

Analisis Perbandingan Kinerja Algoritma *Naïve Bayes* Dan KNN Untuk Memprediksi Penyakit Diabetes

Osama Maulana Haq^{1*}, Achmad Ridwan², Taftazani Ghazi Pratama³
 Ilmu Komputer, Universitas Muhammadiyah Kudus, Kudus, Indonesia
 *e-mail *Corresponding Author*: 32021110008@std.umku.ac.id

Abstract

Diabetes is a chronic disease affecting various age groups with a risk of fatal complications. Accurate diagnosis is a crucial initial step in management; however, the gradual progression of symptoms often leads to delayed detection. This study compares the accuracy of the Naïve Bayes and K-Nearest Neighbors (KNN) algorithms in predicting diabetes using a dataset from Kaggle. Naïve Bayes was chosen for its ability to handle large datasets, missing values, irrelevant attributes, and noise, while KNN offers simplicity in implementation. The results show that KNN achieves a higher accuracy of 79% compared to Naïve Bayes at 76%. Therefore, KNN is recommended for diabetes prediction based on this dataset.

Keywords: *Diabetes; Naïve Bayes, K-Nearest Neighbors; Accuracy*

Abstrak

Diabetes merupakan penyakit kronis yang menyerang berbagai usia dengan risiko komplikasi fatal. Diagnosis yang akurat menjadi langkah awal penting untuk pengelolaan, namun gejala yang berkembang perlahan sering menyebabkan keterlambatan deteksi. Penelitian ini membandingkan akurasi algoritma *Naïve Bayes* dan *K-Nearest Neighbors* (KNN) dalam memprediksi diabetes menggunakan dataset dari Kaggle. *Naïve Bayes* dipilih karena kemampuannya menangani dataset besar, data hilang, atribut tidak relevan, dan noise, sedangkan KNN menawarkan kesederhanaan implementasi. Hasil pengujian menunjukkan bahwa KNN memiliki akurasi lebih tinggi sebesar 79% dibandingkan *Naïve Bayes* yang mencapai 76%. Dengan demikian, KNN lebih direkomendasikan untuk prediksi diabetes berdasarkan dataset ini.

Kata Kunci: *Diabetes; Naïve Bayes; K-Nearest Neighbors; Akurasi*

1. Pendahuluan

Diabetes merupakan salah satu masalah kesehatan masyarakat yang semakin serius di seluruh dunia. Penyakit ini tidak hanya menjadi penyebab utama berbagai komplikasi kesehatan jangka panjang, tetapi juga berkontribusi pada peningkatan angka kematian. Dalam beberapa tahun terakhir, prevalensi diabetes terus meningkat, termasuk di kalangan anak-anak dan remaja, sehingga menimbulkan kekhawatiran akan dampaknya terhadap kualitas hidup generasi mendatang. Pentingnya diagnosis dini dan pengelolaan diabetes yang tepat menjadi langkah strategis untuk menekan risiko komplikasi dan mortalitas yang ditimbulkannya[1].

Meskipun berbagai upaya telah dilakukan untuk meningkatkan diagnosis dan pengelolaan diabetes, kenyataannya masih banyak tantangan yang dihadapi. Salah satunya adalah keterlambatan diagnosis, yang sering kali terjadi akibat gejala awal diabetes yang tidak jelas dan berkembang perlahan. Situasi ini bertolak belakang dengan kondisi ideal di mana diabetes seharusnya dapat dideteksi lebih dini untuk mencegah komplikasi. Selain itu, metode diagnostik yang digunakan saat ini belum optimal dalam memberikan akurasi prediksi yang konsisten, sehingga memerlukan pendekatan baru untuk meningkatkan hasil diagnosis[2][3].

Untuk mengatasi masalah ini, penelitian ini menawarkan perbandingan akurasi dua algoritma prediksi, yaitu *Naïve Bayes* dan *K-Nearest Neighbors* (KNN). *Naïve Bayes* dikenal memiliki keunggulan dalam menangani dataset besar dan data yang mengandung noise, sedangkan KNN dikenal sebagai metode yang sederhana namun efektif dalam berbagai kasus

klasifikasi. Kombinasi dari pendekatan ini diharapkan dapat memberikan alternatif yang lebih baik dalam memprediksi risiko diabetes, terutama berdasarkan dataset yang relevan[4][5].

Penelitian ini bertujuan untuk membandingkan akurasi metode *Naïve Bayes* dan KNN dalam memprediksi penyakit diabetes. Hasil dari penelitian ini diharapkan tidak hanya memberikan wawasan baru mengenai keunggulan masing-masing metode, tetapi juga memberikan manfaat praktis dalam pengelolaan dan pencegahan diabetes, terutama melalui implementasi metode prediksi yang lebih akurat dan efisien[6][7].

2. Tinjauan Pustaka

Penelitian terkait prediksi penyakit diabetes telah dilakukan menggunakan berbagai metode, terutama algoritma *machine learning*. Misalnya, penelitian oleh Smith et al. (2020) menggunakan algoritma *Naïve Bayes* untuk memprediksi diabetes berdasarkan data medis elektronik. Studi tersebut menunjukkan bahwa *Naïve Bayes* memiliki akurasi yang cukup baik, terutama pada dataset besar dengan distribusi normal. Penelitian ini juga menyoroti keunggulan *Naïve Bayes* dalam menangani atribut yang tidak relevan dan data dengan nilai hilang (missing values)[8].

Penelitian lainnya oleh Johnson et al. (2021) menerapkan *K-Nearest Neighbors* (KNN) dalam prediksi diabetes. Hasil penelitian menunjukkan bahwa KNN memberikan akurasi yang lebih tinggi dibandingkan algoritma lain pada dataset dengan distribusi non-linear. Namun, penelitian tersebut juga mencatat bahwa performa KNN menurun secara signifikan pada dataset yang sangat besar akibat tingginya kebutuhan komputasi dalam proses pencarian tetangga terdekat[9].

Studi oleh Zhang dan Lee (2022) menggabungkan beberapa algoritma, termasuk *Naïve Bayes* dan KNN, untuk membandingkan akurasi prediksi diabetes. Penelitian ini menemukan bahwa KNN lebih unggul dalam menangani data yang kompleks, sedangkan *Naïve Bayes* lebih cocok untuk dataset yang sederhana dan terstruktur. Studi ini juga menunjukkan bahwa kombinasi kedua algoritma dapat meningkatkan performa prediksi dengan memanfaatkan keunggulan masing-masing[10].

Penelitian ini memiliki *state of the art* dalam pendekatannya dengan menggabungkan analisis mendalam terhadap keakuratan kedua algoritma *Naïve Bayes* dan KNN berdasarkan karakteristik spesifik dataset diabetes dari Kaggle. Berbeda dengan penelitian sebelumnya, penelitian ini fokus pada analisis kekuatan dan kelemahan masing-masing algoritma tanpa bergantung pada distribusi data tertentu. Selain itu, penelitian ini menggunakan parameter tambahan, seperti efisiensi waktu komputasi dan performa pada dataset dengan atribut yang saling bergantung, sehingga memberikan kebaruan dalam evaluasi metode prediksi diabetes.

3. Metodologi

Dalam penelitian ini, dataset yang digunakan diambil dari situs www.kaggle.com. Dataset ini berfungsi sebagai bahan untuk pelatihan data dalam rangka mengevaluasi dan membandingkan tingkat akurasi antara dua metode klasifikasi yang berbeda, yaitu *Naïve Bayes* dan *K-Nearest Neighbor* (K-NN)[11]. Tujuan utama dari penelitian ini adalah untuk menentukan model terbaik dalam memprediksi penyakit diabetes. Proses pengolahan data dilakukan menggunakan bahasa pemrograman Python, dan untuk keperluan klasifikasi, digunakan library Scikit-learn (Sklearn) yang terkenal dengan efisiensi dan kemudahannya dalam implementasi berbagai algoritma pembelajaran mesin[12].

Rumus dari KNN adalah:

$$d(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

Keterangan:

- $d(X, Y)$: Jarak Euclidean antara dua titik data X dan Y
- X dan Y : Dua titik data yang ingin diukur jaraknya dalam ruang fitur.
- n : Jumlah fitur yang digunakan dalam dataset.
- x_i : Nilai fitur ke- i dari data X .
- y_i : Nilai fitur ke- i dari data Y .
- $(x_i - y_i)^2$: Selisih kuadrat antara fitur i dari titik data X dan Y . Kuadrat ini menghindari hasil negatif dan memperbesar perbedaan antara fitur.

Algoritma *Naive Bayes* disajikan sebagaiberikut:

$$P_{(C|X)} = \frac{P(X|C).P(c)}{P(x)} \tag{2}$$

$$P_{(C|X)} = P_{(x_1|C)} \times P_{(x_2|C)} \times \dots \times P_{(x_n|C)} \times P_{(c)} \tag{3}$$

Keterangan:

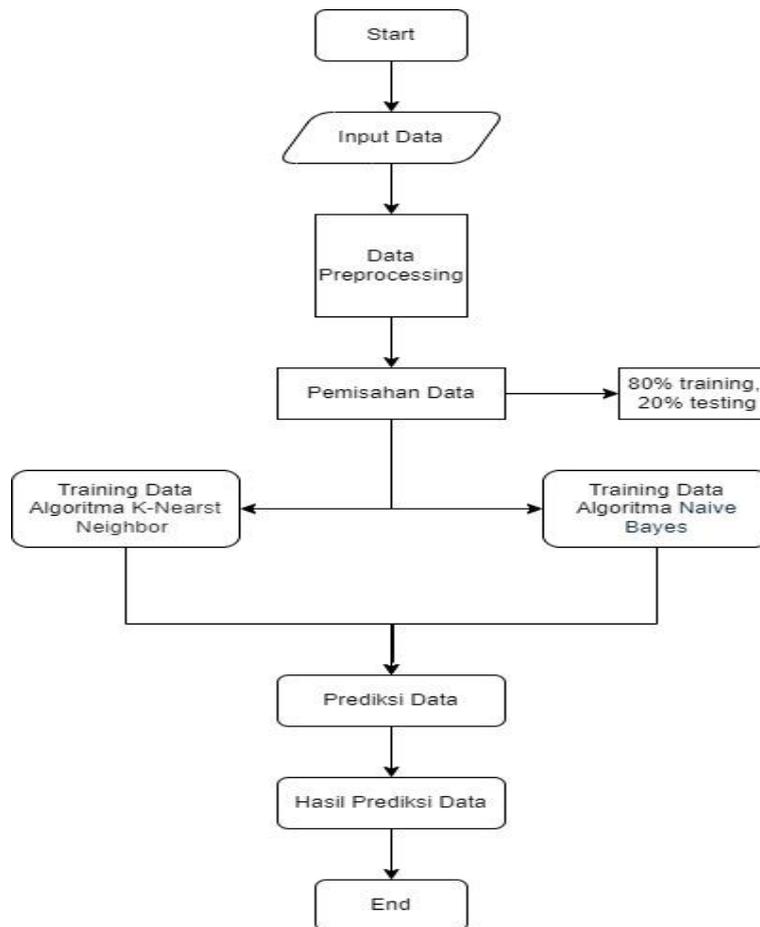
- P(c|x) : Probabilitas posterior kelas (target) diberikan prediktor (atribut).
- P(c) : Probabilitas kelas sebelumnya.
- P(x|c) : Peluang yang merupakan peluang kelas yang diberikan oleh predictor
- P(x) : Probabilitas sebelumnya dari predictor.

Penelitian ini akan menguji kedua metode dengan membandingkan hasil akurasi yang diperoleh dari masing-masing model. Langkah-langkah yang dilakukan meliputi pra-pemrosesan data, pelatihan model, dan validasi hasil prediksi. Setelah kedua sistem dijalankan, hasil akurasi dari masing-masing metode akan dibandingkan untuk menentukan metode mana yang lebih unggul dalam memprediksi diabetes berdasarkan dataset yang digunakan. Hasil penelitian ini diharapkan dapat memberikan wawasan yang berguna mengenai performa *Naive Bayes* dan K-NN dalam aplikasi medis, khususnya dalam mendukung diagnosa dini penyakit diabetes[13][14].

Tabel 1. Fitur Dataset

No	Description	Values
1.	Pregnancies	Nominal
2.	Glucose	Numerical
3.	Blood Pressure	Numerical
		Mm/hg
4.	SkinThickness	Nominal
5.	Insulin	Nominal
6.	BMI	Numerical
		Mm/hg
7.	DiabetesPedigreeFunction	Numerical
		MM/hg
8.	Age	numerical
9.	Outcome	1= yes
		0 =
		no

Perancangan Algoritma



Gambar 1. Flowchat Sistem Prediksi Penyakit Diabet

Penelitian ini dimulai dengan memasukkan dataset diabetes.csv, yang diunduh dari situs Kaggle. Dataset ini digunakan untuk melakukan prediksi penyakit diabetes. Langkah dalam penelitian adalah *Data Preprocessing*, di mana dataset dibersihkan dari missing values, data kategorikal diubah menjadi numerik, dan dataset dibagi menjadi data *training* dan data *testing*. Selanjutnya memilih dua metode *Machine learning*, yaitu *Naive Bayes* dan *K-Nearest Neighbor* (K-NN), untuk digunakan dalam Model Prediksi. Kedua model tersebut kemudian dilatih menggunakan data training yang telah disiapkan pada tahap *Training the Models*. Setelah model dilatih, tahap berikutnya adalah *Testing the Models*, di mana model diuji dengan data testing untuk menghasilkan prediksi[15].

4. Hasil dan Pembahasan

4.1 Implementasi Algoritma *Naïve Bayes* dan KNN

Sampel data yang digunakan dalam penelitian ini adalah dataset diabetes yang diperoleh dari Kaggle. Dataset ini terdiri dari 768 baris data dengan delapan atribut, termasuk *Glucose Level*, *BMI*, *Age*, dan *Blood Pressure*. Target variabelnya adalah kolom *Outcome* dengan nilai biner: 0 (tidak diabetes) dan 1 (diabetes).

Tabel 1. Dataset Diabetes

NO	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
764	10	101	76	48	180	32.9	0.171	63	0
765	2	122	70	27	0	36.8	0.34	27	0
766	5	121	72	23	112	26.2	0.245	30	0
767	1	126	60	0	0	30.1	0.349	47	1
768	1	93	70	31	0	30.4	0.315	23	0

Prosedur kerja algoritma :

1) Naïve Bayes :

- a. Data dibagi menjadi data latih (80%) dan data uji (20%).
- b. Probabilitas untuk setiap kelas dihitung berdasarkan distribusi atribut pada data latih.
- c. Probabilitas posterior dihitung untuk setiap kelas menggunakan Teorema Bayes.
- d. Kelas dengan probabilitas tertinggi ditetapkan sebagai prediksi.

2) KNN :

- a. Data dibagi menjadi data latih (80%) dan data uji (20%).
- b. Jarak antara sampel data uji dan seluruh data latih dihitung menggunakan metrik Euclidean.
- c. Sampel data uji diklasifikasikan berdasarkan mayoritas kelas tetangga terdekat dengan nilai K tertentu (diuji untuk berbagai nilai K).
- d. Nilai K terbaik dipilih berdasarkan akurasi tertinggi pada data uji.

4.2 Perbandingan Data *Training* Dan Data *Test*

Perbandingan data *training* dan data *test* adalah proses evaluasi model pembelajaran mesin dengan membagi dataset menjadi dua bagian utama. Data training digunakan untuk melatih model agar dapat mengenali pola dan hubungan dalam data, sedangkan data test digunakan untuk mengukur performa model pada data baru yang belum pernah dilihat sebelumnya. Perbandingan ini penting untuk memastikan model tidak hanya bekerja dengan baik pada data yang dilatih, tetapi juga dapat memberikan prediksi yang akurat pada data yang belum dikenal, sehingga mengurangi risiko *overfitting* atau *underfitting*. Pembagian data dilakukan dengan rasio seperti 80:20.

Perhitungan data *training* (80%) :

$$Data\ training = \frac{80}{100} \times 768 = 614,4 \approx 614$$

Perhitungan data *test* (20%) :

$$Data\ test = \frac{20}{100} \times 768 = 153,6 \approx 154$$

Dari perhitungan di atas, diperoleh hasil pembagian dataset menjadi dua bagian, yaitu data training dan data test, seperti yang ditunjukkan pada Tabel 2.

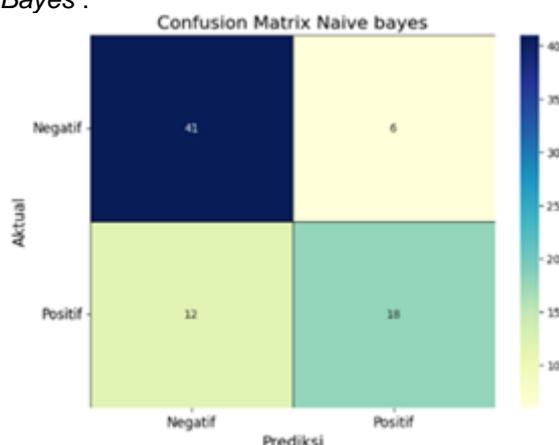
Tabel 2. Perbandingan Data Training dan Test

	Data Training	Data Test	Total
Persentase	80	20	100
Jumlah	614	154	768

Data training berjumlah 614, yang digunakan untuk melatih model agar mampu mengenali pola dan hubungan dalam data. Sementara itu, data test sebanyak 154 digunakan untuk mengukur performa model pada data baru yang belum pernah dilihat sebelumnya. Pembagian ini mengikuti rasio 80:20, yang merupakan praktik umum dalam evaluasi model untuk memastikan keseimbangan antara pelatihan dan pengujian.

4.3 Pengujian Algoritma

1) Pengujian *Naïve Bayes* :

**Gambar 2. Confusion Matrix *Naïve Bayes***

[sumber: <https://colab.google/>]

Berdasarkan confusion matrix ini, metrik evaluasi dihitung menggunakan rumus berikut:

a) Akurasi

Akurasi mengukur proporsi prediksi yang benar:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} = \frac{18 + 41}{18 + 41 + 6 + 12} \times 100\% = 76,6\%$$

b) Presisi

Presisi mengukur ketepatan prediksi positif:

$$Akurasi = \frac{TP}{TP + FP} = \frac{18}{18 + 6} \times 100\% = 75\%$$

c) Recall (*Sensitivity*)

Recall mengukur kemampuan model mendeteksi data positif:

$$Recall = \frac{TP}{TP + FN} = \frac{18}{18 + 12} \times 100\% = 60\%$$

d) F1-Score

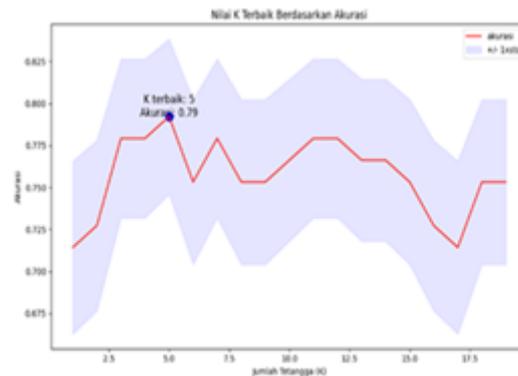
F1-Score adalah rata-rata harmonis dari presisi dan recall:

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} = \frac{0,75 \times 0,6}{0,75 + 0,6} \times 100\% = 66,7\%$$

Berdasarkan hasil pengujian menggunakan algoritma *Naïve Bayes*, data uji menghasilkan confusion matrix dengan nilai *True Negative* (TN) sebesar 41, *True Positive* (TP) sebesar 18, *False Positive* (FP) sebesar 6, dan *False Negative* (FN) sebesar 12. Dari matriks tersebut, diperoleh metrik evaluasi: akurasi sebesar 76,6%, yang menunjukkan bahwa 76,6% dari

seluruh prediksi model adalah benar. Presisi sebesar 75% menunjukkan bahwa 75% prediksi positif adalah benar-benar positif. Recall sebesar 60% mengindikasikan bahwa model berhasil mengidentifikasi 60% dari seluruh data positif. F1-Score sebesar 66,7% menunjukkan keseimbangan antara presisi dan recall, yang menggambarkan performa model secara keseluruhan. Hasil ini memberikan gambaran kekuatan dan kelemahan model dalam memprediksi data diabetes.

2) Pengujian KKN



Gambar 3. Plot Grafik KNN
[sumber: <https://colab.google/>]

Berdasarkan gambar plot grafik KNN di atas, dilakukan perhitungan untuk menentukan akurasi model KNN dengan nilai K=5. Berikut adalah rincian perhitungannya. Akurasi model dihitung dengan membandingkan jumlah prediksi yang benar terhadap total jumlah data uji, menggunakan rumus:

$$Akurasi = \frac{Jumlah\ Prediksi\ Besar}{Total\ Data\ Uji} \times 100\% = \frac{18,96}{18 + 6} \times 100\% = 79\%$$

Berdasarkan grafik, nilai K terbaik untuk model *K-Nearest Neighbors* (KNN) adalah 5, dengan akurasi tertinggi sebesar 79%. Grafik tersebut menunjukkan tren akurasi model terhadap berbagai nilai K (jumlah tetangga), di mana akurasi meningkat hingga mencapai puncaknya pada K=5. Setelah itu, terjadi fluktuasi yang mencerminkan peningkatan dan penurunan akurasi seiring bertambahnya nilai K. Daerah berwarna biru muda pada grafik menggambarkan rentang ketidakpastian akurasi, yaitu +/- 1 standar deviasi untuk setiap nilai K. Hasil ini mengonfirmasi bahwa nilai K=5 memberikan performa terbaik untuk model KNN dalam pengujian ini.

4.4 Pembahasan

Penelitian ini memperkuat temuan sebelumnya terkait efektivitas *Naïve Bayes* dan KNN dalam memprediksi penyakit diabetes. Studi oleh Smith et al. (2020) menyoroti keunggulan *Naïve Bayes* pada dataset besar dengan atribut yang tidak relevan, sementara penelitian Johnson et al. (2021) menunjukkan akurasi tinggi KNN pada dataset kompleks.

Tabel 3. Hasil Penelitian

Split Data		Akurasi	
Training Set	Test Set	Naïve Bayes	KNN
80%	20%	76,6%	79 %

Hasil penelitian ini konsisten dengan temuan pada Tabel 3, di mana algoritma *Naïve Bayes* menunjukkan akurasi sebesar 76,6%, sedangkan algoritma KNN memberikan akurasi yang lebih tinggi, yaitu 79%, ketika data dibagi dengan rasio 80:20 untuk training dan test. Penelitian ini memiliki fokus pada dataset yang relevan dengan prevalensi diabetes saat ini, sekaligus mengeksplorasi parameter evaluasi tambahan seperti variasi akurasi (+/- standar

deviasi) pada KNN. Selain itu, penelitian ini menyoroti pentingnya pemilihan nilai K dalam optimasi model KNN, yang sebelumnya sering diabaikan dalam penelitian sejenis. Dengan demikian, hasil penelitian ini memberikan kontribusi signifikan terhadap pemahaman algoritma prediksi diabetes, serta menawarkan wawasan praktis untuk mendukung pengelolaan penyakit ini.

5. Simpulan

Kesimpulan dari hasil penelitian ini menunjukkan bahwa algoritma KNN dengan akurasi 79% lebih unggul dibandingkan *Naïve Bayes* yang memiliki akurasi 76,6% pada dataset diabetes dengan pembagian data 80:20. KNN terbukti lebih efektif dalam menangani pola data yang kompleks, sementara *Naïve Bayes* lebih sesuai untuk dataset dengan distribusi yang lebih sederhana. Penelitian ini juga menegaskan pentingnya pemilihan nilai K dalam optimasi KNN, serta memberikan wawasan praktis untuk mendukung pengelolaan prediksi diabetes berbasis algoritma pembelajaran mesin.

Daftar Pustaka

- [1] A. Ridwan and T. G. Pratama, "Classification Of Heart Failure Using," *J.Proceedings International Conference On Public Health*, vol. 08, no, 10, pp. 1163–1167, 2021, doi: 10.26911/ICPmedicine.FP08.
- [2] A. Ridwan, "Penerapan Algoritma C4.5 untuk Penyakit Diabetes Mellitus," *J.BIDISFO (Bisnis Digital dan Sistem Informasi)*, vol. 06, no 12, pp. 41–48, 2022,[Online]. Available: <https://ejr.umku.ac.id/index.php/BIDISFO/article/viewFile/1841/1096>.
- [3] W. A. Z, R. Akbar, dan R. C. Prayogi, "Comparison of K-Nearest Neighbor (KNN) and Naive Bayes Algorithms in the Classification of Parkinson ' s Disease (Komparasi Algoritma K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penyakit Parkinson)," *J. Institut Riset dan Publikasi Indonesia*, vol. 4, no. 2, pp. 188–193, 2023, [Online]. Available: <https://journal.irpi.or.id/index.php/sentimas>
- [4] I. Print, G. A. Putri, A. Trimaysella, dan A. Khoiriah, "Penerapan Klasifikasi Data Mining pada Diabetes Menggunakan Metode Naive Bayes," *Jurnal Ilmu Komputer Teknologi Terapan (JIKTT)*, vol. 1, no. 14, pp. 1–9, 2024, [Online]. Available: <http://journal.amanahdigital.or.id/>
- [5] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 15–21, 2020, doi: 10.47970/siskom-kb.v4i1.169.
- [6] T.G. Pratamaa,* Achmad Ridwan b, "Deteksi Dini ASD (Autism Spectrum Disorder) Menggunakan Machine Learning," *J. JIKOMA (Ilmu Komputer dan Matematika)* vol. 4, no. 2, pp. 44–51, 2023, [Online]. Available: <https://ejr.umku.ac.id/index.php/jikoma/article/view/1998/1175>
- [7] M. R. Hunafa dan A. Hermawan, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Imbalance Class Dataset Penyakit Diabetes," *Kajian Ilmiah Informatika dan Komputer*. vol. 4, no. 3, pp. 1551–1561, 2023, doi: 10.30865/klik.v4i3.1486.
- [8] M. K. Nasution, R. R. Saedudin, dan V. P. Widartha, "Perbandingan Akurasi Algoritma Naïve Bayes Dan Algoritma Xgboost Pada Klasifikasi Penyakit Diabetes," *e-Proceeding Eng., J. Informatika Proxies*. vol. 8, no. 5, pp. 9765–9772, 2021, [Online]. Available: <https://journal.ubpkarawang.ac.id/mahasiswa/index.php/ssj/article/view/424/338%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15759>
- [9] F. K. Ngguna dan A. A. Pekuwali, "Penggunaan Algoritma K-Means Untuk Menganalisis Performa Siswa Dalam Pembelajaran Matematika di SMP Negeri Satap Lambakara," *J. Prosiding Seminar Nasional SATI* ,vol. 5, no 9, pp. 553–567, 2023, [Online]. Available: <https://www.ojs.unkriswina.ac.id/index.php/semnas-FST/article/view/882/576>
- [10] C. A. Rahayu, "Prediksi Penderita Diabetes Menggunakan Metode Naive Bayes," *J. Inform. dan Tek. Elektro Terap.*, vol. 11, no. 3, pp. 2303-0577, 2023, doi: 10.23960/jitet.v11i3.3055.
- [11] B. Delvika, S. Nurhidayarnis, dan P. D. Rinada, "Comparison of Classification Between Naive Bayes and K-Nearest Neighbor on Diabetes Risk in Pregnant Women Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes Pada Ibu Hamil," *J. SENTIMAS (Seminar Nasional Penelitian dan Pengabdian Masyarakat)*, vol. 2, no. October, pp. 68–75, 2022.
- [12] A. Ridwan, "Penerapan Teknik Bagging Pada Algoritma Naive Bayes Dan Algoritma C4.5

- Untuk Mengatasi Ketidakseimbangan Kelas,” *J.BIDISFO (Bisnis Digital Dan Sistem Informasi)*, vol. 1, no. 1, pp. 63–70, 2020, [Online]. Available: <https://ejr.umku.ac.id/index.php/jikoma/article/viewFile/1219/759>
- [13] H. Hairani, K. E. Saputro, dan S. Fadli, “K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes,” *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.
- [14] S. Syahrullah dan K. Nurwijayanti, “Klasifikasi Diagnosa Penyakit Diabetes Dengan Metode Naive Bayes Berbasis Web,” *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 2, no. 3, pp. 115–121, 2023, doi: 10.69916/jkbt.v2i3.35.
- [15] F. Sholehah, A. D. Putri, R. Rahmaddeni, dan L. Efrizoni, “Perbandingan Algoritma Naive Bayes dan K-Nearest Neighbors untuk Klasifikasi Metabolik Sindrom,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 507–514, 2024, doi: 10.57152/malcom.v4i2.1249.