POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS

POTERS: Proceedings of Technology, Engineering and Computers

Vol. 1, No. 2, pp. 9 - 16, 2025
p-ISSN. xxxx - xxxx, e-ISSN. 3090 - 3602

PERBANDINGAN ALGORITMA NAIVE BAYES DAN DECISION TREE ID3 DALAM MENGKLASIFIKASIKAN PENYAKIT DIABETES

Calvin Tanujaya[1], Desiyanna Lasut[2*]
[1, 2] Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Buddhi Dharma
*Corresponding Author, email: desiyanna.lasut@ubd.ac.id

ABSTRAK

Diabetes adalah suatu penyakit metabolik yang diakibatkan oleh meningkatnya kadar glukosa atau gula darah, yang seiring waktu dapat menyebabkan terjadinya berbagai komplikasi, seperti penyakit jantung koroner, stroke, obesitas, serta gangguan pada mata, ginjal, dan saraf, pembuluh darah. Berdasarkan data yang didapatkan dari *The International Diabetes Federation (IDF)*, jumlah penderita diabetes pada tahun 2021 di seluruh dunia mencapai 537 juta dan diperkirakan akan terus meningkat menjadi 643 juta di tahun 2030 dan 783 juta pada tahun 2045 dengan peningkatan sebesar 46%. *Data mining* adalah teknik dalam ilmu komputer yang sering digunakan untuk memprediksi kejadian di masa depan, dan menjadi salah satu metode yang banyak diterapkan dalam memprediksi apakah seseorang didiagnosis positif atau negatif diabetes. Penelitian ini bertujuan untuk menentukan metode yang paling baik dalam mengklasifikasikan penyakit diabetes. Metode digunakan untuk prediksi ini adalah algoritma naïve bayes dan ID3. Data untuk penelitian ini diperoleh dari website kaggle dengan 9 atribut dan 768 data. Setelah proses pembersihan data, data dikurangi menjadi 420 data yang digunakan dalam analisis. Hasil pengolahan data mining menunjukkan bahwa algoritma ID3 menghasilkan akurasi sebesar 79.1% dan naïve bayes menghasilkan akurasi sebesar 76.5% dengan menggunakan metode evaluasi 10-fold cross validation. Dari penjelasan di atas, dapat disimpulkan bahwa algoritma naïve bayes dan ID3 terbukti andal dalam memprediksi penyakit diabetes.

Kata kunci: *Data mining*, Diabetes, ID3, Klasifikasi, Naïve bayes

I. PENDAHULUAN

Seiring dengan pesatnya perkembangan zaman dan budaya instan, masyarakat semakin melupakan pentingnya pola hidup yang teratur dan sehat hal ini mengakibatkan resiko diabetes. Diabetes merupakan penyakit metabolik kronis yang ditandai oleh tingginya kadar glukosa dalam darah (Hansen & Hariyanto, 2023). Apabila tidak dikelola dengan baik, diabetes dapat menimbulkan berbagai komplikasi serius seperti penyakit jantung koroner, stroke, obesitas, serta gangguan pada mata, ginjal, dan sistem saraf (Argina, 2020).

Berdasarkan data yang didapatkan dari *The International Diabetes Federation (IDF)*, jumlah penderita diabetes pada tahun 2021 di seluruh dunia mencapai 537 juta dan diperkirakan akan terus meningkat menjadi 643 juta di tahun 2030 dan 783 juta pada tahun 2045 dengan peningkatan sebesar 46% (IDF, 2021).

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 9 Juli 2025.

POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS
POTERS: Proceedings of Technology, Engineering and Computers
Vol. 1, No. 2, pp. 9 - 16, 2025
p-ISSN. xxxx - xxxx, e-ISSN. 3090 - 3602

Selain itu, berdasarkan penelitian menunjukkan bahwa perempuan cenderung memiliki risiko lebih besar terkena diabetes dibandingkan laki-laki. Perbedaan ini dipengaruhi oleh faktor hormonal serta mekanisme metabolisme yang tidak sama antara keduanya (Hardiyanti et al., 2021).

Data mining adalah proses mencari pola atau informasi dari suatu data dengan menggunakan metode tertentu (Pangestu & Noris, 2023). Secara umum, *data mining* adalah proses mengekstraksi pengetahuan dari kumpulan data berukuran besar, yang hasilnya sangat bermanfaat dalam pengambilan keputusan dan pengembangan sistem, termasuk dalam bidang Kesehatan (Yoliadi, 2023).

Dalam proses *data mining*, klasifikasi merupakan salah satu teknik yang digunakan untuk menganalisis data dan menentukan kategori atau kelas dari data tersebut. Metode ini secara otomatis dapat memprediksi kelas dari data baru yang belum memiliki label (Budiarto et al., 2022). Klasifikasi juga dapat disebut sebagai prediksi, yaitu proses untuk memperkirakan apa yang akan terjadi di masa mendatang berdasarkan informasi yang tersedia saat ini dan di masa lalu (Rahman & Sutanto, 2023).Terdapat berbagai jenis algoritma dalam klasifikasi, dan dalam penelitian ini akan digunakan dua di antaranya, yaitu naive bayes dan ID3.

Naive bayes merupakan metode yang didasarkan pada pendekatan probabilitas dan statistika, yang diciptakan oleh ilmuwan Inggris bernama Thomas Bayes, yang disebut sebagai Teorema Bayes (Rino, 2021). Algoritma naïve bayes digunakan untuk memperkirakan kemungkinan di masa depan berdasarkan data atau pengalaman sebelumnya (Renaldy & Putra, 2023). Metode ini menggunakan teorema probabilitas bayes untuk memperkirakan kemungkinan kelas atau label berdasarkan fitur yang diberikan (Timotius & Fenriana, 2024). Sementara itu, ID3 adalah metode pemecahan masalah yang menggunakan struktur pohon keputusan. Konsep dasar dari *decision tree* adalah mengubah data menjadi sebuah pohon keputusan, di mana setiap cabang merepresentasikan aturan atau kondisi tertentu yang telah ditetapkan untuk mengelompokkan data secara sistematis hingga mencapai hasil akhir berupa Keputusan (Nazanah & Jambak, 2023).

Salah satu bahasa pemprograman yang digunakan untuk mengimplementasikan *data mining* adalah python. Python merupakan sebuah bahasa pemrograman yang dapat dimanfaatkan untuk mengelola proses *data mining*

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 9 Juli 2025.

POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS
POTERS: Proceedings of Technology, Engineering and Computers
Vol. 1, No. 2, pp. 9 - 16, 2025
p-ISSN. xxxx - xxxx, e-ISSN. 3090 - 3602

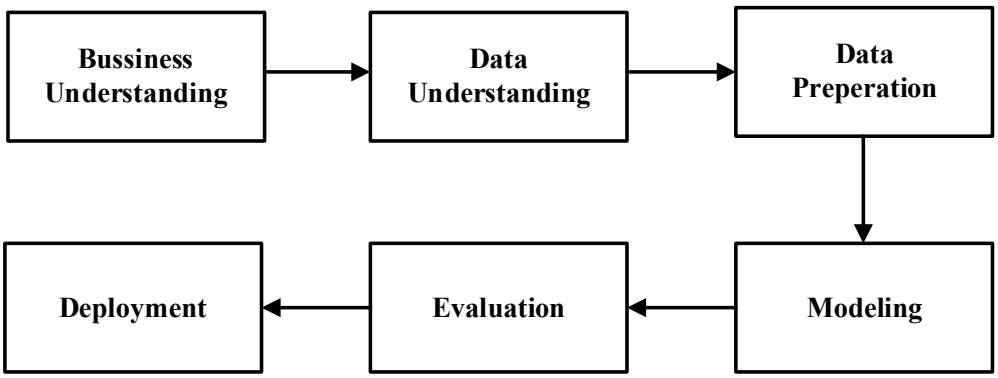
(Bustomi et al., 2023). Selain itu, pengujian juga dibutuhkan untuk mengukur performa metode yang diterapkan. Dalam hal ini, 10-fold cross-validation merupakan teknik yang diterapkan untuk menilai kinerja model, sehingga dapat diketahui tingkat akurasi model tersebut saat melakukan prediksi (Arifin et al., 2021).

Penelitian ini bertujuan untuk membandingkan hasil prediksi dari metode klasifikasi algoritma ID3 dan naive bayes dalam mengklasifikasikan diabetes. Selain itu, manfaat penelitian ini adalah untuk mengetahui metode mana yang memiliki akurasi tertinggi untuk klasifikasi diabetes. Penelitian ini memiliki batasan sebagai berikut:

1. Algoritma yang digunakan adalah naive bayes dan ID3.
2. Data yang digunakan berupa data sekunder yang diperoleh dari kaggle.
3. Objek yang diteliti adalah penyakit diabetes pada pasien Perempuan.
4. Batasan usia pasien adalah 21 - 81 tahun.
5. Atribut yang digunakan pada penelitian ini meliputi kehamilan, gula darah, tekanan darah, ketebalan kulit, insulin, riwayat diabetes, BMI, umur dan hasil.
6. Bahasa pemprograman yang digunakan untuk melakukan proses klasifikasi adalah python.

II. METODOLOGI

Tahapan perancangan model dalam penelitian ini menggunakan pendekatan CRISP-DM (*Cross Industry Standard Process for Data Mining*). Proses ini bertujuan untuk menggali dan menganalisis data dalam jumlah besar agar dapat diubah menjadi informasi yang berguna yang dapat dimanfaatkan dalam pengambilan Keputusan (Lundén et al., 2023).



Gambar 1. Tahap perancangan CRISP-DM

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 9 Juli 2025.

POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS
POTERS: Proceedings of Technology, Engineering and Computers
Vol. 1, No. 2, pp. 9 - 16, 2025
p-ISSN. xxxx - xxxx, e-ISSN. 3090 - 3602

III. HASIL DAN PEMBAHASAN

3.1 Bussiness Understanding

Tujuan dari tahap *business understanding* dalam penelitian ini adalah untuk melakukan klasifikasi penyakit diabetes. Pola yang terbentuk dari proses klasifikasi ini diharapkan dapat digunakan untuk memperkirakan apakah seseorang mengidap diabetes atau tidak. Selain itu, pola tersebut juga dapat membantu individu yang terdeteksi positif untuk segera berkonsultasi dengan tenaga medis guna mendapatkan penanganan lebih lanjut.

3.2 Data Understanding

Dataset yang digunakan dalam penelitian ini diperoleh dari situs www.kaggle.com. Berdasarkan dataset tersebut, terdapat 768 data yang digunakan untuk proses klasifikasi. Dataset ini memiliki 9 atribut berbeda yang berfungsi sebagai variabel untuk mengklasifikasikan penyakit diabetes. Berikut ini adalah penjelasan dari masing-masing atribut.

Tabel 1. Daftar atribut

Atribut	Keterangan	Variabel
Kehamilan	Menunjukkan jumlah kehamilan	*Integer*
Gula Darah	Menunjukkan jumlah gula darah	*Integer*
Tekanan Darah	Menunjukkan jumlah tekanan darah	*Integer*
Ketebalan Kulit	Menunjukkan jumlah ketebalan	*Integer*
Insulin	Menunjukkan jumlah kadar insulin	*Integer*
BMI (Body Mass Index)	Menunjukkan jumlah BMI	*Real*
Riwayat diabetes	Menjelaskan apakah memiliki riwayat penyakit diabetes	*Real*
Umur	Menunjukkan umur	*Integer*
Hasil	Menunjukkan apakah mengidap diabetes	*Binominal*

3.3 Data Preperation

Pada tahap ini, penulis akan melakukan proses pembersihan data yang mencakup penanganan data yang hilang (*missing value*) serta penyeimbangan distribusi data. Langkah ini bertujuan untuk memastikan bahwa data yang

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 9 Juli 2025.

POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS
POTERS: Proceedings of Technology, Engineering and Computers
Vol. 1, No. 2, pp. 9 - 16, 2025
p-ISSN. xxxx - xxxx, e-ISSN. 3090 - 3602

digunakan memiliki kualitas yang optimal, sehingga proses klasifikasi dan analisis dapat berjalan secara akurat dan menghasilkan prediksi yang lebih andal.

Tabel 2. Menghapus data

Program	Keterangan
df = pd.read_csv("diabetes.csv")	Membaca file dari csv
critical_columns = ["Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI", "Age"]	Variabel critical_columns menyimpan daftar nama kolom yang memiliki nilai 0
df_clean = df[(df[critical_columns] != 0).all(axis=1)]	Membersihkan data dengan cara menghapus baris yang memiliki nilai 0, axis=1 berfungsi untuk mengecek apakah semua kolom bernilai true(tidak memiliki nilai 0)

Tabel 3. Menyeimbangkan data

Program	Keterangan
smote = SMOTE(random_state=42)	Menambah data untuk kelas minoritas untuk memastikan kelas positif dan negatif memiliki jumlah yang sama, random_state=42 berfungsi untuk memastikan data yang ditambahkan selalu sama setiap kali program dijalankan
X_train_balanced, y_train_balanced = smote.fit_resample(X_train, y_train)	Menyimpan data yang telah diseimbangkan

3.4 Modeling

Pada tahap ini akan dilakukan pemodelan menggunakan python terhadap dataset diabetes.

Tabel 4. Membaca dan membagi data

Program	Keterangan
df = pd.read_csv("dataset.csv")	Membaca file dari csv
X = df.drop("Outcome", axis=1)	Variabel X menyimpan keseluruhan kolom kecuali kolom outcome
y = df["Outcome"]	Variabel y menyimpan kolom outcome
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)	Membagi keseluruhan data menjadi menjadi 80% untuk data training dan 20% untuk data testing

Tabel 5. Membuat dan melatih model

Program	Keterangan
id3_model=DecisionTreeClassifier (criterion="entropy",random_state=42)	Membuat model decision tree dengan kriteria entropy (ID3)

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 9 Juli 2025.

POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS
POTERS: Proceedings of Technology, Engineering and Computers
Vol. 1, No. 2, pp. 9 - 16, 2025
p-ISSN. xxxx - xxxx, e-ISSN. 3090 - 3602

Program	Keterangan
id3_model.fit(X_train, y_train)	Melatih model ID3 dengan fitur X_train dan label y_train
nb_model = GaussianNB()	Membuat model naive bayes
nb_model.fit(X_train, y_train)	Melatih model naive bayes dengan fitur X_train dan label y_train

3.5 Evaluation

Evaluasi dilakukan untuk mengetahui apakah hasil didapatkan sesuai dengan sasaran yang ingin dicapai dalam tahap business understanding. Berikut hasil pengujian model *data mining* menggunakan metode *10-Fold Cross Validation*.

Tabel 6. Evaluasi 10-*fold cross validation*

Program	Keterangan
scores_id3=cross_val_score(id3_model, X_train,y_train,cv=10,scoring="accuracy") scores_nb=cross_val_score(nb_model, X_train,y_train,cv=10,scoring="accuracy")	Mengevaluasi performa model menggunakan cross validation dengan membagi data *training* menjadi 10 bagian (cv=10), nilai yang dicari adalah akurasi (scoring=”accuracy”)
id3_mean_accuracy=scores_id3.mean()*100 nb_mean_accuracy =scores_nb.mean()*100	Mengambil nilai rata-rata akurasi dari cross validation. Setelah itu nilai dikalikan dengan 100 untuk mengubah nilai desimal menjadi persentase

Tabel 7. Hasil pengujian

Model	Hasil
ID3	79.1%
Naive bayes	76.5%

Berdasarkan hasil pengujian pada tabel 7 menggunakan 10-*fold cross validation* akurasi yang didapatkan dari algoritma ID3 sebesar 79.1% lebih besar dibandingkan naïve bayes sebesar 76.5%.

3.6 Deployment

Penelitian ini bertujuan untuk membandingkan algoritma naive bayes dan ID3 dalam mengklasifikasikan penyakit diabetes. Hasil dari prediksi ini diharapkan dapat mendorong individu untuk lebih waspada terhadap kondisi kesehatannya, serta segera melakukan pemeriksaan medis lebih lanjut apabila terindikasi mengidap diabetes, sehingga penanganan dapat dilakukan sedini mungkin.

POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS
POTERS: Proceedings of Technology, Engineering and Computers
Vol. 1, No. 2, pp. 9 - 16, 2025
p-ISSN. xxxx - xxxx, e-ISSN. 3090 - 3602

IV. SIMPULAN

Berdasarkan hasil pengujian menggunakan python terhadap dataset yang digunakan, diperoleh bahwa algoritma ID3 menghasilkan akurasi sebesar 79,1%, sedangkan algoritma naïve bayes menghasilkan akurasi sebesar 76,5%, dengan menggunakan metode evaluasi 10-Fold Cross Validation. Dari hasil tersebut dapat disimpulkan bahwa algoritma ID3 memiliki performa yang lebih baik dibandingkan Naïve Bayes dalam melakukan klasifikasi penyakit diabetes, sehingga ID3 dapat dijadikan pilihan utama untuk model prediksi pada penelitian ini.

DAFTAR PUSTAKA

Argina, A. M. (2020). Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes. *Indonesian Journal of Data and Science*, *1*(2), 29–33. https://doi.org/10.33096/ijodas.v1i2.11

Arifin, O., Saputra, K., & Fathoni, H. (2021). Implementation of Data Mining using Naïve Bayes Classifier Method in Food Crop Prediction. *Scientific Journal of Informatics*, *8*(1), 43–50. https://doi.org/10.15294/sji.v8i1.28354

Budiarto, E., Rino, R., Hariyanto, S., & Susilawati, D. (2022). Penerapan Data Mining Untuk Rekomendasi Beasiswa Pada SD Maria Mediatrix Menggunakan Algoritma C4.5. *Algor*, *3*(2), 23–34. https://doi.org/10.31253/algor.v3i2.1019

Bustomi, Y., Nugraha, A., Juliane, C., & Rahayu, S. (2023). Data Mining Selection of Prospective Government Employees with Employment Agreements using Naive Bayes Classifier. *Sinkron*, *8*(1), 1–8. https://doi.org/10.33395/sinkron.v8i1.11968

Hansen, & Hariyanto, S. (2023). Perbandingan Algoritma Data Mining Dalam Mengklasifikasi Penyakit Diabetes Menggunakan Model C4.5 Dan Naïve Bayes. *Jurnal Algor*, *4*(2), 1–10. https://jurnal.buddhidharma.ac.id/index.php/algor/index

Hardiyanti, T. O., Wurjanto, A., Kusariana, N., Hestiningsih, R., Epidemiologi dan Penyakit Tropik, P., Kesehatan Masyarakat Universitas Diponegoro, F., Epidemiologi dan Penyakit Tropik, B., & Kesehatan Masyarakat, F. (2021). *HUBUNGAN JENIS KELAMIN DAN BIDANG STUDI DENGAN PRAKTIK*

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 9 Juli 2025.

Vol. 1, No. 2, pp. 9 - 16, 2025
p-ISSN. xxxx - xxxx, e-ISSN. 3090 - 3602

PENCEGAHAN DIABETES MELLITUS TIPE 2 PADA MAHASISWA (Studi Pada Mahasiswa Universitas Diponegoro Semarang). *9*(2). http://ejournal3.undip.ac.id/index.php/jkm

IDF. (2021). *IDF Diabetes Atlas* (10th ed.).

Lundén, N., Bekar, E. T., Skoogh, A., & Bokrantz, J. (2023). Domain Knowledge in CRISP-DM: An Application Case in Manufacturing. *ScienceDirect*, *56*(2), 7603–7608. https://doi.org/10.1016/j.ifacol.2023.10.1156

Nazanah, J. T. M. A., & Jambak, M. I. (2023). Pemanfaatan Algoritma Decision Tree ID3 Bagi Manajemen Bimbel Untuk Menentukan Faktor Kelulusan Pada Sekolah Kedinasan. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *3*(6), 915–924. https://doi.org/10.30865/klik.v3i6.791

Pangestu, R. A., & Noris, S. (2023). Analisa Data Mining Prediksi Lelang Suku Cadang Dengan Metode K-NearestNeighbor (Studi Kasus PT. Parmud Jaya Perkasa). *Jurnal Informatika Multi*, *1*(4), 285–295.

Rahman, R., & Sutanto, F. A. (2023). Data Mining to Predict Gojek's Consumer Satisfaction Level Using Naive Bayes Algorithm. *International Journal of Information System & Technology Akreditasi*, *6*(158), 590–602.

Renaldy, & Putra, D. S. D. (2023). Aplikasi Prediksi Harga Ayam Dengan Metode Naives Bayes Pada Supplier Ayam Potong. *Jurnal Algor*, *4*(2), 141–148. http://repositori.buddhidharma.ac.id/id/eprint/1609

Rino, R. (2021). Comparison of Data Mining Methods Using C4.5 Algorithm and Naive Bayes in Predicting Heart Disease. *Tech-E*, *4*(2), 44–51. https://doi.org/10.31253/te.v4i2.543

Timotius, A., & Fenriana, I. (2024). PERANCANGAN APLIKASI PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN METODE NAÏVE BAYES. *JURNAL ALGOR*, *5*(2), 54–65.

Yoliadi, D. N. (2023). Data mining Dalam Analisis Tingkat Penjualan Barang Elektronik Menggunakan Algoritma K-means. *Insearch (Information System Research) Journal*, *3*(1).

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 9 Juli 2025.