PERBANDINGAN ALGORITMA K-NN DAN RANDOM FOREST DALAM MEMPREDIKSI PENYAKIT LIVER

M. Anhari Arifin[1], M. Rezki Hamdani[2], Willy Millano[3], Rahmaddeni[4], Lusiana Efrizoni[5]
muhammadarifin251218@gmail.com[1], m.rezkihamdani1802@gmail.com[2], willymillano2002@gmail.com[3], rahmaddeni@sar.ac.id[4], lusiana@stmik-amik-riau.ac.id[5]
STMIK AMIK RIAU

ABSTRAK

Kesulitan dalam mendeteksi penyakit liver secara dini merupakan masalah yang umum terjadi. Kini, dengan berkembangnya teknologi, teknik data mining dapat digunakan untuk diagnosis penyakit liver. Data untuk klasifikasi penyakit liver yang optimal menggunakan algoritma K-nearest neighbour dan Random Forest. Metode K-NN di pilih karna prinsipnya cukup sederhana dan mudah untuk di gunakan, namun akurasinya relatif rendah pada beberapa penelitian.sedangkan metode Random Forest memiliki reputasi yang baik untuk memberikan akurasi yang tinggi dalam berbagai tugas klasifikasi dan regresi. Penelitian ini membandingkan dua algoritma K-Nearst Neighbors dan Random forest, bahwasan nya algoritma Random Forest Lebih diunggulkan dari Algoritma K-Nears Neighbors. Dengan menunjukan akurasi 85% pada KNN sedangkan Random Forest 93%Dalam Penelitian ini Random Forest pada perbandangin penyakit liver lebih tinggi score akurat nya dari pada Algoritma K-Nearst Neighbors.

Kata Kunci: K-Nearst Neighbors, Random Forest, Penyakit Liver, Perbandingan.

PENDAHULUAN

Satu masalah suatu kesehatan utama di seluruh dunia adalah penyakit liver. Sangat penting untuk mencegah penyakit ini dan mengelola kondisi kesehatan pasien dengan baik. Teknologi informasi dan Machine learning telah menjadi komponen penting dalam pengembangan model prediktif dalam bidang kedokteran untuk membantu diagnosis penyakit liver.

Algoritma Machine Learning dapat digunakan memprediksi kemungkinan penyakit liver menggunakan data klinis pasien(Haerani & Syafria, 2023). K-NN (K-Nearest Neighbors) dan Random Forest ialah dua algorima yang biasa di gunakan untuk klasifikasi dan prediksi konteks ini (Dwi Prasetya & Sujatmiko, n.d.) Algoritma K-NN mengklasifikasikan sampel dengan menggunakan mayoritas label K tetangga terdekatnya dalam ruang atribut (Zulaikhah et al., 2022a). Namun, Random Forest adalah algoritma yang menggunakan sekumpulan pohon keputusan untuk membuat prediksi. Setiap pohon melakukan pilihan berdasarkan fitur yang dipilih secara acak untuk membuat prediksinya(Erdiansyah et al., 2022).

Proses pengolahan data menjadi sebuah informasiini sangat berguna sebagai dasar pengambilan keputusan yang tepat (Madaerdo Sotarjua et al., n.d.)., Dta yang dimana di kumpulkan dari tahun ke tahun di mana hampir semua data di masukan menggunakan Aplikasi sehingga menumpuk seperti gunung dan tidak berguna dan di buang (Faid et al., 2019). Penelitian lain juga dilakukan dengan melakukan proses preprocessing dengan normalisasi dan mengisi missing value. Normalisasi dilakukan pada data yang telah dipilih atributnya berdasarkan proses seleksi fitur dan missing value dilakukan dilakukan karena adanya data yang kosong atau tidak memiliki nilai atau informasi pada beberapa atributnya. (Zulaikhah et al., 2022b), perbandingan nilai akurasi algorima klasifikasi hal pemrosesan klasifikasi penyakit liver. (Setiawati et al., 2019)

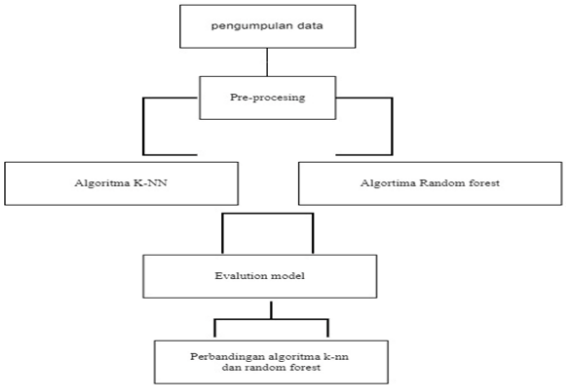
Data mining adalah proses suatu penggalian informasi mengindentifikasi dan data yang berguna agar untuk mengambil suatu keputusan. Beberapa Beberapa diantaranya adalah Metode K-NN,Decession Tree,Random Forest,dan NB (Khomsah, n.d.) .Peneltian yang memiliki niIai akurasi yang Iebih tinggi dibanding menggunakan algoritma neural network(Rekayasa et al., 2022).

Penelitian ini bertujuan untuk membandingkan kinerja kedua algoritma tersebut dalam Metode Machine Learning dapat digunakan untuk memprediksi kemungkinan penyakit liver menggunakan data klinis pasien. (Noviriandini et al., 2019)K-NN (K-Nearest Neighbors) dan Random Fores adalah dua algoritma yang biasa di guna kan untuk Klaslfikasi dan prediksi dalam konteks ini.

Algoritma K-NN mengklasifikasikan sampel dengan menggunakan mayoritas label K tetangga terdekatnya dalam ruang atribut. Namun, Random Forest adalah algoritma yang menggunakan sekumpulan pohon keputusan untuk membuat prediksi. Setiap pohon melakukan pilihan berdasarkan fitur yang dipilih secara acak untuk membuat prediksinya.

METODOLOGI

Metode yang akan digunakan ada yaitu metode K-Nearst Neighbors dan Random Forest, Tahapan penelitian data ini terdiri dari 5 tahap sebagaimana ditunjukan pada gambar 1, Kelima tahap tersebut sebagai berikut:



Gambar 1 Metodologi penilitian

Pengumpulan Data

Pertama pada penelitian ini di mulai pencarian dari situs Kaggle. data digunakan merupakan data public. https://www.kaggle.com/datasets/jeevannagaraj/indian-liver-patient-dataset.

Pre-Procesing

Persiapan awal penelitian yang dilakukan adalah membersihkan data dengan menghilangkan atribut yang tidak digunakan dalam analisis cluster. Pada penelitian ini tidak di dapati atribut yang harus dihapus atau dibuang.(Desiani, 2022) Selanjutnya untuk memudahkan identifikasi, masing-masing atribut diberi nama baru, yaitu:

(Age),(Gender),(Ttal_Bllirubin),(Direct_Bllrubin),(Alkllne_photopase),(Alamin_Aminotrnsferase),(Apartate_Amnotraanferase),(Ttal_Protins),(Albumn),(Albumn_and_Globulln_Ratio),(dataset)

Metode KNN

KNN adalah algoritma klasifikasi yang digunakan untuk Menentukan Klompok Berdasarkan mayoritas pada K tetagga paling dekat. Metode KNN dapat digunakan untuk mengukur jarak antara data uji dan data latih. Nilai jarak menunjukkan seberapa mirip atau dekat keduanya(Mutiara Sari, 2023). Metode KNN adalah metode Pembelajaran terawasi dikueri yang menentukan berdasarkan mayoritas kategori yang di temukan pada metode K-

NN.(Rahmat Rivita & Fikry, 2023)

Random Forest

Random Forest adalah metode kalsifikasi yang terdiri dari kumpulan keputusan yang akandijadikan Suatu vote men-dapatkan hasil terakhir (Fandru Al Rifqi et al., 2022)Random Forest adalah metode Klasfikasi yang terdiri kumpulan pohon keptusan yang akan dijadikan untuk mendapatkan hasil dari berupa data fitu acak independen yang berbeda beda (Madaerdo Sotarjua et al., n.d.)

Evaluasi Mode

Padat tahap ini model algoritma yang telah diterapkan dalam metode pembelajaran klasifikasi dilakukan perhitungan performa model pada algoritma KNN, dan Random Forest.

Untuk menghitung ketepatan, ketepatan, recall, dan skor f1, nilai-nilai tersebut akan dimasukkan ke dalam table yang diberikan. Ketepatan adalah hasil perbandingan antara jumlah prediksi benar pada data positif dengan jumlah total prediksi pada data positif. (Rahmat Rivita & Fikry, 2023)Perhitungan performa model klasifikasi didasarkan pada pengujian objek yang benar dan objek yang salah. Perhitungan performa klasifikasi yang digunakan pada penelitian ini adalah confusion*matrix/ yang berisi perhitungan hasil klasifikasi aktual yang dapat diprediksi (Erdiansyah et al., 2022)Pada Tabel 1 menunjukkan confusion matrix dua kelas.

(1)

$$Accuracy = \frac{TP + TN}{TP + TN + FB + FN}$$

(2)

$$Precision = \frac{TP}{TP + FP}$$

(3)

$$Recall = \frac{TP}{TP + PN}$$

(4)

$$F1\ Score = 2 * \frac{Precision * Recall}{TP + Recall}$$

HASIL DAN PEMBAHASAN

Pengumpulan Data

Dataset yang digunakan merupakan dtaset Publik yang berasal dari situs Kaggle. data digunakan merupakan data public. https://www.kaggle.com/datasets/jeevannagaraj/indian-liver-patient-dataset,Jumlah Dataset 583.

Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Dataset
65	Female	0,7	0,1	187	16	18	6,8	3,3	0,9	1
62	Male	10,9	5,5	699	64	100	7,5	3,2	0,74	1
62	Male	7,3	4,1	490	60	68	7	3,3	0,89	1
.....	.....	.....	.....	.....	.....	.....	.....	....	....	....
72	Male	3,9	2	195	27	59	7,3	2,4	0,4	1

Gambar 2 Dataset

Pre-Processing

Sebelum Proses data dilakukan transformasi ,Transformasi data Merubah data kategori menjadi numeric.Berdasarkan atribut Gender diubah Menjadi 0 dan1,Tabel menunjukan Transformasi data.

Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin	Albumin_and_Globulin_Ratio	Dataset
65	0	0,7	0,1	187	16	18	6,8	3,3	0,9	1
62	1	10,9	5,5	699	64	100	7,5	3,2	0,74	1
62	1	7,3	4,1	490	60	68	7	3,3	0,89	1
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
72	1	3,9	2	195	27	59	7,3	2,4	0,4	1

Gambar 3

Model Prediksi

Setelah Pre-Processing melakukan Transformasi data ,kita akan melakukan model prediksi perbandingan dua Algoritma K-NN dan Random Forest.

A. K-NN.

```
Training score:
 94.21
Test Score:
 85.47
Accuracy:
 0.8547008547008547
[[51 10]
 [ 7 49]]
              precision    recall  f1-score   support

           0       0.88      0.84      0.86        61
           1       0.83      0.88      0.85        56

    accuracy                           0.85       117
   macro avg       0.85      0.86      0.85       117
weighted avg       0.86      0.85      0.85       117
```

Gambar 4

B. Random Forest

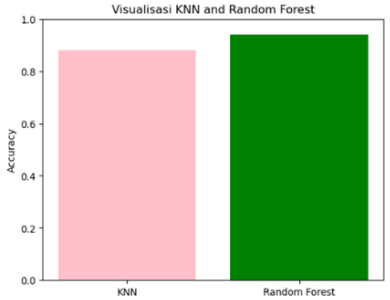
```
Random Forest Training score:
 100.0
Random Forest Test Score:
 94.02
Accuracy:
 0.9401709401709402
[[57  4]
 [ 3 53]]
              precision    recall  f1-score   support

           0       0.95      0.93      0.94        61
           1       0.93      0.95      0.94        56

    accuracy                           0.94       117
   macro avg       0.94      0.94      0.94       117
weighted avg       0.94      0.94      0.94       117
```

Gambar 5

Evaluasi Model



Gambar 6 Evaluasi model

Perbandingan Algoritma

Tabel ini menjelaskan Hasil Perbandingan dua Metode KNN dan Random Forest, Pada Tabel dibawah ini menunjukan Acurracy dan Pecision Pada Random forest diunggulkan sedangkan KNN Unggul pada Recall.

Tabel 1

Method	Acuracy	Precision	Recall
KNN	85%	**0,83**	**0,88**
Random Forest	93%	**0,95**	**0,94**

KESIMPULAN

Penelitian ini membandingkan dua algoritma K-Nearst Neighbors dan Random forest, bahwasan nya algoritma Random Forest Lebih diunggulkan dari Algoritma K-Nears Neighbors. Dengan menunjukan akurasi 85% pada KNN sedangkan Random Forest 93%.

DAFTAR PUSTAKA

Desiani, A. (2022). Perbandingan Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Penyakit Hati. SIMKOM, 7(2), 104–110. https://doi.org/10.51717/simkom.v7i2.96

Dwi Prasetya, W., & Sujatmiko, B. (n.d.). Rancang Bangun Aplikasi dengan Perbandingan Metode K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penderita Penyakit Diabetes.

Erdiansyah, U., Irmansyah Lubis, A., & Erwansyah, K. (2022). Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil. JURNAL MEDIA INFORMATIKA BUDIDARMA, 6(1), 208. https://doi.org/10.30865/mib.v6i1.3373

Faid, M., Jasri, M., & Rahmawati, T. (2019). Perbandingan Kinerja Tool Data Mining Weka dan Rapidminer Dalam Algoritma Klasifikasi. Teknika, 8(1), 11–16. https://doi.org/10.34148/teknika.v8i1.95

Fandru Al Rifqi, M., Dina, M., NKNababan, M., & Aisyah, S. (2022). http://infor.seaninstitute.org/index.php/infokum/index INFOKUM is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC 4.0) COMPARATIVE ANALYSIS OF PHISHING WEBSITE PREDICTION CLASSIFICATION ALGORITHM USING LOGISTIC REGRESSION, DECISION TREE, AND RANDOM FOREST. JURNAL INFOKUM, 10(2). http://infor.seaninstitute.org/index.php/infokum/index

Haerani, E., & Syafria, F. (2023). KLIK: Kajian Ilmiah Informatika dan Komputer Analisis Sentimen Tanggapan Masyarakat Terhadap Kenaikan Biaya Haji Tahun 2023 Menggunakan Metode K-Nearest Neighbor (KNN). Media Online, 4(3), 1562–1569. https://doi.org/10.30865/klik.v4i3.1471

Khomsah, S. (n.d.). Prediksi Harapan Hidup Penderita Hepatitis Kronik Menggunakan Metode-Metode Klasifikasi.

Madaerdo Sotarjua, L., Budhi Santoso, D., Singaperbangsa Karawang Jl Ronggo Waluyo, U. H., Telukjambe Timur, K., Karawang, K., & Barat, J. (n.d.). PERBANDINGAN ALGORITMA KNN, DECISION TREE,*DAN RANDOM*FOREST PADA DATA IMBALANCED CLASS UNTUK KLASIFIKASI PROMOSI KARYAWAN. 7(2), 2022.

Mutiara Sari, R. (2023). PERBANDINGAN METODE KNN DAN MKNN UNTUK DETEKSI DINI DIABETES MELLITUS. In Jurnal MNEMONIC (Vol. 6, Issue 2).

Noviriandini, A., Handayani, P., Bsi, U., & Nusa Mandiri, S. (2019). Prediksi Penyakit Liver Dengan Menggunakan Metode Naïve Bayes Dan K-Nearest Neighbour (KNN). Seminar Nasional Rekayasa Dan Teknologi, 27. http://archive.ics.uci.edu/ml/.

Rahmat Rivita, A., & Fikry, M. (2023). KLIK: Kajian Ilmiah Informatika dan Komputer Klasifikasi Sentimen Masyarakat di Media Sosial Twitter terhadap Calon Presiden 2024 Prabowo Subianto dengan Metode K-NN. Media Online), 3(6), 786–797. https://doi.org/10.30865/klik.v3i6.890

Rekayasa, K. K., Nugraheni, A., Dias Ramadhani, R., Arifa, A. B., & Prasetiadi, A. (2022). Terbit online pada laman web jurnal: http://journal.ittelkom-pwt.ac.id/index.php/dinda Journal of Dinda Perbandingan Performa Antara Algoritma Naïve Bayes dan K-Nearest Neighbour Pada Klasifikasi Kanker Payudara. Data Institut Teknologi Telkom Purwokerto, 2(1), 11–20. http://journal.ittelkom-pwt.ac.id/index.php/dinda

Setiawati, I., Wibowo, A. P., Hermawan, A., Teknologi, M., Universitas, I., & Yogyakarta, T. (2019). IMPLEMENTASI DECISION TREE UNTUK MENDIAGNOSIS PENYAKIT LIVER (Vol. 1, Issue 1).

Zulaikhah, S. H., Aziz, A., & Harianto, W. (2022a). OPTIMASI ALGORITMA K-NEAREST NEIGHBOR (KNN) DENGAN NORMALISASI DAN SELEKSI FITUR UNTUK KLASIFIKASI PENYAKIT LIVER. In Jurnal Mahasiswa Teknik Informatika) (Vol. 6, Issue 2). https://archive.ics.uci.edu/ml/index.php

Zulaikhah, S. H., Aziz, A., & Harianto, W. (2022b). OPTIMASI ALGORITMA K-NEAREST NEIGHBOR (KNN) DENGAN NORMALISASI DAN SELEKSI FITUR UNTUK KLASIFIKASI PENYAKIT LIVER. In Jurnal Mahasiswa Teknik Informatika) (Vol. 6, Issue 2). https://archive.ics.uci.edu/ml/index.php