

ANALISA PERBANDINGAN KLASIFIKASI DATA MINING MENGGUNAKAN ALGORITMA NAÏVE BAYES, K-NEAREST NEIGHBOR (K-NN) DAN RANDOM FOREST UNTUK DIAGNOSIS PENYAKIT JANTUNG

Siti Meilianawati^{*1}, Agung Mulyo Widodo²

^{1,2} Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Ilmu Komputer
Jl. Arjuna Utara No. 9, RT.1/RW.2, Duri Kepa, Kec. Kb. Jeruk,
Kota Jarta Barat, Daerah Khusus Ibukota Jakarta 11510
¹sitmeilia@gmail.com, ²agung.mulyo@esaunggul.ac.id

Abstrak

Salah satu organ terpenting dalam sistem peredaran darah manusia adalah jantung. Penyakit jantung merupakan salah satu penyebab utama kematian terbanyak di Indonesia maupun di belahan dunia lainnya. Organisasi Kesehatan Dunia (WHO) melaporkan pada tahun 2019 bahwa kematian akibat penyakit jantung mencapai sekitar 31 persen dari kematian global. Tujuan penelitian ini untuk membandingkan tingkat akurasi tiga algoritma yaitu algoritma *Naïve Bayes*, *K-NN* dan *Random Forest*, mana di antara ketiga algoritma tersebut yang dinilai memiliki tingkat akurasi atau kemiripan yang tinggi terhadap dataset. Parameter yang digunakan untuk membandingkan ketiga algoritma tersebut merupakan hasil pengukuran kinerja (akurasi, presisi, *recall* dan *f1 score*). Dataset dibagi menjadi dua bagian yaitu 700 data training dan 176 data testing. Parameter klasifikasi terdiri *age*, *sex*, *chestpaintype*, *restingbp*, *cholesterol*, *fastingbs*, *restingecg*, *maxhr*, *exerciseangina*, *oldpeak*, *st_slope*, serta kelasnya *HeartDisease*. Hasil klasifikasi dengan algoritma *K-Nearest Neighbor* memiliki tingkat akurasi terbaik sebesar 86% dengan presisi kelas '1' adalah 86%, dan presisi kelas '0' adalah 86%, *recall* kelas '1' adalah 88%, *recall* kelas '0' adalah 83%, *f1 score* kelas '0' sebesar 84% dan *f1 score* kelas '1' sebesar 87%.

Kata kunci: Penyakit Jantung, Algoritma, Dataset.

Abstract

One of the most important organs in the human circulatory system is the heart. Heart disease is one of the leading causes of death in Indonesia and other parts of the world. The World Health Organization (WHO) reported in 2019 that deaths from heart disease accounted for about 31 percent of global deaths. The purpose of this study is to compare the accuracy of three algorithms, namely the Naïve Bayes, K-NN and Random Forest algorithms, which of the three algorithms are considered to have a high level of accuracy or similarity to the dataset. The parameters used to compare the three algorithms are the results of performance measurements (accuracy, recall and precision). The dataset is divided into two parts, 700 training data and 176 testing data. Classification parameters consist of age, sex, chestpaintype, restingbp, cholesterol, fastingbs, restingecg, maxhr, exerciseangina, oldpeak, st_slope, and the class HeartDisease. The classification results with the K-Nearest Neighbor algorithm have the best accuracy rate of 86% with class '1' precision is 86%, and class '0' precision is 86%, class '1' recall is 88%, class '0' recall is 83%, class '0' f1 score is 84% and class '1' f1 score is 87%.

Keywords: Heart Disease, Algorithm, Data.

1. PENDAHULUAN

Salah satu organ terpenting dalam sistem peredaran darah manusia adalah jantung. Sistem ini memiliki dua fungsi utama yaitu mengedarkan oksigen dan nutrisi ke seluruh tubuh melalui organ tubuh serta mengangkut produk sisa metabolisme. Jantung bertindak sebagai pompa, mengalirkan darah ke seluruh tubuh manusia. Jika jantung mengalami gangguan atau kerusakan, maka akan mengakibatkan gangguan pada seluruh aktivitas organ dalam tubuh manusia [1].

Penyakit jantung merupakan salah satu penyebab utama kematian terbanyak di Indonesia maupun di belahan dunia lainnya. Organisasi Kesehatan Dunia (WHO) melaporkan pada tahun 2019 bahwa kematian akibat penyakit jantung mencapai sekitar 31 persen dari kematian global [2]. Jumlah ini diperkirakan akan terus meningkat hingga mencapai 23,3 juta orang pada tahun 2030 [3]. Berdasarkan data Riset Kesehatan Dasar (Riskesdas) pada tahun 2018 tercatat 1,5% penyakit jantung sesuai diagnosis dokter di Indonesia (Fabiana Meijon Fadul, 2019). Penyakit jantung berkaitan erat dengan faktor risiko seperti tekanan darah tinggi, kolesterol darah, gula darah, asam urat, usia, gaya hidup dan lain-lain [1].

Sebagian besar pasien penyakit jantung tidak menyadari gejala awalnya, dan sejumlah besar orang dengan penyakit jantung meninggal akibat serangan jantung. Kesadaran akan pola hidup sehat dan kekurangan penyakit jantung masih kurang, sehingga masyarakat belum bisa mengenali gejala awalnya. Deteksi dini penyakit jantung memudahkan pemulihan dan mencegah akibat yang lebih fatal. Menentukan kerentanan terhadap penyakit jantung adalah langkah pertama dalam mencegah penyakit jantung yang fatal. Menentukan kerentanan penyakit jantung didasarkan pada informasi tentang faktor risiko seseorang terhadap penyakit jantung [1].

Untuk mendeteksi penyakit jantung dapat dilakukan secara manual yaitu dengan konsultasi dokter spesialis jantung secara langsung dan melakukan beberapa pemeriksaan laboratorium yang kemudian harus dirujuk oleh dokter spesialis jantung tersebut. Hal ini tentunya membutuhkan biaya yang relatif tinggi. Dengan resiko kematian yang sangat tinggi, maka diperlukan suatu sistem yang dapat mendeteksi penyakit jantung pada pasien secara akurat dan biaya yang tidak besar [1].

Perkembangan digitalisasi dalam kehidupan sehari-hari semakin terus berkembang. Dalam bidang pengolahan informasi dan teknologi informasi, terdapat banyak metode, model, dan algoritma yang dapat mempermudah atau mendukung pekerjaan manusia. Hal ini mendorong banyak penelitian tentang penyakit jantung yang salah satunya menggunakan metode komputerisasi. Metode ini dikembangkan dengan menggunakan komputer cerdas yang mampu menangani data dalam jumlah besar [1].

Data mining dan machine learning adalah metode pemrosesan data yang banyak digunakan di berbagai bidang kehidupan. Klasifikasi, clustering, prediksi, peramalan dan asosiasi merupakan fungsi dari data mining. Fungsi klasifikasi adalah fungsi data mining yang akan mengelompokkan sekelompok data ke dalam label tertentu [5]. Proses mengklasifikasikan data masukan ke dalam label yang telah ditentukan dapat menggunakan algoritma machine learning untuk memberikan hasil yang cepat dan akurat. Data masukan dalam proses klasifikasi diubah menjadi data latih dan data uji. Metode klasifikasi yang umum digunakan antara lain Naïve Bayes [6], K-Nearest Neighbors [7], Decision Trees Random Forests [1], dan Support Vector Machines [8].

Penerapan fungsi klasifikasi cukup besar dan hampir di setiap bidang. Fungsi klasifikasi digunakan untuk memprediksi kepuasan pelanggan [9].

Klasifikasi telah banyak diimplementasikan dalam industri medis seperti klasifikasi tekanan darah dengan menggunakan metode modular neural network [10]. Implementasi klasifikasi dengan machine learning untuk diagnosis kanker [11]. Beberapa algoritma machine learning digunakan untuk klasifikasi pasien diabetes [12]. Selain itu algoritma machine learning dan data mining juga dikembangkan dalam bentuk system pakar untuk deteksi penyakit gigi dengan metode Bayes [13], machine learning juga digunakan untuk deteksi awal penyakit tulang belakang[14].

Pada penelitian ini menggunakan model komparatif untuk membandingkan tiga algoritma pembelajaran mesin yang digunakan untuk mengklasifikasikan kerentanan penyakit jantung. Klasifikasi kerentanan penyakit jantung dihasilkan menggunakan tiga algoritma pembelajaran mesin yaitu Naive Bayes, K-Nearest Neighbor (KNN) dan Random Forest. Algoritma Naive Bayes, K-Nearest, dan Random Forest masing-masing memiliki kelebihan dan kekurangan. Oleh karena itu, penelitian ini membuat perbandingan antara ketiga algoritma untuk menentukan algoritma yang paling cocok untuk mengklasifikasikan data penyakit jantung. Parameter yang digunakan untuk membandingkan ketiga algoritma tersebut merupakan hasil pengukuran kinerja (akurasi, presisi, , *recall*, *f1 score*).

2. METODE PENELITIAN

2.1 Pengumpulan Data

Informasi diperoleh melalui teknik pengumpulan data sekunder. Data berasal dari situs web Kaggle Dataset yang dapat diunduh <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>. Dataset tersebut berisi gejala-gejala pada penyakit jantung. Data yang diambil terdiri dari 918 data pasien jantung, di mana masing-masing data memiliki 12 variabel, terdiri atas 11 variabel independen dan 1 variabel dependen yang

merupakan parameter hasil diagnosis penyakit jantung. Variabel independen sebagai berikut :

1. *Age* : Usia pasien
2. *Sex* : Jenis kelamin pasien, yaitu laki-laki (M : 1) dan perempuan (F : 0)
3. *Chest pain type* : Jenis nyeri dada terdapat 4 kategori yaitu
 - 0 : Asimtomatik (ASY)
 - 1 : Angina Atipikal (ATA)
 - 2 : Nyeri Non Angina (NAP)
 - 3 : Tipikal Angina (TA)
4. *RestingBP* : Tekanan darah pasien ketika dalam keadaan istirahat. Satuan yang dipakai adalah mm Hg.
5. *Cholesterol* : Kadar kolesterol dalam darah pasien, dengan satuan mg/dl.
6. *FastingBS* : Kadar gula darah pasien, atribut fbs ini hanya memiliki 2 nilai yaitu :
 - 1 : jika kadar gula darah pasien lebih dari 120 mg/dl.
 - 0 : jika kadar gula darah pasien kurang dari sama dengan 120 mg/dl.
7. *Resting ECG* : Hasil elektrokardiografi terdapat 3 kategori yaitu :
 - 0 : Hipertrovi ventrikel (LVH) keadaan dimana ventricular kiri mengalami hipertropi
 - 1 : Normal
 - 2 : Memiliki kelainan (ST) yaitu keadaan dimana gelombang inversions T dan atau ST meningkat maupun menurun lebih dari 0,5 mV
8. *MaxHR* : Rata-rata detak jantung maksimum dalam satu menit.
9. *ExerciseAngina* : Keadaan dimana pasien akan mengalami nyeri dada apabila berolah raga, mempunyai 2 nilai yaitu :
 - 0 : jika tidak nyeri (No).
 - 1 jika menyebabkan nyeri (Yes).

10. *Oldpeak* : Penurunan segmen *Streets Testing* (ST) akibat olahraga.
11. *STSlope* : Kemiringan puncak latihan segmen *Streets Testing* (ST).

Dalam dataset ini, yang merupakan variabel dependen adalah *HeartDisease* adanya penyakit jantung atau tidak pada pasien. 0 tidak ada penyakit jantung, 1 sakit jantung.

2.2 Pemrosesan Data

a. Data Cleaning

Tahap preprocessing ini dilakukan untuk menangani data-data yang tidak diperlukan, seperti data yang memiliki missing value, outlier, duplikasi data, maupun imbalance data.

b. Transformasi

Pada tahap transformasi data ini dilakukan perubahan data menjadi format data yang sesuai untuk proses data mining. Salah satu teknik yang dapat digunakan dalam proses transformasi data adalah teknik label encoding.

c. Data Mining

Sebelum masuk ketahap penggunaan algoritma machine learning data terlebih dulu melalui proses training dan testing.

Kemudian tahap selanjutnya penggunaan algoritma machine learning. Penelitian ini menggunakan tiga algoritma yaitu algoritma Naïve Bayes, K-Nearest Neighbor dan Random Forest.

d. Evaluasi

Pada penelitian ini evaluasi dilakukan terhadap model yang dihasilkan berupa *confussion matrix*.

2.3 Naïve Bayes

Naive Bayes adalah pengklasifikasi probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari kumpulan data yang diberikan[15].

Menurut Han and Kamber naive Bayes dikenal mengungguli metode klasifikasi yang paling canggih sekalipun. Pengklasifikasi Naive Bayes menunjukkan akurasi dan kecepatan tinggi saat diterapkan pada database besar [15].

Naive merupakan dasar dari statistika Bayes dan dapat diterapkan dalam banyak bidang seperti sains, rekayasa, ilmu ekonomi, teori permainan kedokteran, hukum, dan lain sebagainya. Naive Bayes memiliki persamaan seperti berikut ini :

1. Menentukan Evidence Tunggal (E) dan hipotesa tunggal (H) seperti pada persamaan 2.1.

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \quad (2.1)$$

2. Mencari Evidence tunggal (E) dan hipotesa ganda (H1, H2,...Hn) seperti pada persamaan 2.2.

$$P(H_i|E) = \frac{P(E|H_i) \times P(H_i)}{\sum_{k=1}^n P(E|H_k) \times P(H_k)} \quad (2.2)$$

2.4 K-Nearest Neighbor

Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap obyek berdasarkan data pembelajaran yang jaraknya paling dekat dengan obyek tersebut [16].

Algoritma ini juga merupakan salah satu teknik lazy learning. KNN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing[17].

Jarak yang digunakan adalah jarak Euclidean. Secara umum untuk mendefinisikan jarak antara dua objek x dan y, digunakan rumus jarak Euclidean :

$$d_{yx} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2.3)$$

2.5 Random Forest

Random Forest dimulai dengan teknik dasar data mining, yaitu pohon keputusan (*decision tree*). Pada proses pohon keputusan (*decision tree*), dimana input berupa informasi dibawa ke atas pohon proses berupa akar pohon (*root*), kemudian diturunkan proses berupa daun pohon . untuk menentukan ke kelas mana data input berada[18].

2.6 Pengukuran Klasifikasi

Menurut Prasetyo suatu sistem klasifikasi diharapkan dapat mengklasifikasikan semua dataset dengan benar, namun tidak dapat dipungkiri bahwa kinerja sistem tersebut 100% benar, sehingga sistem klasifikasi juga harus mengukur kinerjanya. Biasanya, kinerja klasifikasi diukur dengan menggunakan matriks konfusi (*confusion matrix*)[19].

Dapat dituliskan persamaan rumus akurasi, presisi, *recall* dan *f1 score* sebagai berikut :

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (2.4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2.5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2.6)$$

$$\text{F1 Score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}} \quad (2.7)$$

3. HASIL DAN PEMBAHASAN

Penelitian ini diawali dengan mengumpulkan data set penyakit jantung, data set yang digunakan berasal dari data set Kaggle dengan total 918 data dengan 12 variabel yang terdiri dari 11 variabel independen dan 1 target atau variabel dependen. Berdasarkan data yang diperoleh, dibuat perbandingan klasifikasi menggunakan algoritma Naive Bayes, K-Nearest Neighbor dan Random Forest.

Sebelum masuk ke tahap pemodelan data menggunakan algoritma machine learning, data harus melalui tahap preprocessing yaitu dilakukan

pembersihan data untuk menangani data-data yang tidak diperlukan, seperti data yang memiliki missing value, outlier, duplikasi data, maupun imbalance data.

Setelah melalui proses cleansing data di transformasi menjadi numerik agar menjadi format data yang sesuai untuk proses data mining.

Tahap selanjutnya data yang akan diujikan dibagi menjadi dua bagian yaitu training dan testing kemudian dianalisis menggunakan software Google Colaboratory. Pembagian data training dan testing pada penelitian ini data training 80% dan data test 20%. Data training memiliki 700 record data dan data testing memiliki 176 record data.

Kemudian selanjutnya yaitu normalisasi data digunakan untuk menskalakan data suatu atribut sehingga berada dalam rentang yang lebih kecil. Pada penelitian ini diperlukan normalisasi dengan metode *min max scaler*. Dari gambar tabel 3.2 dapat diketahui bahwa atribut - atribut penyakit jantung yang digunakan dalam penelitian mempunyai rentang nilai yang lebih kecil dengan bobot yang sama, sehingga dapat memudahkan meningkatkan kualitas data dan meminimalisir kesalahan dalam proses learning terhadap model pembelajaran. Sehingga data yang telah melewati proses normalisasi mempunyai nilai seperti pada gambar 3.1.

```
array([[0., 0., 1., ..., 0., 0., 1.],
       [0., 0., 1., ..., 0., 1., 0.],
       [0., 1., 0., ..., 0., 0., 1.],
       ...,
       [0., 0., 0., ..., 0., 1., 0.],
       [0., 0., 0., ..., 0., 1., 0.],
       [0., 1., 0., ..., 0., 0., 1.]])
```

Gambar 3.1 Data Setelah Dinormalisasi

Setelah itu tahap selanjutnya pengujian terhadap tiga algoritma, yakni algoritma Naive Bayes, algoritma K Nearest Neighbor dan algoritma Random Forest menghasilkan akurasi yang berbeda-beda terhadap kasus penyakit jantung. Kemudian hasil pengujian tersebut diukur untuk mendapatkan tingkat ketepatan dalam mengklasifikasikan penyakit

jantung koroner dengan menghitung rata-rata recall, presisi, dan akurasi dari setiap percobaan menggunakan data testing. Tabel 1 menunjukkan rangkuman dari hasil performance measure terhadap

seluruh hasil pengujian dalam masing-masing algoritma.

Tabel 4.1 Hasil Performance Measure

Algoritma	Performance Measure						
	Presisi '0'	Presisi '1'	Recall '0'	Recall '1'	F1 Score '0'	F1 Score '1'	Akurasi
Naïve Bayes	76%	87%	86%	77%	81%	82%	81%
KNN	86%	86%	83%	88%	84%	87%	86%
Random Forest	84%	85%	81%	86%	82%	85%	84%

Hasil performance measure pada Tabel 4.1 menunjukkan akurasi paling optimal untuk ketiga algoritma. Berdasarkan hasil perbandingan pada Tabel 4.1 hasil klasifikasi dengan algoritma *K-Nearest Neighbor* memiliki tingkat akurasi terbaik sebesar 86% dengan presisi kelas '0' adalah 86%, presisi kelas '1' adalah 86%, *recall* kelas '0' adalah 83%, *recall* kelas '1' adalah 88%, dan *f1 score* kelas '0' adalah 84%, *f1 score* kelas '1' 87%. Hasil akurasi yang diperoleh algoritma *Random Forest* sebesar 84% dengan presisi kelas '0' adalah 84%, presisi kelas '1' adalah 85%, *recall* kelas '0' adalah 81%, *recall* kelas '1' adalah 86%, dan *f1 score* kelas '0' adalah 82%, *f1 score* kelas '1' 85%. Hasil akurasi algoritma *Naïve Bayes* memiliki tingkat akurasi terkecil namun masih dikategorikan sebagai nilai akurasi yang baik yaitu sebesar 81% dengan presisi kelas '0' adalah 76%, presisi kelas '1' adalah 87%, *recall* kelas '0' adalah 86%, *recall* kelas '1' adalah 77%, dan *f1 score* kelas '0' adalah 81%, *f1 score* kelas '1' 82%.

4. KESIMPULAN DAN SARAN

Berdasarkan hasil pengujian dan analisis yang telah dilakukan peneliti dengan model klasifikasi kerentanan penyakit jantung yang dibagi menjadi 700 data *training* dan 176 data *testing*

menggunakan 3 algoritma *machine learning* yaitu *Naïve Bayes*, *K-Nearest Neighbor (KNN)* dan *Random Forest*, maka dapat diambil kesimpulan Berdasarkan perbandingan hasil uji coba, performance measure algoritma *K-Nearest Neighbor* memiliki hasil yang lebih baik dibanding dengan algoritma *Naïve Bayes* dan *Random Forest*. Algoritma *K-Nearest Neighbor* dapat memberikan hasil akurasi sebesar 86% dengan presisi kelas '0' sebesar 86% dan presisi kelas '1' sebesar 86%, *recall* kelas '0' sebesar 83% dan *recall* kelas '1' sebesar 88%, *f1 score* kelas '0' sebesar 84% dan *f1 score* kelas '1' sebesar 87%.

Pada penelitian ini dilakukan uji coba menggunakan data dengan variabel yang umum digunakan, sehingga pada penelitian selanjutnya dapat dikembangkan dengan melakukan uji coba pada dataset dengan variabel yang berbeda. Pada penelitian selanjutnya juga dapat menambahkan algoritma klasifikasi lainnya sehingga mendapat perbandingan yang lebih beragam.

5. REFERENSI

- [1] A. B. Wibisono and A. Fahrurrozi, "Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner," *J. Ilm. Teknol. dan Rekayasa*, vol.

- 24, no. 3, pp. 161–170, 2019, doi: 10.35760/tr.2019.v24i3.2393.
- [2] W. H. Statistics and Overview, “World Health Statistics,” pp. 1–16, 2019.
- [3] M. A. Bianto, K. Kusriani, and S. Sudarmawan, “Perancangan Sistem Klasifikasi Penyakit Jantung Menggunakan Naïve Bayes,” *Creat. Inf. Technol. J.*, vol. 6, no. 1, p. 75, 2020, doi: 10.24076/citec.2019v6i1.231.
- [4] Fabiana Meijon Fadul, “Riset Kesehatan Dasar (Riskesdas),” 2019.
- [5] I. G. I. S. Amna, Wahyuddin S, *Data Mining. Global Eksekutif Teknologi*. [Online]. Available: https://www.google.co.id/books/edition/Data_Mining/xmqvEAAAQBAJ?hl=id&gbpv=1&dq=fungsi+klasifikasi+data+mining&pg=PA32&printsec=frontcover
- [6] D. Sartika and D. I. Sensuse, “Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian,” *Jatiji*, vol. 1, no. 2, pp. 151–161, 2017.
- [7] Hasran, “Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor,” *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 1–4, 2020.
- [8] A. M. Puspitasari, D. E. Ratnawati, and A. W. Widodo, “Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 2, pp. 802–810, 2018, [Online]. Available: <http://j-ptiik.ub.ac.id>
- [9] Fattya Ariani and Andi Taufik, “Perbandingan Metode Klasifikasi Data Mining untuk Prediksi Tingkat Kepuasan Pelanggan Telkomsel Prabayar,” *SATIN - Sains dan Teknol. Inf.*, vol. 6, no. 2, pp. 46–55, 2020, doi: 10.33372/stn.v6i2.666.
- [10] M. Pulido, P. Melin, and G. Prado-Arechiga, “Blood pressure classification using the method of the modular neural networks,” *Int. J. Hypertens.*, vol. 2019, 2019, doi: 10.1155/2019/7320365.
- [11] C. J. Harrison and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction to natural language processing,” *BMC Med. Res. Methodol.*, vol. 21, no. 1, pp. 1–18, 2021, doi: 10.1186/s12874-021-01347-1.
- [12] Y. Shang *et al.*, “The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. Suppl 2, pp. 1–11, 2021, doi: 10.1186/s12911-021-01423-y.
- [13] F. Hadi and Y. Diana, “Sistem Pakar Diagnosa Penyakit Gigi dengan Metode Bayes,” *SATIN - Sains dan Teknol. Inf.*, vol. 5, no. 2, pp. 44–51, 2019, doi: 10.33372/stn.v5i2.513.
- [14] C. Hayat and Angelina A Latuny, “Rancang Bangun Aplikasi Informasi Awal Penyakit Tulang Belakang dengan Metode Forward Chaining,” *SATIN - Sains dan Teknol. Inf.*, vol. 6, no. 1, pp. 89–97, 2020, doi: 10.33372/stn.v6i1.617.
- [15] Mustika; and Y. dkk Ardhilla, “DATA MINING DAN APLIKASINYA - Google Books,” *Widina Bhakti Persada Bandung*. pp. 3–243, 2021. [Online]. Available: [https://www.google.co.id/books/edition/DATA_MINING_DAN_APLIKASINYA/53FXEAAAQBAJ?hl=id&gbpv=1&dq=Data+mining,+sering+juga+disebut+Knowledge+Discovery+In+Database+\(KDD\),+adalah+kegiatan+yang+mencakupi+pengumpulan,+pemakaian+data+historis+untuk+menemukan+kete](https://www.google.co.id/books/edition/DATA_MINING_DAN_APLIKASINYA/53FXEAAAQBAJ?hl=id&gbpv=1&dq=Data+mining,+sering+juga+disebut+Knowledge+Discovery+In+Database+(KDD),+adalah+kegiatan+yang+mencakupi+pengumpulan,+pemakaian+data+historis+untuk+menemukan+kete)

- [16] G. Setiawan, "Algoritma KNN," *Gastron. ecuatoriana y Tur. local.*, vol. 1, no. 69, pp. 5–24, 2019.
- [17] Suparyanto dan Rosad (2015, "Algoritma K-Nearest Neighbor (KNN)," *Suparyanto dan Rosad (2015*, vol. 5, no. 3, pp. 248–253, 2020.
- [18] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis J. Ilm. Ekon. dan Bisnis*, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-bisnis.v13i2.247.
- [19] A. Heriyanto, "Penerapan Metode K-Nearest Neighbor (KNN) Untuk Klasifikasi Stanting Pada Balita," *Publ. Ilm. Univ. Muhammadiyah Jember*, 2021.