

## KLASIFIKASI ALGORITMA *NAIVE BAYES* DAN *K-NEAREST NEIGHBOR* PADA PENDERITA DIABETES

Feri Irawan<sup>1)</sup>, Tati Suprapti<sup>2)</sup>, Agus Bahtiar<sup>3)</sup>

<sup>1,2,3)</sup> Teknik Informatika, STMIK IKMI Cirebon, Kota Cirebon, Jawa Barat  
E-mail: <sup>1)</sup>feriirawann110@gmail.com, <sup>2)</sup>tatisuprapti112004@gmail.com, <sup>3)</sup>agusbahtiar038@gmail.com

### ABSTRAK

Diabetes Melitus merupakan penyakit kronis yang disebabkan oleh gangguan metabolisme dimana glukosa tidak dapat digunakan dengan baik dan ditandai dengan hiperglikemia. Diabetes adalah penyakit yang terjadi ketika gula darah terlalu tinggi dan tubuh tidak lagi merespon hormon insulin dampak yang ditimbulkan jika penyakit diabetes tidak ditangani dengan cepat dapat menimbulkan komplikasi hingga kematian. Tujuan penelitian ini adalah untuk memprediksi dengan menggunakan klasifikasi untuk menentukan algoritma apa yang cocok dalam mendiagnosa penyakit diabetes. Data yang diambil merupakan data dari *source kaggle* yang bersifat publik. Pada penelitian ini menggunakan aplikasi *Orange* dalam pengklasifikasiannya. Penelitian ini menggunakan dua algoritma *Naive Bayes* dan *K-Nearest Neighbor*. Hasil yang diperoleh dari klasifikasi *Naive Bayes accuracy* sebesar 76.6 *precision* sebesar 76.8, *recall* sebesar 76.7 sedangkan *K-Nearest Neighbor* mendapatkan *accuracy* sebesar 92.6, *Precision* sebesar 92.6, *recall* sebesar 92.6.

**Kata kunci** : diabetes, klasifikasi, *Naive Bayes*, *K-Neares Neighbor*

### ABSTRACT

*Diabetes Mellitus is a chronic disease caused by metabolic disorders in which glucose cannot be used properly and is characterized by hyperglycemia. Diabetes is a disease that occurs when blood sugar is too high and the body no longer responds to the insulin hormone. The impact if diabetes is not treated quickly can cause complications and even death. The purpose of this study is to predict using classification to determine which algorithm is suitable for diagnosing diabetes. The data taken is data from the public Kaggle source. In this study using the application of orange in its classification. This study uses two algorithms Naive Bayes and K-Nearest Neighbor. The results obtained from the classification of Naive Bayes accuracy of 76.6 precision of 76.8, recall of 76.7 while K-Nearest Neighbor gets accuracy of 92.6, Precision of 92.6, recall of 92.6. Based on the classification results above, the K-Nearest Neighbor algorithm has the better results in diagnosing diabetes.*

**Keywords:** *diabetes, classification, Naive Bayes, K-Neares Neighbor*

### 1. PENDAHULUAN

Diabetes melitus (DM) merupakan penyakit gangguan metabolisme kronis, dimana glukosa darah tidak dapat digunakan dengan baik, sehingga menyebabkan keadaan hiperglikimia[1]. Diabetes melitus (DM) merupakan kelainan endokrin yang terbanyak di

jumpai. Terdapat banyak faktor penyebab resiko terjadinya penyakit (DM) tersebut, diantaranya adalah pengaruh beberapa hormon didalam tubuh yang dapat mengganggu kerja dari insulin. Diabetes melitus adalah penyakit kronis yang disebabkan oleh kadar glukosa yang terlalu tinggi mengakibatkan kekurangan insulin. Diabetes adalah masalah kesehatan di seluruh

dunia yang mempengaruhi sekitar 120 juta penderita. Angka ini akan meningkat ketika masyarakat tidak mengetahui faktor-faktor yang dapat menyebabkan penyakit diabetes [2].

Berdasarkan etiologinya, diabetes dibagi menjadi tiga jenis: diabetes tipe 1, diabetes tipe 2, dan diabetes gestasional. Diabetes tipe 1 disebabkan oleh kerusakan autoimun pada sel-sel di pankreas, yang menyebabkan hilangnya produksi insulin. Sementara itu, diabetes tipe 2 menyebabkan resistensi insulin, yaitu ketidakmampuan tubuh manusia untuk merespon hormon insulin [3]. Diabetes tipe 3 atau diabetes gestasional merupakan salah satu jenis diabetes yang terjadi akibat oleh perubahan hormon pada saat kehamilan [4].

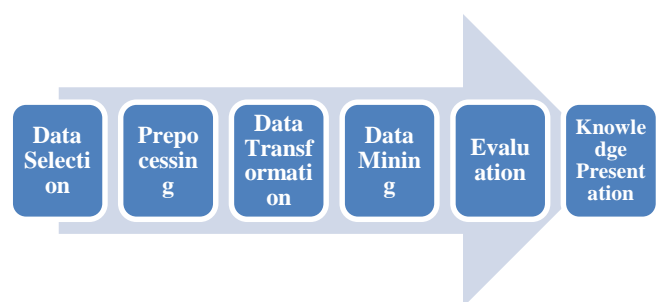
Diabetes adalah salah satu penyebab utama dari 10 kematian di seluruh dunia dan meningkat setiap tahun. Menurut International Diabetes Federation 2021, jumlah penderita diabetes di dunia mencapai 537 juta pada tahun 2021, dan diperkirakan akan meningkat menjadi 643 juta pada tahun 2030 dan 783 juta pada tahun 2045 [5]. Penyakit diabetes melitus adalah penyakit yang memiliki komplikasi tinggi, identifikasi awal merupakan satu-satunya cara untuk menghindari komplikasi [6]. Adapun faktor penyebab terjadinya diabetes yaitu kadar glukosa darah yang tinggi berinteraksi dengan faktor lain, seperti usia tua, obesitas fisik yang rendah dan penyakit lainnya, dapat memodulasi respons imun dan inflamasi [7]. Oleh karena itu, penanganan yang tepat dan optimal faktor risiko terjadinya obesitas, dengan menjaga pola makan dan aktivitas fisik dan olahraga secara teratur dapat membantu penggunaan insulin oleh tubuh dan menghindari obesitas [8].

Selain itu, hal yang harus diwaspadai dari diabetes melitus adalah bahaya komplikasi yang timbul jika diabetes melitus tidak terkontrol [9]. Diagnosis dini menjadi titik awal untuk mencegah terjadinya diabetes melalui pola hidup sehat, sehingga tidak muncul komplikasi. Jika seseorang yang menderita diabetes tidak terdiagnosis dan tidak diobati dalam waktu yang lama, kemungkinan besar kesehatan tubuhnya akan memburuk menyebabkan komplikasi hingga kematian [10].

Dengan melihat permasalahan di atas diperlukannya penanganan untuk diagnosa penyakit diabetes dengan cepat dengan mengklasifikasikan agar dapat menggambarkan dan membedakan kelas data dan konsep sehingga model dapat digunakan untuk memprediksi diabetes dengan menggunakan algoritma *Naïve Bayes* dan *K-Nearest Neighbor*. Dilihat dari tingkat akurasi penelitian terdahulu, *K-Nearest Neighbor* memiliki tingkat akurasi sebesar 77% dan *Naïve Bayes* memiliki tingkat akurasi sebesar 89%. Tujuan peneliti ini untuk mengklasifikasikan data diabetes menggunakan kedua algoritma tersebut mana yang lebih baik dan akurat menggunakan *machine learning* [11].

## 2. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini menggunakan proses *Knowledge Discovery Database (KDD)*, yang bertujuan untuk menggali dan menganalisis data yang sangat besar untuk mendapatkan informasi yang bermanfaat. Dari data yang akan dilakukan klasifikasi sebanyak 2000 data, terdiri dari 9 atribut sebagai berikut ; kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, berat badan, peluang mengalami diabetes, umur, hasil terkena diabetes atau tidak. Adapun tahapan-tahapan KDD yang akan dilakukan dapat dilihat pada Gambar 1 :



Gambar 1. Tahapan KDD *Data Mining*

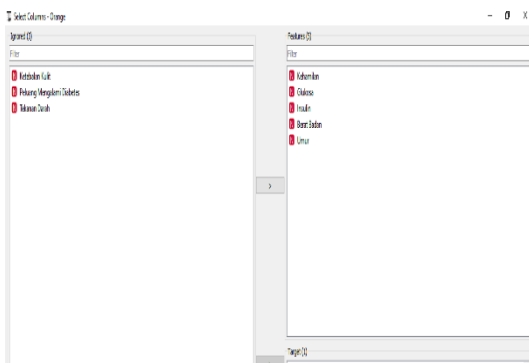
### 2.1. Data Selection

Kumpulan data diagnosa penyakit diabetes didapatkan dari *source kaggle* yang akan diteliti berjumlah 2000 data, terdiri dari 9 atribut sebagai berikut : kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, berat

badan, peluang mengalami diabetes, umur, hasil diabetes atau tidak.

## 2.2. Preprocessing

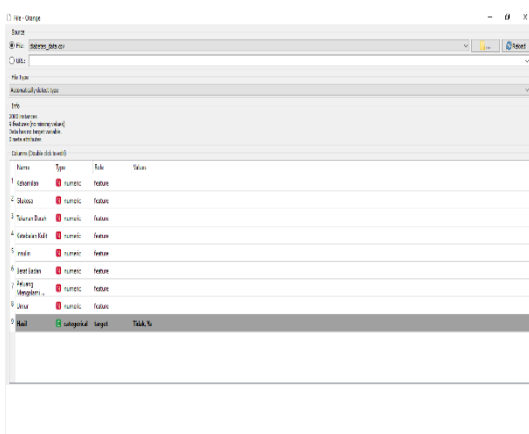
Proses ini dilakukan untuk data diabetes agar terhindar *missing value* pada data yang digunakan, dengan pemilihan atribut, karena nilai yang hilang akan mengganggu proses klasifikasi [12]. Dalam tahapan ini pemilihan atribut dilakukan berdasarkan faktor yang mempengaruhi terkena diabetes dan *widget* yang digunakan dalam pemilihan atribut yaitu *select columns* dilihat pada Gambar 2.



Gambar 2. Preprocessing

## 2.3. Transformation

Tahapan ini dilakukan setelah tahapan transformasi sesuai dengan data pengelompokan menjadi dua kategori variabel yaitu variabel prediktor terdiri 8 atribut yaitu dari kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, berat badan, peluang mengalami diabetes, umur dan variabel target yaitu terkena diabetes atau tidak dilihat pada Gambar 3.



Gambar 3. Transformation

## 2.4. Data Mining

*Data mining* adalah semi otomatis yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstrak dan mengidentifikasi informasi potensial yang berguna dan informasi berguna yang disimpan dalam *database* yang jumlahnya lebih banyak [13]. Pada langkah ini *data mining* yang digunakan yaitu *Naive Bayes* dan *K-Nearest Neighbor*.

### a. Naive Bayes

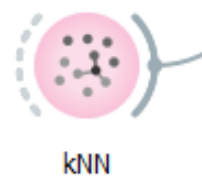
*Naive Bayes* merupakan metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik. Algoritma ini bisa digunakan dalam pengklasifikasian [14]. Widget *Naive Bayes* dilihat pada Gambar 4.



Gambar 4. Naive Bayes

### b. K-Nearest Neighbor

*K-Nearest Neighbor* merupakan metode yang melakukan klasifikasi berdasarkan kedekatan lokasi atau jarak suatu data dengan data yang lain [15]. Widget *K-Nearest Neighbor* dilihat Pada Gambar 5.



Gambar 5. K-Nearest Neighbor

## 2.5. Evaluation

Tahapan ini dilakukan setelah proses tahapan klasifikasi algoritma *Naive Bayes* dan *K-Nearest Neighbor* selesai menghasilkan data yang terkena diabetes dan tidak terkena diabetes kemudian akan dihitung akurasi, presisi dan *recall*-nya dengan menggunakan rumus berikut :

Rumus accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (1)$$

Rumus Precision

$$Precision = \frac{TP}{TP+FP} * 100\% \quad (2)$$

Rumus Recall

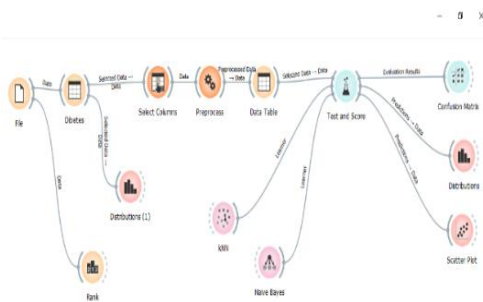
$$Recall = \frac{TP}{TP+FN} * 100\% \quad (3)$$

Keterangan :

- TP : True Positive
- TN : True Negative
- FP : False Positive
- FN : False Negative

## 2.6. Knowledge Presentation

Knowledge Presentation menunjukkan dan menjelaskan tahapan-tahapan dan proses klasifikasi yang dilakukan menggunakan aplikasi Orange menggunakan dua algoritma Naïve Bayes dan K-Nearest Neighbor dapat dilihat pada Gambar 6.



Gambar 6. Model Klasifikasi

## 3. HASIL

Hasil yang diperoleh dari proses penelitian, diantaranya :

### 3.1 Penentuan Kriteria

Dalam menentukan faktor yang mempengaruhi terkena diabetes menggunakan lima kriteria yaitu : glukosa , umur, berat badan, insulin, kehamilan.

Hasil ini di dapatkan dari widget Rank yang lebih mempengaruhi faktor terkena diabetes dilihat pada Gambar 7

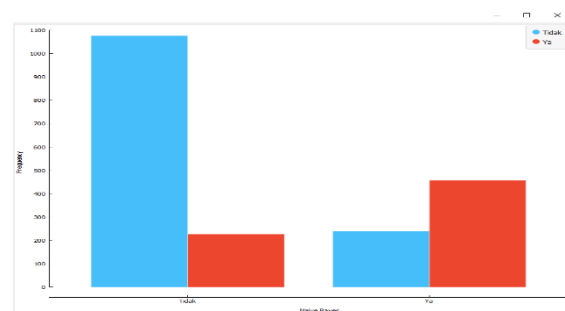
	#	Gain ratio	Gini
1	N Glukosa	0.082	0.097
2	N Umur	0.041	0.049
3	N Berat Badan	0.040	0.044
4	N Insulin	0.028	0.028
5	N Kehamilan	0.021	0.026

Gambar 7. Hasil Widget Rank

Hasil ini di dapatkan dari widget Rank yang lebih mempengaruhi faktor terkena diabetes dilihat pada Gambar 7

### 3.2 Distributions

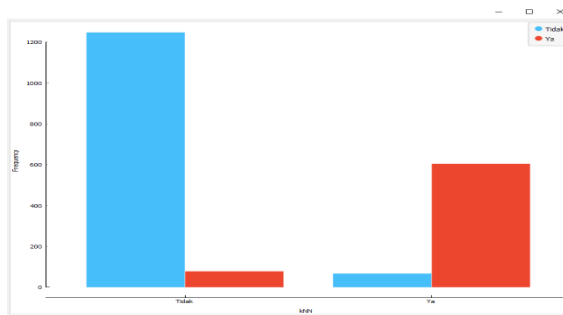
1. Hasil distributions menggunakan algoritma Naive Bayes



Gambar 8. Naïve Bayes

Gambar 8 menunjukkan data yang dinyatakan Negative berjumlah 1.303 setelah melakukan prediksi menyatakan 1.076 Negative, 227 Positive dan yang dinyatakan Positive berjumlah 697 setelah melakukan prediksi menyatakan 240 Negative dan 257 dinyatakan Positive.

2. Hasil *distributions* menggunakan algoritma *K-Nearest Neighbor*



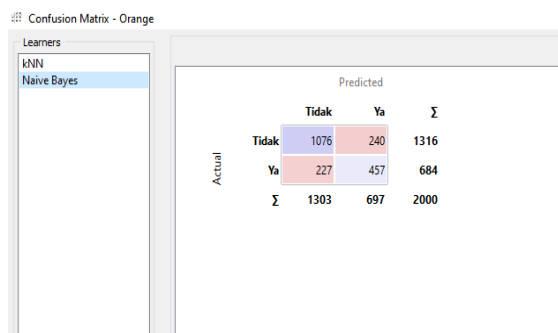
Gambar 9. *K-Nearest Neighbor*

Gambar 9 menunjukkan data yang menyatakan *Negative* berjumlah 1.327 setelah melakukan prediksi menyatakan 1.248 *Negative*, 79 *Positive* dan yang dinyatakan *Positive* berjumlah 673 setelah dilakukan prediksi menyatakan 68 *Negative* dan 605 dinyatakan *Positive*.

3.3 *Confusion Matrix*

*Confusion Matrix* merupakan salah satu cara untuk mengukur akurasi, presisi dan *recall* dari setiap algoritma yang digunakan

1. Hasil *confusion matrix* menggunakan algoritma *Naive Bayes*



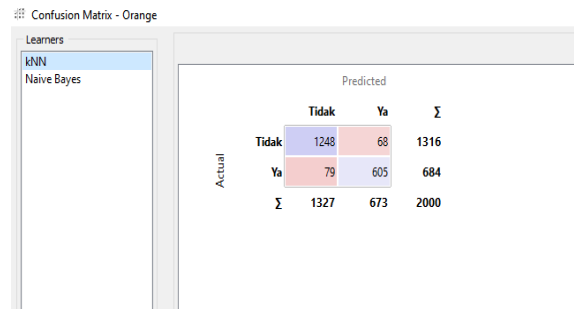
Gambar 10. Hasil *Confusion Matrix* *Naive Bayes*

Berdasarkan Gambar 10 dapat disimpulkan bahwa

**Pertama**, untuk *True Negative* (TN) sebanyak 240 *Treu Positive* (TP) sebanyak 1.076. Dari prediksi yang tidak terkena diabetes 1.316

**Kedua**, untuk *True Negative* (TN) sebanyak 227, *True Positive* (TP) sebanyak 457. Dari prediksi yang terkena diabetes 684

2. Hasil *confusion matrix* menggunakan algoritma *K-Nearest Neighbor*



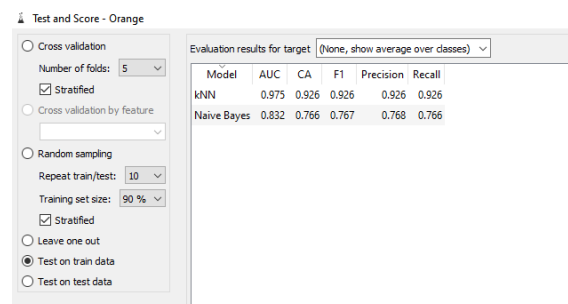
Gambar 11. Hasil *Confusion Matrix* KNN

**Pertama**, untuk *True Negative* (TN) sebanyak 78 *True Positive* (TP) sebanyak 1.248. Dari prediksi yang tidak terkena diabetes 1.316

**Kedua**, untuk *True Negative* (TN) sebanyak 79, *True Positive* (TP) sebanyak 605. Dari prediksi yang terkena diabetes 684

3.4 *Test and Score*

Tahapan *test and score* merupakan hasil setelah diproses klasifikasi dengan menggunakan data training, 1 atribut sebagai target, 5 atribut numerik kehamilan, glukosa, insulin, berat badan, umur maka diperoleh hasil *test and score* pada Gambar 12.



Gambar 12. Hasil *Test and Score*

Berdasarkan hasil perhitungan di atas dengan *confusion matrix* diperoleh nilai perbandingan *Accuracy*, *Precision* dan *Recall*.

**Tabel 1.** Hasil Akurasi

Metode	Accuracy	Precision	Recall
Naive Bayes	76.6	76.8	76.6
KNN	92.6	92.6	92.6

Dilihat dari hasil perbandingan kedua algoritma diatas yaitu *Naive Bayes* dan *K-Nearest Neighbor* yang memiliki *accuracy* tertinggi adalah *K-Nearest Neighbor*.

#### 4. KESIMPULAN DAN SARAN

Dari hasil klasifikasi yang menggunakan algoritma *Naive Bayes* dan *K-Nearest Neighbor* untuk memprediksi penyakit diabetes, *Naive Bayes* menghasilkan *accuracy* 76.6, *precision* 76.8, *recall* 76.6 dan *K-Nearest Neighbor* menghasilkan *accuracy* 92.6, *precision* 92.6 *recall* 92.6. Hasil kedua algoritma diatas dapat disimpulkan bahwa algoritma yang memiliki *accuracy* lebih baik adalah algoritma *K-Nearest Neighbor*. Saran buat peneliti selanjutnya menggunakan metode lain sehingga bisa melihat akurasi mana yang lebih baik lagi.

#### 5. DAFTAR PUSTAKA

[1] L. Adam dan M. B. Tomayahu, "TINGKAT STRES DENGAN KADAR GULA DARAH PADA PASIEN DIABETES MELITUS," *Jambura Health and Sport Journal*, vol. 1, no. 1, hlm. 1–5, Feb 2019, doi: 10.37311/jhsj.v1i1.2047.

[2] S. U. Putri, E. Irawan, dan F. Rizky, "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4. 5," *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, vol. 2, no. 1, hlm. 39–46, 2021.

[3] N. P. Husain, "Best First Feature Selection Dan Radial Basis Function Untuk Klasifikasi Penyakit Diabetes,"

ILTEK : Jurnal Teknologi, vol. 16, no. 01, pp. 29–33, Apr. 2021, doi: 10.47398/iltek.v16i01.41.

[4] P. D. Rinanda, B. Delvika, S. Nurhidayarnis, N. Abror, dan A. Hidayat, "Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil: Comparison of Classification Between Naive Bayes and K-Nearest Neighbor on Diabetes Risk in Pregnant Women," *Malcom: Indonesian Journal of Machine Learning and Computer Science*, vol. 2, no. 2, hlm. 68–75, 2022.

[5] S. Y. Hananto, S. T. Putri, dan A. P. W. Puspita, "Studi Kasus: Penatalaksanaan Diabetes Self Management Education (DSME) terhadap Kadar Glukosa Darah pada Pasien Diabetes Melitus Tipe 2," *Jurnal Keperawatan*, vol. 20, no. 4, hlm. 128–137, 2022.

[6] D. N. Anisa, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma Naive Bayes," *Dinamika Informatika*, vol. 14, no. 1, 2022.

[7] H. E. Ardiani, T. A. E. Permatasari, and S. Sugiatmi, "Obesitas, Pola Diet, dan Aktifitas Fisik dalam Penanganan Diabetes Melitus pada Masa Pandemi Covid-19," *Muhammadiyah Journal of Nutrition and Food Science (MJNF)*, vol. 2, no. 1, p. 1, Jul. 2021, doi: 10.24853/mjnf.2.1.1-12.

[8] A. M. Putri, Y. Hasneli, dan S. Safri, "Faktor-Faktor Yang Mempengaruhi Derajat Keparahan Neuropati Perifer Pada Pasien Diabetes Melitus: Literature Review," *Jurnal Ilmu Keperawatan*, vol. 8, no. 1, hlm. 38–53, 2020.

[9] S. Syaipuddin dan S. Nurbaya, "Pengabdian Masyarakat Melalui Penyuluhan Upaya Pencegahan Penyakit Diabetes Melitus Di Kelurahan Antang

- Kecamatan Manggala Kota Makassar,” *Community Engagement and Emergence Journal (CEEJ)*, vol. 3, no. 1, hlm. 78–82, 2021.
- [10] R. P. Fadhillah, R. Rahma, A. Sepharni, R. Mufidah, B. N. Sari, dan A. Pangestu, “Klasifikasi Penyakit Diabetes Mellitus Berdasarkan Faktor-Faktor Penyebab Diabetes menggunakan Algoritma C4. 5,” *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 7, no. 4, hlm. 1265–1270, 2022.
- [11] N. Nurdiana, S. F. Rodiyansyah, dan A. Algifari, “Studi Komparasi Algoritma ID3 dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus,” *INFOTECH journal*, vol. 6, no. 2, hlm. 18–23, 2020.
- [12] H. Hozairi, A. Anwari, dan S. Alim, “IMPLEMENTASI ORANGE DATA MINING UNTUK KLASIFIKASI KELULUSAN MAHASISWA DENGAN MODEL K-NEAREST NEIGHBOR, DECISION TREE SERTA NAIVE BAYES,” *Network Engineering Research Operation*, vol. 6, no. 2, hlm. 133, Nov 2021, doi: 10.21107/nero.v6i2.237.
- [13] A. Nugraha, O. Nurdiawan, dan G. Dwilestari, “PENERAPAN DATA MINING METODE K-MEANS CLUSTERING UNTUK ANALISA PENJUALAN PADA TOKO YANA SPORT,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 6, no. 2, hlm. 849–855, Nov 2022, doi: 10.36040/jati.v6i2.5755.
- [14] Alvina Felicia Watratan, Arwini Puspita. B, dan Dikwan Moeis, “Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia,” *Journal of Applied Computer Science and Technology*, vol. 1, no. 1, hlm. 7–14, Jul 2020, doi: 10.52158/jacost.v1i1.9.
- [15] I. Yolanda dan H. Fahmi, “Penerapan Data Mining Untuk Prediksi Penjualan Produk Roti Terlaris Pada PT. Nippon Indosari Corpindo Tbk Menggunakan Metode K-Nearest Neighbor,” *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, vol. 3, no. 1.1, hlm. 9–15, 2020.