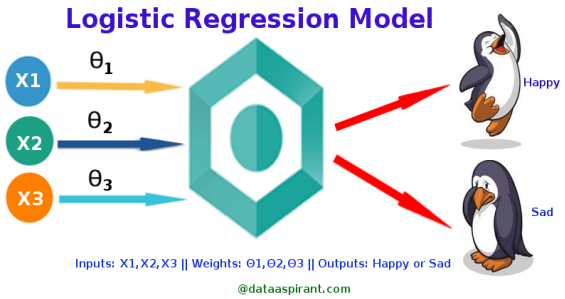
Naive Bayes adalah metode klasifikasi yang berasal dari *teorema Bayes*. Metode ini mengandalkan Probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris bernama Thomas Bayes, bertujuan untuk meramalkan kemungkinan di masa mendatang berdasarkan pengalaman yang terjadi di masa sebelumnya. Karena sifatnya yang mengasumsikan independensi antara fitur-fitur, metode ini disebut sebagai *Naive Bayes* (Watratan et al., 2020).



Sumber : (analytics vidhya, 2022)

Gambar 4. Algoritma *Naïve Bayes*

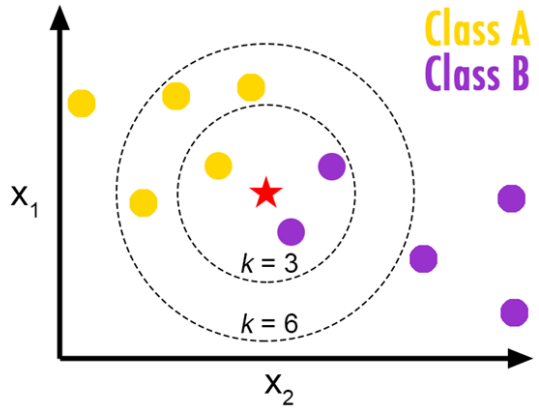
Logistic Regression merupakan bagian dari metode data mining yang digunakan untuk menganalisis data yang menjelaskan hubungan antara satu variabel *respons (dependen)* dengan satu atau lebih variabel predictor (Reviantika et al., 2021).



Sumber : (Mining, 2020)

Gambar 5. Model *Logistic Regression*

Metode KNN adalah algoritma yang dipakai dalam mengklasifikasikan data dengan jarak yang dekat, dan merupakan teknik *lazy learning* yang populer dalam kategori pembelajaran berbasis instansi (Puspita & Widodo, 2021).



Sumber : (suprianto, 2020)

Gambar 6. Model *KNN*

HASIL DAN PEMBAHASAN

Dari penelitian yang sudah dilakukan berikut adalah hasil yang didapatkan :

1. Dataset

Dataset yang diterapkan dalam penelitian ini adalah Ada 1025 data pasien dengan penyakit jantung, di mana 80% atau 820 data telah dipilih sebagai data pelatihan, sementara 20% sisanya digunakan sebagai data uji. Atribut-atribut yang dimasukkan meliputi usia (dalam tahun), jenis kelamin, cp, trestbps (tekanan darah istirahat pada saat masuk rumah sakit), chol (kolesterol serum), fbs (gula darah puasa > 120 mg/dl), restecg, dan thalach (denyut jantung maksimum yang dicapai), dan *target (diagnosis of heart disease).* Dataset dapat dilihat pada gambar 7.

age	sex	cp	tresbps	chol	Fbs	restecg	thalach	target
52	1	0	125	212	0	1	168	0
53	1	0	140	203	1	0	155	0
34	0	1	118	210	0	1	192	1
51	0	2	140	308	0	0	142	1
58	1	2	140	211	1	0	165	1
60	1	2	140	185	0	0	155	0
67	0	0	106	223	0	1	142	1

Sumber : (UC irvine Machine Learning Repostory, 1988)

Gambar 7. Dataset Pasien Penyakit Jantung

Dari gambar di atas dapat dilihat bahwa target dari dataset yang sudah dikumpulkan berdasarkan atribut yang sudah dijelaskan adalah terdiagnosis penyakit jantung dengan nilai 1, dan tidak terdiagnosis penyakit jantung dengan nilai 0.

2. Arsitektur Model *Machine Learning*

Terdapat tiga model *machine learning* yang digunakan dalam penelitian ini, yakni *Naïve Bayes, Regresi Logistik,* dan KNN .

Pada tahap awal menggunakan algoritma *Naïve Bayes* dengan menggunakan aplikasi *Python*¸dimulai dengan melakukan *import* pada *numpy*,*matplot*, *pandas*, dll.

1. Metode *Naïve Bayes*

```
[ ] print(classifier.score(X_test2, y_test2))

    0.8341463414634146

[ ] y_true = y_test2
    y_pred2 = predicted

    print(accuracy_score(y_true, y_pred2))

    0.8341463414634146

[ ] print(confusion_matrix(y_test2, y_pred2))
    print(classification_report(y_test2, y_pred2))

    [[76 21]
     [13 95]]
                  precision    recall  f1-score   support

               0       0.85      0.78      0.82        97
               1       0.82      0.88      0.85       108

        accuracy                           0.83       205
       macro avg       0.84      0.83      0.83       205
    weighted avg       0.84      0.83      0.83       205
```

Sumber : Peneliti (2024)

Gambar 8. Metode *Naïve Bayes*

Dari pengujian data pasien penyakit jantung menggunakan metode *Naïve Bayes*, diperoleh tingkat akurasi sebesar 83% dari total 1025 data.

2. Metode Logistic Regression

```
[ ] from sklearn.naive_bayes import GaussianNB
    nb = GaussianNB()
    nb.fit(x_train.T, y_train.T)

    acc = nb.score(x_test.T,y_test.T)*100
    accuracies['Naive Bayes'] = acc
    print("Accuracy of Naive Bayes: {:.2f}%".format(acc))

    Accuracy of Naive Bayes: 86.89%
```

Sumber : Peneliti (2023)

Gambar 9. Metode *Logistic Regression*

Dari pengujian yang sudah dilakukan pada data pasien penyakit jatung dengan menggunakan metode *Logistic Regression* diperoleh hasil akurasi sebesar 87% dari 1025 data.

3. Metode KNN (*k-nearest neighbour)*

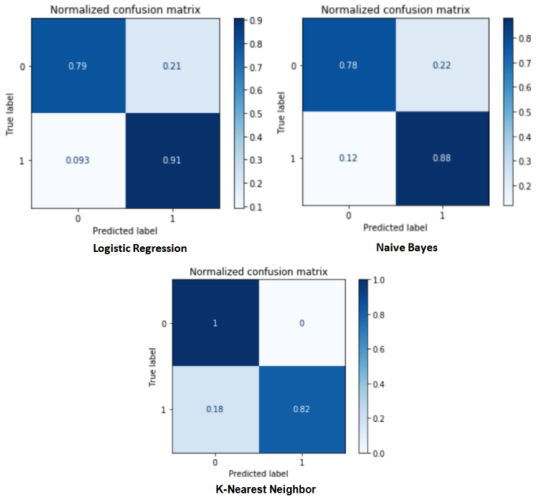
```
[ ]  knn.score(X_test3,y_test3)

     0.9073170731707317

[ ]  y_pred3 = knn.predict(X_test3)
     print(confusion_matrix(y_test3, y_pred3))
     print(classification_report(y_test3, y_pred3))

     [[97  0]
      [19 89]]
                   precision    recall  f1-score   support

                0       0.84      1.00      0.91        97
                1       1.00      0.82      0.90       108

         accuracy                           0.91       205
        macro avg       0.92      0.91      0.91       205
     weighted avg       0.92      0.91      0.91       205
```

Sumber : Peneliti (2024)

Gambar 10. Metode *KNN*

Dari pengujian yang sudah dilakukan pada data pasien penyakit jatung dengan menggunakan metode *K-NN* (*k-nearest neighbour)* diperoleh hasil akurasi sebesar 91% dari 1025 data.

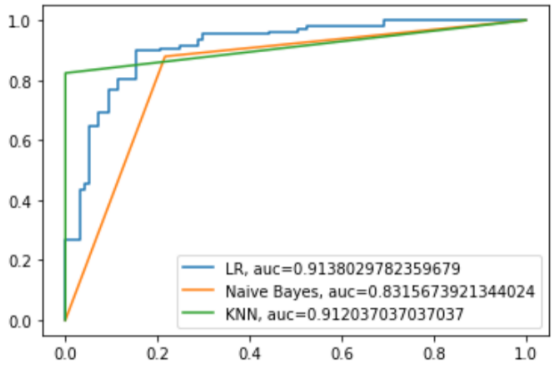
Dari matriks kebingungan yang dihasilkan, terbukti bahwa algoritma KNN memiliki akurasi tertinggi dalam hasil klasifikasi, mencapai 0,90731 dengan nilai K yang sesuai dengan jumlah target, yaitu 2. Algoritma Logistic Regression mendapatkan nilai akurasi sebesar 0,84878, sementara naive bayes mencatat tingkat akurasi terendah, yaitu 0,83414.



Sumber : Peneliti (2024)

Gambar 11. Matrix Model Algoritma

Gambar 11 Tampilan hasil confusion matrix yang telah dinormalisasi untuk setiap algoritma dipaparkan. Evaluasi dimulai dengan menggunakan Area Under Curve (AUC) sebagai metrik untuk mengukur akurasi dari ketiga algoritma yang berbeda. Logistic Regression, KNN, dan Naïve Bayes, yang memiliki tingkat yang berbeda. Algoritma Logistic Regression dan KNN menunjukkan peningkatan akurasi, sedangkan Naïve Bayes tidak mengalami peningkatan.



Sumber : Peneliti (2024)

Gambar 12. Nilai AUC dari ketiga algoritma

Pada Gambar 12 Nilai AUC dari ketiga algoritma ditampilkan, dengan rentang nilai antara 0 dan 1. Sebuah nilai 1 menunjukkan bahwa model memiliki kemampuan sempurna dalam membedakan kelas, sedangkan nilai 0,5 menunjukkan kinerja yang setara dengan pengacakan kelas.

Dibawah ini merupakan hasil perbandingan dari ketiga model algoritma yang dimanfaatkan. Tabel 1 mencerminkan kinerja yang dihasilkan oleh ketiga algoritma, di mana algoritma K-Nearest Neighbor menunjukkan akurasi tertinggi, mencapai 91%, sementara metode Naïve Bayes mencatat akurasi terendah, yaitu 83%.

Tabel 1. Hasil Perbandingan Kinerja Algoritma

Parameter Kinerja	**Logistic Regression**	**K-Nearest Neighbor**	**Naïve Bayes**
Accuracy	0.85	0.91	0.83
Precision	0.83	1	0.82
Recall	0.91	0.82	0.88
AUC	0.91	0.91	0.83

Sumber: Peneliti (2024)

KESIMPULAN

Kesimpulan yang diperoleh dari penelitian yang telah dilakukan dengan menggunakan data pasien penyakit jantung dari data *public* adalah algoritma yang tertinggi dengan menggunakan metode *k-nearest neighbour* (KNN) yaitu akurasi sebesar 91%, sedangkan dengan algoritma *logistic regression* akurasinya sebesar 85%, dan *naive bayes* akurasinya sebesar 83%.

Oleh karena itu, kesimpulan yang dapat diambil adalah bahwa algoritma k-nearest neighbour (KNN) menghasilkan akurasi tertinggi, mencapai 91%, diikuti oleh Logistic Regression dan Naive Bayes.

REFERENSI

analytics vidhya. (2022). *No Title*.

Annisa, R., Studi, P., Informasi, S., Kampus, A., Pontianak, K., Teknologi, F., Bina, U., Informatika, S., Barat, K., & Forest, R. (2019). *ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING*. *3*(1).

Bimo, P., Setio, N., Retno, D., Saputro, S., & Winarno, B. (2020). *Klasifikasi dengan Pohon Keputusan Berbasis Algoritme*. *3*, 64–71.

Caballe, S. (2019). *No Title*. Web Page.

Cnn, D. M., Arsal, M., Agus, B., & Anggraini, D. (2020). *Jurnal Nasional Teknologi dan Sistem Informasi Face Recognition Untuk Akses Pegawai Bank Menggunakan Deep Learning*. *01*, 55–63.

Damuri, A., Riyanto, U., Rusdianto, H., & Aminudin, M. (2021). *Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako*. *8*(6), 219–225. https://doi.org/10.30865/jurikom.v8i6.3655

DataIndonesia.id. (2023). *No Title*. https://dataindonesia.id/kesehatan/detail/kematian-akibat-penyakit-jantung-di-indonesia-terus-meningkat

dqlab.id. (2021). *No Title*.

Informatika, J., & Informasi, S. (2020). *INFORMASI (Jurnal Informatika dan Sistem Informasi) Volume 12 No.1 / Mei/ 2020*. *12*(1), 67–80.

Journal, C., Bianto, M. A., Informatika, M. T., & Yogayakarta, U. A. (2019). *Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes*. *6*(1).

Mining, D. (2020). *No Title*.

Musu, W., Ibrahim, A., Studi, P., Informatika, T., Makassar, U. D., Studi, P., Informatika, M., & Makassar, U. D. (2021). Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4 . 5. *PROSIDING SEMINAR ILMIAH SISTEM INFORMASI DAN TEKNOLOGI INFORMASI*, *X*(1), 186–195.

Prasetio, R. T., & Ripandi, E. (2019). *Optimasi Klasifikasi Jenis Hutan Menggunakan Deep Learning Berbasis Optimize Selection*. *6*(1), 100–106.

Prasetyo, R., Nawawi, I., & Fauzi, A. (2021). *Komparasi Algoritma Logistic Regression dan Random Forest pada Prediksi Cacat Software*. *06*(Siringoringo 2017), 275–281.

Pratama, R. R. (2020). *Analisis Model Machine Learning Terhadap Pengenalan Aktifitas Manusia*. *19*(2), 302–311.

Puspita, R., & Widodo, A. (2021). *Perbandingan Metode KNN , Decision Tree , dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS*. *5*(4), 646–654.

Putry, N. M., Sari, B. N., Kom, M., Informatika, T., & Karawang, U. S. (2022). *KOMPARASI ALGORITMA KNN DAN NAÏVE BAYES UNTUK KLASIFIKASI DIAGNOSIS PENYAKIT DIABETES MELITUS*. *10*(1).

Reviantika, F., Informatika, J. T., Malang, U. M., Azhar, Y., Informatika, J. T., Malang, U. M., Marthasari, G. I., Informatika, J. T., & Malang, U. M. (2021). *Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression*. *04*(03), 155–160.

Rizki, M., Basuki, S., & Azhar, Y. (2020). *Implementasi Deep Learning Menggunakan Arsitektur Long Short Term Memory Untuk Prediksi Curah Hujan Kota Malang*. *2*(3), 331–338.

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). *Pemanfaatan Machine Learning dalam Berbagai Bidang : Review paper*. *5*(April), 75–82.

suprianto, dodit. (2020). *No Title*.

UC irvine Machine Learning Repostory. (1988). *No Title*.

Utomo, D. P. (2020). *Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung*. *4*(April), 437–444. https://doi.org/10.30865/mib.v4i2.2080

Watratan, A. F., B, A. P., Moeis, D., Informasi, S., & Makassar, S. P. (2020). *JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST) Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia*. *1*(1), 7–14.