JURNAL ILMIAH SISTEM INFORMASI (JUSI)
VOL 5, No. 1, Januari 2026, pp. 229 - 241
p-ISSN : 2803-1507 (media cetak) | e-ISSN : 2803-1531 (media online)
Link Publikasi: https://ejurnal.provisi.ac.id/index.php/JUISI/issue/view/169

Application of KNN & Decision Tree Algorithms in Predicting Diabetes Using Rapid Miner

Zaehol Fatah*[1], Baitul Anam[2]
[1,2] Program Studi Teknologi Informasi, Universitas Ibrahimy Situbondo
Jl. KHR. Syamsul Arifin, Sukorejo, Sumberejo, Kec. Banyuputih, Kabupaten Situbondo, Jawa Timur
e-mail: zaeholfatah@email.com*[1], baitulanam190603@email.com[2]

ARTICLE INFO

Article History:

Received: 28-10-2025
Revised: 29-11-2025
Accepted: 12-12-2025
Published: 21-01-2026

ABSTRACT

Diabetes prediction plays a crucial role in supporting early detection and preventing long-term complications associated with chronic diseases. This study aims to compare the performance of the K-Nearest Neighbor (KNN) and Decision Tree algorithms in predicting diabetes using the Pima Indian Diabetes dataset within the RapidMiner environment. The dataset comprises 768 records, with eight key health attributes related to diabetes risk. The research methodology includes data preprocessing, normalization, handling missing values, and model evaluation using K-fold cross-validation. The results indicate that the Decision Tree algorithm achieves an accuracy of 74.68%, outperforming KNN, which achieves 68.18%. The superior performance of the Decision Tree is attributable to its ability to capture data patterns more effectively and to provide clearer interpretability through its decision rule structure. This study contributes to the field of medical data analytics by providing empirical evidence on algorithmic comparisons and demonstrating the usefulness of RapidMiner in developing predictive models for early diabetes detection.

Keywords: *Diabetes Prediction, K-Nearest Neighbor, Decision Tree, Machine Learning, RapidMiner.*

Abstrak

Prediksi diabetes merupakan langkah penting dalam mendukung deteksi dini serta pencegahan komplikasi jangka panjang yang disebabkan oleh penyakit kronis. Penelitian ini bertujuan membandingkan kinerja algoritma *K-Nearest Neighbor* (KNN) dan *Decision Tree* dalam memprediksi diabetes menggunakan dataset Pima Indian Diabetes pada aplikasi *RapidMiner*. Dataset yang digunakan terdiri dari 768 data dengan delapan atribut kesehatan utama yang berkaitan dengan risiko diabetes. Metode penelitian meliputi preprocessing data, normalisasi, penanganan missing value, serta evaluasi model menggunakan *K-Fold Cross Validation*. Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* memperoleh akurasi sebesar 74,68%, lebih tinggi dibandingkan KNN yang hanya mendapatkan akurasi 68,18%. Keunggulan *Decision Tree* disebabkan kemampuannya membaca pola data dengan lebih baik serta menghasilkan struktur keputusan yang mudah diinterpretasikan. Penelitian ini memberikan kontribusi dalam bidang analitika kesehatan dengan menghadirkan bukti empiris perbandingan algoritma serta menunjukkan efektivitas *RapidMiner* dalam pengembangan model prediksi untuk deteksi dini penyakit diabetes.

Kata Kunci: *Prediksi Diabetes, K-Nearest Neighbor, Pohon Keputusan, Pembelajaran Mesin, RapidMiner.*

1. PENDAHULUAN

DOI: https://doi.org/10.51903/37naet22

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Perkembangan teknologi informasi dan ilmu data telah memberikan kontribusi signifikan [1], [2], [3], khususnya dalam bidang kesehatan, dalam membantu proses diagnosis dan prediksi penyakit secara lebih cepat dan akurat [4]. Salah satu penyakit yang menjadi perhatian serius di dunia medis adalah diabetes mellitus, yaitu gangguan metabolik kronis yang disebabkan oleh peningkatan kadar gula darah akibat gangguan produksi atau kerja insulin. Berdasarkan data *World Health Organization* (WHO), jumlah penderita diabetes terus meningkat setiap tahunnya, termasuk di Indonesia yang menempati peringkat tinggi di kawasan Asia Tenggara [4]. Diabetes adalah penyakit kronis yang jumlah penderitanya meningkat setiap tahun. Deteksi dini sangat penting untuk mencegah komplikasi seperti penyakit jantung, gagal ginjal, dan kerusakan saraf. Di era digital, machine learning telah banyak diterapkan dalam analisis data kesehatan untuk membantu memprediksi risiko penyakit berdasarkan pola data historis [5]. Perkembangan teknologi kecerdasan buatan (Artificial Intelligence) dan data mining memberikan peluang besar dalam mengembangkan sistem prediksi berbasis data [6].

Salah satu pendekatan yang banyak digunakan adalah algoritma *Machine Learning* (ML) [7]. Algoritma K-Nearest Neighbor (KNN) dan Decision Tree merupakan dua metode klasifikasi populer yang banyak digunakan dalam penelitian kesehatan karena kemudahannya dalam penerapan dan interpretasi. Dengan menggunakan dataset *Pima Indian Diabetes*, penelitian ini bertujuan membandingkan performa kedua algoritma tersebut menggunakan aplikasi RapidMiner [8]. *RapidMiner Studio* merupakan salah satu perangkat lunak yang mendukung implementasi algoritma ML secara visual tanpa perlu menulis kode pemrograman. Dengan fitur *drag-and-drop, RapidMiner* memungkinkan peneliti untuk membangun model KNN dan DT dengan efisien [9]. Penelitian ini bertujuan untuk membandingkan performa kedua algoritma dalam memprediksi penyakit diabetes menggunakan *RapidMiner*. Kontribusi penelitian ini adalah memberikan analisis empiris terhadap performa KNN dan DT serta menunjukkan efektivitas penggunaan RapidMiner dalam penelitian prediktif [10].

2. TINJAUAN PUSTAKA

2.1. *Prediksi Diabetes dan Metode Klasifikasi dalam Machine Learning*

Penelitian terkait prediksi diabetes telah banyak dilakukan dengan berbagai pendekatan machine learning. Berbagai studi menunjukkan bahwa peningkatan kasus diabetes secara global mendorong pengembangan model prediktif berbasis data untuk membantu proses diagnosis medis [10]. Kondisi ini memperkuat urgensi pengembangan model prediksi yang mampu memberikan hasil akurat dan dapat menjadi dasar dalam pengambilan keputusan medis [11]. [12] menerapkan algoritma K-NN dengan delapan atribut kesehatan dan memperoleh akurasi 93% dalam mendeteksi risiko diabetes pada populasi masyarakat desa. Namun, sebagian besar penelitian terdahulu hanya memanfaatkan satu algoritma, dan tidak melakukan komparasi menyeluruh untuk menentukan metode terbaik. Padahal, perbandingan langsung antar algoritma pada dataset yang sama memberikan pemahaman lebih dalam tentang kelebihan dan keterbatasan masing-masing model. Selain itu, penelitian terdahulu juga cenderung menggunakan dataset dengan jumlah atribut terbatas, sehingga tidak mengakomodasi faktor risiko yang lebih kompleks seperti pola tidur, tingkat stres, kebiasaan konsumsi makanan cepat saji, atau riwayat keluarga [13]. Dengan pendekatan ini, penelitian diarahkan untuk menemukan algoritma yang paling unggul dalam memprediksi diabetes secara akurat, sekaligus memberikan kontribusi ilmiah berupa benchmarking komprehensif dari tiga metode populer dalam klasifikasi kesehatan [14].

2.2. *Implementasi Algoritma K-Nearest Neighbor (KNN) untuk Prediksi Diabetes*

Beberapa penelitian melaporkan bahwa pendekatan *K-Nearest Neighbor* (KNN) mengklasifikasikan data dengan mempertimbangkan kedekatan antara data uji dan sejumlah data latih terdekat yang ditentukan oleh nilai K. Dalam prediksi diabetes, KNN banyak digunakan karena konsepnya sederhana dan mampu bekerja baik pada data kesehatan yang memiliki banyak variabel seperti glukosa, tekanan darah, dan indeks massa tubuh (BMI) [15]. Penelitian ini menggunakan aplikasi RapidMiner untuk membangun model KNN. Proses dimulai dengan pembersihan dan normalisasi data, karena KNN sensitif terhadap perbedaan skala antar atribut. Selain itu, dataset Pima Indian Diabetes yang digunakan memiliki distribusi kelas tidak seimbang, sehingga diterapkan teknik SMOTE-ENN untuk membuat data lebih seimbang sebelum proses klasifikasi dilakukan [16]. Sementara itu, penelitian lain oleh [17] membandingkan beberapa algoritma dan menemukan bahwa KNN cenderung sensitif terhadap skala atribut, sehingga memerlukan normalisasi yang tepat untuk menghasilkan prediksi yang optimal. Secara keseluruhan, penelitian ini menunjukkan bahwa dengan

p-ISSN : 2803-1507 e-ISSN : 2803-1531

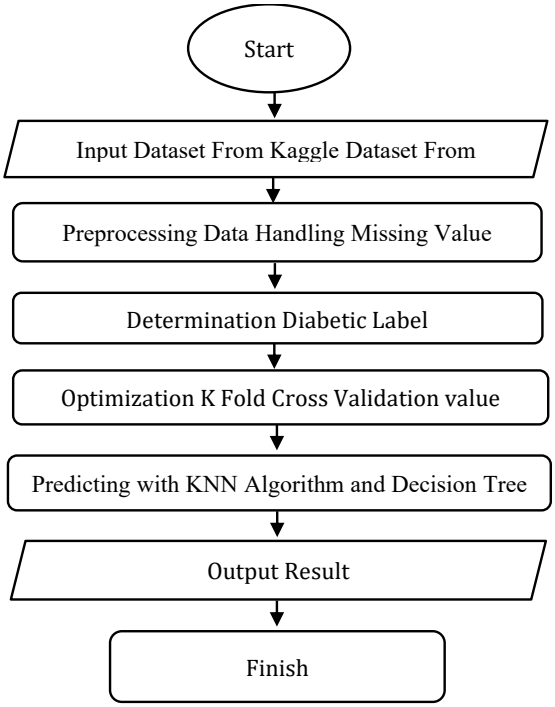
pemilihan nilai K yang tepat dan proses praproses data yang benar, algoritma KNN dapat menjadi metode yang efektif untuk membantu memprediksi risiko diabetes menggunakan RapidMiner [18].

2.3. *Prediksi Diabetes Menggunakan Decision Tree dan Algoritma C4. 5*

Selain KNN, algoritma *Decision Tree* juga banyak digunakan dalam klasifikasi medis karena kemampuannya menghasilkan model yang mudah diinterpretasikan. Penelitian oleh [19] menunjukkan bahwa C4.5 berhasil mengidentifikasi atribut paling signifikan dalam prediksi diabetes melalui perhitungan entropy dan information gain. Penelitian ini menggunakan algoritma C4.5, yaitu versi pengembangan dari ID3 yang lebih akurat. C4.5 memilih atribut terbaik dengan menghitung entropy dan information gain, kemudian membentuk pohon keputusan berdasarkan nilai tersebut. Semakin tinggi gain suatu atribut, semakin besar kemungkinan atribut itu menjadi penentu dalam proses klasifikasi [20]. Data pasien dikumpulkan dan kemudian diproses menggunakan *RapidMiner*. Hasil yang diperoleh dari perhitungan manual dan hasil *RapidMiner* menunjukkan pohon keputusan yang sama. Dari pohon keputusan tersebut terlihat bahwa faktor seperti kadar gula darah tinggi, berat badan tertentu, dan tekanan darah memiliki pengaruh besar dalam menentukan apakah diabetes bersifat turunan atau tidak. Secara umum, penelitian ini menunjukkan bahwa *Decision Tree* dan algoritma C4.5 dapat digunakan sebagai alat bantu sederhana untuk memprediksi risiko diabetes dan memberikan gambaran faktor-faktor utama yang memengaruhi hasil prediksi [21].

3. METODOLOGI PENELITIAN

Penelitian ini menggunakan metodologi penelitian yang sistematis agar proses penelitian dapat dilakukan secara terarah dan setiap tahapan pengerjaannya jelas. Metodologi penelitian ini disusun untuk memastikan bahwa setiap langkah yang dilakukan sesuai dengan tujuan penelitian yang telah ditetapkan. Selain itu, penggunaan metodologi yang terstruktur membantu peneliti dalam mengelola data, menerapkan algoritma, serta melakukan evaluasi hasil secara objektif. Alur metodologi penelitian yang digunakan pada penelitian ini dapat dilihat pada Gambar 1**.**



Gambar 1. Metodologi Penelitian

3.1. *Dataset*

Dataset yang digunakan adalah *diabetes.csv* berisi 768 data dengan 8 atribut:

1. *Pregnancies*
2. *Glucose*
3. *BloodPressure*

p-ISSN : 2803-1507 e-ISSN : 2803-1531

4. *SkinThickness*
5. *Insulin*
6. *BMI*
7. *DiabetesPedigreeFunction*
8. *Age*
9. *Outcome* (0 = Tidak Diabetes, 1 = Diabetes)

Hasil statistik awal dataset (ringkas):

1. Total data: 768
2. Atribut paling berpengaruh: Glucose, BMI, Age
3. Distribusi kelas tidak seimbang (*Outcome* 1 sekitar 34.8%)

3.2. *Preprocessing Data*

Preprocessing data adalah tahap penting untuk melakukan analisis data, yang melibatkan persiapan dan prubahan pada data mentah menjadi lebih terstruktur hingga siap dilakukan analisis. Fase untuk menjaga kualitas serta keakuratan data dan memastikan data siap digunakan secara efektif dalam tools [22]. *Cleansing* data adalah proses untuk menghilangkan atau membersihkan data dari yang tidak terpakai atau data yang salah. Peneliti menggunakan dataset dari Kaggle dimana pada dataset tersebut sudah dikelompokan, sehingga sudah dapat dilakukan penelitian, akan tetapi pada penelitian ini peneliti menemukan beberapa data yang kosong atau belum terisi, sehingga perlu dilakukan *tools handling missing value* yang berfungsi untuk mengisi data yang kosong[23]

3.3. *Diabetic Label*

Diabetic Label adalah proses pelabelan yang dilakukan oleh peneliti untuk menghasilkan target dari penelitian yang akan dilakukan oleh peneliti, pelabelan dilakukan berdasarkan kolom yang dapat mengidentifikasi dari kolom-kolom yang lain atau kategori data yang lainnya, misalnya pada penelitian ini untuk menentukan pasien yang terjangkit penyakit diabetes, namun ada beberapa kategori yang mendukung terjangkitnya penyakit tersebut contohnya usia, konsumsi makanan, dll.[24]

3.4. *K-Fold Cross Validation*

K-Fold Cross Validation dapat digunakan evaluasi dari tahap suatu algoritma dengan membagi data sampling secara random kemudian menggabungkannya sebanyak dari nilai K. Tujuan pengujiannya adalah untuk menemukan model dengan nilai akurasi terbaik[25]. Pada penelitian ini peneliti menggunakan *K-Fold* cross validation dengan langkh sebagai berikut: mempersiapkan data latih dan data uji, membagi data menjadi K=5 sampai K=10, menerpkan algorotma yang dugunakan, dan dilakukan perhitungan akurasi dari K=5 sampai dengan K=10.

3.5. *Algorithm Classification*

Algorithm Clasification adalah metode yang digunakan untuk mengelompokkan data ke dalam kelas atau kategori berdasarkan atribut atau sifatnya [26] *algorithm* yang diterapkan pada penelitian ini meliputi *K-Nearest Neighbor,* serta metode *Decision Tree,* berikut penjelasan algoritma yang digunakan pada penelitian ini: *KNearest Neighbor* merupakan *algorithm* yang digunakan unutk mendeteksi serta pembanding dari kondisi sebelumnya dengan yagng baru. memahami terkait data yang dievaluasi dengan data yang dikenal sebagai K-NN[10]. K-NN dapat menghasilkan akurasi yang cukup tinggi dalam memprediksi penyakit diabetes, apalagi pada data kesehatan yang tidak terstruktur, K-NN juga mudah diimplementasikan, namun jika tidak diatur maka akan terjadi *overfitting.* Untuk menghindari terjadinya *overfitting* peneliti melakukan *tunning* pada *dataset* yaitu dengan mengubah nilai K dengan cara menentukan jumlah tetangga terdekat dalam memprediksi kelas [10]. Perhitungan jarak pada algoritma K-NN menggunakan rumus jarak *Euclidean* yang ditunjukkan pada Persamaan (1).

$$dis = \sqrt{\sum_{i=0}^{n}(X_1 - X_2)^2} \quad (1)$$

Keterangan:
Dis = *distance*
i = *variable*
n = dimensi data
X1 = dimensi sampel
X2 = data uji

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Decision tree merupakan algoritma dari klasifikasi yang populer karena mudah dipahami orang dan termasuk pada metode klasifikasi data mining. Menurut [27], *Decision tree* merupakan gagasan *flowchart* struktur *tree*, di mana setiap node mewakili atribut, serta cabang menunjukkan hasil pengujian atau hasil nilai atribut, sedangkan daunny menunjukkan kelas. Dalam proses pembentukan pohon keputusan, pemilihan atribut terbaik dilakukan berdasarkan nilai *Information Gain*. Perhitungan nilai *Information Gain* ditunjukkan pada Persamaan (2), sedangkan perhitungan nilai *Entropy* sebagai ukuran ketidakpastian data ditunjukkan pada Persamaan (3).

$$Gain\,(S, A) = Entrophy\,(s) - \sum_{i=1}^{n} \frac{Si}{S} * Entrophy \quad (2)$$

$$Entrophy\ (s) = \sum_{i=1}^{n} -EEdd * \text{Log}_2\ \text{pi} \quad (3)$$

Keterangan :
S = Himpunan *Case*
A = *Atribute*
n = Jumlah Banyak *Atribut* A
Si = Jumlah *Case* pada banyaknya ke-i
S = Jumlah *Case* pada S
Pi = Proposi Si pada S

3.6. *Output Result*

Tahap ini merupakan bagian akhir dari proses penelitian, yaitu melakukan perbandingan terhadap performa ketiga algoritma yang digunakan dalam memprediksi penyakit diabetes. Melalui hasil pengujian tersebut, peneliti dapat menentukan algoritma mana yang memiliki tingkat akurasi paling tinggi sehingga dianggap paling efektif untuk digunakan. Selain akurasi, peneliti juga mempertimbangkan nilai *Recall* dan *Precision* sebagai indikator tambahan [27]. *Recall* menunjukkan seberapa banyak data positif yang berhasil diprediksi dengan benar dibandingkan jumlah data positif sebenarnya. Sementara itu, *Precision* mengukur seberapa banyak prediksi positif yang benar-benar tepat dari seluruh prediksi yang dinyatakan positif. Adapun *F1-Score* merupakan nilai gabungan dari Precision dan Recall, yaitu rata-rata harmonik yang memberikan gambaran lebih seimbang mengenai performa model [28]. Perhitungan nilai *Recall, Precision,* dan *F1-Score* masing-masing ditunjukkan pada Persamaan (4), Persamaan (5), dan Persamaan (6).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (5)$$

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (6)$$

4. HASIL DAN PEMBAHASAN

Penelitian ini menyajikan hasil serta pembahasan mengenai proses prediksi penyakit diabetes menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Decision Tree* pada aplikasi *RapidMiner*. Permasalahan diabetes saat ini semakin meningkat, tidak hanya dipengaruhi oleh konsumsi gula, namun juga berbagai faktor lain seperti usia, berat badan, tekanan darah, riwayat keluarga, hingga pola hidup. Oleh karena itu, diperlukan suatu metode prediksi yang dapat membantu proses diagnosis awal secara lebih akurat. Pada penelitian ini, dua algoritma diterapkan dan dibandingkan untuk mengetahui metode mana yang menghasilkan performa paling baik dalam memprediksi penyakit diabetes.

4.1. *Dataset*

Dataset yang digunakan adalah dataset Pima Indian Diabetes yang terdiri dari 768 data dengan 8 atribut utama, yaitu: *Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, dan Age*. Seluruh atribut tersebut berperan dalam menentukan apakah seorang

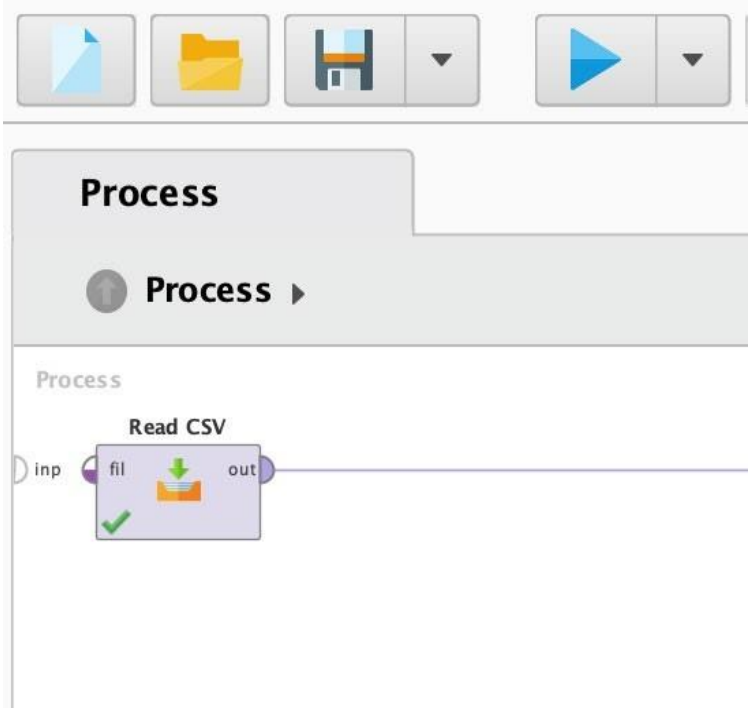
p-ISSN : 2803-1507 e-ISSN : 2803-1531

pasien terindikasi diabetes atau tidak. Tabel 1 berikut menggambarkan beberapa data awal dari dataset yang digunakan.

Tabel 1. 1-8 Dataset Prediksi Diabetes Berdasarkan Pengumpulan Data pada Kaggle

No	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Age
1	6	148	72	35	0	33.6	50
2	1	85	66	29	0	26.6	31
3	8	183	64	0	0	23.3	32
4	1	89	66	23	94	28.1	21
5	0	137	40	35	168	43.1	34
…	…	…	…	…	…	…	…
768	2	99	60	17	120	36.8	42

Dataset ini kemudian digunakan sebagai data utama dalam proses pelatihan dan pengujian model machine learning. Dataset yang telah dikumpulkan kemudian dilakukan sebuah proses lanjutan pada tools *RapidMiner*, pada *RapidMiner* menggunakan operator *Read CSV*, operator tersebut digunakan untuk men-ekstrak file *Excel* kedalam *RapidMiner*. Proses dapat ditunjukkan pada Gambar 2.

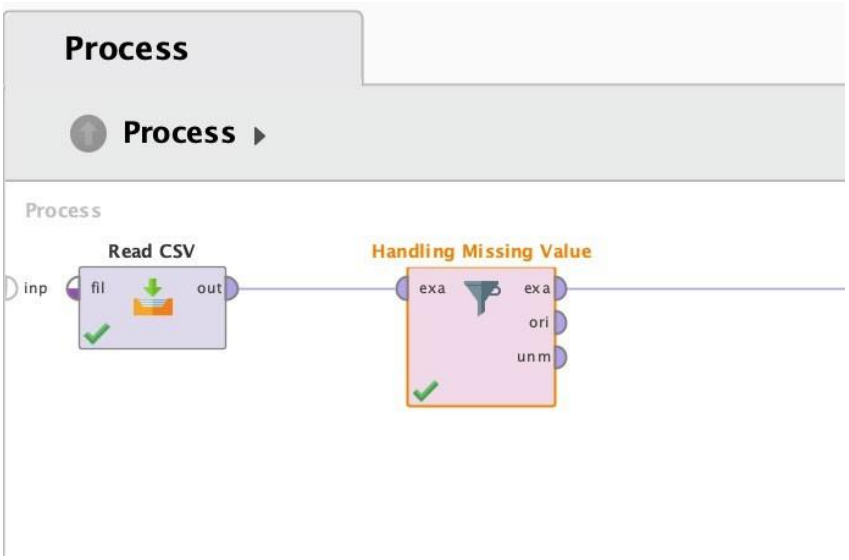


Gambar 2. Proses Ekstrak File Excel ke RapidMiner

4.2. *Preprocessing Data*

Tahapan preprocessing diperlukan untuk memastikan seluruh data telah bersih dan siap untuk dianalisis. Pada dataset ini dilakukan normalisasi nilai menggunakan operator *Normalize (Z-Transformation)* pada *RapidMiner* agar seluruh atribut berada pada skala yang seragam. Hal ini diperlukan karena algoritma KNN sangat sensitif terhadap perbedaan skala data. Selain itu, juga dilakukan pemeriksaan terhadap nilai kosong (missing value). Meskipun dataset *Pima Indian Diabetes* tergolong bersih, proses pengecekan tetap dilakukan untuk memastikan tidak ada data yang mengganggu performa model. Proses ini ditunjukkan menggunakan operator seperti *Replace Missing Value* yang mengisi nilai kosong dengan nilai median atau mean sesuai tipe atribut. Pada *dataset* yang digunakan peneliti terdapat beberapa kesalahan atau data yang kosong sehingga peneliti menggunakan *missing value* untuk mengisi data kosong tersebut agar *dataset* yang digunakan terstruktur dan selaras dengan datadata laiinnya, hal tersebut mengakibatkan *dataset* yang digunakan dapat dilakukan penelitian untuk prediksi penyakit diabetes. Oleh karena itu *d*apat ditunjukan pada Gambar 3.

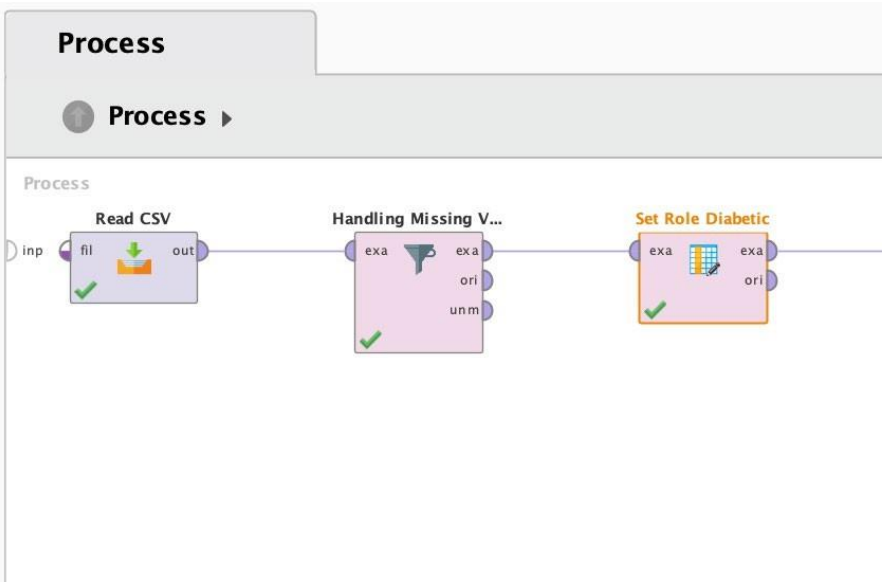
p-ISSN : 2803-1507 e-ISSN : 2803-1531



Gambar 3. *Handling Missing Value* Digunakan untuk *Cleaning Dataset*

4.3. *Diabetic Label*

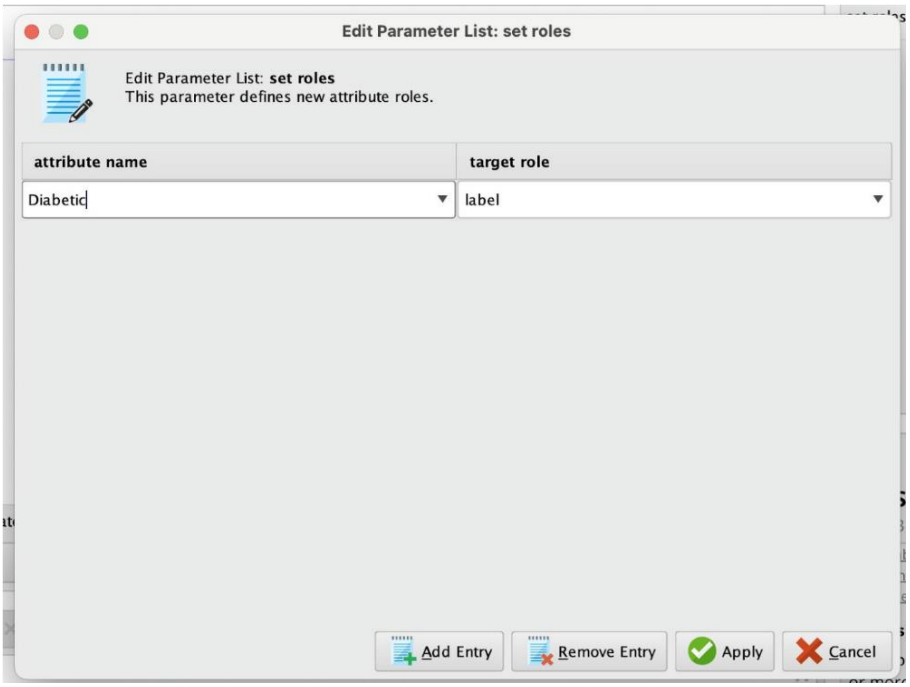
Pada penelitian ini, atribut *Outcome* digunakan sebagai label atau kelas yang menjadi acuan prediksi. Proses pelabelan dilakukan menggunakan operator *Set Role* pada *RapidMiner*, yang berfungsi menetapkan kolom mana yang menjadi label, attribute, atau id. Penentuan label ini sangat penting karena menjadi dasar bagi algoritma dalam melakukan proses klasifikasi diabetes. Jika label tidak ditentukan dengan benar, maka sistem tidak dapat menghasilkan model prediksi yang akurat. Proses pelabelan dataset menggunakan operator *Set Role* pada *RapidMiner* ditunjukkan pada Gambar 4**.**



Gambar 4. Pelabelan Dataset

Pada tahap pelabelan menggunakan *operator Set Role*, dilakukan penentuan atribut mana yang akan dijadikan sebagai kelas. Dalam penelitian ini, atribut *Diabetic* ditetapkan sebagai label utama pada dataset diabetes. Penentuan label ini sangat penting karena menjadi dasar bagi proses klasifikasi, yaitu untuk membedakan apakah seorang pasien termasuk kategori diabetes atau tidak. Tanpa penetapan label yang tepat, proses prediksi tidak dapat berjalan dan model tidak akan mampu menghasilkan keluaran yang valid. Proses penetapan label tersebut dapat dilihat pada Gambar 5.

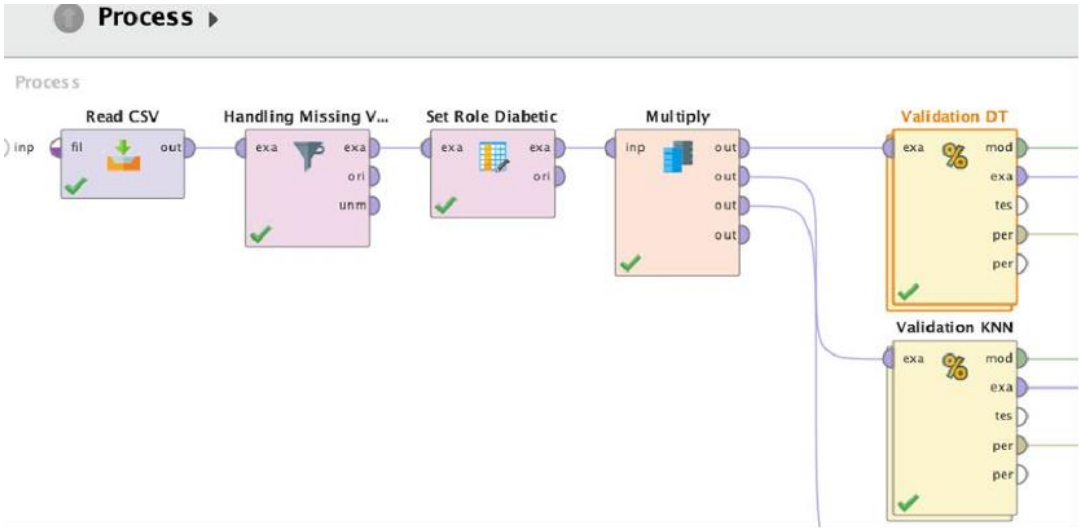
p-ISSN : 2803-1507 e-ISSN : 2803-1531



Gambar 5. Kelas yang Digunakan sebagai Label

4.4. *K-Fold Cross Validation*

Untuk menghindari terjadinya *overfitting* dan memastikan model memiliki generalisasi yang baik, penelitian ini menggunakan metode *K-Fold Cross Validation* dengan nilai K = 10. Dengan pendekatan ini, dataset dibagi menjadi 10 bagian di mana setiap bagian secara bergantian digunakan sebagai data uji, sementara bagian lainnya digunakan sebagai data latih. Metode ini juga memberikan evaluasi yang lebih stabil karena performa model tidak hanya bergantung pada satu kali pembagian data, tetapi pada 10 kali iterasi. Proses ini dilakukan untuk memisahkan data latih dan data uji menggunakan operator *K-Fold Cross Validation* pada setiap algoritma yang diterapkan. Selain itu, operator *Multiply* digunakan untuk memungkinkan perbandingan performa antar metode. Peneliti menerapkan nilai K mulai dari 5 hingga 10 guna meminimalkan risiko underfitting maupun *overfitting*, serta mengurangi bias pada data. Dengan rentang K tersebut, model diharapkan mampu menghasilkan akurasi yang lebih stabil dan representatif [29]. Proses pengujian dataset menggunakan operator *K-Fold Cross Validation* ditunjukkan pada Gambar 6.



Gambar 6. Operator K-Fold Cross untuk Pengujian Dataset

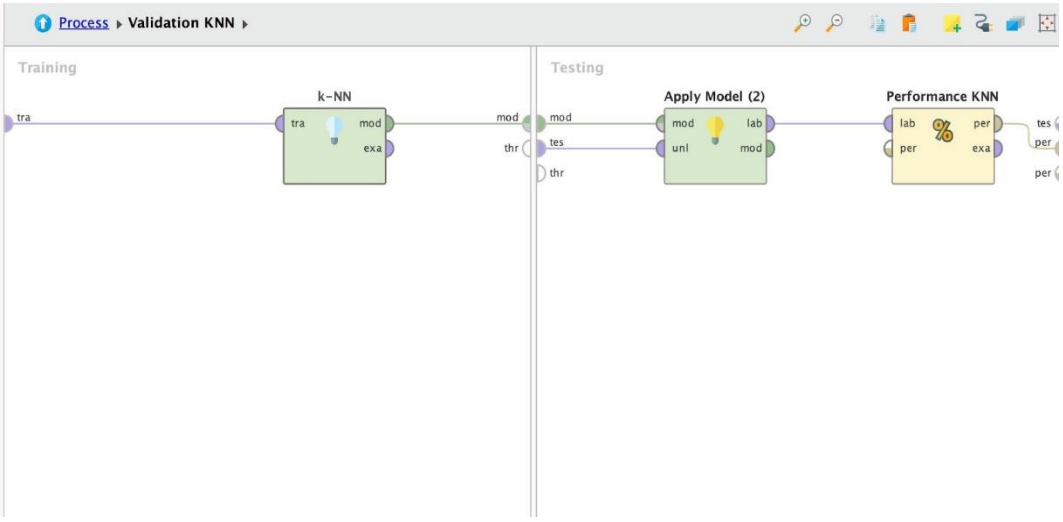
p-ISSN : 2803-1507 e-ISSN : 2803-1531

4.5. *Penerapan Algoritma KNN dan Decision Tree*

Pada tahap ini dilakukan penerapan algoritma menggunakan operator yang tersedia di *RapidMiner*.

4.5.1. *Penerapan Algoritma KNN*

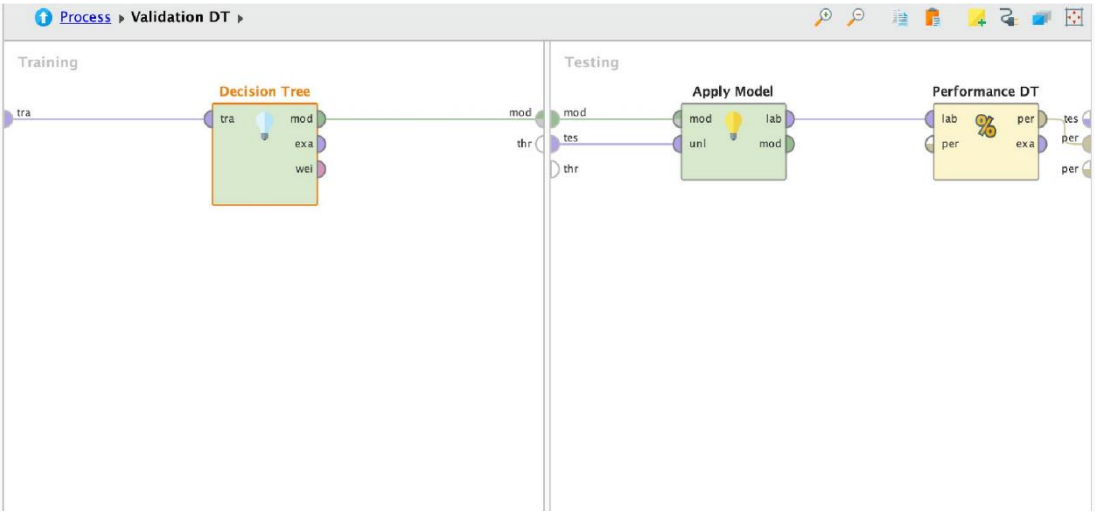
Proses ini dilakukan sebagai tahap penerapan algoritma dalam penelitian menggunakan tools *RapidMiner*. Pada metode K-NN digunakan operator pendukung seperti *Apply Model* dan *Performance* untuk menghasilkan evaluasi akurasi, *precision, recall*, dan *f1-score* sebagai indikator kinerja model. Hasil evaluasi performa algoritma K-NN tersebut dapat dilihat pada Gambar 7.



Gambar 7. Penerapan Algoritma KNN pada RapidMiner

4.5.2. *Penerapan Algoritma Decision Tree*

Decision Tree digunakan menggunakan operator *Decision Tree*, kemudian dihubungkan dengan operator Apply Model dan Performance agar dapat dihitung performa model secara keseluruhan. hasil dari performa pada metode K-NN dilihat berdasarkan akurasi, *recall,* dan, *precision.* Dapat diltunjukkan pada Gambar 8.



Gambar 8. Penerapan Algoritma Decision Tree pada RapidMiner

4.6. *Output Result*

Tahap ini merupakan hasil akhir dari seluruh proses pengujian algoritma untuk memprediksi diabetes. Berdasarkan proses pengujian, diperoleh hasil sebagai berikut:

p-ISSN : 2803-1507 e-ISSN : 2803-1531

4.6.1. *Hasil Algoritma KNN*

1. Akurasi: 68.18%
2. Precision: 56.25%
3. Recall: 49.09%
4. F1-Score: 52.43%

Model KNN menunjukkan performa yang kurang stabil, terutama pada nilai *recall* yang rendah. Hal ini mengindikasikan bahwa KNN sering gagal mendeteksi pasien yang benar-benar diabetes, sehingga metode ini kurang direkomendasikan untuk prediksi pada dataset ini.

4.6.2. *Hasil Algoritma Decision Tree*

1. Akurasi: 74.68%
2. Precision: 62.50%
3. Recall: 72.73%
4. F1-Score: 67.23%

Decision Tree memberikan hasil yang lebih baik dan stabil dibandingkan KNN. Nilai *recall* yang lebih tinggi menunjukkan bahwa Decision Tree lebih mampu mengidentifikasi kasus diabetes secara akurat. Selain itu, struktur pohon keputusan mempermudah proses interpretasi hasil.

4. DISKUSI

Hasil pengujian menunjukkan bahwa algoritma *Decision Tree* menghasilkan akurasi sebesar 74,68%, lebih tinggi dibandingkan *K-Nearest Neighbor* yang hanya mencapai 68,18%. Perbedaan ini juga terlihat pada nilai *recall*, di mana *Decision Tree* mampu mengenali 72,73% kasus diabetes aktual, sementara *KNN* hanya mencapai 49,09%. Nilai *recall* yang lebih tinggi menunjukkan bahwa *Decision Tree* lebih efektif dalam mendeteksi pasien yang benar-benar mengalami diabetes, yang merupakan aspek krusial dalam konteks prediksi medis. Temuan ini menjawab tujuan penelitian yang menekankan perbandingan kemampuan kedua algoritma dalam mengidentifikasi kasus diabetes secara akurat.

Jika dibandingkan dengan penelitian [10] yang menggunakan dataset dan platform serupa, *Decision Tree* juga menunjukkan performa yang lebih stabil dibandingkan *KNN* dalam klasifikasi diabetes. Hasil ini sejalan dengan temuan [19] yang melaporkan bahwa algoritma C4.5 memiliki kemampuan pemilihan atribut yang lebih adaptif melalui mekanisme *information gain*. Sebaliknya, performa *KNN* yang lebih rendah pada penelitian ini menguatkan laporan [17] bahwa algoritma tersebut sangat sensitif terhadap distribusi data dan ketidakseimbangan kelas. Dengan demikian, hasil penelitian ini memperkuat kecenderungan bahwa algoritma berbasis pohon keputusan lebih sesuai untuk dataset kesehatan yang memiliki variabilitas atribut tinggi.

Nilai *precision* dan *F1-score* juga menunjukkan pola yang konsisten, di mana *Decision Tree* memperoleh *F1-score* sebesar 67,23%, lebih tinggi dibandingkan *KNN* sebesar 52,43%. Perbedaan ini mengindikasikan bahwa *Decision Tree* mampu menjaga keseimbangan antara ketepatan prediksi dan kemampuan mendeteksi kasus positif secara lebih baik. Rendahnya *recall* pada *KNN* menunjukkan adanya kecenderungan model gagal mengenali pasien diabetes, meskipun telah dilakukan normalisasi dan *cross validation*. Kondisi ini mencerminkan keterbatasan algoritma berbasis jarak dalam menangani data medis yang memiliki tumpang tindih nilai antar kelas, sebagaimana juga dilaporkan pada penelitian lain di bidang kesehatan [8].

Dari sisi implikasi, struktur pohon keputusan yang dihasilkan *Decision Tree* memungkinkan interpretasi yang lebih jelas terhadap peran atribut seperti kadar glukosa, BMI, dan usia dalam proses klasifikasi. Karakteristik ini sejalan dengan kebutuhan sistem pendukung keputusan medis yang menuntut transparansi dan kemudahan pemahaman oleh tenaga kesehatan [21]. Sementara itu, *KNN* tidak menyediakan mekanisme interpretasi internal karena keputusan klasifikasi hanya bergantung pada kedekatan data. Selain itu, penelitian ini memiliki keterbatasan pada ukuran dataset yang relatif kecil serta distribusi kelas yang tidak seimbang, yang berpotensi memengaruhi stabilitas performa kedua model.

5. KESIMPULAN DAN SARAN

6.1. *Kesimpulan*

p-ISSN : 2803-1507 e-ISSN : 2803-1531

Berdasarkan hasil penelitian yang telah dilakukan mengenai penerapan algoritma *K-Nearest Neighbor* (KNN) dan Decision Tree dalam memprediksi penyakit diabetes menggunakan dataset Pima Indian Diabetes pada platform RapidMiner, dapat disimpulkan bahwa algoritma *Decision Tree* memberikan performa klasifikasi yang lebih baik dibandingkan KNN. Model *Decision Tree* yang diuji pada penelitian ini mampu mencapai akurasi sebesar 74,68% dengan nilai recall sebesar 72,73%, yang menunjukkan kemampuan model dalam mendeteksi pasien yang benar-benar terindikasi diabetes secara lebih efektif. Sementara itu, algoritma KNN hanya menghasilkan akurasi sebesar 68,18%, yang menunjukkan bahwa model tersebut kurang stabil dan lebih sensitif terhadap skala data serta ketidakseimbangan kelas dalam dataset. Selain itu, Decision Tree memiliki keunggulan dalam hal interpretabilitas, karena struktur pohon keputusan yang dihasilkan dapat dengan mudah dipahami sehingga membantu dalam mengetahui faktor-faktor yang paling berpengaruh terhadap prediksi, seperti kadar glukosa, BMI, tekanan darah, dan usia. Temuan ini menunjukkan bahwa *Decision Tree* tidak hanya menghasilkan prediksi yang lebih akurat, tetapi juga menawarkan transparansi dalam pengambilan keputusan model. Meskipun demikian, penelitian ini memiliki keterbatasan terkait ukuran dataset yang relatif kecil serta ketidakseimbangan kelas yang tidak sepenuhnya ditangani, sehingga performa model berpotensi berubah jika diterapkan pada dataset lain yang lebih kompleks atau bervariasi. Oleh karena itu, evaluasi lanjutan terhadap kualitas generalisasi model masih diperlukan sebelum sistem dapat diimplementasikan secara lebih luas.

6.2. *Saran*

Berdasarkan hasil penelitian dan analisis yang diperoleh, beberapa saran dapat diberikan untuk pengembangan penelitian selanjutnya. Pertama, penelitian perlu menggunakan dataset yang lebih besar dan lebih variatif agar model memiliki kemampuan generalisasi yang lebih baik terhadap populasi pasien dengan karakteristik yang berbeda. Kedua, untuk mengatasi permasalahan ketidakseimbangan kelas, disarankan menerapkan teknik penyeimbangan data seperti SMOTE atau metode sejenis sehingga performa model, terutama pada deteksi kasus diabetes positif, dapat meningkat. Ketiga, algoritma lain seperti *Random Forest*, *Gradient Boosting*, SVM, atau *Neural Network* dapat dieksplorasi sebagai pembanding guna mengetahui model mana yang memberikan hasil lebih optimal pada kasus prediksi diabetes. Keempat, proses tuning hyperparameter yang lebih mendalam dapat dilakukan baik pada KNN maupun *Decision Tree* agar performa model mencapai tingkat optimal. Selain itu, penelitian selanjutnya dapat mengembangkan sistem prediksi diabetes ke dalam bentuk aplikasi berbasis web atau mobile yang terintegrasi dengan data rekam medis sehingga dapat membantu tenaga kesehatan dalam melakukan deteksi dini secara cepat dan akurat. Terakhir, penambahan lebih banyak fitur kesehatan seperti riwayat keluarga, pola hidup, aktivitas fisik, dan parameter klinis lainnya akan memberikan konteks prediksi yang lebih kaya dan meningkatkan akurasi model secara keseluruhan. Pendekatan ini memastikan bahwa kemajuan teknologi tetap membawa manfaat tanpa menimbulkan risiko yang merugikan[30].

DAFTAR PUSTAKA

[1] A. P. Sari and K. R. Salman, "Pengaruh Pengetahuan Pajak, Norma Subjektif, dan Penghargaan Finansial Terhadap Minat Karir Mahasiswa Menjadi Konsultan Pajak dengan Self-Efficacy Sebagai Variabel Moderasi," *Kompak J. Ilm. Komputerisasi Akunt.*, vol. 18, no. 1, pp. 412–422, Jul. 2025, doi: 10.51903/kompak.v18i1.2276.

[2] S. Arkha, A. Nugroho, and A. Wibowo, "Evaluating Digital Transformation within Integration Limitations using Desk-Based Analytical Case Study," *J. Technol. Informatics Eng.*, vol. 4, no. 2, pp. 289–299, Aug. 2025, doi: 10.51903/jtie.v4i2.365.

[3] S. U. Anggono and S. Supriyanto, "Peluang dan Tantangan Sistem Pelatihan Berbasis AI dalam Meningkatkan Keterampilan Tenaga Kerja di Era Digital," *J. Ilm. Sist. Inf.*, vol. 4, no. 2, pp. 224–231, May 2025, doi: 10.51903/6mbmzk12.

[4] A. Astuti, A. A. Mane, and M. Y. Saleh, "Peran Perkembangan Teknologi Digital Terhadap Strategi Pemasaran Dan Distribusi Umkm Kota Makassar," *Indones. J. Bus. Manag.*, vol. 6, no. 1, pp. 175–180, 2023, doi: 10.35965/jbm.v6i1.3788.

[5] Priyatna, B. S., Wahyudin, and Widiyasari, "Hubungan Pengetahuan dan Sikap dengan Perilaku Pencegahan Diabetes Mellitus pada Usia 45–54 di Sukagumiwang," *Jurnal Kesehatan Holistic*, vol. 9, no. 2, pp. 267–276, 2025, doi: 10.33377/jkh.v9i2.262.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

[6] S. Masrichah, "Ancaman dan Peluang Artificial Intelligence (AI)," *Khatulistiwa: Jurnal Pendidikan dan Sosial Humaniora*, vol. 3, no. 3, pp. 83–101, 2023, doi: 10.55606/khatulistiwa.v3i3.1860.

[7] A. S. Sitio and F. A. Sianturi, "Penerapan Algoritma Machine Learning dalam Analisis Pola Perilaku Penggunaan Internet," *DIKE J. Ilmu Multidisiplin*, vol. 2, no. 2, pp. 46–51, 2024, doi: 10.69688/dike.v2i2.102.

[8] T. Brian, A. N. A. Maulidhia, E. N. Sholikhah, and S. Wibowo, "Perbandingan Model Decision Tree, Support Vector Machine dan K-Nearest Neighbors untuk Memprediksi Kualitas Air Minum," *INTEGER J. Inf. Technol.*, vol. 10, no. 1, pp. 95–104, 2025, doi: 10.31284/j.integer.2024.v10i1.7523.

[9] Fauziah, A. S. R., and A. Sri Ramadhantya, "Penggunaan Rapidminer untuk Memprediksi Kelulusan Mahasiswa dengan Algorithm Naive Bayes," *Jurnal Teknologi Informasi*, vol. 10, no. 1, pp. 52–60, 2024, doi: 10.52643/jti.v10i1.4671.

[10] M. R. Ardianto and R. Rushendra, "Prediksi Penyakit Diabetes Berdasarkan Perbandingan Klasifikasi Metode K-Nearest Neighbor, Naïve Bayes, dan Decision Tree Menggunakan Rapid Miner," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 2, pp. 973–985, 2025, doi: 10.29100/jipi.v10i2.6079.

[11] I. S. Muallif, H. Budiman, and N. Ransi, "Penerapan Data Mining untuk Prediksi Pergerakan Harga Saham Menggunakan Algoritma K-Nearest Neighbor," *Prosiding Seminar Nasional Pemanfaatan Sains dan Teknologi Informasi*, vol. 1, no. 1, pp. 297–306, 2023.

[12] Nazori Suhandi, Rendra Gustriansyah, and A. Destria, "Klasifikasi Penyakit TBC menggunakan Metode UMAP dan K-NN," *bit-Tech*, vol. 7, no. 3, pp. 843–852, 2025, doi: 10.32877/bt.v7i3.2227.

[13] B. Novalina Sitorus and Harmayani, "Implementasi Algoritma K-Nearest Neighbor (KNN) untuk Klasifikasi Pelayanan Publik," *Journal of Science and Social Research*, vol. 8, no. 1, pp. 314–320, 2025. [Online]. Available: https://jurnal.goretanpena.com/index.php/jssr/article/download/2370/1526

[14] R. Alfiatul Karima and Z. Fatah, "Implementasi Metode K-Nearest Neighbor untuk Klasifikasi Penyakit Paru-Paru pada Anak," *Jurnal Ilmiah Multidisiplin Ilmu*, vol. 1, no. 6, pp. 13–14, 2024, doi: 10.69714/yx4smf68.

[15] F. M. N. Akbar, "Metode KNN (K-Nearest Neighbor) untuk Menentukan Kualitas Air," *J. Tekno Kompak*, vol. 18, no. 1, p. 28, 2024, doi: 10.33365/jtk.v18i1.3241.

[16] Idham, H. Rosika, and Yuliadi, "Implementasi RapidMiner untuk Clustering Data Penjualan Pakaian Menggunakan Metode K-Means," *JUTECH: Journal Education and Technology*, vol. 5, no. 1, pp. 221–231, 2024, doi: 10.31932/jutech.v5i1.3642.

[17] T. T. Herdyawan, D. C. Saputra, G. C. Aldosion, S. S. Nurhidayat, and Sukrinah, "Implementasi Algoritma K-Nearest Neighbor (KNN) menggunakan RapidMiner untuk Prediksi Penyakit Diabetes berdasarkan Dataset Pima Indian," *Journal of Information Technology and Informatics Engineering*, vol. 1, no. 1, pp. 36–40, 2025. [Online]. Available: https://journal.jci.co.id/jitie/id/article/view/111

[18] Pahlevi, I. R., "Penerapan Naive Bayes untuk Prediksi Penyakit Diabetes dengan Menggunakan Rapid Miner," *J-CEKI : Jurnal Cendekia Ilmiah*, vol. 4, no. 4, pp. 313–322, 2025, doi: 10.56799/jceki.v4i4.8531.

[19] F. Elfaladonna and A. Rahmadani, "Analisa Metode Classification-Decission Tree dan Algoritma C.45 untuk Memprediksi Penyakit Diabetes dengan Menggunakan Aplikasi Rapid Miner," *SINTECH (Science Inf. Technol. J.*, vol. 2, no. 1, pp. 10–17, 2019, doi: 10.31598/sintechjournal.v2i1.293.

[20] C. R. Aditya Nugroho and T. Kristiana, "Penerapan Algoritma C4.5 untuk Kepuasan Pelanggan Toko Online Parfume Chantik," *J. Algoritm.*, vol. 3, no. 1, pp. 10–21, 2022, doi: 10.35957/algoritme.v3i1.3169.

[21] R. Rofiani, L. Oktaviani, D. Vernanda, and T. Hendriawan, "Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru menggunakan Algoritma C4.5," *Jurnal Tekno Kompak*, vol. 18, no. 1, pp. 126–139, 2024, doi: 10.33365/jtk.v18i1.3525.

p-ISSN : 2803-1507 e-ISSN : 2803-1531

[22] A. Wahyu and R. Rushendra, "Klasterisasi Dampak Bencana Gempa Bumi Menggunakan Algoritma K-Means di Pulau Jawa," *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 1, p. 174, 2022, doi: 10.26418/jp.v8i1.52260.

[23] Intan Permata and Esther Sorta Mauli Nababan, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam*, vol. 2, no. 2, pp. 65–71, 2023, doi: 10.55606/jurrimipa.v2i2.1336.

[24] Y. Wang *et al.*, "Impact of Gold-Standard Label Errors on Evaluating Performance of Deep Learning Models in Diabetic Retinopathy Screening: Nationwide Real-World Validation Study," *J. Med. Internet Res.*, vol. 26, pp. 1–12, 2024, doi: 10.2196/52506.

[25] H. Hafid, "Penerapan K-Fold Cross Validation untuk Menganalisis Kinerja Algoritma K-Nearest Neighbor pada Data Kasus Covid-19 di Indonesia," *Journal of Mathematics, Computations, and Statistics*, vol. 6, no. 2, pp. 161–168, 2023, doi: 10.35580/jmathcos.v6i2.53043.

[26] D. K. Doni, Yuhandri, and A. Ramadhanu, "Penerapan Algoritma Haar Cascade Clasifier dan Computer Neural Network Sebagai Presensi Karyawan," *J. KomtekInfo*, vol. 11, no. 4, pp. 398–408, 2024, doi: 10.35134/komtekinfo.v12i1.565.

[27] D. Septhya *et al.*, "Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 15–19, 2023, doi: 10.57152/malcom.v3i1.591.

[28] A. I. Putri, Y. Syarif, P. Jayadi, F. Arrazak, and F. N. Salisah, "Implementation of Decision Tree and Support Vector Machine (SVM) Algorithm for Stunting Risk Prediction," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 349–357, 2023, doi: 10.57152/malcom.v3i2.1228.

[29] R. Kurniawan, A. T. Martadinata, and S. D. Cahyo, "Klasifikasi Tingkat Kematangan Buah Sawit Berbasis Deep Learning dengan Menggunakan Arsitektur Yolov5," *J. Inf. Syst. Res.*, vol. 5, no. 1, pp. 302–309, 2023, doi: 10.47065/josh.v5i1.4408.

[30] Zaehol Fatah, M. K., *TIK dan Masyarakat*, PT Penamuda Media, 2025. [Online]. Available: https://bacabuku.com/book/bk115964/computers-technology-tik-dan-masyarakat