Indonesian Journal of Digital Business

Journal homepage: https://ejournal.upi.edu/index.php/IJDB

Komparasi Algoritma C4.5, Naive Bayes, K-Nearest Neighbor, Random Forest Untuk Prediksi Faktor Penyebab Penyakit Diabetes

Muhammad Alfian Fadillah[1], Evi Dewi Sri Mulyani[2], Ruuhwan[3]

[1,2,3]Teknik Informatika, Universitas Perjuangan Tasikmalaya, Jawa Barat, Indonesia
Correspondence: E-mail: ruuhwan@unper.ac.id

ABSTRAK

Diabetes adalah penyakit metabolik kronis yang ditandai dengan peningkatan kadar glukosa darah (gula darah), yang seiring waktu menyebabkan kerusakan serius pada jantung, pembuluh darah, mata, ginjal, dan saraf. Penyakit diabetes menjadi salah satu jenis penyakit yang mematikan di dunia. Pengklasifikasian secara tepat orang-orang yang memiliki hasil tes laboratorium apakah positif atau negatif memiliki penyakit diabetes penting dilakukan untuk memperoleh penanganan yang tepat. Penelitian ini dataset yang digunakan bersumber dari komunitas global pada studi kasus Electronic Health Record (EHRs). Data yang diperoleh sebanyak 10.000 record, memiliki delapan atribut dan satu atribut dengan status pasien sebagai label (class) yang menyatakan 859 data pasien yang menderita penyakit diabetes, 9141 data pasien yang tidak menderita penyakit diabetes. Tujuan dari penelitian ini adalah untuk mengkomparasikan algoritma C4.5, Naive bayes, K-Nearest Neighbor dan Random Forest dalam penentuan klasifikasi data pasien diabetes. Hasil penelitian ini dilakukan dengan membagi data testing dan data training dengan perbandingan 90 : 10, 80 : 20, dan 70 : 30. Hasil penelitian menunjukan bahwa secara keseluruhan komparasi algoritma C4.5, Naive bayes, K-Nearest Neighbor dan Random Forest, dari percobaaan dengan pembagian data training : data testing 90 : 10, 80 : 20, 70 : 30. Jika dibandingkan dengan nilai accuracy algoritma Naïve Bayes dan K-Nearest Neighbor, nilai accuracy dengan menggunakan algoritma klasifikasi

INFO ARTIKEL

Riwayat Artikel:
Received 18 January 2024
First Revised 27 January 2024
Accepted 18 March 2024
Available online 25 March 2024
Publication Date 31 March 2024

Kata Kunci:
Diabetes,
Decision Tree,
C4.5,
Naïve Bayes,
K-Nearest Neighbor,
Random Forest.

C4.5 dan Random Forest adalah yang terbesar pada percobaan data training 90% : data testing 10% dan percobaan data training 70% : data testing 30%. Sedangkan evaluasi menggunakan ROC curve, Algoritma Random Forest menjadi yang tertinggi pada percobaan data training 70% : data testing 30% dan data training 80% : data testing 20% dengan nilai mendekati 1.000 yaitu 0.972 dan 0.970. Dari hasil keseluruhan pengujian model dapat disimpulkan bahwa kinerja C4.5 dan Random Forest hampir sama bagusnya, baik itu dilihat dari tingkat accuracy maupun AUC nya.

1. PENDAHULUAN

Seiring perkembangan zaman ilmu medis kini bukan hanya mengandalkan teknik analisa konvensional saja, melainkan telah menggabungkan perkembangan teknologi didalamnya. Salah satu contoh penggabungan dunia teknologi dengan dunia medis adalah dengan cara memprediksi diabetes. Ada banyak cara untuk memprediksi diabetes, salah satunya adalah menggunakan ilmu yang bernama data mining. Data mining salah satu bidang ilmu dalam komputer yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar, yang nantinya kumpulan informasi data tersebut akan diolah dan akan menghasilkan informasi yang lebih akurat atau menampilkan informasi yang terbaik (Hidayat 2015). Sedangkan menurut Han dan Kamber (Ha, Kambe, and Pe 2011) *data mining* adalah proses menemukan pola yang belum diketahui sebelumnya atau informasi yang berguna dari data mentah. Definisi ini mencerminkan konsep umum bahwa data mining melibatkan penggunaan algoritma dan teknik statistik untuk menggali pengetahuan yang tersembunyi atau pola yang berharga dari data besar.

Penelitian yang sama dilakukan oleh (Alzubaidi, Halawani, and Jarrah 2023), menekankan bahwa menggunakan metode pengujian k-fold cross-validation untuk memvalidasi kinerja model yang diusulkan dan confusion matrix untuk mengukur performa klasifikasi. Penelitian ini menghasilkan temuan berdasarkan pengukuran accuracy dari setiap algoritma, yang meliputi Random Forest, Logistic Regression, XGBoost. Dari ketiga model yang diuji, Model random forest mencapai nilai accuracy sebesar 95.06%, Logistic Regression mendapatkan nilai accuracy sebesar 97.01% dan XGBoost mendapatkan nilai accuracy sebesar 97.16%. Berdasarkan hasil di atas, penelitian ini menunjukkan bahwa teknik pembelajaran mesin yang digunakan mampu memberikan prediksi yang akurat untuk diagnosis diabetes melitus. XGBoost, secara khusus, muncul sebagai algoritma yang paling efektif dalam studi ini. Namun, keunggulan margin keakuratan XGBoost atas Logistic Regression relatif kecil.

Penelitian serupa dilakukan oleh Mohammed Layth Zubairi Alkaragole (Layth, Alkaragole, and Sefer Kurnaz 2019), menyarankan untuk membandingkan efektivitas kinerja dari beberapa algoritma klasifikasi dalam memprediksi diabetes berdasarkan faktor resiko. Metode yang digunakan dalam penelitian ini adalah Naive bayes, Decision Tree, dan Support Vector Machine (SVM). Dalam analisis algoritma untuk klasifikasi data, algoritma Naive bayes menunjukkan accuracy sebesar 90%, sedangkan algoritma Decision Tree mencapai accuracy 86% dengan Area Under Curve (AUC) sebesar 0,500, menandakan performa yang kurang baik

dalam membedakan kelas. Sementara itu, algoritma Support Vector Machine menjadi yang terbaik dengan accuracy tertinggi mencapai 91%.

Penelitian serupa dilakukan oleh Monalisa Panda (Panda *et al*. 2022), berkonsentrasi untuk menciptakan model prediksi diabetes yang efektif dan presisi tinggi. Metode yang digunakan dalam penelitian ini melibatkan empat algoritma pembelajaran mesin, yaitu Logistic Regression, K-Nearest Neighbor, Support Vector Machine, dan Extreme Gradient Boosting (XGBoost). Dalam penelitian prediksi penyakit diabetes menggunakan empat algoritma pembelajaran mesin, hasil menunjukkan bahwa algoritma XGBoost mendominasi dengan accuracy tertinggi sebesar 81,25%, diikuti oleh Logistic Regression dengan accuracy 80,73%. Meskipun Logistic Regression dan Support Vector Machine memiliki precision yang hampir sama, yaitu 76,60% dan 74,00% secara berturut-turut, XGBoost tetap unggul dengan recall sebesar 62,90%. Di sisi lain, K-Nearest Neighbor menunjukkan performa yang lebih rendah dengan accuracy 78,13% dan recall 51,61%.

(Suherman, Ruuhwan, and Sudiarjo 2023), berkonsentrasi untuk menerapkan Algoritma C4.5 dalam memprediksi preferensi jurusan siswa di SMK Laboratorium Jakarta. Penelitian ini juga membandingkan penggunaan aplikasi RapidMiner dan Python dalam implementasi algoritma Decision Tree C4.5. Hasil penelitian menunjukkan bahwa aplikasi RapidMiner mencapai accuracy sebesar 94,44%, precision 81,37%, dan sensitivity 74,00%. Sementara itu, implementasi dengan Python mencapai accuracy 93% karena pembulatan otomatis, tetapi tidak ada perbedaan yang signifikan dibandingkan dengan RapidMiner. Hal ini menunjukkan bahwa algoritma C4.5 efektif dan akurat dalam memprediksi pola penjurusan siswa di SMK Laboratorium Jakarta.

(Mulyani *et al*. 2020), berfokus pada penerapan Algoritma Naïve Bayes Classifier untuk menentukan kelayakan pemberian kredit kepada nasabah. Penelitian ini bertujuan untuk membantu para analis kredit dalam menilai signifikansi dan kelayakan pemberian kredit. Hasil penelitian menunjukkan bahwa dengan menggunakan 16 atribut, algoritma Naïve Bayes Classifier menghasilkan tingkat accuracy sebesar 59,00%. Ini merupakan peningkatan dari penelitian sebelumnya yang menggunakan 9 atribut dengan tingkat accuracy 56,00%. Dengan demikian, penggunaan lebih banyak atribut dalam model klasifikasi membantu meningkatkan accuracy dalam menentukan kelayakan pemberian kredit kepada nasabah.

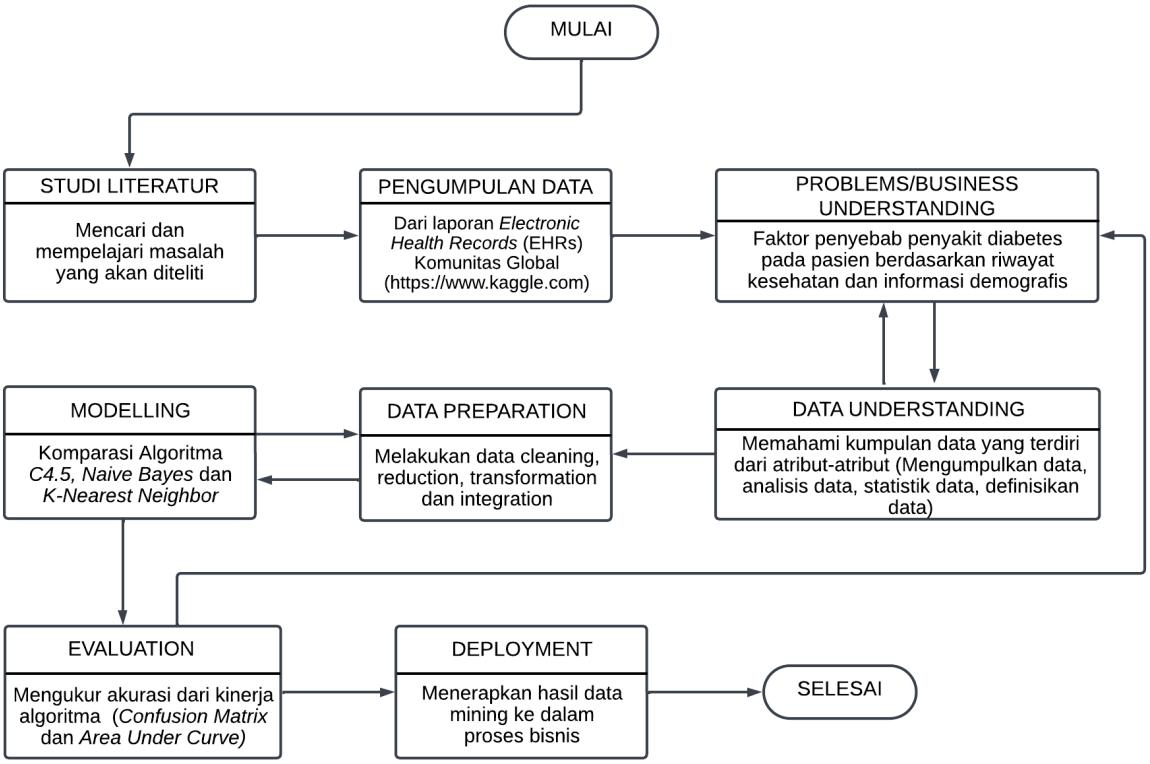
Penggunaan data mining dengan algoritma seperti C4.5, Naive bayes, K-Nearest Neighbor, Deep Learning, Artificial Neural Network (ANN), Support Vector Machine dan Generalized Linear Model (GLM). Namun dari algoritma tersebut yang paling banyak digunakan penelitian adalah algoritma C4.5 dan algoritma Naive bayes. mereka berfokus untuk mencari nilai accuracy dari algoritma yang mereka gunakan serta tidak menggunakan validasi data. Padahal jika menggunakan pengujian accuracy pada masing-masing algoritma akan lebih akurat algoritma mana yang terbaik untuk menghasilkan pola. Sebagai pilihan untuk diagnosa penyakit diabetes dapat menjadi alternatif pilihan yang tepat, tetapi sampai saat ini belum diketahui algoritma yang paling akurat dalam memprediksi penyakit diabetes.

Berdasarkan latar belakang yang telah dijelaskan di atas, maka pada penelitian ini akan dilakukan komparasi data mining algoritma C4.5, Naive bayes, K-Nearest Neighbor (K-NN) dan Random Forest untuk mengetahui algoritma mana yang memiliki accuracy yang terbaik dalam mendeteksi penyakit diabetes.

2. METODE

Metode dasar yang digunakan dalam penelitian ini adalah metode kuantitatif. Tujuannya yaitu membangun model klasifikasi dan mengukur kinerja pada masing-masing model untuk

mendapatkan kinerja yang terbaik dari model yang dihasilkan dengan algoritma C4.5, Naive bayes, K-Nearest Neighbor (K-NN) dan Random Forest. Gambar 1 memberikan gambaran secara umum mengenai siklus hidup dalam proses penelitian.



Gambar 1. Siklus Hidup Proses Penelitian.

3. HASIL DAN PEMBAHASAN

Pada tahapan pengolahan data awal, eksperimen yang digunakan pada penelitian ini menggunakan model Cross-Standard Industry for Data mining (CRISP-DM) (IBM 2021). Seperti yang telah dijelaskan sebelumnya.

Data understanding, data yang didapat berasal dari laporan Electronic Health Records (EHRs) adalah sumber data utama untuk kumpulan data prediksi diabetes yang melibatkan pengumpulan data medis dan demografi dari pasien yang telah didiagnosis atau berisiko terkena diabetes. Jumlah data yang digunakan sebanyak 10.000 record, memiliki delapan atribut dan satu atribut dengan status pasien sebagai label (class) yang menyatakan 859 data pasien yang menderita penyakit diabetes, 9141 data pasien yang tidak menderita penyakit diabetes. Dapat dilihat pada tabel 1.

Dalam penelitian ini, eksperimen dilaksanakan dengan tujuan untuk menentukan tingkat akurasi terbaik di antara algoritma C4.5, Naive Bayes, K-Nearest Neighbor (K-NN), dan Random Forest. Keempat algoritma ini dibandingkan untuk menemukan yang paling efektif. Setelah pembuatan model, pengujian dilakukan dengan menggunakan 10-fold cross validation. Perbandingan antara data training dan data testing yang digunakan adalah sebagai berikut: 90% data training dan 10% data testing, 80% data training dan 20% data testing, serta 70% data training dan 30% data testing.

Pada tabel 2 merupakan hasil komparasi algoritma berdasarkan data training dan data testing.

Tabel 1. Sampel Data Training.

Gender	Age	Hypertension	Heart_Disease	Smoking_History	BMI	HbA1c_Level	20	Diabetes
Female	80	0	1	Never	25.2	6.6	140	0
Female	54	0	0	No Info	27.3	6.6	80	0
Male	28	0	0	Never	27.3	5.7	158	0
Female	36	0	0	Not Current	23.5	5	155	0
Male	76	1	1	Not Current	20.1	4.8	155	0
Female	20	0	0	Never	27.3	6.6	85	0
Female	44	0	0	Never	19.3	6.5	200	1
Female	79	0	0	No Info	23.9	5.7	85	0
Male	42	0	0	Never	33.6	4.8	145	0
Female	32	0	0	Never	27.3	5	100	0
Female	53	0	0	Never	27.3	6.1	85	0
Female	54	0	0	Former	54.7	6	100	0
Female	78	0	0	Former	36.1	5	130	0
Female	67	0	0	Never	25.7	5.8	200	0
Female	76	0	0	No Info	27.3	5	160	0
Male	78	0	0	No Info	27.3	6.6	126	0
Male	15	0	0	Never	30.4	6.1	200	0
Female	42	0	0	Never	24.5	5.7	158	0
Female	42	0	0	No Info	27.3	5.7	80	0

Tabel 2. Hasil Komparasi Algoritma Berdasarkan Data Training dan Data Testing.

Algoritma	*Data training*	*Data testing*	*Accuracy*	ROC/AUC
C4.5	90	10	97.21%	0.874
	80	20	97.15%	0.900
	70	30	97.31%	0.732
Naïve Bayes	90	10	95.64%	0.953
	80	20	95.56%	0.952
	70	30	95.57%	0.955
K-Nearest Neighbor	90	10	94.91%	0.866
	80	20	94.86%	0.872
	70	30	94.84%	0.868
Random Forest	90	10	97.29%	0.972
	80	20	97.21%	0.970
	70	30	97.33%	0.972

Dari Tabel 2, kita dapat melihat hasil komparasi dari keempat algoritma yang digunakan dalam penelitian ini, yaitu C4.5, Naive Bayes, K-Nearest Neighbor, dan Random Forest. Komparasi dilakukan berdasarkan pembagian data training dan data testing dengan perbandingan 90:10, 80:20, dan 70:30. Pada pembagian data training sebesar 90% dan data testing sebesar 10%, algoritma Random Forest dan C4.5 menunjukkan nilai accuracy tertinggi. Random Forest mencapai accuracy 97.29%, sedangkan C4.5 mencapai 97.21%. Hal ini lebih tinggi dibandingkan dengan Naive Bayes yang mencapai 95.64% dan K-Nearest Neighbor dengan akurasi terendah 94.91%. Selain itu, nilai Area Under Curve (AUC) dari Random Forest juga menunjukkan performa terbaiknya dibandingkan ketiga algoritma lainnya, yaitu C4.5, Naive Bayes, dan K-Nearest Neighbor. Untuk pembagian data training 80% dan data testing 20%, Random Forest kembali menunjukkan nilai accuracy dan AUC yang terbesar. Sedangkan yang terakhir, untuk percobaan perbandingan data training 70% data testing 30% tidak ada perbedaan dengan percobaan kedua yang dimana menghasilkan nilai accuracy dan AUC tertinggi adalah algoritma Random Forest memiliki nilai accuracy sebesar 97.33%, nilai AUC sebesar 0.972. Untuk rata-rata keseluruhan percobaan dapat dilihat pada tabel 4 dibawah ini:

Tabel 3. Rata-rata Hasil Komparasi Algoritma Berdasarkan Data Training dan Data Testing.

Algoritma	*Data training*	*Data testing*	*Accuracy*	ROC/AUC
Decision Tree C4.5	90	10	97.21%	0.874
	80	20	97.15%	0.900
	70	30	97.31%	0.732
	Rata-rata		97.22%	0.835
Naïve Bayes	90	10	95.64%	0.953
	80	20	95.56%	0.952
	70	30	95.57%	0.955
	Rata-rata		95.59%	0.953
K-Nearest Neighbor	90	10	94.91%	0.866
	80	20	94.86%	0.872
	70	30	94.84%	0.868
	Rata-rata		94.87%	0.869
Random Forest	90	10	97.29%	0.972
	80	20	97.21%	0.970
	70	30	97.33%	0.972
	Rata-rata		97.28%	0.971

Dari Tabel 3, rata-rata akurasi dari algoritma Random Forest mencapai 97.28%. Ini merupakan rata-rata akurasi tertinggi dibandingkan dengan algoritma C4.5, Naive Bayes, dan K-Nearest Neighbor. Selain itu, untuk rata-rata nilai Area Under Curve (AUC), Random Forest juga menunjukkan performa terbaik dengan nilai AUC sebesar 0.971. Model confusion matrix akan menghasilkan matriks yang terdiri dari empat bagian: true positif, false positif, true negatif, dan false negatif. Berikut dibawah ini merupakan hasil confusion matrix dari algoritma klasifikasi C4.5, Naive bayes, K-Nearest Neighbor dan Random Forest untuk data training 90% data testing 10% sebagai accuracy yang paling tinggi didapatkan.

accuracy: 97.21% +/- 0.69% (micro average: 97.21%)

	true Tidak	true Ya	class precision
pred. Tidak	8218	242	97.14%
pred. Ya	9	531	98.33%
class recall	99.89%	68.69%	

Gambar 2. Confussion matrix Algoritma C4.5 (Data training 90%, Data testing 10%).

Penjelasan dari gambar 2 menunjukan bahwa, diketahui terdapat sebanyak 531 jumlah data yang diprediksi diabetes dan pada kenyataannya memang menderita penyakit diabetes, 8218 data diprediksi tidak diabetes dan pada kenyataannya memang tidak menderita penyakit diabetes, 9 data yang diprediksi diabetes tetapi kenyataannya tidak menderita penyakit diabetes, dan 242 data diprediksi tidak menderita diabetes tetapi kenyataannya diabetes. Berdasarkan gambar 2 menunjukan bahwa, tingkat accuracy dengan menggunakan algoritma C4.5 untuk perbandingan data training dan data testing 90% : 10% adalah sebesar 97.21%.

Model confussion matrix yang kedua dengan menggunakan algoritma klasifikasi Naive bayes, Untuk perbandingan data training dan data testing 90% : 10% sehingga didapatkan hasil pada gambar 3 sebagai berikut:

accuracy: 95.64% +/- 0.76% (micro average: 95.64%)

	true Tidak	true Ya	class precision
pred. Tidak	8089	254	96.96%
pred. Ya	138	519	79.00%
class recall	98.32%	67.14%	

Gambar 3. Confussion matrix Algoritma Naive bayes (Data training 90%, Data testing 10%).

Pada gambar 3, menunjukan bahwa, sebanyak 8.089 jumlah data yang diprediksi tidak menderita penyakit diabetes dan pada kenyataannya memang tidak menderita, 519 data diprediksi diabetes dan pada kenyataannya memang diabetes, 254 data yang diprediksi tidak menderita penyakit diabetes tetapi kenyataannya diabetes, dan 138 data diprediksi diabetes tetapi kenyataannya tidak menderita penyakit diabetes. Berdasarkan gambar 3 menunjukan bahwa, tingkat accuracy dengan menggunakan algoritma Naïve Bayes untuk perbandingan data training dan data testing 90% : 10% adalah sebesar 95.64%.

Model confussion matrix yang ketiga dengan menggunakan algoritma K-Nearest Neighbor,

Untuk perbandingan data training dan data testing 90% : 10% sehingga didapatkan hasil pada gambar 4 sebagai berikut:

accuracy: 94.91% +/- 0.65% (micro average: 94.91%)

	true Tidak	true Ya	class precision
pred. Tidak	8171	402	95.31%
pred. Ya	56	371	86.89%
class recall	99.32%	47.99%	

Gambar 4. Confussion matrix Algoritma K-Nearest Neighbor (Data training 90%, Data testing 10%).

Pada gambar 4 menunjukan bahwa, sebanyak 8.171 jumlah data yang diprediksi tidak menderita penyakit diabetes dan pada kenyataannya memang tidak menderita, 371 data diprediksi diabetes dan pada kenyataannya memang diabetes, 402 data yang diprediksi tidak menderita penyakit diabetes tetapi kenyataannya diabetes, dan 56 data diprediksi diabetes tetapi kenyataannya tidak menderita penyakit diabetes. Berdasarkan gambar 4, menunjukan bahwa, tingkat accuracy dengan menggunakan algoritma K-Nearest Neighbor untuk perbandingan data training dan data testing 90% : 10% adalah sebesar 94.91%.

Model confussion matrix yang keempat dengan menggunakan algoritma Random Forest, Untuk perbandingan data training dan data testing 90% : 10% sehingga didapatkan hasil pada gambar 5 sebagai berikut:

accuracy: 97.29% +/- 0.35% (micro average: 97.29%)

	true Tidak	true Ya	class precision
pred. Tidak	8223	240	97.16%
pred. Ya	4	533	99.26%
class recall	99.95%	68.95%	

Gambar 5. Confussion matrix Algoritma Random Forest (Data training 90%, Data testing 10%).

Pada gambar 5, menunjukan bahwa, sebanyak 8.223 jumlah data yang diprediksi tidak menderita penyakit diabetes dan pada kenyataannya memang tidak menderita, 533 data diprediksi diabetes dan pada kenyataannya memang diabetes, 240 data yang diprediksi tidak menderita penyakit diabetes tetapi kenyataannya diabetes, dan 4 data diprediksi diabetes tetapi kenyataannya tidak menderita penyakit diabetes. Berdasarkan gambar 5, menunjukan bahwa, tingkat accuracy dengan menggunakan algoritma Random Forest untuk perbandingan data training dan data testing 90%: 10% yang paling tertinggi adalah sebesar 97.29%.

4. KESIMPULAN

Dalam penelitian ini dilakukan pengujian model dengan membandingkan empat metode data mining yaitu algoritma C4.5, Naïve Bayes, K-Nearest Neighbor dan Random Forest. Penelitian ini dataset yang digunakan bersumber dari komunitas global pada studi kasus Electronic Health Record (EHRs) bertujuan untuk menyelidiki berbagai faktor yang berhubungan dengan kesehatan dan keterkaitannya untuk mengklasifikasikan diabetes pada pasien berdasarkan riwayat kesehatan dan informasi demografis mereka yang telah didiagnosis atau berisiko terkena diabetes. Data yang diperoleh sebanyak 10.000 record, memiliki delapan atribut dan satu atribut dengan status pasien sebagai label atau kelas (class) yang menyatakan 859 data pasien yang menderita penyakit diabetes, 9141 data pasien yang tidak menderita penyakit diabetes. Model yang diuji akan menghasilkan nilai accuracy, dan Area Under Curve (AUC) dari setiap algoritma.

Berdasarkan hasil evaluasi dan validasi, komparasi algoritma C4.5, Naive Bayes, K-Nearest Neighbor, dan Random Forest menunjukkan performa yang berbeda-beda tergantung pada pembagian data training dan data testing. Pada percobaan dengan pembagian data training 90% dan data testing 10%, serta data training 70% dan data testing 30%, algoritma C4.5 dan Random Forest menunjukkan nilai accuracy yang paling tinggi dibandingkan dengan Naive Bayes dan K-Nearest Neighbor. Sedangkan untuk evaluasi menggunakan ROC curve berdasarkan nilai AUC, Random Forest menunjukkan performa terbaik pada percobaan data training 70% dan data testing 30%, serta data training 80% dan data testing 20%, dengan nilai mendekati 1.000, yaitu 0.972 dan 0.970.

Dari hasil keseluruhan pengujian model, dapat disimpulkan bahwa kinerja algoritma C4.5 dan Random Forest hampir sama bagusnya, baik dilihat dari tingkat akurasi maupun nilai AUC-nya.

7. REFERENSI

Alzubaidi, Abdulaziz A., Sami M. Halawani, and Mutasem Jarrah. 2023. "Towards a Stacking Ensemble Model for Predicting Diabetes Mellitus Using Combination of Machine Learning Techniques." *International Journal of Advanced Computer Science and Applications* 14(12): 348–58.

Bramer, Max. 2016. *Introduction to Data Mining*.

Ha, Jiawei, Micheline Kambe, and Jian Pe. 2011. Data Mining: Concepts and Techniques *Data Mining: Concepts and Techniques*.

Herdiana, O., Maulani, S., and Firdaus, E. A. (2021). Strategi Pemasaran Produk Industri Kreatif Menggunakan Algoritma K-Means Clustering Berbasis Particle Swarm Optimization. Nuansa Informatika, 15(2), 1-13.

Hermawan, Y., Ahman, E., and Sundari, R. S. (2022). THE EFFECTS OF ENTREPRENEURSHIP EDUCATION, SELF-EFFICACY AND ORIENTATION TOWARD ENTREPRENEURSHIP INTENTION AND ITS'IMPLICATION ON ENTREPRENEURSHIP COMPETENCE. Central Asia & the Caucasus (14046091), 23(1).

Hidayat, Muhammad Mahaputra. 2015. "Data Mining Data Mining." *Mining of Massive Datasets* 2(January 2013): 5–20. https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part.

IBM. 2021. "Modeler CRISP-DM Guide."

Jin, Ziwei et al. 2020. "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12343 LNCS: 503–15.

Layth, Mohammed, Zubairi Alkaragole, and Asst Sefer Kurnaz. 2019. "Comparison of Data Mining Techniques for Predicting Diabetes or Prediabetes by Risk Factors." *International Journal of Computer Science and Mobile Computing* 8(3): 61–71. www.ijcsmc.com.

Mulyani, Evi Dewi Sri Mulyani et al. 2020. "Klasifikasi Penentuan Kelayakan Pemberian Kredit Menggunakan Metode Naive Bayes Classifier Classification of Determination of Credit Worthiness Using the Naive Bayes Classifier Method." *Jurnal VOI (Voice Of Informatics)* 9(2): 81–92. https://voi.stmik-tasikmalaya.ac.id/index.php/voi/article/view/226.

Panda, Monalisa, Debani Prashad Mishra, Sopa Mousumi Patro, and Surender Reddy Salkuti. 2022. "Prediction of Diabetes Disease Using Machine Learning Algorithms." *IAES International Journal of Artificial Intelligence* 11(1): 284–90.

Suherman, Nurisya Rahma, Ruuhwan Ruuhwan, and Aso Sudiarjo. 2023. "Implementation of Data Mining at Laboratory Vocational High School Using The C4.5 Algorithm to Predict Students Major Preferences." *Innovation in Research of Informatics (INNOVATICS)* 5(2): 65–70.

Zhang, Shichao, Chengqi Zhang, and Qiang Yang. 2003. 17 Applied Artificial Intelligence *Data Preparation for Data Mining*.